University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Warburton, Katherine Ann

Title:
A Computational Analysis of the Structure and Evolution of Hierarchical Category Systems

Date:
2026

Persistent Link:
https://hdl.handle.net/11343/371548

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

A Computational Analysis of the Structure and Evolution of Hierarchical Category Systems

by

Katherine Warburton

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

School of Psychological Sciences
University of Melbourne

Graduate Department of Computer Science
University of Toronto

© Copyright 2027 by Katherine Warburton

A Computational Analysis of the Structure and Evolution of Hierarchical Category Systems

Katherine Warburton
Doctor of Philosophy

School of Psychological Sciences
University of Melbourne

Graduate Department of Computer Science
University of Toronto

2027

Abstract

From folk taxonomies of animals to institutional systems like the Dewey Decimal Classification, hierarchical category systems shape how we think, communicate, and access information. Because these systems must compress vast amounts of data into meaningful but manageable categories, they inevitably reflect the needs, perspectives, and experiences of their creators. In this thesis, I present a large-scale computational analysis of the structure and evolution of hierarchical category systems, demonstrating how they encode bias through unequal representation of the things they classify, and how the order in which the environment is encountered can shape the structure and fit of category systems.

First, I examine gender bias in two institutional category systems, focusing on asymmetries in the existence of gendered subcategories and whether they reflect psychological principles of typicality. I compare these asymmetries to those in gendered language use, examining whether they can be explained by gender base rates and whether they have changed over time. Next, I develop a framework that conceptualizes how bias is encoded in the hierarchical arrangement of items and categories within a system. This framework measures Western and gender bias in two US-based library classification systems with millions of categories and is applied to animal concepts in English WordNet to demonstrate its broader applicability to cultural category systems. Finally, I propose that one reason hierarchical category systems can become biased is that they are formed incrementally with incomplete knowledge of the space of items they must categorize. Using a combination of computational modelling and behavioural experiments, I show that locally optimal decisions can produce suboptimal hierarchies depending on stimulus order. The model adapts

a hierarchical clustering objective and outperforms existing rational models in predicting human behaviour, highlighting the need for intrinsically hierarchical models.

Overall, this thesis shows how hierarchical category systems can encode skewed representations of the world through both the biases embedded in their structure and the incremental processes by which they evolve. The methods developed here provide a foundation for future work on bias mitigation in both human and machine-generated classification systems and contribute to a better understanding of categorization in general.

Acknowledgements

This PhD spanned two universities, two continents, three supervisors, five years, and contained enough paperwork to last a lifetime. This means that I have been lucky enough to meet many, many wonderful people at opposite ends of the world. They have made my time during the PhD so much more enjoyable.

First, I want to thank my lovely supervisors, Lea Frermann, Charles Kemp, and Yang Xu, for their continued support throughout the entire PhD process and their patience with my many emails, especially during the writing of this thesis. Our differing time zones meant we sometimes met at unusual hours, and they were always willing to make it work. Yang was my first introduction to the PhD program and always made our lab a fun and engaging place to work. I will always be grateful to him for suggesting this joint PhD when we first met and for introducing me to this unique opportunity. Charles always has thoughtful advice to offer and made time for many enjoyable lunchtime conversations. He helped me stay (relatively) calm throughout this process and always reminded me to try the simple solution first. Finally, Lea gave me a place to land when I first arrived in Melbourne and made me feel welcome from the very beginning. I am deeply grateful for all of her guidance and support throughout this PhD. I would also like to thank my thesis examiners, Barend Beekhuizen, Robert Goldstone, and Kenny Smith, for taking the time to read this thesis and for their valuable feedback and discussion.

From Toronto, I would like to thank Aida Ramezani, Zhewei Sun, Jade Yu, and everyone in the Cognitive Lexicon Laboratory, the LCC group, and the broader CL community for making DL Pratt a welcoming place despite the lack of windows. I would also like to thank Jai Aggarwal and Jules Watson for letting me distract them from across the hall, and finally John Xu, who was my joint PhD counterpart and was with me at both universities. He made moving from Toronto to Melbourne much less daunting.

From Melbourne, I would like to thank Viola Pucci, who was my first friend in the city and always ensured I took breaks and explored this new place. Uyen Doan and Merrick Giles, with whom I had many lovely, long coffee breaks; Manikya Allister, Violet Chae, Allegra Micomonaco, Bel Moore, Dyana Muller, and Zoë Wilson, for being the best desk neighbours; Jasmin Stariolo for being a wonderful roommate; Jon Januar for always being up for baking; and Redrick Gromove and Liam Timms for many wonderful games of D&D. I would also like to thank Bryan Chen, Damian Curran, Matteo Guida, Chunhua Liu, Yulia Otmakhova, and everyone in Lea's lab group who were worth walking down five flights of stairs for. Finally, I would like to thank Simon De Deyne, Courtney Hilton, Frank Mollica, Ella Woodford, and the broader CHDH and MSPS community for being so warm and welcoming, and for the many great lunches and conversations.

Beyond the PhD, I would like to thank my friends and family back home in Canada, including Loraine Diza, Val Kobilaski, Ben McKerry, Sam McPhail, Haadia Mufti, Emily Poon, Tessa and Meghan Warburton, Andena Xhiku, and Nancy Yarnell, for the many texts, phone calls, and Zoom calls that made living abroad feel less lonely. I would especially like to thank my parents, JJ and Richard, who will always read whatever I send them.

Contents

1 Introduction **1**

1.1 What are Hierarchical Category Systems? 1

1.2 Why Study Hierarchical Category Systems? 3

1.3 Open Questions in Understanding Hierarchical Category Systems 5

1.3.1 What Makes a Hierarchical Category System Effective? 5

1.3.2 How Do Hierarchical Category Systems Evolve? 6

1.4 Core Research Questions 7

1.5 Main Contributions 8

1.6 Thesis Overview 9

1.7 Summary of Published Work 11

2 Background **12**

2.1 The Structure of Category Systems 14

2.1.1 Hierarchy in Category Systems 14

2.1.2 Hierarchy & Optimal Category Systems 18

2.1.3 Biased Category Systems 25

2.2 The Evolution of Category Systems 31

2.2.1 Psychology: Emergence & Development of Cultural Systems 32

2.2.2 Library & Information Science: The History of Institutional Systems 34

2.2.3 Computer Science: Incremental Learning & Models of Categorization 35

3 Quantifying Bias in the Labels of Institutional Category Systems **37**

3.1 Introduction 37

3.2 Gender Bias in Institutional Category Systems 39

3.2.1 Data and Methods 39

3.2.2 Results 41

3.3 Base Rates and Gender Bias 43

3.3.1 Data and Methods . . . 43
3.3.2 Results . . . 44
3.4 Gender Bias Across Time . . . 45
3.4.1 Data and Methods . . . 45
3.4.2 Results . . . 46
3.4.3 Why has Wikipedia changed but not the LCSH? . . . 47
3.5 Discussion . . . 47
3.6 Conclusion . . . 48

4 Quantifying Structural Bias in Hierarchical Category Systems 49
4.1 Library Classification Systems . . . 53
4.1.1 Representing Library Classification Systems as Trees . . . 54
4.2 Defining Structural Biases . . . 54
4.2.1 Category Bias . . . 54
4.2.2 Item Bias . . . 56
4.3 Study 1: Western Category Bias . . . 58
4.3.1 Methods . . . 59
4.3.2 Results . . . 60
4.3.3 Summary . . . 66
4.4 Study 2: Item Gender Bias . . . 66
4.4.1 Methods . . . 66
4.4.2 Results . . . 67
4.4.3 Summary . . . 70
4.5 Discussion . . . 71
4.5.1 Why Library Classification Systems are Biased . . . 72
4.5.2 Beyond Library Classification . . . 73
4.5.3 Limitations . . . 74
4.5.4 Future Work . . . 75
4.6 Conclusion . . . 76

5 The Evolution of Hierarchical Category Systems 77
5.1 Introduction . . . 77
5.1.1 Evolution & Development of Hierarchies . . . 78
5.1.2 Computational Models of Categorization . . . 80
5.1.3 The Present Work . . . 81
5.2 Sequence Effects in Category Evolution . . . 83

5.2.1 The Evolution of a Simple Category System for Shells 83
5.2.2 Computational Model: Ordered CKMM 84
5.2.3 Hypotheses 86
5.3 Experiment 87
5.3.1 Methods 87
5.3.2 Results 92
5.4 Model Comparison 102
5.5 Discussion 104
5.5.1 When Does the Model Diverge from Human Behaviour? 107
5.5.2 Limitations & Future Work 109
5.6 Conclusion 110

6 Conclusion **112**
6.1 Main Findings 113
6.2 Implications 114
6.2.1 Implications for the Basic-Level Advantage 114
6.2.2 Practical Implications for Library Classification 115
6.2.3 Implications for NLP, Computational Technologies, & Beyond 118
6.3 Limitations 118
6.4 Future Directions 119
6.4.1 Biases Across Cultural, Institutional, and Synthetic Systems 119
6.4.2 Evolution of Institutional Systems 120
6.4.3 Emergence and Development of Cultural Systems 121
6.5 Final Remarks 122

A Supplementary Material for Chapter 4 **123**
A.1 LCC and DDC Category Examples 123
A.2 Computational Details 123
A.3 Distributional Bias Node Data 125
A.4 Distributional Bias: Plug-in Shannon Entropy 125
A.5 Domestic Bias in WordNet 125

B Supplementary Material for Chapter 5 **128**
B.1 Model Details 128
B.1.1 Ordered CKMM Objective 128
B.1.2 Model Implementation 129

B.1.3 CKMM Simulations 132
B.2 Bayesian Analyses 133
B.2.1 Sequence Effects 133
B.2.2 Priming Effects 136
B.2.3 Hierarchy Effects 137
B.3 Experiment Block Design 137
B.4 Extended Experiment Results 137
B.4.1 Hypothesis 1a(iii): Controlling for Depth in Participant Data 137
B.4.2 Priming Effects: Including Distractors in Participant Data 139
B.5 Cross-Validation Details 140

Bibliography **142**

List of Tables

1.1 Space of current and future research on hierarchical category systems 10

2.1 Summary of research on category systems across disciplines: Psychology, library and information science, and computer science 13
2.2 Functions of hierarchy in category systems 18

3.1 Datasets used to compile a list of jobs and identities. 41
3.2 Proportional odds ordinal regression BIC scores 45

4.1 Summary of circulation statistics for the LCC and DDC 62
4.2 Probability that a random non-Western node is deeper in the classification than a Western one 64
4.3 Descendant bias in the LCC and DDC 65
4.4 Results of the record-matching process between MARC and VIAF 67
4.5 2007 circulation statistics for male versus female authors 68
4.6 Category bias analysis for domestic versus wild mammal species in English WordNet 3.0 74

5.1 Multinomial regressions predicting category assignment from order. 95
5.2 Likelihood ratio tests for the effect of order on category assignment in human and simulated data 96
5.3 Jensen–Shannon divergences of the sequence conditions compared to the all-at-once baseline 100
5.4 Mean CKMM Scores relative to the minimum score across order conditions 101

6.1 Space of current and future research on hierarchical category systems revisited.112

A.1 "Handicrafts. Arts and Crafts" (TT1-999) subcategory names in the LCC and "Handicrafts" (745.5) subcategory names in the DDC. 123
A.2 Nodes in the LCC and DDC not used in the distributional bias analysis . . 125
A.3 Wikipedia pages scraped to obtain Wikipedia IDs for mammal species . . . 127

B.1 Grid search parameter values for fitting ordered CKMM models to participant data 132
B.2 Bayesian multinomial regressions predicting category assignment from order 134
B.3 Leave-one-out cross-validation comparison of Bayesian multinomial models 135
B.4 Experiment trial blocks 138
B.5 Distribution of first-item category assignments as a function of the number of subsequent same-category assignments 140
B.6 Grid search parameter values for fitting rational models of categorization to participant data 141
B.7 Model comparison using k-fold cross-validation 141

List of Figures

1.1 Types of hierarchical category systems 2

3.1 Types of gendered categories in the LCSH, Wikipedia, and Google Ngram . 42
3.2 Distribution of actual and perceived gender base rates of various jobs across cases W, WM, and M in the LCSH, Wikipedia, and Ngram-D 44
3.3 Cumulative proportions of gendered categories in the LCSH, Wikipedia, and Google Ngram over time 46

4.1 An example of how Religion is categorized in (a) the LCC and (b) the DDC. 53
4.2 Illustration of category count, level, and descendant bias 55
4.3 Illustration of item count, level, and distributional bias 56
4.4 The distributions of male and female authors across "Handicraft" subcategories in the LCC and DDC 57
4.5 Category count bias in the LCC and DDC 61
4.6 Category level bias in the LCC and DDC 63
4.7 Item level bias in the LCC and DDC 69
4.8 Distributional Bias in the LCC and the DDC 70

5.1 The evolution of the categorization of fungi by the Dewey Decimal Classification 78
5.2 Illustration of the three factors affecting category evolution: location, order, and depth 83
5.3 Left and right extremes of the four stimulus sets presented to participants . 88
5.4 Visualization of the hierarchical category system setup at the start of a trial 90
5.5 Example displays from "One-at-a-Time" and "All-at-Once" trials 91
5.6 Mean proportion of items in categories L, X, R by location 93
5.7 Mean proportion of items in categories L, X, R by stimulus set 93
5.8 Mean proportion of items in categories L, X, R by location and order . . . 94
5.9 Distribution over number of same category assignments for novel items in a sequence trial 99

5.10 Log-likelihood gain over random baseline for CKMM and RMC models on held-out folds 104

A.1 Distributional bias in the LCC and the DDC 126

B.1 Distribution over number of same category assignments for all items in a sequence trial 139

Chapter 1

Introduction

Every day, humans rely on hierarchical category systems that exist both implicitly in our minds and explicitly in our environments to make sense of the world. For example, when I walk into the library, a sign tells me that call numbers 700-999 are on the second floor and 000–699 are on the third. This sign refers to the Dewey Decimal System, a library classification system that organizes books by subject. Because of it, I know that books on psychology will be found on the third floor and history books on the second. These systems project a structure onto our environment that shapes how we interpret and interact with it. Now, if I were to walk into the same library 100 years ago, the arrangement of books would look rather different. For example, books on fungi used to be classified under plants, reflecting the then-dominant view that they were a subset of the plant kingdom. As our understanding changed, the classification system became inconsistent with our knowledge, and fungi were eventually reclassified to reflect their status as a unique kingdom. These systems do not form all at once; instead, they evolve over time as we interact with and learn more about our social and physical environments. In this thesis, I study the structure and evolution of hierarchical category systems to better understand how people represent the world around them, and why the systems we use come to look the way that they do.

1.1 What are Hierarchical Category Systems?

A **category** is a group of things that are treated as equivalent in some way (Mervis & Rosch, 1981). Members of a category can share functions (e.g., cutlery), features (e.g., birds), and event structures (e.g., celebrations), among other forms of similarity. A **category system** is a division of the world into a set of categories that are arranged in relation to one another. In a hierarchical category system, categories are organized into parent–child relationships, where more general categories contain more specific ones (Mervis & Rosch, 1981). There is some debate about whether hierarchical relationships between categories are explicitly represented in the mind (Collins & Quillian, 1969), or if these hierarchical relationships are

an emergent property of mental operations on other forms of representation (Rips et al., 1973; T. T. Rogers & McClelland, 2004). However, whether in the mind or not, hierarchy shows up cross-culturally as a common structure of human category systems (Berlin, 1992; Holman, 2005).

Hierarchical category systems can be constructed for any set of people, places, things, concepts, or actions, and they can represent different types of relationships. For example, they may capture taxonomic relationships: my tabby is not only a cat, but also a pet, an animal, and, more broadly, a living thing (Kay, 1971). They may also capture part–whole relationships: a finger is part of a hand, which is part of an arm (Markman, 1981; McClure, 1975; Tversky, 1989). For simplicity, I will refer to the things categorized by a system as **items**. In practice, human category systems are not strict hierarchies in which every subcategory has only one direct parent. A dandelion, for instance, is both a flower and a weed. Nevertheless, such systems consist of items that are categorized at multiple levels of specificity.

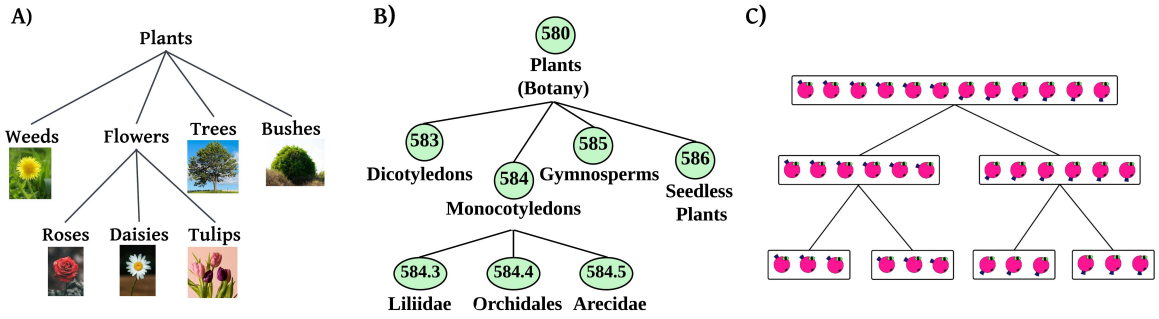


Figure 1.1: Types of hierarchical category systems. A) is an illustration of a plant folk taxonomy.[1] B) is an illustration of the categorization of plants in the Dewey Decimal Classification. C) is a synthetic category system constructed for a task involving the categorization of artificial alien organisms.

For any given set of items, multiple category systems can exist, reflecting differing environments and goals (e.g., Boster & D'Andrade, 1989; Bowker & Star, 1999; Medin et al., 1997; Tanaka & Taylor, 1991). These systems appear in two broad forms: **cultural** systems, which reflect the categories we carry in our heads and often express through language (often called folk taxonomies; Berlin et al., 1966; Holman, 2005), and **institutional** systems, which are large-scale formal systems created by institutions to serve specific purposes (Glushko et al., 2008). Examples of each type of system are illustrated for plants in A) and B) of Figure 1.1. To return to the cat example, in scientific taxonomy, a domestic cat (*Felis catus*) is a member of the *Felis* genus, the *Felidae* family, and the *Carnivora* order, reflecting a category system that differs from how cats are categorized in everyday

[1]Plant images in A) are free to use under the Unsplash License. From top left to bottom right, the photographers are Thammy Kolb, Mandeep Kaur, Edgar Louis, Ameen Fahmy, Tobias Rademacher, and Earl Wilcox.

language and thought. Unlike cultural category systems, institutional systems are often explicitly documented because they are too large to fit in any one person's head.

Figure 1.1 also includes a third kind of category system, which I call **synthetic** systems. Unlike cultural and institutional category systems, which arise naturally through everyday human activity, synthetic systems are constructed by participants from artificial stimuli during experiments (e.g., Murphy & Smith, 1982; Sokal & Rohlf, 1980). These systems have played an important role in psychological research on categorization. Experiments involving cultural category systems often require underlying category structures to be inferred from behaviour, and conclusions about how people learn and form categories can be confounded by existing category boundaries. By contrast, studies of synthetic systems allow researchers to manipulate category structure directly and examine categorization processes under controlled conditions where the influences of existing category boundaries are minimized. Cultural, institutional, and synthetic systems are all examples of human categorization and can provide complementary perspectives on how people construct categories and organize the world around them.

1.2 Why Study Hierarchical Category Systems?

Categorization is an effective cognitive tool that allows us to compress vast amounts of information into meaningful and manageable units. In a hierarchical category system, knowing the relationships between categories allows us to infer the properties of a subcategory from its membership in a parent category (Collins & Quillian, 1969; Markman & Callanan, 1984; Murphy, 2002; Murphy & Lassaline, 1997). Learning that a persimmon is a fruit, for instance, tells me it is edible, probably sweet, and has seeds. These systems help us reason quickly about new information without having to learn everything from scratch. But because information is compressed, these systems can also skew how we represent the world by emphasizing certain distinctions while ignoring others (Bowker & Star, 1999). This is not necessarily problematic, as the distinctions emphasized by a category system often reflect the needs of its users. For example, landscapers may classify trees according to practical functions, such as whether they provide shade or serve ornamental purposes, while botanists may focus on structural and biological characteristics (Medin et al., 1997). However, the distinctions encoded by a category system can also perpetuate more harmful biases.

Bias has many definitions. For example, it can be defined as a "tendency or preference", or as "an inclination for or against something" (American Psychological Association, 2018). Notably, Friedman and Nissenbaum (1996, p. 332) differentiate between general definitions of bias that see it as a "slant" in system design or outcomes, and more specific ones that associate these slants with "moral meaning" leading to systematic social discrimination. In this work, I take the former, broader perspective and define bias as a skew, tendency, or preference in how information is represented and encoded by a category system. This can

manifest as an asymmetry in how two groups are represented relative to one another (e.g., men and women), how a system represents items relative to an underlying similarity space, or simply as deviations in how different category systems represent the same set of things.

The links between categorization and bias have been studied in two main ways. First, research has demonstrated that categorization biases perception, with category learning affecting judgments of similarity and difference (Beale & Keil, 1995; Dubova & Goldstone, 2021; Goldstone, 1994; Goldstone & Hendrickson, 2010; Kuhl, 1991), and category labels influencing how perceptual information is interpreted (Huttenlocher et al., 2000; Lupyan, 2012; Lupyan & Spivey, 2010; Roberson et al., 2005). Second, category systems can also reflect and reinforce social biases by structuring how groups are mentally represented. For example, people tend to form more fine-grained categories for in-group members than out-group members, and they tend to perceive out-group members as more homogeneous (Judd & Park, 1988; Mackie & Worth, 1989; Park & Rothbart, 1982; Rubin & Badea, 2012). More broadly, implicit associations link social categories with particular traits and attributes (Greenwald et al., 1998).

These biases can arise from multiple sources. They may emerge through the process of categorization itself, or reflect biases present in a person's environment. However, regardless of their origin, research on categorization and bias tends to focus on the biases associated with individual categories, examining their labels, definitions, and boundaries. These bodies of work rarely consider how the structure of a category system as a whole can encode bias. Hierarchical organization in particular has received little attention in this context, despite evidence that it shapes perception in meaningful ways. For example, hierarchical organization of topics influences perceptions of what is considered general versus niche (Loehrlein, 2012), and in the Dewey Decimal Classification, humans occupy the same categorization depth as other broad groups of mammals like carnivores and non-human primates, implying an equivalence that does not necessarily reflect reality. The structure of a system can be a source of bias.

Studying the structure of hierarchical category systems therefore has two purposes. First, it provides insight into how people organize and understand their environment across cultures, as the structure of a system can reveal what is frequent, salient, or socially relevant to a group. Second, it helps identify biases that may be encoded and reinforced by the arrangement of items, categories, and labels within a system. These structural biases are particularly important to uncover when they reflect or reinforce social biases. Calls have been made for the study of large-scale naturally occurring datasets like institutional category systems to uncover psychological principles as they occur out in the real world (Goldstone & Lupyan, 2016). For example, analyzing large text corpora provides insights about gender and racial stereotypes encoded in our language (Garg et al., 2018). Institutional systems are much more accessible than the ones we carry around in our heads due to their explicitly recorded structures. This thesis therefore focuses on analyses of bias in the structure of

institutional category systems before turning to experiments and simulations that explore one process through which such systems might become biased and come to be structured as they are.

Beyond their importance for understanding how people and institutions organize knowledge, hierarchical category systems also form the basis of many information systems and datasets that underlie modern technologies. For example, databases are often structured using hierarchical B-trees to facilitate efficient retrieval. WordNet organizes concepts into hierarchical relationships through hyponym and hypernym relations (G. A. Miller, 1994) and underlies ImageNet, a dataset used to train image recognition systems, with downstream effects on the technologies built using these systems (Luccioni & Rolnick, 2023). Understanding how these structures perpetuate bias is an essential first step in addressing and mitigating the biases present in the many technologies that facilitate access to information and resources.

1.3 Open Questions in Understanding Hierarchical Category Systems

Despite extensive research on categorization, an important question remains: why do hierarchical category systems look the way they do? As established, the structure of a category system is dependent on the perspectives of the people who create and use it. However, the processes through which these perspectives become embedded in hierarchical structures are not fully understood. Two related questions are particularly important. First, what properties determine whether a category system is a good representation of the things it categorizes? Second, how do category systems change over time, and how do these changes influence the quality of the resulting representation?

1.3.1 What Makes a Hierarchical Category System Effective?

Multiple category systems can exist for a single set of resources; however, not all category systems are created equal. Librarians have preferences for different classification systems (Lund & Agbaji, 2018a), and research into human categorization has demonstrated that, even though all objects can be classified at multiple levels of abstraction, people learn certain levels of classification more easily than others (Mervis & Rosch, 1981). There exist some properties that make one classification system more optimal than another. These properties could be domain-specific, i.e., what makes an institutional category system optimal is entirely independent of what makes a classification system that organizes clothes optimal, but similar trends emerge across literatures.

According to psychology, a good category system efficiently represents the things it categorizes. Optimal systems group similar items together and dissimilar items apart,

balancing simplicity and informativeness to support efficient communication about one's environment (Kemp et al., 2018; Rosch, 1978). A system will therefore be structured around what is present and relevant to the people who create and use it. If birds are ecologically relevant to a community, they will likely be elaborately categorized, and if fish are not, they may receive far less attention. An optimal hierarchy should reflect the underlying structure of the domain it represents. However, domains can have multiple underlying structures, and different category systems capture different aspects of the same domain, emphasizing some relationships and structures while obscuring others (Bowker & Star, 1999; Weinberger, 2007). A system may therefore be a good representation of a particular environment for one purpose while being a poor representation of that same environment for another purpose, or the world more broadly.

These ideas are made more explicit in library and information science through the notion of warrants. These are the principles that justify which categories should exist, how they should be named, and how they should relate to one another (Svenonius, 2000). For example, literary warrant holds that the system should be justified by the things it classifies, use warrant that it should align with how users search and think about a domain, and structural warrant that it should maintain a coherent internal organization. These warrants may trade off against one another, reflecting different priorities in the design of a category system. Together, they reflect the intuition that a good category system accurately represents the space of things it organizes, while efficiently compressing information and facilitating retrieval or communication about these resources. This also means, however, that a system that represents its domain well will reflect any biases present within it. If the things being categorized are themselves skewed such that they over-represent some perspectives, topics, or groups at the expense of others, then a system designed to optimally represent them will reproduce that skew. A category system can therefore be biased because the things it represents are.

1.3.2 How Do Hierarchical Category Systems Evolve?

Category systems can also become biased through the process by which they are built. This bias is internal to the system and can arise in two ways: through cognitive constraints that favour certain structures and shape how we interpret the world, or through the dynamics of category formation itself. The order in which stimuli are encountered, for instance, influences the categories we construct (Carvalho & Goldstone, 2017; Egré et al., 2013) and some of the resulting categories may represent the world better than others. A system can therefore become biased simply because of the history of its construction and development.

Hierarchical category systems do not form all at once, but rather develop gradually over time. As they develop, they are shaped by both individual experience with an environment (Keil, 1989), and the transmission of knowledge between and across generations (Fer-

dinand & Perfors, 2020; Kirby et al., 2014). Typically, category evolution is studied in the context of language with a focus on how a set of words evolves to optimally divide a space of meanings in a way that supports efficient communication (Carr et al., 2017; Kemp et al., 2018). However, this work often treats category systems as flat, overlooking the hierarchical structure that characterizes human categorization. Similarly, research on sequence effects and category learning has focused on flat systems, and while the developmental literature addresses how hierarchies might emerge in children (Mervis & Crisafi, 1982; Waxman, 1990), it does not address how those hierarchies are subsequently adapted. Work like that of Chi (1992) and Keil (1981) offers some insight into how hierarchies are revised to accommodate new information, but the hierarchical structures being modified are contained in people's heads and can never be directly observable. The changes to the structure can therefore only be inferred.

Institutional category systems provide well-documented examples of hierarchical category system evolution. When looking at how these systems change, it is easy to find examples where category systems become less optimal with time. For example, cancer is still today grouped with tuberculosis and leprosy in the Dewey Decimal System, reflecting early associations between three diseases that were poorly understood at the time, even though they are now understood as being quite distinct. This suggests that a category system that was warranted in the past may become less so as the environment and our understanding of it changes, and as new things have to be accommodated into an existing system. Because these systems are built incrementally, their current structures reflect past decisions. Thus, to understand the structure of hierarchical category systems and how they can become biased, this thesis will also study how they evolve, with a focus on the conditions that cause them to become suboptimal or skewed representations of the things they categorize.

1.4 Core Research Questions

In this thesis, I study the structure of hierarchical category systems as a whole and how they evolve to accommodate new things in the environment. More specifically, I define three key research questions.

RQ1. Are the biases in institutional category systems consistent with psychological principles of categorization?

RQ2. How does the structure of a hierarchical category system encode bias, and what kinds of biases does it encode?

RQ3. How does a hierarchical category system evolve in response to sequentially presented stimuli?

Altogether, by exploring how bias can manifest in the structure of a hierarchical category system and what factors influence how they evolve, I aim to understand why category systems are structured the way that they are, and, in particular, how they can become skewed representations of the world, not necessarily providing an accurate picture of the items they are meant to represent.

1.5 Main Contributions

This thesis investigates the structure and evolution of hierarchical category systems and sits at the intersection between computer science and cognitive psychology. From a cognitive perspective, I use analyses of institutional category systems and traditional experimentation as windows into how humans internally represent the world and use them to study how they develop and how they can be skewed or biased. From a computational perspective, I treat category systems as structured objects that can be represented, analyzed, and modelled. Institutional category systems are translated into computationally accessible formats, requiring decisions about how to represent hierarchical relationships and what information to preserve or compress. This enables large-scale analyses that would not be feasible by hand. I make the following four contributions.

First, I demonstrate the utility of using institutional category systems as datasets to test psychological principles of categorization at scale. I draw from computational linguistic approaches to demonstrate how principles of typicality manifest in real-world category systems and examine the unequal labelling of subcategories for men and women in the Library of Congress Subject Headings, Wikipedia, and written language. I explore the extent to which unequal labels are the result of unequal percentages of men and women in various groups versus an implicit people=men bias. Moreover, across all the systems I examine, I demonstrate the pervasiveness of gender bias in the structure and labels of hierarchical category systems and show how it has persisted across time.

Second, I introduce a novel framework for conceptualizing and measuring bias in the structure of a hierarchical category system, and show how it can be applied to both institutional and cultural category systems. The framework distinguishes between bias encoded in the arrangement of categories (internal nodes) within a hierarchy and bias in the distribution of items at its leaves, allowing for a more detailed understanding of how bias can arise in different parts of a system. Using this framework, I quantitatively demonstrate unequal representation of non-Western topics and women authors in both the Library of Congress and Dewey Decimal classification systems. While bias in library classification systems has been discussed qualitatively, this framework enables measurable comparisons across classification systems and clearer identification of where bias is most pronounced.

Third, I introduce a new experimental paradigm that tests the effects of stimulus presentation order on the development of hierarchical category systems. Unlike existing work,

which typically has participants learn a category system from scratch, I ask participants to contribute to an existing category system while also making its hierarchical structure explicit during the task. This allows me to study sequence effects in explicitly hierarchical contexts. With this paradigm, I replicate existing findings on sequence effects in human categorization and extend them to show how they influence the evolution of hierarchical category systems over time. I also connect short-term sequence effects in categorization to longer-term changes in category system structure, as both reflect incremental and incomplete access to the environment during system construction.

Finally, I developed a computational model of categorization that simulates how a hierarchical category system evolves in response to a sequentially encountered environment. The model connects ideas of category optimality from cognitive science with existing computational clustering objectives, showing how a system that is locally optimizing the fit of its categories to observed items can still produce globally suboptimal structures as new items are encountered. Decisions that are optimal at the time become suboptimal as more of the environment is revealed or discovered. The model better accounts for human behaviour in the categorization task than existing flat categorization models, providing an alternative hierarchical account of categorization in settings where the structure is explicitly hierarchical. The model also takes advantage of dynamic programming to compute the computationally intensive clustering objective in a reasonable amount of time, thus making the model feasible to run on larger-scale datasets.

Overall, this work reinforces the need to consider hierarchy in analyses and models of categorization and demonstrates how category systems create skewed representations of the world through their encoding of bias and their incremental formation. This has implications for how both humans and machines represent the world.

1.6 Thesis Overview

In this thesis, I introduce a body of research investigating the structure and evolution of hierarchical category systems across cultural, institutional, and experimental contexts. What is addressed by this work and what gaps remain are outlined in Table 1.1. The rest of the thesis is structured as follows.

In Chapter 2, I review existing literature on categorization and category systems across psychology, library and information science, and computer science. In particular, I examine how each one addresses ideas of bias, optimality, and category evolution and highlight shared ideas about what makes a good category system across the three fields. I use library and information science and computer science to fill in some of the gaps in categorization research that psychology does not address.

In Chapter 3, I study bias in the existence (or non-existence) of explicitly labelled subcategories within a category system. I use this to explore how principles of typicality

Table 1.1: Space of current and future research on hierarchical category systems. Category systems are explored along two dimensions: the type of system being studied (cultural, institutional, synthetic) and whether it is the structure or evolution of the system that is being studied. Dots represent areas of research not addressed by this thesis.

	Cultural	**Institutional**	**Synthetic**
Structure	Chapter 4	Chapter 3 & 4	•
Evolution	•	Chapter 3	Chapter 5

manifest in real-world category systems. More specifically, I build on the idea that typical category members often go unmarked (we say "flightless birds" but not "flighted birds") and that typicality can reflect stereotypes, and use this to measure asymmetries in the existence of gendered subcategories (e.g. having a category for "female scientists" but not "male scientists") across two institutional category systems: the Library of Congress Subject Headings and Wikipedia's category system. I then determine the extent to which this asymmetry corresponds to the perceived or actual frequency of men in parent categories and begin to explore the evolution of institutional category systems by exploring how these asymmetries have changed with time.

In Chapter 4, I extend my analyses of bias to the structure of institutional category systems by introducing a framework for measuring bias in the structure of hierarchical category systems and applying it to two library classification systems: the Dewey Decimal System and the Library of Congress Classification. Here, I represent the two library systems as trees, with categories as internal nodes and books as the items at the leaves. I explicitly define bias as a preference for one group over another, such that there are asymmetries in how the groups are represented by a category system. I then distinguish between category bias — or bias in the number and arrangement of internal nodes — and item bias, which reflects bias in the number and distribution of items among the leaves of a category system. I demonstrate the utility of this framework by performing a large-scale analysis of Western category biases and item gender biases, measuring the extent to which they occur in each library classification system. Finally, to demonstrate that these methods can be extended to the structure of cultural category systems, I briefly discuss English WordNet and apply the bias framework to mammal concepts within it, demonstrating a bias toward domestic over wild mammals.

In Chapter 5, I move away from the structure of hierarchical category systems and instead investigate their evolution using a novel experiment and computational model. Motivated by observations of order effects in library classification systems over time, I introduce a computational model that simulates and predicts these effects in the evolution of a hierarchical category system by sequentially placing each item in the category location that maximizes within-category similarity and between-category dissimilarity. I then de-

sign an experiment that elicits these effects in participants by having them incrementally incorporate novel stimuli into an explicitly hierarchical category system. The model better accounts for participant behaviour than the existing model, and together they demonstrate that the order in which items are presented can influence the resulting category structures, sometimes leading to less optimal divisions of the item space.

Finally, in Chapter 6, I summarize the findings of this thesis and reiterate how they contribute to our understanding of the structure and evolution of human category systems across cultural, institutional, and experimental contexts. I then discuss future work that can both reinforce our understanding of the filled-in entries of Table 1.1 and be used to fill in the remaining gaps. I discuss the significance of this work to psychology, computer science, and library and information science.

1.7 Summary of Published Work

Portions of this thesis have been published separately and are listed below.

- **Chapter 3**: The analysis of labels and gender bias was published as Warburton, K., Kemp, C. & Frermann, L. (2025). Scientists & Women Scientists: Exploring Gender Biases in Institutional Category Systems. *Proceedings of the Annual Meeting of the Cognitive Science Society, 47*.
- **Chapter 4**: The bias framework and analyses are published as Warburton, K., Kemp, C., Xu, Y. & Frermann, L. (2024). Quantifying Bias in Hierarchical Category Systems. *Open Mind: Discoveries in Cognitive Science, 8*, 102–130. A preliminary version was published as Warburton, K., Kemp, C., Xu, Y. & Frermann, L. (2023). Quantifying Bias in Library Classification Systems. *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*.
- **Chapter 5**: The experiment was published as Warburton, K., Frermann, L., Xu, Y. & Kemp, C. (2026). An Experimental Study of the Evolution of Hierarchical Category Systems. *Proceedings of the Annual Meeting of the Cognitive Science Society, 48*. An extended manuscript including the computational model is currently in preparation.

Chapter 2

Background

Categorization is a fundamental part of human cognition that has been extensively studied by psychologists and cognitive scientists. These disciplines have focused on cultural category systems, which reflect the categories people carry around in their heads and often express through language (e.g., folk taxonomies). These systems are often hierarchical, dividing the world into categories at increasing levels of abstraction. Although psychologists and cognitive scientists have primarily studied cultural category systems, hierarchical category systems also appear as organizing structures across a range of other disciplines. The two most relevant to the present work are library and information science and computer science. Library and information science is concerned with institutional category systems such as the Dewey Decimal Classification and thus provides insight into hierarchical category systems constructed deliberately by humans and made visible through explicit documentation. Subfields of computer science, such as machine learning and natural language processing, use hierarchical structures to organize and represent patterns in data and have developed formal tools for evaluating the quality of category hierarchies.

Across psychology, library and information science, and computer science, similar themes emerge about the role of hierarchy in categorization, what makes a category system optimal, how such systems can encode and perpetuate bias, and how they evolve over time. Table 2.1 summarizes the space of literature reviewed in this chapter and lists key sources, organized by discipline and theme. Although I refer to the first column of the literature as psychology for conciseness, this category also includes work from cognitive anthropology and linguistics. These fields similarly examine the structure and evolution of cultural category systems. The remainder of this chapter discusses each theme in turn. I begin by reviewing work on the structure of hierarchical category systems (Section 2.1), highlighting how hierarchy (2.1.1), optimality (2.1.2), and bias (2.1.3) are conceptualized across disciplines. I then review work related to the evolution of category systems (Section 2.2), addressing cultural systems studied by psychology (2.2.1), institutional systems studied by library and information science (2.2.2), and computational models from computer science (2.2.3).

Table 2.1: Summary of research on category systems across disciplines: Psychology, library and information sciences, and computer science. Columns are divided by discipline and rows labels correspond to themes and sub themes of the literature.

	Psychology	**Library & Information Science**	**Computer Science**
Structure			
Hierarchy	Kay (1971) C. H. Brown (1976) Rosch et al. (1976) Markman (1981) Tanaka and Taylor (1991) Berlin (1992) Atran et al. (1997) Medin et al. (1997)	Fernandez and Eastman (1990) Jacob (2004) Green (2006) Hajibayova (2013) Glushko et al. (2016) Lund and Agbaji (2018b)	Rosenberg (1979) Heller and Ghahramani (2005) Blundell et al. (2011) Murtagh and Contreras (2012, 2017)
Optimality	Collins and Quillian (1969) Rosch (1978) Parducci (1983) Chi et al. (1989) Corter and Gluck (1992) Regier et al. (2007) Kemp and Tenenbaum (2008) Kemp et al. (2018) Carr et al. (2020)	Cutter (1876) Lancaster (1970) Rodriguez (1984) Losee (1992, 1993) Kwasnik (2000) Svenonius (2000) Green and Panzer (2014) Martínez-Ávila and Budd (2017)	Huffman (1952) Quinlan (1986) Li et al. (2015) Cohen-Addad et al. (2019) Naumov et al. (2022) Blockeel et al. (2023)
Bias	Park and Rothbart (1982) Judd and Park (1988) Mackie and Worth (1989) Greenwald et al. (1998) Goldstone et al. (2001) Rubin and Badea (2012) Timeo et al. (2017) Waxman (2021) Schimmack (2021) Misch et al. (2022)	Berman (1971) Hupp (1991) Intner and Futas (1996) Olson and Ward (1997) Bowker and Star (1999) Olson (2002) Olson (2004) Knowlton (2005) B. Quinn (2012) Loehrlein (2012) Mullender (2025)	Friedman and Nissenbaum (1996) Kirkpatrick (2016) Caliskan et al. (2017) Shankar et al. (2017) Garg et al. (2018) De Vries et al. (2019) Suresh and Guttag (2021) Bailey et al. (2022) Bai et al. (2025)
Evolution	Carey (1985) Mervis (1987) Keil (1989) Waxman (1990) Medin and Bettger (1994) Kirby et al. (2014) Carstensen et al. (2015) Carvalho and Goldstone (2015) Mack and Palmeri (2015) Stöttinger et al. (2016) Hawkins et al. (2018) Silvey et al. (2019)	Crovisier and Intner (1987) Suchecki et al. (2012)	Palmeri (1999) Ribert et al. (1999) Sahoo et al. (2006) Verheyen et al. (2008) Canini and Griffiths (2011) Gepperth and Hammer (2016) Hawkins et al. (2018)

2.1 The Structure of Category Systems

Category systems are studied across many disciplines. In this chapter, I review literature on category systems from psychology, library and information science, and computer science. I highlight themes that emerge across these disciplines, focusing on discussions of hierarchy, optimality, bias, and the ways these concepts intersect.

2.1.1 Hierarchy in Category Systems

Psychology

There is an extensive body of literature examining the structure of human category systems in psychology. One line of work looks at the structure of individual categories and whether they are defined by necessary and sufficient features (Bruner et al., 1986), prototypes (Osherson & Smith, 1981; Rosch, 1975, 1978), stored exemplars (Medin & Schaffer, 1978; Nosofsky, 1986), or causal–explanatory theories about the domain being categorized (Carey, 1985; Keil, 1989; Murphy & Medin, 1985). This thesis, however, is less concerned with the structure of individual categories and more with the structure of category systems as a whole, focusing on how categories are organized hierarchically in relation to one another.

From animals, clothes, tools (Rosch et al., 1976), and body parts (Majid, 2010), to places (Basso, 1984), emotions (Hosoya et al., 2017), and beyond, researchers have documented hierarchical organization in human category systems. Within these systems, two main kinds of relations are typically distinguished: taxonomic and partonomic relationships (Markman, 1981; Tversky, 1989; Tversky & Hemenway, 1984). Taxonomies involve "kind-of" relations, where a child category is a kind of the parent (Kay, 1971); for example, a "car" is a kind of "vehicle". Partonomies, by contrast, involve "part-of" relations, where the child is a component of the parent, such as a "button" being part of a "shirt" (C. H. Brown, 1976; McClure, 1975). In practice, the same items can be organized in both ways. For example, a "thumb" is a kind of "body part" but also a part of the "hand", and buttercream can be part of a "cake" while also being a kind of "frosting". Whether organized taxonomically or partonomically, hierarchical category systems are widespread and have been a major focus of research in psychology and anthropology.

One major focus of psychological research on hierarchical category systems concerns levels of abstraction within these hierarchies. Psychologists often distinguish between three main levels: superordinate, basic, and subordinate categories (R. Brown, 1958; Rosch et al., 1976). The basic level occupies a privileged position because it is generally the most cognitively efficient level of categorization at which items are categorized most quickly and recalled most easily (Mervis & Rosch, 1981; Murphy & Smith, 1982; Rosch et al., 1976). The superordinate level consists of more general categories that encompass basic-level ones, while the subordinate level consists of more specific categories that subdivide them. For

example, if "cat" is the basic level, then "animal" is superordinate and "Norwegian Forest Cat" is subordinate. However, the basic level is not fixed. It can vary with expertise, such that experts' basic-level categories are often more fine-grained (Johnson & Mervis, 1997; Tanaka & Taylor, 1991), and it can also vary with context, depending on a person's goals and environment (Barsalou, 1991). For example, categories like "Norwegian Forest Cat" may function as the basic level for veterinarians. Although the specific categories that function as the basic level vary across individuals and contexts, the principles underlying this level of categorization are often taken to reflect general cognitive constraints (Rosch et al., 1976).

Alongside psychological research on hierarchical category systems, cognitive anthropologists have examined the organization of folk category systems across cultures and domains of expertise. Folk category systems are the everyday naming systems through which people organize the world. A substantial body of work has examined the structure of folk taxonomies and partonomies and how they vary cross-culturally (Basso, 1984; Berlin et al., 1966; Boster & D'Andrade, 1989; C. H. Brown, 1976; Ellen, 1977; Holman, 2005; McClure, 1975). This literature overlaps with psychological research on category hierarchies as both identify a privileged intermediate level in hierarchical category systems. In cognitive anthropology, this is referred to as the folk-generic level, which corresponds to the level of categorization most commonly used in everyday reference and naming (Atran et al., 1997; Berlin, 1992). Like the basic level, its position is not fixed and can shift with culture, experience, and expertise (Boster, 1985; Boster & Johnson, 1989; Medin et al., 2006). Evidence suggests that the overall structure of a hierarchy and the criteria by which items are sorted into categories can also vary under these conditions (Medin et al., 1997).

Although hierarchical category systems are prevalent, there is debate about whether such structures are explicitly represented in the mind. While some symbolic models, such as Collins and Quillian (1969), treat hierarchy as an explicit structure stored in memory, others instead argue that hierarchical relations are computed from set comparisons between stored concept representations (Rips et al., 1973). More modern distributed accounts propose that hierarchical effects emerge from patterns of activation rather than stored symbolic relations (T. T. Rogers & McClelland, 2004). From a developmental perspective, Rakison (2000) argues that there is insufficient evidence to conclude that infants form taxonomies or possess adult-like basic-level categories, questioning their existence as an innate structure in the mind. More generally, some work has questioned the role of hierarchical structure in memory and cognition, since feature inheritance relationships between parent and child categories are not consistently used in conceptual reasoning tasks (Hampton, 1982; Sloman, 1998). Still, these critiques do not deny that people can and do organize concepts hierarchically. Whether or not hierarchical structure is explicitly represented in the mind, it remains pervasive in how people organize and talk about the world.

Despite the breadth of work on hierarchical category systems, much of the literature focuses on basic or folk-generic levels of categorization rather than the structure of the system as a whole. In addition, cultural category systems traditionally studied in psychology are often difficult to access because they exist primarily within the minds of their users. As a result, they are typically investigated through tasks such as sorting pictures or names into categories, listing features, making similarity judgments, or learning artificial category structures. However, these methods can remove categorization from its natural context. To address this limitation, I turn to research in library and information science, which analyzes real-world hierarchical classification systems as they "naturally occur" (Goldstone & Lupyan, 2016).

Library & Information Science

Hierarchical category systems are not limited to everyday cultural categorizations. Humans also construct hierarchical category systems to organize knowledge and resources in more formalized, institutional contexts (Glushko et al., 2008). We use library classification systems to organize books, scientific taxonomies to illustrate the genetic and evolutionary relationships between species, and the International Classification of Diseases to align medical diagnoses and terminology across the world. Institutional category systems can categorize broadly, like Wikipedia, whose categories are meant to organize articles that cover the space of human knowledge, or they can cover relatively niche domains like the Aarne–Thompson–Uther Index, which organizes folktales. They serve a variety of important but often invisible functions, governing how we view and access information (Bowker & Star, 1999)

Like cultural systems, institutional systems vary across social and environmental contexts such that multiple systems can exist for the same set of resources. The use of one system over another often signals membership within a particular group, discipline, or belief system (Bowker & Star, 1999). For example, within the US, both the Library of Congress Classification and the Dewey Decimal System are used to organize books in libraries, with usage often divided between academic and public institutions (Lund & Agbaji, 2018a, 2018b). Similarly, there are at least four ornithological taxonomies, signalling different beliefs about how to divide birds into species (Tomer, 2019). Ideas related to the basic level have also been explored in institutional category systems such as library classifications, thesauri, semantic networks, and databases, where the basic level is sometimes understood as the level most frequently accessed or indexed and as relatively stable across different information systems (Fernandez & Eastman, 1990; Green, 2006; Hajibayova, 2013).

Institutional systems do not have the same constraints as cultural systems, as they do not have to fit in any one person's brain. Some systems, like library classification systems, have constraints such that the system must facilitate organization within a physical space,

while other systems are entirely digital. Institutional category systems are generally believed to be grounded in cultural categorization, but they also influence cultural systems through their organization of knowledge (Glushko et al., 2008). At the same time, they differ from everyday cultural systems in important ways: they are typically larger, more rigid, and almost always explicitly documented (Jacob, 2004).

Institutional category systems, despite their central role in organizing knowledge in real-world contexts, have received comparatively little attention within psychology and cognitive science. This raises broader questions about the principles that shape hierarchical category systems and about how institutional systems relate to patterns observed in human categorization more generally. I address this in Chapter 3, examining how principles of categorization play out in two institutional category systems. These similarities and differences with cultural systems make them an important source of insight into the structure of human categorization.

Computer Science

Hierarchical category structures extend beyond human category systems and are also used in computer science. Here, they can be used to represent categorical relationships between data elements, including both human-defined categories (e.g., WordNet, which encodes relationships among lexical concepts) and categories generated through computational methods (e.g., hierarchical clustering of text embeddings). Although some computational hierarchies resemble human category systems, others capture relationships that emerge from patterns in the data and may not correspond to human-like groupings.

A tree is an abstract data type made up of nodes. These nodes are connected such that each node can have any number of child nodes and every child can have exactly one parent, except for the node at the top that has no parents (Rosenberg, 1979). This is often referred to as the root node. The nodes that lie at the bottom of the tree with no children are called leaves, and all other nodes are internal nodes. Trees are often binary such that each parent has at most two children, but when trees contain nodes that have more than two children, these trees are referred to as non-binary. In this thesis, I represent category systems as trees where internal nodes are categories and items are the leaves. Under this framework, depth refers to the number of descendants between a given node and the root, and width to the number of direct children a node has. Tree data structures are used to store information in databases (e.g., Comer, 1979), and are the outcome of algorithms aiming to learn underlying patterns in unstructured data.

Hierarchical clustering algorithms are used in machine learning to find underlying patterns in data through the arrangement of items into trees (Murtagh & Contreras, 2012, 2017; Sokal, 1963; Yim & Ramdeen, 2015). The methods can be agglomerative, such that items are successively merged into clusters until only one remains at the top, or, less com-

monly, divisive, where one large cluster is successively split until individual items remain at the leaves (Roux, 2018). Because these algorithms ultimately result in the creation of trees, they can also be binary or non-binary, although non-binary methods are much more computationally intensive and therefore less common (Blundell et al., 2011; Heller & Ghahramani, 2005). These computationally created categories are then compared to existing, human-generated groupings or are used to discover latent relationships in the data.

Hierarchical category systems can be difficult to analyze. Computer science, with its reliance on tree structures in many algorithms and databases, has useful tools for representing and simulating these systems so that they can be analyzed at scale.

2.1.2 Hierarchy & Optimal Category Systems

Simon (1962) argues that hierarchy is a general solution to the problem of complexity. Any system that is made up of a large number of parts, whether it is a body, a university, a language, or a mind, is best organized through a hierarchical decomposition into interrelated subsystems, where each subsystem is recursively decomposed into smaller subsystems until the most basic components are left. In this view, hierarchy is not merely a feature of category systems but a general structural principle of the world, arising from its efficiency as an organizing scheme. In a similar vein, hierarchical organization appears across cognitive science, library and information science, and computer science as a shared solution to the problem of efficiently organizing, retrieving, and generalizing over large structured information spaces. In this section, I review work from the three domains that address the function of hierarchy. The common themes that emerge are described in Table 2.2. Because multiple hierarchical organizations can exist for the same set of items, I also review measures of optimality that explain why certain systems are better than others.

Table 2.2: Functions of hierarchy in category systems. These ideas are not mutually exclusive and can be applied across psychology, computer science, and library and information science.

Function	Definition
Storage	The system is hierarchical because it enables efficient compression of items into a format that retains maximal information while being a manageable size.
Retrieval	The system is hierarchical because it allows for efficient search across connected levels of abstraction.
Generalization	The system is hierarchical because it enables inference about novel items and categories by projecting properties across levels of abstraction.
Structural Fidelity	The system is hierarchical because it reflects natural patterns or structures in the set of things being categorized.

Psychology

According to Rosch (1978), "the task of a category system is to provide maximum information with the least cognitive effort" (p. 92). Related to this idea, a substantial body of work has examined how languages partition a space of meanings into categories, demonstrating that human category systems exist at some optimal trade-off between being simple enough that they are learnable, yet informative enough to support effective communication about the environment (Carr et al., 2020; Kanwal et al., 2017; Kemp & Regier, 2012; Kemp et al., 2018; Levinson & Meira, 2003; Regier et al., 2007; Zaslavsky et al., 2018, 2019).

In the context of language, simplicity is often operationalized as the average code length of the labels assigned to categories (i.e. how many characters they have), but it can also be defined in terms of the number of categories in a system, where simpler systems have fewer categories and more complex ones have more. Informativeness is defined in terms of how much information is gained by knowing an object belongs to one category over another (Carr et al., 2020; Parducci, 1983). It is often measured in terms of how likely an object is to be referred to (need) and the amount of additional information that is needed to distinguish it from other within-category members (category size) (Carr et al., 2020; Kemp & Regier, 2012). But it can more broadly be defined in terms of the number of shared features (similarity) between items in a category, as the more features items share, the more predictable those features are given they belong to a certain category (Corter & Gluck, 1992). Category systems with lots of smaller categories tend to be more informative but also more complex, whereas systems with few large categories are simple but not as informative.

A related principle to both simplicity and informativeness is **compactness**. This is defined as "the extent to which similar meanings belong to the same semantic categories and dissimilar meanings belong to different semantic categories" (Carr et al., 2020, p. 4). Although not using this terminology, Mervis and Rosch (1981) similarly proposed that categorization is most cognitively efficient at the level where within-category similarity is maximized, and between-category similarity is minimized and Carr et al. (2020) further states that increasing compactness improves both simplicity and informativeness, suggesting that optimal category systems simultaneously reduce within-category variation and increase between-category distinctiveness.

Simplicity, informativeness, and compactness are useful ways to understand and measure category system optimality; however, most applications of these principles focus on flat divisions of the space or on picking out individual levels of abstraction. As mentioned in Section 2.1.1, Rosch et al. (1976) identifies the basic level as the most cognitively efficient, and accordingly, much of the analysis of category system optimality in the context of hierarchy has focused on comparing the optimality of different levels of abstraction, or using measures of optimality to identify which level is basic (Corter & Gluck, 1992),

rather than assessing the optimality of the system as a whole. Levinson and Meira (2003) suggest that cross-linguistic variation in spatial category systems reflects successive partitioning of a shared conceptual space, with broader spatial categories gradually splitting into more specific distinctions, and Berlin and Kay (1991) propose similar trajectories for colour terminology. However, these accounts again focus primarily on how systems differ in the granularity of their partitions, rather than on systems that simultaneously contain multiple levels of abstraction. They do not specifically address why systems are structured hierarchically. More abstract superordinate categories in a system might favour simplicity, while more specific ones favour informativeness, but why both exist is unclear.

While simplicity and informativeness explain why categories partition meaning in certain ways, they do not by themselves explain why category systems are hierarchically structured. However, many of the proposed functions of hierarchy can be understood as extensions of these same principles. Tackling hierarchy more directly, Collins and Quillian (1969) suggests that humans organize concepts hierarchically because this supports efficient storage of information in memory. In this view, knowledge is structured through class inheritance, where information stored at higher levels need not be redundantly encoded at lower levels. To make this more concrete, if a person encounters a novel item that they believe to be a cat, then they do not need to store that it has whiskers, purrs, and hunts mice; they just need to store it as an instance of the cat category. Hierarchy functions as an optimal compression of the item space, preserving maximal information while making the system a reasonable size for limited-capacity brains.

Hierarchy has also been argued to support efficient retrieval of information from memory (Collins & Loftus, 1975). In hierarchical semantic networks, concepts are linked through relations of class membership and property inheritance, allowing retrieval processes to operate by traversing structured paths between concepts rather than searching memory exhaustively. Closely related concepts are therefore easier to access because they share more of the same representational structure. Under this view, hierarchy improves the efficiency with which information can both be stored and accessed. Because compact categories group together items with many shared features, category membership narrows the space of likely properties and associations, making relevant information easier to access. Under this view, the same structural properties that increase simplicity and informativeness also improve the efficiency with which information can be retrieved.

In psychology, hierarchy is additionally understood as enabling efficient generalization (Chi et al., 1989; Murphy, 2002). Learning a property of a broader category allows that property to be extended to more specific subcategories unless contradicted by experience. Knowing that something is a cat, for instance, allows a person to infer many likely properties before directly observing them. More generally, hierarchical systems support inductive reasoning because knowledge acquired at one level of abstraction can be projected to related categories at other levels. Hierarchy therefore enables category systems not only

to represent information efficiently, but also to support prediction about novel instances. From this perspective, generalization can be understood as a consequence of informativeness. The more information category membership provides about an item's properties, the more reliably knowledge can be extended to novel instances. Compactness increases informativeness not only for communication but also for prediction and inference about novel instances.

A final account suggests that hierarchical structure reflects stable patterns in the external environment. Rosch (1978) states that the basic level of categorization corresponds to natural discontinuities in the perceived world and best captures its correlational structure, suggesting that category systems are constrained by how the world is organized at least at one level of abstraction. Anderson (1991) similarly concludes that the mind has the structure it does because the world does. More directly, hierarchical organization has been modelled as inference over latent structure in the environment (Katz et al., 2008; Kemp & Tenenbaum, 2008). In a domain-specific account of biological classification, Atran (1998) similarly argues that the hierarchical structure of folk taxonomies reflects evolved cognitive constraints that reflect ecologically stable patterns of relationships between living things. Thus, hierarchy is prevalent in the systems we create because it is a prevalent structure in nature. Under this view, an optimal system closely aligns its structure with the structure of the natural world. The function of hierarchy is to represent real patterns of increasing similarity and dissimilarity between items in our environments.

Hierarchy can serve many functions related to the optimality of a category system and its effectiveness as a system of representation. Although many measures of optimality have been proposed, what remains less clear is how to measure the optimality of hierarchical category systems as whole systems. Computer Science has answers to this that will be discussed later, and in Chapter 5, I attempt to address this by adapting a clustering objective that scores the optimality of a whole system. However, before turning to computer science, I next discuss how similar ideas about the function of hierarchy and the optimality of systems emerge in the library and information sciences.

Library & Information Science

In library and information science, hierarchical structure is an explicit design choice meant to facilitate the organization of resources and knowledge, and access to them (Bowker & Star, 1999; Glushko et al., 2016). Such hierarchical arrangements are not arbitrary but reflect underlying design principles that determine how categories are formed, related, and labelled. Library classification systems are typically justified through warrants, which specify what categories should be made and how they should be named (Glushko et al., 2016; Svenonius, 2000).

The three most common warrants are literary warrant, use warrant, and structural warrant. Literary warrant is the idea that the classification system must be justified by the things that it is categorizing (Rodriguez, 1984; Svenonius, 2000). For example, if suddenly there are lots of new books in a newly emerging environmental science field and they tend to fall into two groups, one being more related to geography and the other to biology, then it would be justified to split this category into two subgroups. Use warrant is the idea that the system should be aligned with the language, categorization, and search behaviour of its users (Cutter, 1876; Lancaster, 1970; Svenonius, 2000). Finally, structural warrant is the idea that the labels and categories of a system can be introduced to serve particular structural functions (Svenonius, 2000). Examples of this include the DDC's 10-class requirement, such that every category must be divided into no more than 10 subgroups and the introduction of "general" and "miscellaneous" categories into systems to ensure every piece of literature has a place even if it's not explicitly covered. In practice, a category system negotiates trade-offs among competing warrants. Overemphasizing bottom-up use and literary warrant can weaken structural coherence, while an overly rigid top-down structural warrant can reduce usability (Svenonius, 2000).

Through discussions of these warrants and different approaches to measuring system quality, the same functions of hierarchy summarized in Table 2.2 emerge. Losee (1993) emphasizes use warrant, arguing that a successful system minimizes the time patrons spend browsing. To evaluate system quality, the hierarchical structure is flattened into a linear arrangement of books along a hypothetical shelf, and optimality is assessed in terms of book similarity (Losee, 1992, 1993). In this framework, an optimal ordering places similar books closer together and dissimilar books farther apart, such that proximity in the sequence corresponds to degrees of similarity (Losee, 1992, 1993). The hierarchical structure is thus indirectly optimized for maximal within-category similarity and minimal between-category similarity (compactness), since items that are near each other in the ordering belong to more closely related categories than those that are distant. This organization supports retrieval by grouping similar books while separating dissimilar ones, and the specific role of hierarchy is to capture graded levels of similarity. For example, closely related subfields of chemistry remain nearer to each other than to distant domains such as religion or sports because of how the hierarchical organization is flattened across a shelf. In this sense, the role of hierarchy is to facilitate retrieval (browsing), as it ensures that any location on a shelf will contain related literature, and thus finding one book enables discovery of related books. Hierarchy also enables efficient access or retrieval for people with different levels of knowledge, with more general categories for lay people and more specific ones for experts.

Ideas that hierarchy provides efficient storage of information emerge in discussions of literary and structural warrant. Green and Panzer (2014) use the distribution of books across topics to justify adjusting the granularity of classification, arguing that when a category grows due to increasing numbers of assigned works, it should be subdivided into finer

categories, whereas sparsely populated categories may be collapsed into one or removed as subcategories entirely. In this way, literary warrant functions as a mechanism of compression within a hierarchical system, creating a pressure to distribute literature such that the structure remains efficient and categories do not become either overly dense or overly sparse. A related form of efficiency appears under structural warrant, where the design of the classification notation itself encodes a trade-off between informativeness and simplicity: more descriptive call numbers increase expressive specificity by embedding more detailed information in the notation, while shorter notations prioritize efficiency and ease of use by compressing that information into more compact identifiers (Svenonius, 2000). The idea that the hierarchy allows for compressed representations through class inheritance also emerges (Kwasnik, 2000). Across both cases, hierarchy functions as a mechanism for compressing information while preserving as much meaningful structure as possible, thereby facilitating the efficient storage of knowledge within the classification system.

Literary warrant can additionally be understood as the idea that the role of hierarchy is to reflect the structure of the things it is classifying. If the "structure of the world" is taken to mean the structure inherent in the things being classified, then literary warrant attempts to capture and reflect that underlying structure through the organization of the classification system. In this sense, hierarchy in library classification can be understood as a mechanism of structural fidelity when a classification system aims to preserve or reflect the structure of the domains it organizes. An additional form of warrant relevant here is scientific or philosophical warrant (Bliss, 1933; Glushko et al., 2016; Martínez-Ávila & Budd, 2017), which concerns the extent to which the language, categories, and structure of a classification system correspond to those used by scientists or domain experts. Zins and Santos (2011), for example, apply the principle of scientific warrant in evaluating existing library classifications by examining how adequately different domains of knowledge are represented within them. From this perspective, hierarchy functions to maintain structural fidelity because the structure of the classification system is intended to mirror the structure of the disciplines it represents, or even the structure of the natural world itself.

Generalization provides a final potential justification for hierarchy in library classification and knowledge systems. While this function is less commonly discussed in library and information science, the idea that hierarchical classification systems support inference across levels of abstraction nevertheless emerges. Kwasnik (2000), in a discussion of the appeal of hierarchy, notes that because of property inheritance, "hierarchy allows reasoning from incomplete evidence" (pp. 5). The placement of a subject within a hierarchy allows users to situate unfamiliar items within a broader conceptual context and to project expectations from more general categories onto more specific ones.

Like in psychology, hierarchy can be understood as serving many roles, and the optimality of a system is also understood in terms of the similarity of items within and across categories, and the simplicity and informativeness of the system. However, much like in

psychology, measures of optimality often focus on flat representations of the hierarchical system (like with Losee's work). Additionally, there are competing pressures that determine the extent to which the systems match the structure of the items being classified, the world, or how people structure the domain, and this thesis attempts to formalize some of the comparisons and analyze them at scale.

Computer Science

In computer science, hierarchical data structures are primarily used to enable efficient information retrieval and compact data representation. For example, Huffman encodings generate binary trees in which more frequent symbols are placed closer to the root and less frequent symbols deeper in the tree, allowing information to be compressed into shorter, lossless codes (Huffman, 1952). The quality or optimality of a system is thus defined in terms of code length (simplicity) while also trying to preserve the maximum amount of information, as in psychology. Similarly, balanced binary search trees organize ordered data hierarchically so that items can be searched in efficient $O(\log n)$ time (Sleator & Tarjan, 1985). B-trees extend this principle by trading off depth for width, enabling databases to remain efficient to search while also aligning well with the constraints of computer storage systems (Comer, 1979).

Machine learning, specifically, uses hierarchy for the purposes of generalization and structural fidelity. Decision trees are used in supervised learning to classify unseen data by hierarchically partitioning the feature space into increasingly specific regions based on decision rules (Blockeel et al., 2023). They enable generalization by learning patterns from labelled training data, allowing the category of novel items to be predicted (Quinlan, 1986). During inference, decision trees follow a sequence of feature-based decisions from the root to a leaf node to determine the final class label. Hierarchical clustering methods, by contrast, are a form of unsupervised learning used when the underlying category structure of the data is unknown. By constructing a tree-like structure (dendrogram), they aim to reflect the intrinsic organization of the data, thereby supporting structural fidelity and providing a hierarchical representation of similarity relations between items in a dataset. This representation can be partitioned at different levels, with each cut of the dendrogram corresponding to a different level of abstraction in the resulting categorization.

Decision trees are often framed in information-theoretic terms, where each split or decision is meant to maximize the amount of information gained (Blockeel et al., 2023; Quinlan, 1986), thus judging the quality of a hierarchy based on its informativeness as a data structure. In agglomerative clustering, hierarchical clustering algorithms typically use linkage criteria such as average linkage, which defines cluster similarity in terms of pairwise distances between clusters (Sokal & Michener, 1958; Yim & Ramdeen, 2015), to decide which clusters to merge. In contrast, objective-based hierarchical clustering methods construct

a hierarchy (often represented as a binary tree T over a set of n items) by optimizing a global objective defined over possible tree structures (Cohen-Addad et al., 2019; Naumov et al., 2022). These objectives provide a way of evaluating the quality of an entire hierarchy according to a predefined notion of optimality, rather than relying on local merge heuristics. Intuitively, this corresponds to a pressure for similar items to be grouped together at lower levels of the hierarchy, while more dissimilar items are separated higher in the tree. Although hierarchical clustering is traditionally used to construct binary tree structures, it can also be extended to non-binary settings (Blundell et al., 2011; Naumov et al., 2022) and used to evaluate or analyze the structure of existing human-defined category systems.

Across psychology, library and information science, and computer science, optimal category systems group similar things together and dissimilar things apart to support efficient storage, retrieval, and generalization, or to accurately represent the underlying structure of the domain. In hierarchical systems, deeper categories correspond to increasingly specific groupings of related items. Although these systems operate under different constraints across domains, similar structural principles emerge, and computational approaches in computer science provide a way to operationalize and evaluate notions of optimality at scale.

2.1.3 Biased Category Systems

Hierarchical category systems vary with needs, expertise, and environments such that we can effectively interact with the world around us. Because hierarchical category systems by nature involve compression of information, they can create skewed or distorted representations of the world, in which some distinctions are emphasized, and others are ignored or minimized (Bowker & Star, 1999). Biases can arise through these mismatched or unequal representations. Across psychology, computer science, and library and information science, two broad categories of bias emerge, which I will refer to as **internal bias** and **external bias**. Internal bias is bias imposed by the category system onto the items being classified, whereas external bias is bias already present in the set of things to be classified that may or may not be picked up and represented by the system. I will discuss how these ideas emerge across the three disciplines.

Psychology

Categorization allows humans to efficiently organize information about the world, but it also skews what we learn about the world and how we perceive it. In psychology, these skewed perceptions that might arise from categorization are often discussed in terms of different forms of bias. Some biases arise from the assumptions and constraints that make learning categories from limited data possible and thus might be the reason for skewed representations, while others emerge from the social and perceptual consequences of categorization itself.

In the psychology literature, inductive bias can be used to refer to the assumptions and constraints that allow humans to generalize from limited data while also shaping the kinds of representations that are learned (Griffiths et al., 2010). These inductive biases have been explored in the context of categorization to determine what kinds of category structures children and adults learn more readily and what assumptions they may hold (Griffiths et al., 2008; Keil, 1981; Markman, 1989; Shepard et al., 1961). For example, some researchers have suggested that humans possess an inductive bias toward hierarchical category representations. Children appear predisposed to interpret nouns as representing taxonomic relationships (Waxman, 1990; Waxman & Kosowski, 1990), and the proposal by Atran (1998) that humans possess an evolved tendency to organize living things taxonomically suggests a similar bias. Although inductive biases are central to categorization, this thesis focuses on a different use of the term bias. Here, I broadly define bias as a systematic skew or tendency in how a category system represents aspects of the world, such that certain groups, environments, or other features are "preferenced" or represented unequally by the system. In this sense, our focus is on biases encoded within category systems rather than on the inductive biases that guide categorization itself.

One domain in which biases in category systems have been extensively studied is social categorization. Psychologists have investigated how social categories such as racial and gender categories are formed, and how these categories in turn influence how people perceive and behave toward members of these categories (Misch et al., 2022; Timeo et al., 2017; Waxman, 2021). A canonical example is out-group bias, where individuals tend to hold more positive attitudes toward members of their own group and more negative attitudes toward out-group members (Brewer, 1999; Greenwald & Pettigrew, 2014). These biases can be explicit stereotypes that people are conscious of and act upon, or implicit attitudes that influence behaviours (Greenwald et al., 1998; Schimmack, 2021). One such example is the out-group homogeneity effect, which finds that people tend to perceive members of their out-groups as being more similar to one another than members of their in-group (Judd & Park, 1988; Mackie & Worth, 1989; Park & Rothbart, 1982; Rubin & Badea, 2012).

However, such category effects are not limited to social groups. Humans hold a more general categorization bias in which they tend to overestimate within-category similarity and between-category dissimilarity (Ashby & Zeithamova, 2022; Dubova & Goldstone, 2021; Ester et al., 2020; Goldstone, 1994; Goldstone & Hendrickson, 2010; Goldstone et al., 2001). For example, people discriminate more easily between colours that cross linguistic colour-category boundaries than between equally spaced colours within the same category (Roberson et al., 2005). People are much better at discriminating between sounds that cross categorically distinct phoneme boundaries in their own languages than in others (Kuhl, 1991). The mere act of assigning objects to categories can alter how they are perceived relative to one another. Additionally, the perception of objects also tends to be

influenced by category labels (Huttenlocher et al., 2000; Lupyan, 2012; Lupyan & Spivey, 2010).

The social biases, although likely influenced by the inductive constraints that guide categorization, can be seen as a kind of external bias. They are often the result of our environments and the people we interact with. For example, exposure to members of other races can reduce prejudice toward the out-group, suggesting in part that prejudices and biases can arise from unequal representation of different groups in one's environment (Zebrowitz et al., 2008). The perceptual biases can be thought of as internal biases imposed by the category itself. These categories skew perception and impose a biased structure onto the environment.

Both internal and external forms of bias discussed above focus primarily on how individual categories influence people's perceptions, expectations, and judgments of the things being categorized. They concern biases that arise from the properties and associations of individual categories, and from the process of assigning objects to categories. However, these forms of bias do not consider how biases can emerge from the structure of a category system as a whole, including how categories and their members are organized in relation to one another. Library and information science and computer science both consider more structural aspects of bias in category systems and will be discussed below.

Library & Information Science

Library scientists have been discussing biases in library classification systems since at least the 1970s. Berman (1971) highlighted issues with the language used in the Library of Congress Subject Headings (LCSH), a controlled vocabulary distinct from the Library of Congress Classification, which is used to assign topics to books. Berman found that the subject headings systematically favoured specific groups (e.g., White, Western, Christian men) through unequal labelling of various topics. For example, the subject heading "Intelligence Levels" included subheadings for Chinese, Jewish, and Black people but not groups like European or White people, implying that race or ethnicity affected intelligence and that certain groups deviated from the norm. Headings such as "Women as Accountants," with no corresponding heading for men, suggested one gender as the default, and similarly, "God" was only applied to books on the Christian deity, with marked terms like "God (Sikhism)" being used for books on other deities. Since then, researchers have continued to explore biases in the LCSH, finding that, although significant changes have been made, biases still remain (Adler et al., 2017; Angell & Price, 2012; Hobart, 2020; Howard & Knowlton, 2018; Knowlton, 2005; Mullender, 2025) towards women (Hyde, 2002; M. N. Rogers, 1993), immigrants (Lo, 2019), and Indigenous peoples (Kam, 2007), to name just a few.

Beyond the language used in subject headings, researchers have also examined biases in the coverage, placement, and distribution of topics relating to various social groups in

library classification systems (Higgins, 2016; Hilles, 2023; Intner & Futas, 1996; Olson, 2002; Olson & Schlegl, 1999; Westenberg, 2022). For example, in the DDC, topics related to women are often restricted to very specific or limited sets of categories (Olson & Ward, 1997), and categories for Christianity make up six of the ten direct subcategories of religion while the remaining religions of the world are mostly contained in one labelled "Other Religions" (Zins & Santos, 2011).

Some of these biases, called collection biases, arise externally from the books being classified (Dilevko & Grewal, 1997; England, 2023; Hupp, 1991; B. Quinn, 2012). Because of literary warrant, what is represented in a classification system reflects what exists in the literature, and thus, biases in the books are encoded in library classification systems. In contrast, researchers like Olson (2002) discuss biases that are internal to the system itself, emerging from classification decisions that go beyond what is present in the items and instead shape how knowledge is organized and made visible. In such cases, biases already present in the literature may be exaggerated by the system, and biases not present in the collection (although likely present in society) may be encoded by the system. For example, Dewey's ten-class requirement reflects a form of structural warrant but also imposes a rigid organizational constraint that forces all domains into a uniform structure, regardless of whether dividing something into 10 categories is meaningful for the literature being classified. This represents an internally imposed bias that is not a reflection of the underlying literature (Olson, 2002). More broadly, systems designed to be universal inevitably collapse individual differences, encoding bias through the assumption that certain perspectives or worldviews are general or normative (Olson, 2002). Although library systems are often perceived as objective (Bowker & Star, 1999; Olson, 2002), they reflect specific design choices, including which dimensions of similarity are considered relevant and which are ignored (Olson, 2001).

Hierarchy is another structural mechanism that encodes bias. By organizing classification systems into strict hierarchical structures, specific parent–child relationships are made explicit while others are excluded or left implicit (Olson, 2004). In terms of hierarchical structure, Loehrlein (2012) suggests that hierarchy introduces bias by reinforcing the assumption that items higher in the structure are more general and those lower down are more specific, regardless of whether this reflects the actual relationships between items. Lerner and Lomi (2018) find that articles assigned to categories lower in the Wikipedia category hierarchy have lower popularity and quality due to a lack of editing that is a consequence of their lack of visibility. In much of the literature on category labels, hierarchy is also implicit in discussions of bias. Subcategories (e.g., female doctors) are often understood as specialized or marked variants of a parent category (doctors), while unmarked categories (e.g., doctors) are typically read as default or more general, universal examples of the parent category (Berman, 1971; M. N. Rogers, 1993). While hierarchy can support efficient storage and retrieval, it also imposes an artificial structure that requires decisions about what to

include, exclude, or subordinate at each level. As such, hierarchical organization can embed bias beyond what is present in the items being classified.

Biases can therefore show up in the category labels a system uses, in the decisions about what to include and exclude, in the arrangement of categories themselves, and in the distribution of literature across the categories of the library classification system. In Chapter 3, I compare biases in institutional category labels to biases in the cultural categories contained in language. Although hierarchy is explored in some library and information science contexts, it is mostly discussed as a theoretical construct with reference to particular examples (Olson, 2002, 2004) or in experimental contexts demonstrating how hierarchy can induce biases in participants (Loehrlein, 2012). To complement this work, I introduce a large-scale computational analysis of how hierarchy can encode and perpetuate bias in Chapter 4 and discuss the extent to which these biases could be external, arising from the literature that is classified, or internal and imposed on the literature by the structure of the system itself. I draw from the existing accounts of biases in these systems to guide our framework and the biases I choose to focus on.

Computer Science

Bias in computer science has largely been discussed in the context of biases in machine learning. In computational linguistics, many works look at how word embeddings, which form the basis of large language models (LLMS), encode gender, ethnic, and racial biases (Bolukbasi et al., 2016; Garg et al., 2018; Manzini et al., 2019; Papakyriakopoulos et al., 2020; Rozado, 2020; Swinger et al., 2019; Zhao et al., 2019), or how the outputs and generated text of these LLMS can therefore reflect and reinforce biased worldviews (Bai et al., 2025; Hofmann et al., 2024; Watson et al., 2023). Those looking at image classification and facial recognition discuss the unequal abilities of algorithms to identify and label women, racial minorities, and objects in non-Western cultures and contexts (Buolamwini & Gebru, 2018; De Vries et al., 2019; Klare et al., 2012). These works largely attribute the biases in these machine learning models to biases in the datasets they are trained on, and investigations of the image and language datasets these algorithms are often trained on find clear evidence of Western and gender biases (Harrison et al., 2023; Liu et al., 2021; Luccioni & Rolnick, 2023; Shankar et al., 2017; Stock & Cisse, 2018). For example, Visual Genome, a dataset that contains labelled images, has fewer images of women and girls than it does of men and boys (Harrison et al., 2023). By encoding and reinforcing biases in the data, these algorithms can perpetuate these biases and cause real-world harm such as wrongful convictions (Moy, 2021). They must therefore be acknowledged and addressed.

Other works on bias look at how machine learning algorithms, design choices, and outputs can enforce biases that go beyond what is contained in the data (Friedman & Nissenbaum, 1996; Kirkpatrick, 2016). The policy that determines order in recommender

algorithms affects what people pay attention to (Lerman & Hogg, 2014). For example, an algorithm using measures of popularity to recommend media and websites to users will end up enforcing rich-get-richer schemes where items at the top of a recommendation list get viewed more, inflating their popularity and thus making them increasingly more likely to appear at the top (Ciampaglia et al., 2018). More generally, learning bias refers to biases introduced by the learning process itself, including the choice of model class, objective function, or inductive assumptions. These design choices shape what patterns the model prioritizes during training and can therefore produce systematic distortions in the learned outputs beyond what is directly present in the data (Suresh & Guttag, 2021). Aggregation bias, or the ecological fallacy, suggests that relationships learned from mixed populations may not hold uniformly across all subgroups (Suresh & Guttag, 2021). It is often attributed to biases in the data (Mehrabi et al., 2021), but it can also arise from algorithmic design, where optimizing a single global objective over heterogeneous populations privileges majority patterns and obscures subgroup-specific structure. An algorithm that fails to account for individual differences or whose purpose is to generalize to new and unseen data may therefore create bias from its attempts to find a general solution for a population with substantial variability. Similar to ideas discussed in library and information science, systems that attempt to create universal solutions will result in biases towards individuals who deviate from the norm.

Researchers in this field often divide these biases into higher-level groupings and, although not using the same language, explicitly distinguish between biases that are a result of the specific algorithm or structure imposed on the data (internal bias) versus bias already present in the data that is used to train an algorithm or agent (external bias; Mehrabi et al., 2021). For example, Friedman and Nissenbaum (1996) differentiates between preexisting biases that exist independently of the system but are nonetheless captured by it, technical biases that arise from the limitations, constraints, and assumptions of the architecture used to model them, and emergent bias which arises from a mismatch between the system and its users due to change over time or incorrect assumptions about who the system is for. The first two map relatively cleanly onto the notions of internal and external biases, with preexisting biases being external and technical being internal; however, the third type of bias suggests an interesting dynamic where the system might be justified at the time it was created, but as our knowledge changes or different groups of people start to use the system, it becomes inadequate.

Across psychology, library and information science, and computer science, researchers distinguish between biases that are external and biases that are internal to the system. In both Chapters 3 and 4, I discuss the extent to which the biases I examine in institutional systems reflect each of these two kinds. In Chapter 3 in particular, I compare biases in the labels of institutional systems to biases in everyday language use and the extent to which these biases are reflected, mitigated, or exaggerated by these systems.

More generally, computer science offers large-scale quantitative techniques for evaluating the prevalence of social biases in online communities, datasets, and category systems. Beyond looking at the biases inherent in machine learning algorithms and datasets, computational analyses have also been used to identify and measure biases in institutional category systems like Wikipedia (Falenska & Çetinoğlu, 2021; Lerner & Lomi, 2018; Reagle & Rhue, 2011; Sun & Peng, 2021; Wagner et al., 2015). And natural language processing methods specifically have been used to detect and measure social biases in various linguistic communities at a larger and more naturalistic scale than a lab experiment may be able to (Bailey et al., 2022, 2025; Bhatia & Bhatia, 2021; Caliskan et al., 2017; Durrheim et al., 2023; Govindarajan et al., 2023). Thus, this thesis joins an existing body of work that uses computational techniques to explore how social biases are encoded in the datasets, systems and technologies people use.

Finally, like psychology and library and information science, discussions of bias, whether the result of the system, the language, or the data, have mostly been discussed in the context of flat categories or not in the context of categorization at all. Research on aggregation effects has demonstrated that changing the size of a grouping category can significantly alter statistical results (Gehlke & Biehl, 1934), suggesting that hierarchical structures may encode bias through the differing scales of aggregation represented by categories in the hierarchy. The hierarchical structure of a system can encode bias, and thus acknowledging how bias can be encoded in hierarchical structure and parent-child relationships between categories is important to understanding the ways bias can show up in the output of many machine classification techniques, like hierarchical clustering, and in the representations that underlie a lot of databases.

2.2 The Evolution of Category Systems

As demonstrated in the previous section, it is possible to assess the goodness of an existing category system in terms of its optimality and bias at a specific point in time. However, even with this knowledge, creating real-world systems that are optimal is difficult because these systems develop over time with incremental updates that adjust for changes in the real world. As suggested by Friedman and Nissenbaum (1996), systems that were once warranted for the things they classify can become biased or suboptimal as the items, environment, or people who use them change. In this section, I review work from psychology, library and information science, and computer science that relates to this idea, highlighting how sequential encounters with an environment and past choices can constrain the systems that appear today. Although the term evolution is often used in the context of cultural evolution across broad time scales, I take it to generally refer to any change in a category system that happens incrementally over time. This includes experimental timescales,

developmental timescales within a lifespan, institutional timescales, and the longer, more traditional evolutionary timescales that span generations.

2.2.1 Psychology: Emergence & Development of Cultural Systems

Human languages evolve through cultural transmission to divide spaces of meaning into structured semantic categories (Carstensen et al., 2015; Ferdinand & Perfors, 2020; Griffiths & Kalish, 2007; Kirby et al., 2008, 2014). This has often been studied using iterated learning paradigms, where category systems are passed from one generation to the next, either through simulations involving computational agents (Griffiths & Kalish, 2007; K. Smith et al., 2003) or through laboratory experiments with human participants (Carstensen et al., 2015; Kirby et al., 2008). Findings from this work suggest that semantic category systems emerge from competing pressures: languages must be simple and structured enough to be learnable across generations while also being expressive enough to support efficient communication within generations (Carr et al., 2017; Kirby et al., 2015). Other findings suggest that, although communication provides a pressure for expressivity such that communication facilitates the creation of more fine-grained categories, cultural transmission alone is sufficient to create structured, learnable categories that are useful for communication (Silvey et al., 2019). In particular, communication can lead to the creation of non-optimal category systems shared between learners (Silvey et al., 2013). Thus, transmission of knowledge across generations leads to the conceptual systems we express through language.

However, two assumptions are implicit in much of the cultural evolutionary work. First, category systems are treated as flat divisions of a space. When hierarchy is addressed, the focus is on how different environmental needs lead to the evolution of categories at different levels of specificity (Hawkins et al., 2018), rather than the development of the system as a whole. Second, these accounts often treat the set of things to categorize as static and do not address how these systems are adapted to accommodate novel stimuli. Work on semantic change addresses how new concepts are incorporated into human lexicons through processes such as word reuse (Xu et al., 2024), but does not address how the broader organization of categories, such as their hierarchical structure and relationships among categories, changes as new items are introduced. How hierarchies emerge and adapt over time remains an open question.

Researchers often draw parallels in how developmental trajectories emerge both within a person along a developmental time scale and within a culture across an evolutionary timescale. For example, it has been suggested that folk generic categories or basic level categories emerge first in the evolution of a folk taxonomy (Berlin, 1977), and, as will be discussed below, the order of emergence for levels of abstraction within a child's category system follows similar trajectories. The development and learning literature deals more di-

rectly with category hierarchy and thus provides insights into how such systems are acquired and how they can change over time.

The development of hierarchical category systems appears early in childhood and becomes more structured with age. Evidence that children represent inclusion relationships between specific and general categories is present from as early as preschool, although their ability to articulate or use these relations is not always consistent (C. L. Smith, 1979). Some research suggests that children tend to learn basic-level categories before superordinate and subordinate ones (Mervis & Crisafi, 1982; Waxman et al., 1991), while others suggest that superordinate categories emerge first during infancy, followed by more specific categories (Mandler & McDonough, 1998; P. C. Quinn & Johnson, 2000). Although it is not certain whether infants have adult-like hierarchies (Rakison, 2000), preschoolers display evidence of such categories (Mervis, 1987). Moreover, evidence suggests that children's expectations about lexical categories bias them towards interpreting nouns as referring to higher-level taxonomic categories and adjectives as referring to more specific, lower-level categories (Waxman, 1990; Waxman & Kosowski, 1990; Waxman et al., 1991). Thus, through experience with language, children are able to start placing concepts into hierarchical systems.

These hierarchical category systems do not stay the same once acquired. Work on conceptual change argues that concepts are embedded within intuitive theories that become reorganized through development (Carey, 1985; Keil, 1989). In particular, conceptual development may involve both assimilating new information into existing conceptual structures and restructuring the system itself when existing structures become insufficient (Carey, 2000; Chi, 1992). Rather than development consisting solely of refinement of existing categories, these accounts suggest that category boundaries and hierarchical organization can also change over development (Keil, 1981, 1989). From this perspective, hierarchical category systems are constructed and reorganized over time as new information is encountered.

Looking beyond the development of category systems, the category learning literature suggests a key phenomenon in the construction of category systems: sensitivity to the order in which items are encountered. Early and late exemplars can disproportionately influence category boundaries (Medin & Bettger, 1994), and categorization more generally is shaped by the sequence of experiences (Carvalho & Goldstone, 2015; Clapper, 2015; Stewart et al., 2002; Zaki & Salmi, 2019). For example, the order of stimulus presentation influences where a category boundary is drawn along continuous feature dimensions like colour (Egré et al., 2013; Stöttinger et al., 2016) and can influence the level of categorization participants most readily apply to items (Mack & Palmeri, 2015). Together, these findings suggest that early category decisions constrain future ones and that the full history of a system matters for understanding why it now looks the way that it does.

A limitation of these works is that they mostly rely on computational models and experiments meant to simulate how synthetic categories are learned and evolve out in the

world. As a result, they are often divorced from their natural contexts and time scales. Well-documented institutional category systems offer a window into how category systems evolve out in the wild. In the next section, I review literature related to the evolution of these institutional systems.

2.2.2 Library & Information Science: The History of Institutional Systems

Institutional systems have the requirement to maintain integrity such that the meaning of their categories is not lost across time, but also to be flexible such that they can accommodate an ever-changing body of information and resources (Glushko et al., 2016). Systems like the LCC and the DDC have been around since the late 1800s and are continually updated in an attempt to better reflect the changing body of literature they are meant to classify. However, these systems can be difficult to change, and often their structures reflect their history. For example, in 1987, the LCC's categorization of astronomy literature did not reflect the current needs or practices of the field and many of the subdivisions in category QB (astronomy) were based on outdated divisions of knowledge (Crovisier & Intner, 1987). The classification of fungi as a plant persisted long after its recognition as a distinct kingdom, as its classification was only updated in more recent editions of the DDC (Dewey, 2003). Systems that were warranted at the time they were created seem to become less warranted with time, and updating them becomes difficult. Early category distinctions persist even when they do not make the most sense.

The evolution of Wikipedia's category system has been studied more directly. Wang et al. (2013) represents Wikipedia's category system as a network and explores how edges between categories form over time, finding a rich-get-richer effect where well-connected categories are more likely to acquire additional edges. Suchecki et al. (2012) represents it as a hierarchy and notes that before major reorganization, the quality of the system declines such that the links between pages are more frequently between than within top-level categories. What they do not explain is the process by which these systems become suboptimal, thus requiring these major updates.

When looking at the development of institutional systems, the prevailing idea is that systems can become worse with time as early choices constrain future ones. Eventually, there may be major restructuring to account for these changes, but it is a difficult process that lags behind, such that institutional category systems seem to always be playing catch-up. Of course, the critiques are likely, in part, due to the fact that systems like Wikipedia and the DDC have the impossible task of categorizing an ever-changing body of literature that in theory spans the spectrum of human knowledge. There is no feasible way that such a system will be equally as good for everything while still being kept to a size and granularity that is feasible to maintain. Nevertheless, there still seem to be temporal aspects to these critiques, where once appropriate structures no longer adequately serve their intended purpose.

Although these systems are not on the same scale as cultural category systems that are learned and updated by humans, they reflect a similar process: systems are created with limited knowledge and subsequently adapted to accommodate new knowledge, potentially resulting in structures that become increasingly less warranted by the items they classify. Accounts of this process in library and information science have so far been mainly descriptive. It remains unclear how and why these systems become suboptimal, or what threshold of misalignment between a system and its items warrants major restructuring. I address the first of these questions in Chapter 5 and leave the latter as an open question for future research.

2.2.3 Computer Science: Incremental Learning & Models of Categorization

Computer science algorithms have been used to formalize models of categorization. Computational models of category learning formalize different assumptions about how category structure is represented and learned. Exemplar-based models simulate incremental learning through trial-by-trial updates to similarity-based representations and support categorization across levels of abstraction, but do not explicitly encode category systems as structured hierarchies or account for how such systems are reorganized over time (Palmeri, 1999; Verheyen et al., 2008). Rational models, which can be formalized as nonparametric Bayesian inference (Anderson, 1991; Sanborn et al., 2010), account for categorization through probabilistic clustering of observations and capture sequential effects via incremental updating, but similarly assume flat category systems form without explicit hierarchical organization. In contrast, hierarchical Bayesian models, such as those developed by Canini and Griffiths (2011), infer taxonomic structure from labelled examples, but primarily focus on structure induction rather than how new items are sequentially incorporated into an existing hierarchical system. Models that encode hierarchy while still capturing these sequential updating effects are needed.

Sequence effects are not only relevant to the development of human category systems, but also to the development of computationally created ones. Incremental learning research in computer science investigates how systems update categories sequentially over time, under memory and computational constraints, while balancing the incorporation of novel information with the preservation of previously learned structure (Gepperth & Hammer, 2016). Such methods are often used in cognitive science to simulate human learning. For example, as discussed previously, Sanborn et al. (2010) uses particle filtering, a sequential Bayesian inference method that incrementally updates probabilistic hypotheses about the category space as new observations are encountered, to simulate category learning. Other incremental learning models are applied to classification and clustering problems in which categories are built and updated over time as new data arrive, with new categories sometimes created when novel patterns emerge in the data. Together, these models formalize how

early representational choices constrain later organization while still permitting adaptation to changing environments.

With incremental learning methods, there is often a need to make locally optimal decisions that ideally lead to a globally optimal solution structure. For example, in k-means clustering, the final clusters can depend strongly on the initial choice of centroids. While each iteration improves the clustering objective, the algorithm may converge to a local optimum rather than the globally optimal solution. (Bubeck et al., 2012). Probabilistic approaches attempt to address this by making non-deterministic choices weighted by their optimality or likelihood of being correct (e.g. particle filtering; Sanborn et al., 2010), though these methods still do not guarantee an optimal overall solution. Because such methods build solutions sequentially without full knowledge of the solution space, they can still converge to suboptimal outcomes.

Incremental learning methods have been applied to hierarchical clustering of text documents (Sahoo et al., 2006), and in some approaches allow hierarchical structures to be updated incrementally as new data arrive, reducing the need to recompute the full hierarchical clustering, which is computationally intensive (Ribert et al., 1999). The same need to make locally optimal choices that lead to globally optimal solutions exists in the construction of hierarchical category systems, and there is the risk that a certain sequence can lead to less ideal solutions.

The history of a system matters with respect to its current structure. How this plays out in a hierarchical category system and what mechanisms might facilitate it remains an open question. While this thesis will not tackle why or how hierarchy emerges, it will explore how such systems are products of a sequential process through the use of computational models in Chapter 5.

Chapter 3

Quantifying Bias in the Labels of Institutional Category Systems

3.1 Introduction

Some members of a category are more typical than others (Rosch, 1973), and typicality is influenced by social biases and stereotypes (Lakoff, 1987). For example, a study of human naming choices in English found that photos of male athletes tended to be tagged with the sport they were playing, whereas photos of female athletes were more often tagged with their gender (Harrison et al., 2023). This pattern reflects the *people=men* bias, which suggests that many cultures have masculine defaults and perceive men as more generic or typical examples of humans than women (Bailey et al., 2019; Cheryan & Markus, 2020; Hamilton, 1991; Merritt & Kok, 1995; Silveira, 1980; Van Berkel et al., 2017).

The perception of women as atypical leads to asymmetries in how people are categorized. While typical category members are often unmarked and can even stand in for the category as a whole (e.g., mankind meaning humankind), atypical members are often explicitly labelled or qualified with the features that make them deviate from the norm (Brekhus, 1998; Lakoff, 1987). So while men in science might generally be categorized as *scientists*, women are more likely to be categorized as *women scientists*.

Here we explore asymmetries in gendered categories provided by two institutional category systems: the Library of Congress Subject Headings (LCSH)[1] and Wikipedia. The LCSH is a controlled vocabulary of terms such as *Linguistic change* or *Odors in the Bible* that can be applied to books. Updates are centrally reviewed by "cataloguing policy specialists" in the Library of Congress (Library of Congress, n.d.). Wikipedia is a free online encyclopedia that can be edited by anyone and that incorporates a category system for grouping related articles. As such, both systems are designed to organize and help navi-

[1]Note that the LCSH is distinct from the hierarchical Library of Congress Classification that is used by libraries to shelve books.

gate a large body of information. Although institutional category systems can reflect and reinforce the biases of their creators and the items they categorize (Bowker & Star, 1999), systems like the LCSH and Wikipedia are often perceived as objective, which makes it important to document where and how they perpetuate bias.

Both the LCSH (Berman, 1971; Hobart, 2020; Hyde, 2002; M. N. Rogers, 1993) and Wikipedia (Falenska & Çetinoğlu, 2021; Sun & Peng, 2021; Wagner et al., 2015) have been previously criticized for being biased against women in several respects. Here we focus on just one aspect of these systems, and our first goal is to quantify the extent to which these systems include many categories that mark women (e.g. *women scientists*) in the absence of similar categories for men. If present, this asymmetry demonstrates a bias against women by treating them as the "odd ones out" whose membership in a category (e.g. *scientists*) is not assumed but must instead be explicitly labelled.

Although we expect to find more categories that mark women overall, some categories such as *nurse* are more strongly associated with women than men, and are likely to be marked accordingly (Duffy & Keir, 2004; Garg et al., 2018; White & White, 2006). Our second goal is to test whether category base rates (i.e. the proportions of men and women belonging to the category) affect how categories are marked. We hypothesize that base rates will override the *people=men* bias, leading men to be marked in female-dominated categories, while women remain marked in male-dominated categories or categories with near-equal representation of men and women.

Our third goal is to examine whether gender biases have decreased over time. Bias has been documented in the LCSH since at least 1971 (Berman, 1971) and in Wikipedia since 2011 (Antin et al., 2011). In the years since, there have been dedicated efforts to reduce bias. For example, the Cataloging Lab aims to improve accuracy and reduce bias in the LCSH by facilitating proposals and revisions of headings, and The Women in Red are a group of volunteers who seek to reduce systemic gender bias in Wikipedia. By measuring changes in bias over time, we aim to determine how much progress has been made and how much work remains to be done.

From one perspective, our work is an exercise in applied cognitive science that aims to understand how principles of typicality and categorization play out in two important real-world settings (the LCSH and Wikipedia). It contributes to a body of research that seeks to deepen our understanding of categorization "in the wild" (Glushko et al., 2008) and our understanding of how social biases are encoded in the datasets, systems and technologies we use (Bailey et al., 2022; Watson et al., 2023). From another perspective, our project uses institutional category systems to develop large-scale tests of theoretical ideas about categorization, such as the influence of base rates on typicality. While both perspectives may be useful, we mostly emphasize the former.

3.2 Gender Bias in Institutional Category Systems

Our first analysis explores the extent to which gender asymmetries exist in the categories of the LCSH and English Wikipedia. In addition to comparing these systems with each other, we compare both with usage frequency data derived from the Google Ngram corpus, which is made up of millions of published books. Although Google Ngram is not necessarily representative of the full range of works underlying the LCSH and Wikipedia, it provides a useful point of comparison because both systems are intended to reflect the language used in published works. For the LCSH, this connection is direct. The principle of literary warrant means that its subject headings should reflect the language used in the literature it classifies (Library of Congress, Policy and Standards Division, 2016). For Wikipedia, the connection is more indirect. Its categories reflect the topics covered in articles, whose content is expected to be based on reliable published sources (Wikipedia:Verifiability, 2025). We do not expect Google Ngram to provide an unbiased baseline, as the language in its corpus also reflects biases present in the literature. Rather, comparing these systems with Google Ngram helps us assess whether gender asymmetries in the LCSH and Wikipedia mirror, mitigate, or exaggerate those present in written English language use, using a corpus that likely resembles the literature underlying these systems.

We start by compiling a broad set of gender categories that aim to be relatively comprehensive, then focus on two subsets of particular interest — categories related to jobs (e.g., *Male musicians*) and categories related to ethnicities and nationalities (e.g. *Yoruba men* and *German women*). As discussed later, considering these two subsets allows us to probe the influence of category base rates and of the *people=men* bias on gender asymmetry.

3.2.1 Data and Methods

LCSH and Wikipedia.

Our LCSH data are derived from the official LCSH data dump released by the Library of Congress, which contains digital records for headings created as recently as 2024.[2] For Wikipedia, we use a preprocessed version of English Wikipedia's 2020 category dump (Lu, 2020). To collect gendered headings, we first search for any category that contains the words 'men', 'male', 'women', and 'female'. The singular terms 'man' and 'woman' are excluded because they are not commonly used in Wikipedia (Falenska & Çetinoğlu, 2021) and the LCSH, and when they do appear, they are often part of a proper name (e.g. *Isle of Man*, *Spider-Man*), or used to refer to human beings in general (e.g. *Fall of man*).

Next, we remove any category that includes words for men and women (e.g. *European Men's and Women's Team Badminton Championships*). Following Falenska and Çetinoğlu (2021), we then identify pairs such as (*Canadian women*, *Canadian men*) and (*Women*

[2]See https://id.loc.gov/authorities/subjects.html

scientists, *Male scientists*) that are identical except for the inclusion of a different gendered noun or adjective.[3] We then group the terms into three classes: W includes unpaired categories that exist only in a female version (e.g. *Female doctors* exists but *Male doctors* does not), M includes categories like *Male caregivers* that exist only in a male version, and WM includes categories that appear in both male and female versions. The total number of gendered categories is $W + M + WM$, and this total equals 2,668 for the LCSH and 15,955 for Wikipedia.

Google Ngram Data

We derive usage frequencies from version 3 of English Google Ngram (Michel et al., 2011), which includes Part-of-Speech (POS) tags (Lin et al., 2012). Using Google Ngram data, we collected bigrams and trigrams such as 'Australian_ADJ men', 'women in science_NOUN', and 'male facial_ADJ attractiveness_NOUN'. Each n-gram included exactly one gender word, and for consistency we used the same gender words as for LCSH and Wikipedia. If the gender word is 'female' or 'male', then the gender word must occur before a noun such that it modifies that noun. The words 'women' and 'men' can occur before or after nouns because, as in the LCSH and Wikipedia, 'women' is often used in place of 'female' as a modifier.

We applied four additional rules designed to identify the kinds of category labels that frequently appear in Wikipedia and the LCSH. First, an n-gram may not include punctuation. Second, each n-gram must have a complete set of POS tags. Third, 'in', 'of', 'for', 'and', or 'by' may not appear as the first or last word of an n-gram. And fourth, all other words must have more than two letters and must be tagged as a noun, proper noun, or adjective, except for the last word, which cannot be tagged as an adjective. The final set of n-grams included 273,608 bigrams and 584,488 trigrams. We ignored case when collecting these n-grams, and the frequency of each n-gram was defined as its total aggregate frequency between 2010 and 2019.

To simplify the dataset, we replace all occurrences of the words 'men' and 'male' with the token $\langle M \rangle$ and 'female' and 'women' with the token $\langle W \rangle$. We then lemmatize the remaining words in each n-gram based on their POS tag and aggregate frequencies across n-grams that now have the same form. Finally, we pair n-grams (e.g. ('Australian_ADJ men', 'Australian_ADJ women')) as we did for the LCSH and Wikipedia. For each pair, we computed the percentage of gendered occurrences that mention women ($\%W$). We removed all pairs where the frequency of either member of the pair was less than 100. The total number of gendered pairs was 17,642.

[3] *Women* is paired with *male* because in the LCSH, Wikipedia, and Google Ngram *women* often appears as a modifier (instead of female), but the same is rarely true for *men*. In our data sets, the distinction between female/male and women/men does not reliably indicate a difference between sex and gender.

Table 3.1: Datasets used to compile a list of jobs and identities.

Domain	Datasets
Identities	List of contemporary ethnic groups; List of adjectival and demonymic forms for countries and nations
Jobs	Detailed occupation by sex education age earnings 2019; Gender occupation percentages (Garg et al., 2018); Thesaurus of job titles; Table 3. Gender perception of role nouns (Misersky et al., 2014).

To make our Ngram data comparable to our LCSH and Wikipedia data, we classify each n-gram pair as W, M, or WM based on the value of $\%W$. Given lower and upper thresholds l and u, pairs with $\%W < l$ are classified as M, pairs with $l \leq \%W \leq u$ are classified as WM, and pairs with $\%W > u$ are classified as W. We create three classifications, each of which uses different values of l and u. Ngram-D is a default classification that sets $l = 0.4$ and $u = 0.6$. Ngram-L and Ngram-W preserve the property that l and u are equidistant (or nearly so) from 0.5, but adjust the distance between l and u to allow the proportion of WM pairs to match the LCSH and Wikipedia classifications, respectively. The values of the thresholds for Ngram-L and Ngram-W are reported in Figure 3.1.

Jobs and Identities

For simplicity, we refer to both ethnicities and nationalities as identities. We compiled a list of jobs and identities from sources listed in Table 3.1. Given these lists, we used string matching to find job categories of the form *(male|female|women|men) JOB* and identity categories of the form *IDENTITY (women |men)* in the LCSH, Wikipedia, and Google Ngram. For the LCSH, 11.28% of the gendered categories correspond to jobs and 7.68% to identities. For Wikipedia, these percentages are 1.03% and 1.40%, respectively, and for Ngram data 2.64% and 1.72%. Because the n-grams are lemmatized, we used lemmatized versions of the jobs and identities for Ngram data only.

3.2.2 Results

Figure 3.1 shows the relative proportions of W, WM, and M classes for the three data sets. The top row shows that the LCSH has a very high proportion of W categories and very small proportions of WM and M categories overall. This trend holds for both jobs and identities, showing that the LCSH has a strong gender bias against women. Wikipedia also shows this gender bias overall and for jobs; however, in both these cases the proportion of WM categories is higher than for the LCSH, indicating less bias than the LCSH. Wikipedia shows almost no gender bias for identities, because almost all identity categories are classified as WM.

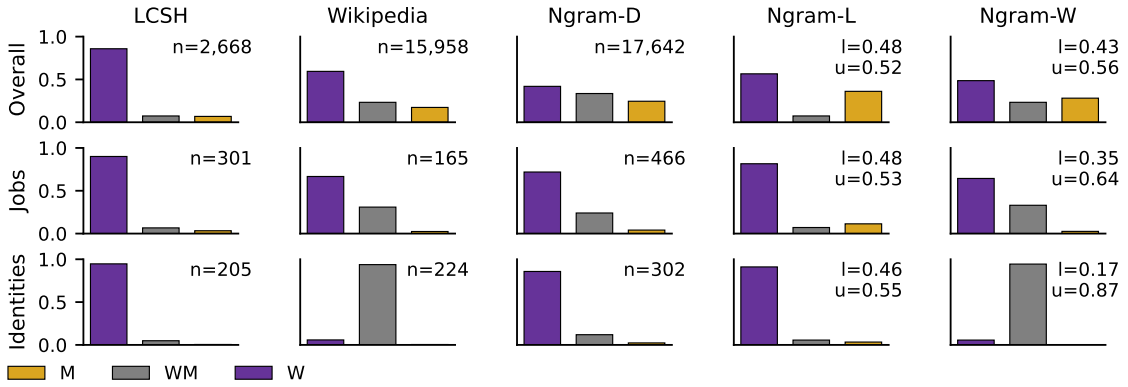


Figure 3.1: Types of gendered categories in the LCSH, Wikipedia, and Google Ngram. The top row shows overall results, and the middle and bottom rows show results for jobs and identities. As described in the text, the thresholds for the Ngram-L and Ngram-W classifications are chosen so that the *WM* proportions match those for LCSH and Wikipedia respectively.

To compare the LCSH and Wikipedia to Ngram-D, we performed a proportional odds ordinal regression with classification ordered as $M < WM < W$ ($M = -1$, $WM = 0$, $W = 1$) as the dependent variable and data set (LCSH, Wikipedia, or Ngram) as a predictor. Compared with Ngram-D, the LCSH showed a significant positive shift in classification toward the W end of the ordinal scale ($\beta = 2.05$, 95% CI [1.94, 2.16], $p < .001$). For Wikipedia, we found a smaller but still significant shift toward W ($\beta = 0.63$, 95% CI [0.59, 0.67], $p < .001$). Overall, both Wikipedia and the LCSH have more W classes and fewer M and WM classes than Ngram-D.

We used the same ordinal regression approach to compare Ngram-D to Wikipedia and the LCSH in the domains of jobs and identities. For jobs, LCSH showed a significant shift toward the W end of the ordinal classification scale relative to Ngram-D ($\beta = 1.24$, 95% CI [0.81, 1.66], $p < .001$) whereas Wikipedia did not differ significantly from Ngram-D ($\beta = 0.21$, 95% CI [-0.58, 0.17], $p = .28$). Wikipedia is more similar to Ngram-D than to the LCSH, as both have more *WM* classes. Finally, for identities, LCSH was associated with higher cumulative log odds of receiving a classification toward W relative to Ngram-D ($\beta = 1.10$, 95% CI [0.41, 1.78], $p = .002$), whereas Wikipedia was associated with lower cumulative log odds of a classification toward W ($\beta = -4.09$, 95% CI [-4.64, -3.54], $p < .001$). The LCSH is much more similar to Ngram-D than Wikipedia, largely because Wikipedia has such a high proportion of *WM* categories.

When comparing the LCSH to Ngram-L, we find that overall as well as for jobs and identities, the LCSH has more W categories and fewer M categories than Ngram. This pattern is supported by an ordinal regression indicating that, overall, LCSH is associated with higher cumulative log odds of receiving a classification toward W relative to Ngram-L ($\beta = 1.59$, 95% CI [1.48, 1.70], $p < .001$). The corresponding effect for jobs is also significant ($\beta = 0.75$, 95% CI [0.31, 1.19], $p = .001$), whereas for identities the difference is

not significant ($\beta = 0.57$, 95% CI [-0.16, 1.29], $p = .125$). These results suggest that gender asymmetries are more pronounced overall for the LCSH than for language use, and that this conclusion is robust for jobs but not identities.

A similar comparison between Wikipedia and Ngram-W suggests that Wikipedia is more biased overall, with a significant shift toward the W end of the ordinal scale ($\beta = 0.49$, 95% CI [0.45, 0.53], $p < .001$). For jobs, we find no significant difference between Wikipedia and Ngram-W, suggesting that Wikipedia's gender asymmetries for jobs reflect asymmetries in language use ($\beta = 0.10$, 95% CI [-0.27, 0.47], $p = .60$). For identities, Wikipedia is again not significantly different from Ngram-W ($\beta = -0.05$, 95% CI [-0.78, 0.69], $p = .91$). However, the extreme thresholds $l = 0.17$ and $u = 0.87$ required to match the WM proportion across Wikipedia and Ngram-W suggest that Wikipedia gives balanced treatment to pairs of categories that are skewed in usage.

Our results confirm that the LCSH and Wikipedia both display gender bias, and that this bias is more extreme for LCSH than for Wikipedia. The top row of Figure 3.1 also suggests that the global bias in both systems is more extreme than the bias found in usage data, but for identities, Wikipedia seems less biased than usage data. The following section considers factors that may contribute to these biases.

3.3 Base Rates and Gender Bias

The LCSH bias for identities cannot be explained by base rates, because the number of Canadian women (say) is almost identical to the number of Canadian men. Instead, the asymmetry suggests that the LCSH incorporates the default *people=men* bias, which means that categories involving women need to be explicitly marked. For jobs, however, the base rates of men and women are often different, and our second analysis asks if these differences shape the bias we see in job categories. As mentioned earlier, we hypothesize that the *people=men* bias is overridden in the case of jobs with a higher proportion of women. We consider both actual and perceived base rates for the jobs in our data to see which is the better predictor of bias.

3.3.1 Data and Methods

We derived base rates for our list of jobs by combining US 2019 census data[4] with 2015 US jobs base rates from Garg et al. (2018). Perceived base rates were drawn from Table 2 of Misersky et al. (2014), which includes results from an experiment in which people were asked to "estimate [...] the extent to which the presented social and occupational groups actually consisted of women and men." For LCSH, Wikipedia, and Ngram-D, the number of categories for which we have actual and perceived base rates is shown in Figure 3.2.

[4]Available at: https://www.census.gov/data/tables/2022/demo/acs-2019.html

3.3.2 Results

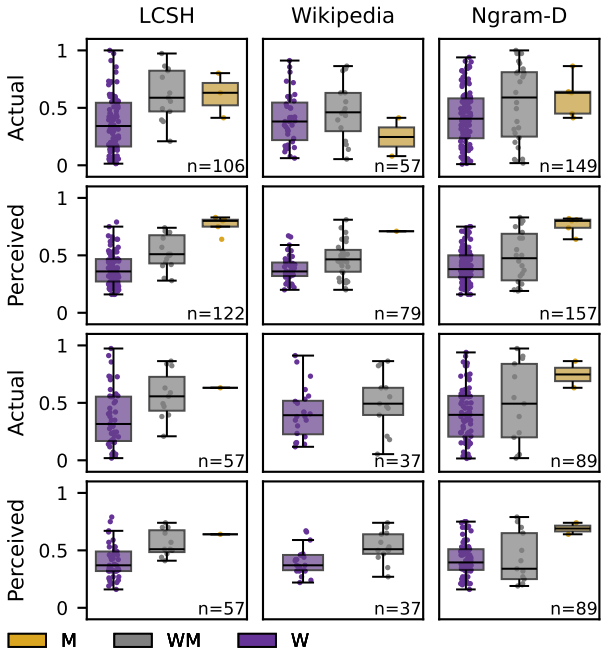


Figure 3.2: Distribution of actual and perceived gender base rates of various jobs across cases W, WM, and M in the LCSH, Wikipedia, and Ngram-D. The top two rows include categories for which we have actual (top row) or perceived (second row) base rates. The bottom rows include categories for which we have both actual and perceived base rates.

The top two rows of Figure 3.2 show the distributions of actual and perceived base rates for W, WM, and M categories in each classification. Across all three data sets, the purple boxes (W categories) are lower than the grey boxes (WM categories), indicating that women are marked in male-dominated categories. The yellow boxes (M categories) are consistently higher than the grey boxes for the second row but not the first row, suggesting that perceived base rates account better than actual base rates for gender bias across the three data sets.

To confirm this impression, we again used proportional odds ordinal regression models with gender classification (ordered as $M < WM < W$) as the dependent variable. We compared three models for each data set: the first was a null model with an intercept only, and the second and third included actual and perceived base rates as predictors. To enable these models to be compared, we included only categories for which we have both actual and perceived base rates, and the bottom two rows of Figure 3.2 show distributions for these categories.

BIC scores for the regressions are reported in Table 3.2. For both LCSH and Wikipedia, the model based on perceived base rates accounts best for the data. Following the conventions for interpreting BIC scores proposed by Raftery (1995), the magnitudes of the BIC scores provide positive but not strong evidence that the perceived base rates model is the

Table 3.2: BIC scores for each proportional odds ordinal regression model. Rows correspond to the model type and columns to the dataset. BICs are reported as offsets from the smallest entry in each column, which means that 0 indicates the best model in each column.

	LCSH	Wiki	Ngram-D
Null	7.92	4.51	**0**
Actual	5.85	6.68	1.95
Perceived	**0**	**0**	3.45

best of the three. For Ngram-D, however, the null model scores slightly better than the two models that incorporate base rates.

The results just described suggest that gender asymmetries in LCSH and Wikipedia are influenced by base rates, and that perceived base rates account for these asymmetries better than do actual base rates. Overall, our results for identities and jobs show that asymmetry in the LCSH is shaped by both base rates and the *people=men* bias, whereas for Wikipedia we have evidence only of the role of base rates.

3.4 Gender Bias Across Time

Gender bias has become less extreme in Western society over time (Bhatia & Bhatia, 2021), and this section explores whether the asymmetries we have documented in the LCSH, Wikipedia, and Google Ngram have become weaker over the past few decades.

3.4.1 Data and Methods

We use the date a digital LCSH record was created for the heading as a proxy for its date of addition to the system. Because the system was not digitized until 1986, any heading added before then has a date of addition of 1986. Digitizing all existing headings may have taken some time, and we therefore only consider changes in gendered subject headings between 1990 and 2019 (the most recent Ngram year).

The Wikipedia dataset does not contain the date a category was added to the system. Most Wikipedia categories, however, have an associated web page, and we approximate a category's date of addition by extracting the date on which its associated page was first revised. This approach yielded addition dates for 94% of the gendered Wikipedia categories. We consider the period from 2004 to 2019 because Wikipedia's category system did not exist prior to 2004.

For Google Ngram, we use the default Ngram-D thresholds and create gendered phrase pairs using the same method as Study 1, except for two key differences. First, the phrase pairs are created for each year between 1990 and 2019. Second, the minimum frequency for each phrase in a gendered phrase pair is 20 as opposed to 100. A lower threshold is

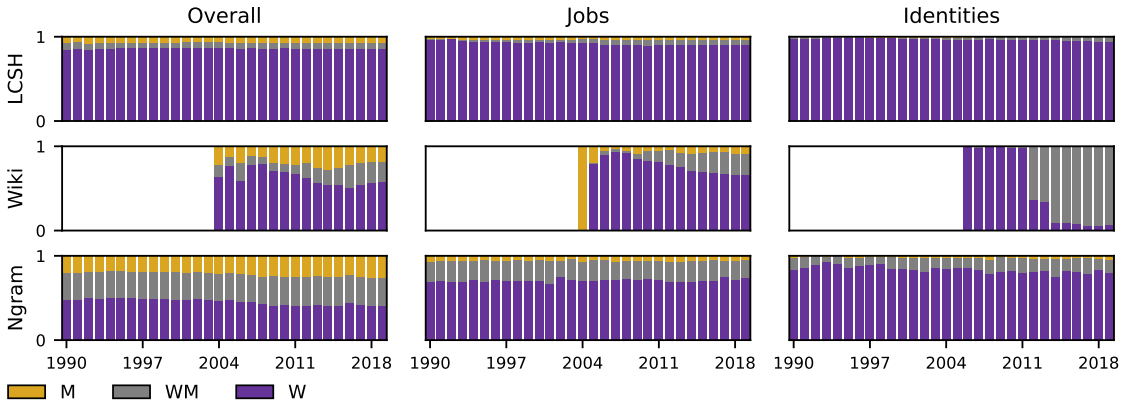


Figure 3.3: Cumulative proportions of gendered categories in the LCSH, Wikipedia, and Google Ngram over time. The top row shows overall results, and the middle and bottom rows show results for jobs and identities. Here, Ngram uses the same W, WM, and M category boundaries as Ngram-D.

appropriate because the counts for each pair are now based on the frequency of gendered phrases in a single year as opposed to an aggregate frequency across ten years.

For each of the three data sets, we compute the proportion of W, WM and M categories for each year. We use the default frequency boundaries of 0.40 and 0.60 when classifying n-grams as W, WM, or M.

3.4.2 Results

Figure 3.3 shows how the proportions of W, WM, and M categories have changed over time. The LCSH shows no overall change despite the addition of 1,960 gendered headings between 1990 and 2019. For jobs and identities, there appears to be a gradual increase in the proportion of WM, although this change is very small.

For Wikipedia, the picture is very different. The overall plot reveals an increase in WM categories since 2004, and this pattern is even clearer for jobs and especially dramatic for identities. In 2004, all categories were classified as W, but in 2012 there was a sharp increase in the proportion of WM categories, and by 2019, the categories added over the previous seven years left the complete set of identities relatively balanced.

For Ngram-D there are no discernible trends for jobs and identities, and the proportions of W, WM, and M appear stable over time. For the full set of gendered n-grams, however, the proportion of M categories gradually increases between 1990 and 2019, although this increase is not very large.

Overall, our results suggest that the gender asymmetries in the LCSH and Google Ngram have been relatively stable over the past three decades. In contrast, Wikipedia asymmetries for jobs and identities have decreased since the early 2010s, and this decrease is especially striking in the case of identities.

3.4.3 Why has Wikipedia changed but not the LCSH?

Wikipedia has policies specifying how the intersection between gender and ethnicity should be approached, and the page outlining these policies (Wikipedia:Categorizing articles about people, 2025) was created in 2005, before the 2012 change we saw for identities in Figure 3.3. The 2012 change may reflect the publication in the early 2010s of papers highlighting gender bias in Wikipedia (Antin et al., 2011; Reagle & Rhue, 2011) and the emergence of groups such as the Women in Red (Wikipedia:WikiProject Women in Red, 2025) dedicated to reducing these biases.

For the LCSH, gender biases have been criticized since 1971 (Berman, 1971) if not before, and although groups have emerged with the goal of reducing these biases (e.g. the Cataloging Lab), our data suggest that gender asymmetries have remained stable since 1990. One reason why Wikipedia alone has changed in recent decades could be the effort required before a new LCSH heading is approved. Proposals for new headings must be approved by the Library of Congress and must be accompanied by evidence that they meet the requirement of literary warrant (Svenonius, 2000). In contrast, Wikipedia follows a collaborative editing model that allows for very rapid changes.

3.5 Discussion

We explored gender asymmetries in two institutional category systems and found three key results. First, although both the LCSH and Wikipedia contain gender bias, the LCSH tends to exaggerate bias found in language use, whereas Wikipedia mirrors it (jobs) or mitigates it (identities). Second, while the gender asymmetries in both the LCSH and Wikipedia are influenced by perceived base rate, only the LCSH displays evidence of marking behaviour consistent with the *people=men* bias. Third, gender asymmetries have remained relatively unchanged in the LCSH and Ngram over the past three decades; however, asymmetries in Wikipedia's jobs and identities have decreased with time, most notably in the case of identities.

Our results demonstrate how institutional category systems can encode people's biases through unequal labelling of various groups, and that ideas of category typicality, which have mostly been investigated in the context of cultural categories that we carry around in our heads, extend to these systems. In most cases, the intent behind marking female categories is probably not to cause harm, and explicit labelling can address historical exclusion (e.g. there was an explosion of LCSH categories for women in the 1970s; Hyde, 2002). Categories, however, still shape how we perceive the world (Goldstone et al., 2001), and biased systems can reinforce a picture of women as the "odd ones out" whose belonging is not assumed but specified (Brekhus, 1998).

Our results also demonstrate that these biases are not always inevitable. Work can be done to alleviate them in institutional systems, as in the case of identities in Wikipedia. Although we speculate that this is at least in part due to Wikipedia's more collaborative and accessible editing model, future work is needed to better understand the factors affecting change and provide a better understanding of when de-biasing efforts succeed.

Our analysis is limited in several respects. First, we did not focus on the subset of categories shared by the LCSH, Wikipedia, and Google Ngram, and some observed differences may reflect variation in the concepts represented in each system. Nonetheless, we believe the differences we find are meaningful, as they illustrate how gender is marked within each system as a whole. Second, our automatic text processing may have influenced our results, as our use of string matching to align jobs and identities with categories prioritized precision over recall. Moreover, we excluded morphologically gendered terms (e.g., actress), which led to a few unpaired cases (e.g., male actor). These are rare in our data and unlikely to affect our overall assessment of asymmetry. Third, our conclusions are based on asymmetries in English n-grams and Western category systems, which may not generalize to other cultural or linguistic contexts. Finally, we base our analysis on an oversimplified binary view of gender due to the very small fraction of categories pertaining to non-binary or gender-diverse people in all systems included in our work. Our methods could be used in future studies that take a more nuanced view of gender.

3.6 Conclusion

In summary, we demonstrate how gender asymmetries manifest in institutional category systems and how principles of categorization and typicality can manifest out in the real world. The stability of LCSH asymmetries versus the decline of some of Wikipedia's suggests that structural factors —such as editorial accessibility and collaborative editing—may play a critical role in bias reduction. Understanding these mechanisms is essential for developing more balanced classification systems. Although the LCSH has not changed much in the past three decades, we hope that the next three will lead to a more balanced presentation of gender.

Chapter 4

Quantifying Structural Bias in Hierarchical Category Systems

Categories inevitably reflect the needs, perspectives, and experiences of the people who create them. Consider, for example, Steinberg's famous depiction of the *View of the World from 9th Avenue*[1], which devotes half of the page to three New York City blocks but shows China, Russia, and Japan as tiny blobs on the horizon. A *View of the World from Tiananmen Square* would look rather different, and might include separate categories for Shānxī and Shǎnxī provinces while making no distinction between Washington State and Washington, D.C.

Although systems of categories are often subjective, the distinctions that they encode or fail to encode can have important consequences (Crawford & Paglen, 2021). For example, (Gould, 1990) points out that the United States' categorization of drugs as legal or illegal results in some addictive drugs being advertised on TV, and others carrying life sentences. The categorization of an animal or plant population as a distinct species, as opposed to a variant of an existing species, can affect conservation efforts and biodiversity research (Freeman & Pennell, 2021; Thomson et al., 2021). Finally, categories can also lead to harmful stereotypes, especially when coarse categories are used for members of out-groups in contrast to the finer-grained categories used for members of one's in-groups (Park & Rothbart, 1982). Because category systems can encode and reinforce stereotypes, it is important to ensure that the biases encoded by these systems are acknowledged as such instead of treated as ground truth.

Social, developmental and cognitive psychologists have previously explored how categories (e.g. racial and gender categories) arise and how these categories influence behaviour (Brewer, 2007; Misch et al., 2022; Timeo et al., 2017; Waxman, 2021). Existing work highlights two key connections between categorization and bias. First, people tend to have positive attitudes towards in-group categories and negative attitudes towards out-

[1]See *View of the World from 9th Avenue*

group categories, and methods such as the Implicit Association Test (Greenwald et al., 1998; Schimmack, 2021) attempt to measure these attitudes. Second, categories can bias how people perceive individual members of in-groups and out-groups. For example, there is a tendency to perceive members of one's out-group as being more similar to one another than members of one's in-group (Judd & Park, 1988; Mackie & Worth, 1989; Park & Rothbart, 1982; Rubin & Badea, 2012). The link between categorization and biased perception has been explored more generally (Dubova & Goldstone, 2021; Goldstone et al., 2001), and the tendency to overestimate both within-category similarity and between-category distinctiveness is sometimes referred to as categorization bias (Ashby & Zeithamova, 2022) or categorical bias (Ester et al., 2020). Here we focus on a third connection between categorization and bias, and consider ways in which the structure of a category system (i.e. the extensions of the categories that it includes) can reflect bias. A canonical example is that a category system may include fine-grained categories in areas that are deemed valuable or important, and coarser categories in areas deemed less worthy of attention.

The term "bias" has been used to refer to distinct concepts in the literature. For example, "inductive bias" refers to constraints or expectations that guide learning (Griffiths et al., 2010; Markman, 1989), and is not directly relevant to our study. We define bias as preferential treatment for one group (e.g. Western individuals) over another (e.g. non-Western individuals). As we discuss later, this bias can either reflect external biases that have shaped the items to be categorized, or can be internal to a category system and imposed on the items by this system. To study this notion of bias, we develop methods to measure how different groups are represented in a category system. Establishing that a system is biased requires us to demonstrate that the system departs from an unbiased alternative, and it is not always clear how an unbiased system should weigh different groups. We therefore begin with the simplifying assumption that an unbiased system should give roughly equal weight to each group, but return to this assumption later and discuss the extent to which it is appropriate for the specific oppositions that we consider (Western vs non-Western, and male vs female and non-binary).

Much of the previous work on bias in categorization has focused on biases in individual categories or in flat category systems. However, items can be categorized at multiple levels of abstraction (Mervis & Rosch, 1981) and natural categories are often organized into conceptual hierarchies. For example, "flower" is a subcategory of "plant" which is a subcategory of "living things". In addition, many formalized systems of categories such as biological taxonomies, medical ontologies, and library classification systems have hierarchical structures. Here we consider several kinds of biases that can occur in hierarchical category systems. Some of these biases have counterparts in flat systems: for instance, cat lovers could have more fine-grained category divisions for cat breeds than they do for dog breeds. Other biases, however, are distinctive to hierarchical systems. For example, (Loehrlein, 2012) demonstrated that people are biased towards concepts located near the

top of a hierarchical system such that these concepts are perceived as being more important than those at the bottom.

Our methods are general and could potentially be used to measure bias in any hierarchical category system. Laboratory methods commonly used to elicit hierarchical category systems include successive pile sorting and hierarchical clustering based on judgments of similarity (Medin et al., 1997, 2006), and our approach could be applied to hierarchies generated by any such method. Another example is WordNet (G. A. Miller, 1994), a lexical database that organizes nouns and verbs into hierarchies. WordNet is both an influential theory of human lexical memory and a resource used to develop and test many other theoretical contributions in cognitive science, and understanding bias in WordNet is therefore important. WordNet aims for very broad coverage of the lexicon, and a separate research tradition aims to document folk taxonomies of specific semantic domains including plants, animals, artifacts, diseases, and soils (Holman, 2005). Our methods could be used to identify areas given more and less weight by these taxonomies — for example, we could measure the extent to which an animal taxonomy privileges domesticated animals ahead of wild animals, and compare the strength of this bias across cultures. Our approach may therefore be broadly useful as a tool for studying the ways in which people's category systems are influenced by cultural and individual biases, and for exploring the idea that category systems are not exclusively shaped by intrinsic properties of the things that they categorize, but are instead heavily influenced by human needs and values.

Although our approach has many possible applications, here we take library classification as a case study and apply our methods to the Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC). These library classification systems are large-scale, hierarchical examples of human categorization that are directly accessible and much more amenable to computational analysis than the category systems that all of us carry around in our heads. Focusing on library classification also allows us to connect our approach with a large body of existing work in the library and information sciences devoted to uncovering and mitigating bias in the LCC (Angell & Price, 2012; Howard & Knowlton, 2018; Intner & Futas, 1996; Kam, 2007; M. N. Rogers, 1993), the DDC (Higgins, 2016; Kua, 2004; Olson & Ward, 1997; Westenberg, 2022), or both (Mai, 2010; Zins & Santos, 2011). Category systems, especially more formal systems like library classifications, are often perceived as neutral or objective, making it all the more important to develop methods that enable us to quantify and thus acknowledge and address the biases that may be implicit in these systems. As such, studying formal systems like these is valuable in its own right and can also contribute to a better understanding of categorization in general (Glushko et al., 2008).

Previous work on bias in the LCC and DDC has documented that the language used to label categories, and the location of topics and books in the classification schemes can encode harmful bias. For example, in the LCC, unglossed religious terms like "God," and

"devotional literature" refer to these concepts only in the context of Christianity (Knowlton, 2005). Similarly, subject headings such as "engineers" that have subheadings such as "women engineers" but not "male engineers" assume men as the default (M. N. Rogers, 1993). In general, the LCC and DDC systems are biased and unsystematic in their coverage of non-Western religions and racial groups (Westenberg, 2022; Zins & Santos, 2011) and are biased in their categorization of non-Western languages and literatures (Higgins, 2016; Howard & Knowlton, 2018; Kua, 2004). In addition, both systems struggle to represent topics related to women and women's studies, and these topics are often restricted to limited sets of categories that are scattered across the classification scheme (Intner & Futas, 1996; Olson & Ward, 1997). We thus apply our methods to two case studies of bias. The first case study measures Western bias, or bias in favour of Western culture, in the categories of the LCC and DDC. The second measures gender bias and compares the representation of books written by women to books written by men in both systems.

Our work systematically quantifies the extent of bias within the two library classification systems that we consider. For institutional category systems such as these, quantifying bias is important because a quantitative measure can be used to identify the parts of a system that show the strongest bias and are therefore most important to consider when proposing future improvements to the system. Quantifying bias in category systems is also important because categories can influence perception and behaviour (Goldstone et al., 2001; Loehrlein, 2012), and it is therefore important to understand the extent to which a system might bias its users' understanding of the items that it categorizes. A third benefit of a quantitative approach is that it allows for the comparison of bias across two or more related classification systems, and we illustrate this by comparing the LCC and the DDC. Finally, our quantitative approach can be applied at a relatively large scale, and therefore allows us to analyze many more items and categories than a single researcher would be able to process on their own.

We begin in the next section by providing background on the LCC and DDC. Next, we distinguish between category bias, or bias reflected in the internal nodes of a hierarchical system, and item bias, or bias reflected in the distribution of items to be classified. We then focus on Western category bias and item gender bias in particular and present quantitative analyses that measure and compare the extent to which these biases are present in the LCC and DDC. We finish by demonstrating how our approach generalizes to other category systems through a case study of category bias in WordNet. We discuss limitations and future directions, including a broader perspective of how our methods could be used to quantify cultural and individual differences in diverse category systems.[2]

[2]Code for this project can be found at https://osf.io/wd9c6/.

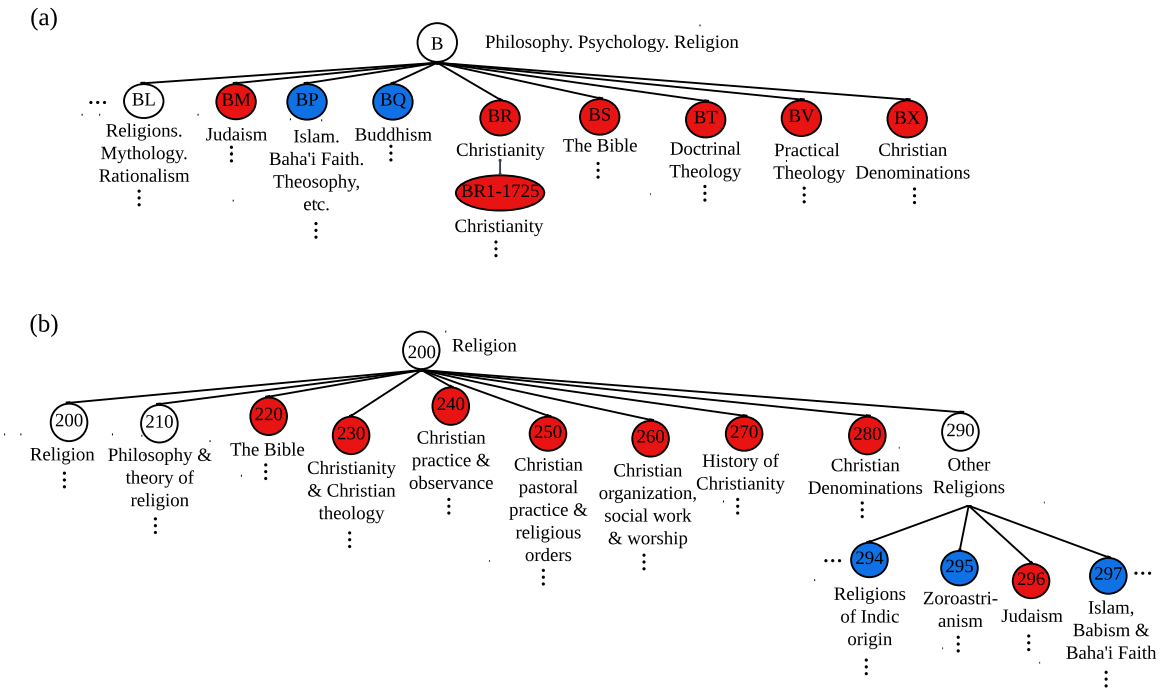


Figure 4.1: An example of how Religion is categorized in (a) the LCC and (b) the DDC. Red categories are tagged as Western and blue as non-Western. White categories cannot be tagged as solely Western or non-Western. For each library classification system, only a subset of the relevant categories is shown.

4.1 Library Classification Systems

The LCC and DDC are both hierarchical systems that contain a set of main classes, each corresponding to a different discipline. These main classes are recursively subdivided into increasingly more specific subcategories that classify smaller and smaller subsets of the literature. In the LCC, there are 21 main classes, and classification numbers are alphanumeric. There is no formal limit on the number of subcategories a category can have. Figure 4.1a illustrates this system with the classification of religious literature. The DDC has 10 main classes, and classification numbers are entirely numeric. Each category can have a maximum of 10 children. Figure 4.1b illustrates the classification of religious literature in the DDC. The category hierarchy is represented by the position of the digit that differentiates a category. The DDC has stronger structural constraints than the LCC as it enforces the strict upper limit on the number of subcategories (Svenonius, 2000). As a result, the LCC tends to be flatter and the DDC deeper.

4.1.1 Representing Library Classification Systems as Trees

We used a tree structure representation to capture the hierarchical structure of the LCC and the DDC. Every node in the tree represents a different category in the classification, storing a category's name, label, and the books it classifies. The category label indicates the range of classification numbers that fall under it. The branches between a parent and its children represent the hierarchical relationship between a category and its subdivisions.

We used the OhioLINK Circulation Data[3], a large publicly available data set of books and their circulation, to represent the books in our analysis of bias in library classification systems. OhioLINK contains 6.78 million MARC bibliographic records[4] for books and manuscripts in the Ohio academic libraries (OhioLINK Collection Building Task Force et al., 2011). These bibliographic records include the LCC and DDC classifications assigned to a book. Only books that had both an LCC and a DDC classification were kept, resulting in 3.31 million books. These books were placed into the DDC and LCC tree structures using their relevant classification numbers. For each book, we found the most specific category associated with its classification number, and then recursively added it to each parent category until the top of the tree was reached. This ensured that each parent category contained all the books of its subcategories. For each book, we stored its author and circulation statistics. See Appendix A.2 for more details.

4.2 Defining Structural Biases

As suggested earlier, bias can be defined as a preference for one group over another. This definition can be applied separately to the categories or internal nodes that lie within a hierarchical system and to the items that are found at its leaves, and we refer to these two cases as **category bias** and **item bias**, respectively. For us, the items are books, but in general an item could be any entity (e.g. a person, thing, concept, or action) that a classification system categorizes.

4.2.1 Category Bias

Assume that blue and red represent distinct but comparable labels that can be applied to the internal nodes of a hierarchical classification system. For example, in Figure 4.1, red categories are related to Western topics and blue categories are related to non-Western topics. Category bias occurs when the system gives preferential treatment to one group of nodes (e.g. red nodes) ahead of the other. We assume for now that an unbiased system treats red and blue categories identically.

[3]This work contains information from OhioLINK Circulation Data which is made available by OCLC Online Computer Library Center, Inc. and OhioLINK under the ODC Attribution License.

[4]The MARC (machine-readable cataloguing) 21 Format for bibliographic data is a digital format used to describe items catalogued by libraries.

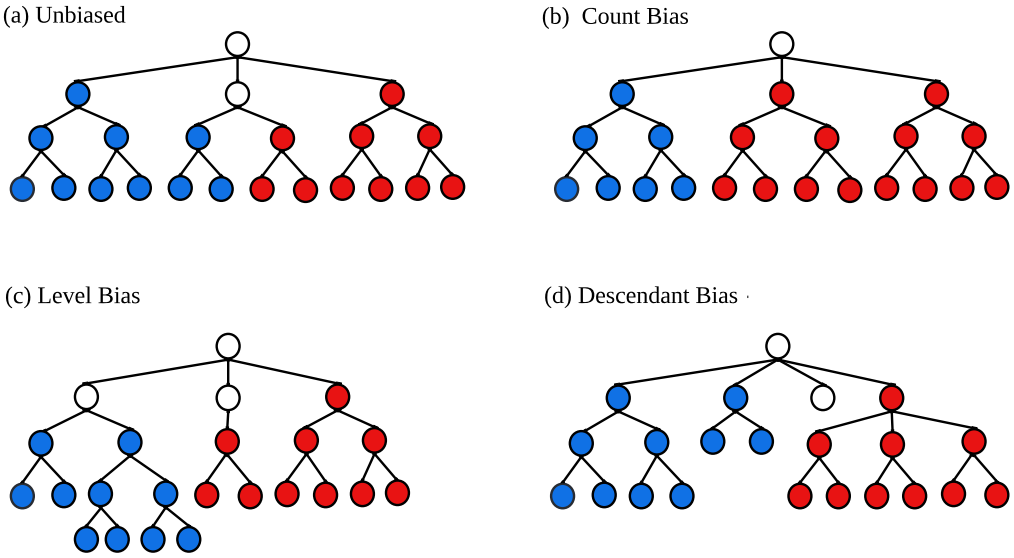


Figure 4.2: An illustration of (a) an unbiased classification, (b) a classification with count bias, (c) a classification with level bias, and (d) a classification with descendant bias. The three category biases demonstrate bias against blue categories and in favour of red categories. White categories cannot be categorized as either red or blue.

Three kinds of category biases are illustrated in Figure 4.2: count bias, level bias, and descendant bias. Category count bias (Figure 4.2b) occurs when there are more red than blue categories in a classification scheme. Thus, more classification space is devoted to red categories. Category level bias (Figure 4.2c) occurs when red starting categories occur higher in a classification structure than blue starting categories. A **starting category** (or starting node) is the first category in a classification sub-tree that can be labelled as red or blue. Starting categories that are higher in the classification scheme are conceptualized as more general or important than those that are deeper. Finally, descendant bias (Figure 4.21d) occurs when red starting categories have more descendants than blue starting categories on average. In other words, red categories are privileged by having more fine-grained category divisions.

The three biases in Figure 4.2 may often be correlated in practice — for example, if there are more red categories (category count bias), it is likely that red starting categories will have more descendants (descendant bias). The biases, however, are conceptually distinct and can be separated in principle. For example, Figure 4.2d shows that even when node counts are held constant for red and blue, it is possible to observe level bias (in favour of blue) and descendant bias (in favour of red). We therefore propose that considering the three biases individually is worthwhile, as they highlight different aspects of category bias.

4.2.2 Item Bias

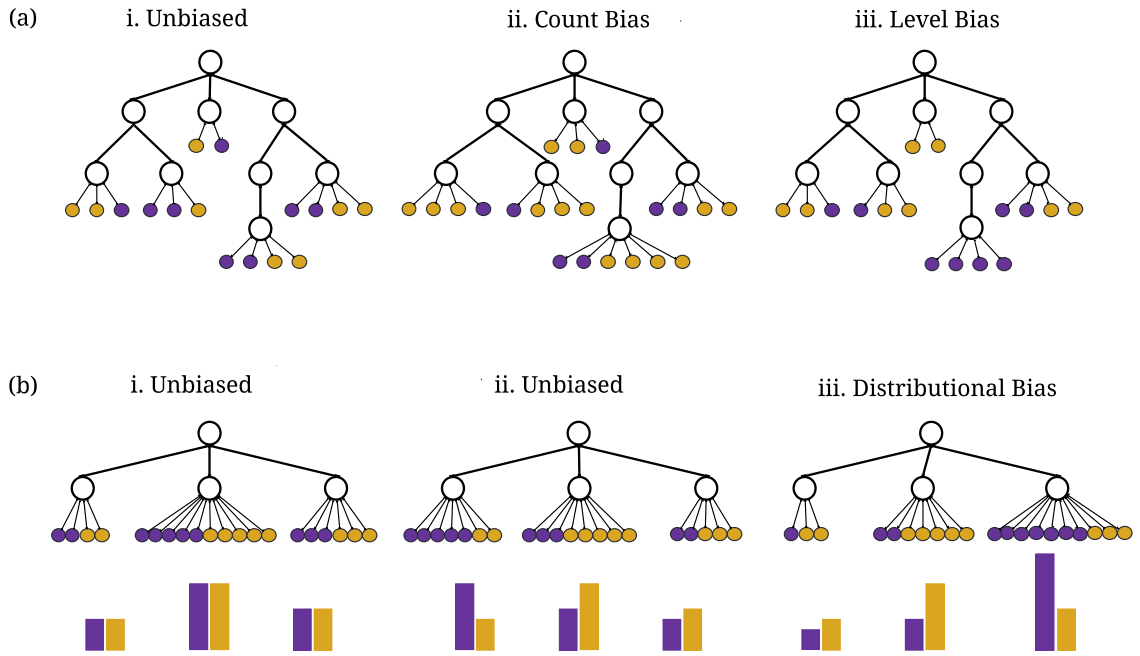


Figure 4.3: An illustration of (a) item count bias and level bias, and (b) distributional bias. In (a), there is (i) a classification with no level or count bias; (ii) a classification with count bias; and (iii) a classification with level bias. In (b), there are (i, ii) two examples of categorizations without distributional bias, and (iii) an example of distributional bias. The shapes of the distributions of purple and gold items across the three lowest categories in the tree are shown with bar graphs. All 3 item biases demonstrate bias against purple items and in favour of gold items.

Instead of assigning the internal nodes of a hierarchical system to groups, assume now that purple and gold represent distinct but comparable labels that can be applied to a set of items. For example, gold items could be books written by men and purple items could be books written by women and non-binary people. Figure 4.3 shows several examples in which the items are shown as small circles at the leaves of a classification hierarchy. Item bias occurs when the system gives preferential treatment to one group of items (e.g. gold items) ahead of the other. As before, we assume that an unbiased system would give equal treatment to gold and purple items.

Three kinds of item biases are illustrated in Figure 4.3: count bias, level bias, and distributional bias. Item count bias (Figure 4.3aii) occurs when there are more gold than purple items classified by a system. Item level bias (Figure 4.3aiii) is similar to category level bias, and occurs when gold items tend to be found higher in the classification tree than purple items. Finally, distributional bias (Figure 4.3biii) occurs when gold items are distributed more broadly across the classification system than are purple items. In other

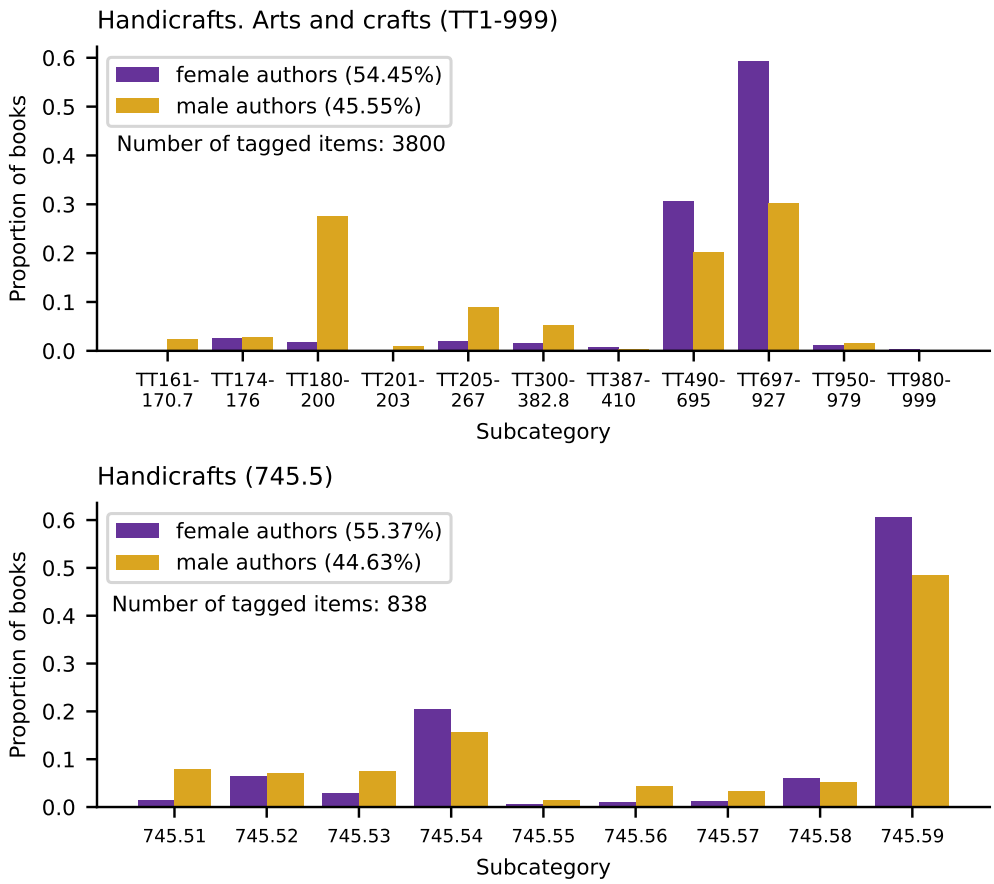


Figure 4.4: The distributions of male and female authors across "Handicraft" subcategories in the LCC and DDC. Specifically, these are subcategories of "Handicrafts. Arts and Crafts" in the LCC (top) and "Handicrafts" in the DDC (bottom). The number of tagged items is the number of books that have an author with a known gender in the dataset. See Appendix A.1 for the full set of subcategory names.

words, purple items are more restricted to a limited part of the classification scheme than are gold items.

Distributional bias can be diagnosed by comparing the shape of the distributions of gold items to the shapes of the distributions of purple items. In Figure 4.3biii, the distribution of the gold items across the three categories at the lower level of the hierarchy is relatively flat, but the purple distribution is concentrated on the third of the three categories. In contrast, Figure 4.3bi and Figure 4.3bii show distributions of purple and gold authors that do not suffer from distributional bias. In Figure 4.3bi, the shape of the distribution of purple authors is identical to the shape of the distribution of gold authors. In Figure 4.3bii, although the distributions are not identical, they are shuffled versions of one another and therefore equally as flat.

Figure 4.4 shows what distributional bias can look like in library classification systems. In the LCC, the books by men are more evenly spread across the subcategories of

"Handicrafts. Arts and crafts," than are the books by women. The books by women are predominantly restricted to two subcategories, "Home arts. Homecrafts" (TT697-927) and "Clothing manufacture. Dressmaking. Tailoring" (TT490-695). In the DDC's equivalent category, "Handicrafts," the difference between the shape of the distribution of books by men and books by women does not appear to be as big.

Item biases can arise for at least two different reasons. They can be attributed to how the classification system distributes items, to external social biases found in the set of items to be classified, or to both. We therefore distinguish between two kinds of item biases: external biases and internal biases. External biases are biases found in the set of items to be classified. These biases are the result of broader social forces and are not imposed by the classification system itself. Internal biases are biases in the ways items are assigned to categories by the classification system. These biases are imposed on the items by the classification system itself.

In the context of gender bias in library classification systems, item count bias is a clear example of external bias as it comes from unequal numbers of purple and gold authors in the set of items to be classified. This bias could arise because society provides more opportunities for men to write books than women, or because libraries are more likely to acquire books written by men (B. Quinn, 2012), or both. How to characterize distributional bias and level bias is less clear. For example, in the case of distributional bias in a library classification system, it might be that purple and gold authors write about equally diverse sets of topics, but that the interests of purple authors are given limited space in the classification system. Thus, it could be that the distributional bias is an example of an internal bias. It could also be that the topics addressed by purple authors are genuinely less diverse than the topics addressed by gold authors because of social pressures that encourage purple authors to specialize in a limited set of areas. Thus, the distributional bias could also be an example of an external bias. Similarly, level bias against purple items could be the result of the classification system placing topics of interest to purple authors lower in the tree (internal bias), or social pressures pushing purple authors to write in smaller, more niche categories (external bias). Although the origins of level bias and distributional bias may not be clear, both biases are worth investigating as they can provide insight into how different groups are represented in a classification scheme, regardless of whether this difference in representation is imposed by the system itself or the result of external forces.

4.3 Study 1: Western Category Bias

One of the most well-documented biases in the LCC and the DDC is Western bias, or bias in favour of Western culture. In this study, we use our definitions of category bias to quantify the degree of Western bias in both systems.

4.3.1 Methods

To study Western category bias, we selected main classes in the classification trees that related to history, religion, language, and literature as these categories tended to have category names that could be identified as Western or non-Western. Other topics such as philosophy, although having the potential to exhibit Western biases, did not have categories that could clearly be identified as Western or non-Western. For the LCC, we used main classes B, D, E, F, and P, which are named "Psychology, Philosophy, Religion," "World History," "History of the Americas," "History of the Americas (local)", and "Language & Literature" respectively. For main class B, we only used the categories that classified religious literature. For the DDC, we used main classes 2, 9, 4, and 8, which are named "Religion," "History," "Language", and "Literature." We combined Language and Literature into a single topic because in the LCC many of the categories in main class P did not allow for easy separation of the two. Similarly, we grouped together all main classes related to history. We thus investigated Western bias across three topics: religion, history, and language & literature.

We manually tagged the categories selected for each topic as Western or non-Western, drawing on distinctions that have been previously suggested in the literature.[5] Still, the tagging process is inevitably subjective, and in cases where a label of Western or non-Western was unclear, we left the category untagged, aiming for precision over recall. This somewhat limits the results, as there might be cases where a country or language, etc. falls into a category with a clear label in the LCC but not the DDC or vice versa.

The classes related to history tended to be divided into categories based on geographical and political divisions such as country or continent. We therefore used a list of Western countries that were defined based on a cultural definition of Western as opposed to a political, economic, or geographical definition (Hall, 2018; Lamo de Espinosa, 2017; Trubetskoy, 2017). For example, Australia tends to be considered a Western country despite not being geographically in the Western hemisphere. 68 countries, about 35% of the world's countries, were included in the list of Western countries, and we assumed that countries left off the list were non-Western. For each history-focused main class, we considered all categories associated with a country and tagged them as Western or non-Western based on the list. The tagged category became a starting category. If a category represented a group of countries (i.e. a category for a continent or a region) and all the categories beneath it shared the same tag, then that broader category became the starting category and inherited the tag. Similarly, every category under a starting category inherited the starting category's tag.

In the language and literature-related classes, some categories were related to regional divisions, like the history-focused classes, so we based our tagging on the list of Western countries used previously. Examples of these categories include "German literature" and "Languages and literature of Eastern Asia, Africa, Oceania." Some categories were related

[5] A list of category tags can be found in the linked code repository.

to language families, so we considered where these languages or language groups originated from to make the tagging choice. "Romance languages" is one example. The main deviation from the tagging method used for history was how we tagged Indigenous languages and literature from North America, South America, and Oceania. Consistent with our cultural definition of the Western concept (Hall, 2018), we tagged them as non-Western even if they originated from a country or region that is listed as Western.

Finally, for the main classes covering religion, we mostly tagged Abrahamic religions as Western and other religions as non-Western. The few exceptions included tagging Scientology as Western and Islam as non-Western. Islam is an Abrahamic religion, but we made the conservative decision to tag Islam as non-Western, because the opposite decision would likely increase any Western bias that we might find. The categories "Doctrinal Theology" and "Practical Theology" were tagged as Western because they have only been used to classify literature on Christianity (Zins & Santos, 2011).

4.3.2 Results

In total there were 3,009 categories on the topics of religion, language & literature, and history in the LCC, and 13,536 in the DDC.[6] Based on the tagging method, 86.3% of categories could be tagged as either Western or non-Western in the LCC, and 91.4% in the DDC. We refer to tagged categories as "nodes" to be consistent with our use of a tree representation.

Count Bias

To compute category count bias, we compared the percentage of nodes tagged as Western to the percentage of nodes tagged as non-Western in both classification systems. In the LCC, 2598 nodes were tagged (62.24% Western). In the DDC, there were 12,372 tagged nodes (69.35% Western).

Figure 4.5 shows analogous results for each of the three individual topics. For each topic, there is a higher percentage of nodes tagged as Western than non-Western. In the LCC, religion has the highest percentage of Western nodes. In the DDC, history and religion had percentages that were almost equally high. For all topics, the DDC had a higher percentage of Western nodes than the LCC. To test the statistical significance of this result, we randomly assigned all nodes a Western or non-Western label with equal probability. We repeated the process 10, 000 times, using the proportion of the times the absolute difference between Western and non-Western node counts was greater than or equal to the observed absolute difference as the p-value. For all topics in the DDC, and religion and history in

[6]In the LCC, 571 of the categories were from the topic Religion, 920 from Language & Literature, and 1518 from History. In the DDC, 2420, 7758, and 3358 categories were from Religion, Language & Literature, and History, respectively.

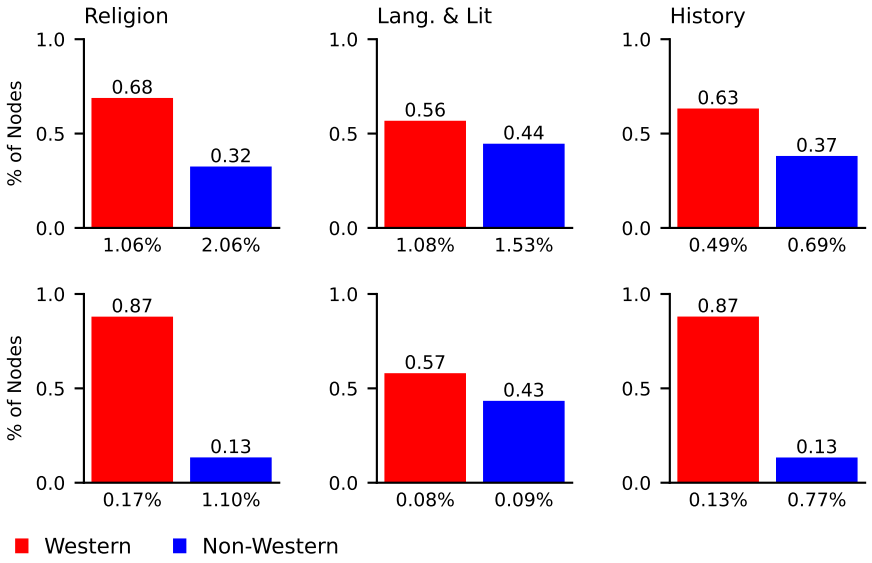


Figure 4.5: Category count bias results for religion, language & literature, and history in the LCC (top) and the DDC (bottom). The bars are labelled on top with the proportion of nodes they represent. The labels below the x-axis are the mean % of books per node.

the LCC, $p < .001$. For language & literature in the LCC, $p = .003$. All category count biases were therefore statistically significant.

We have conservatively assumed that an unbiased system has an equal number of Western and non-Western nodes, but this assumption could be adjusted using statistics such as population sizes or the percentage of Western countries. If anything, these statistics tend to suggest that an unbiased system should devote more space to non-Western than to Western nodes. For example, Africa and Asia accounted for 75% of the world's population in 2022 (United Nations, DESA, Population Division, 2022). Western category count bias is substantial relative to a conservative 50-50 baseline, and would be even stronger relative to a baseline favouring non-Western nodes.

Library classification systems follow the principle of literary warrant, which means that their structures are derived from and justified by the body of literature that they classify (Svenonius, 2000). Based on this principle, it could be argued that there are more Western categories because there are more Western books that need to be classified. To test this idea, we calculated the mean rate of books per Western node and non-Western node in each system. These rates are reported as labels below the x-axis of Figure 4.5. An unbiased system might be expected to have relatively equal rates of books per node. We found that language & literature in the DDC and history in the LCC have relatively equal rates of books per node for Western and non-Western nodes. Otherwise, there tend to be more books per non-Western node than per Western node. The difference in rates is most pronounced for religion (0.17% vs. 1.10%) and history (0.13% vs. 0.77%) in the DDC. These

Table 4.1: Summary of circulation statistics for the LCC and DDC overall (O), and for topics religion (R), language & literature (LL), and history (H). Circulation statistics are from 2007.

		% of books in circulation.		**% of books taken out**		**Rate of circulation**	
		W	NW	W	NW	W	NW
Overall	LCC	0.92	0.97	0.27	0.27	2.78	2.62
	DDC	0.92	0.97	0.25	0.27	2.71	2.70
Religion	LCC	0.93	0.98	0.28	0.44	2.69	3.13
	DDC	0.93	0.98	0.28	0.44	2.65	3.12
Lang. & Lit.	LCC	0.92	0.97	0.25	0.19	2.78	2.05
	DDC	0.92	0.97	0.23	0.18	2.72	2.08
History	LCC	0.92	0.97	0.28	0.28	2.86	2.68
	DDC	0.93	0.98	0.29	0.31	2.79	2.89

findings suggest that in some cases, especially in the DDC, the higher Western category count cannot entirely be accounted for by literary warrant.

Similarly, it could be argued that there are more Western nodes because Western books are in higher demand than non-Western books. To explore this idea, we compared the circulation of books classified in Western nodes to those in non-Western nodes. Circulation statistics were drawn from the OhioLINK circulation data, and for each book we extracted three pieces of information: (i) whether the book was in circulation (i.e. available for borrowing) in 2007, (ii) whether the book was borrowed in 2007, and (iii) how often the book was borrowed in 2007. Mean values of all three variables are shown in Table 4.1. Across all topics and for both classification systems, a larger percentage of books classified under non-Western nodes are in circulation than books classified under Western nodes. For religion, a larger percentage of circulating non-Western books were taken out than circulating Western books in 2007 for both the LCC and DDC. In addition, among all religion books that were taken out, non-Western books had a higher mean rate of circulation. The opposite was true for language & literature, where a larger percentage of circulating Western books were taken out, and Western books had a higher mean rate of circulation. For history, these statistics varied slightly but were relatively similar for Western and non-Western books. Overall, the circulation statistics do not seem to justify the large discrepancy between Western and non-Western node counts.

Level Bias

To measure category level bias, we first compared the mean depth of Western starting nodes (starting categories) to the mean depth of non-Western starting nodes. If a group of nodes

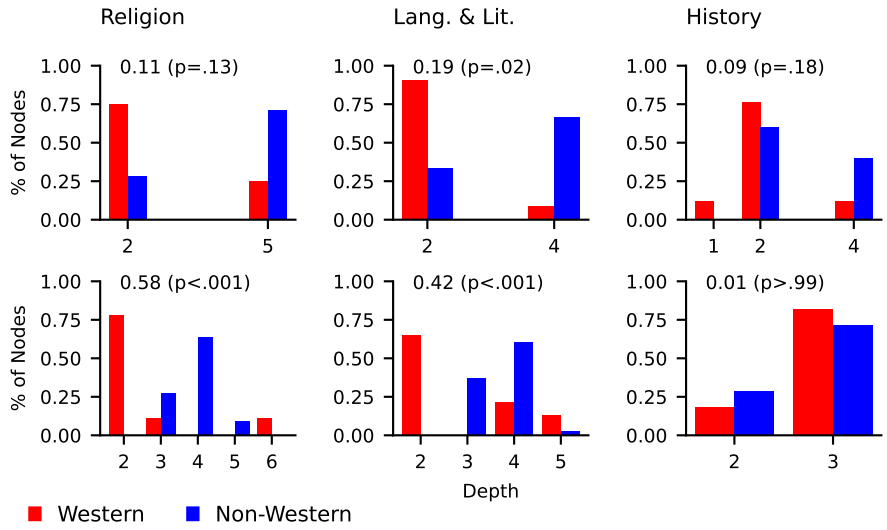


Figure 4.6: Level bias statistics for the topics religion, language & literature, and history in the LCC (top) and the DDC (bottom). Each plot is annotated with the JSD between the Western and non-Western distributions and the significance of the JSD in brackets. Significance was calculated with a permutation test.

has a greater mean starting depth, then this implies that, on average, the nodes occur deeper in the tree than a group with a lesser mean starting depth. The LCC has a mean Western starting node depth of 2.28 and a mean non-Western starting node depth of 3.35. In the DDC, these depths are 2.77 and 3.57.

For each topic, Figure 4.6 shows the distributions of starting nodes over classification tree depths. To quantify the difference in distributions over Western and non-Western starting depths, we computed the Jensen-Shannon Divergence (JSD) between these distributions. To test the statistical significance of the results, we performed permutation tests. For each topic, the depth labels were shuffled among the Western and non-Western nodes to create randomized depth distributions. This shuffling was carried out 10,000 times, and the proportion of times the JSD between the randomized Western and non-Western depth distributions was greater than or equal to the actual JSD was used as the p-value.

For all topics in the LCC, the mean depth of non-Western starting nodes is greater than the mean depth of Western starting nodes. The same finding applies to religion and to language & literature in the DDC. The one exception is history in the DDC, where the mean Western depth is slightly greater than the mean non-Western depth. For the two systems, the greatest divergences between Western and non-Western depth distributions occur for language & literature (LCC) and religion (DDC). In both classification systems, the divergence for history is not statistically significant ($p = .18$ and $p > .99$ in the LCC and DDC, respectively). In the LCC, the divergence statistic was significant only for language & literature ($p = .02$), and for the DDC both religion ($p < .001$) and language & literature

Table 4.2: The probability that a randomly selected non-Western starting node (NW) is deeper in the classification than a randomly selected Western starting node (W), given the two nodes are not equal.

	Probability		**p**	
	LCC	**DDC**	**LCC**	**DDC**
Overall	0.87	0.82	$< .001$	$< .001$
Religion	0.89	0.89	.11	.02
Lang. & Lit.	0.95	0.76	.01	.01
History	0.86	0.36	.10	.72

($p < .001$) produced significant results. Religion did not demonstrate significant category level bias in the LCC ($p = .13$).

The depths for the LCC and the DDC are not directly comparable, because the DDC tends to have a larger set of starting node depths than the LCC (e.g., the starting nodes for religion are spread across 5 different tree depths in the DDC versus just 2 in the LCC). We therefore computed an alternative measure of category level bias that uses the probability that a randomly selected non-Western starting node is deeper in a classification system than a randomly selected Western starting node. A Western and a non-Western starting node were randomly sampled 10,000 times. The depths of the two nodes were compared to determine the number of times the non-Western one was deeper than the Western one and vice versa (ties were ignored). The resulting statistic measured the probability that a non-Western starting node would be deeper in a tree than a Western starting node, given they were not of the same depth. The results are shown in Table 4.2. For every topic except history in the DDC, it is more likely that a randomly selected non-Western starting node is deeper in the tree than a randomly selected Western node. Western nodes for history in the DDC have a higher chance of starting deeper in the tree. Based on this statistic, the LCC displays a stronger level bias than does the DDC. To test for significance, we performed a permutation test by randomly shuffling the depths among the Western and non-Western nodes and recalculating the probability that a non-Western node was deeper in the tree. This was repeated 10,000 times, and the proportion of times the absolute value of the difference between 50% and the recalculated probability was greater than the actual difference was used as the p-value. The results are in Table 4.2. The significance by topic mirrored the significance of the initial divergence statistic for category level bias.

Descendant Bias

We measure descendant bias by comparing the mean number of descendants per Western starting node to the mean number of descendants per non-Western starting node. We also recorded the number of starting nodes and the mean percentage of books per starting

Table 4.3: Summary of descendant bias for the LCC and DDC overall (O), and for topics religion (R), language & literature (LL), and history (H).

		Start node count		**% of books per start node**		**Mean descendants per start node**	
		W	NW	W	NW	W	NW
Overall	**LCC**	36	26	2.8	3.8	43.9	36.7
	DDC	43	56	2.3	1.8	198.5	66.7
Religion	**LCC**	8	7	12.5	14.3	44.5	23.3
	DDC	9	11	11.1	9.1	196.7	22.5
Lang. & Lit.	**LCC**	11	9	9.1	11.1	29.2	27.9
	DDC	23	38	4.3	2.6	183.7	82.2
History	**LCC**	17	10	5.9	10.0	53.2	54.1
	DDC	11	7	9.1	14.3	231.0	51.9

node. All statistics were computed for the LCC and DDC overall, as well as separately for religion, language & literature, and history. The results are shown in Table 4.3. To test for significance, the Western and non-Western tags were randomly shuffled among the starting nodes and the absolute difference in Western and non-Western descendant means was recomputed. This process was repeated 10,000 times, keeping track of the number of times the recomputed difference in means was greater than the absolute value of the observed difference in means.

There is evidence of descendant bias in the DDC but not in the LCC. In the LCC, the difference in Western and non-Western descendant means was minimal and not statistically significant ($p = .44$ for religion, .92 for language & literature, and .97 for history). In contrast, all topics in the DDC had, on average, at least double the number of descendants per Western starting node than per non-Western starting node. Descendant bias was statistically significant for the topics language & literature ($p = .02$) and religion ($p < .001$) but not for history ($p = .62$). Although history had the largest absolute difference in means, its lack of significance can be attributed to the fact that it has only 18 starting nodes, and one of them ("History of the North Americas") has a descendant count that is much larger than the other descendant counts.

The LCC has more Western than non-Western starting nodes. The DDC exhibits the opposite trend except in the case of history. The fact that there are more non-Western starting nodes for language & literature and religion could account for the lower non-Western descendant means in the DDC. However, the permutation test keeps the number of each type of starting node constant. Descendant bias is thus not an artifact of the number of starting nodes. The mean percentage of books per starting node is roughly equal between Western and non-Western ones. The biggest difference is found for history in the DDC.

However, the difference shows more books per non-Western node than Western (14.3% vs 9.1%). Thus, in the DDC, descendant bias is not fully explained by the idea that Western starting nodes have more books than non-Western starting nodes

4.3.3 Summary

We found evidence of all three kinds of Western category bias in both the LCC and the DDC. Our results also suggest that the DDC encodes a higher degree of category bias than does the LCC, as the DDC has a higher category count and descendant bias, but the LCC has higher level bias. Of the three topics considered, religion tended to have the highest amount of bias in both systems. The amount of bias in history and language & literature varied across metrics. Language & literature did not display much count bias but displayed significant level bias in both systems, and displayed significant descendant bias in the DDC. History displayed significant descendant bias in the DDC and significant count bias in both systems. However, level bias for history was minimal, especially in the DDC where level bias favoured non-Western nodes as opposed to Western ones.

Our finding that both the LCC and DDC show Western category bias is expected given previous work on biases in library classification (Knowlton, 2005; Zins & Santos, 2011), but our approach departs from previous work in attempting to systematically quantify the nature and extent of this bias. For example, religion is known to be a topic that shows substantial Western bias (Fox, 2019), but to our knowledge previous studies have not systematically quantified the level of bias observed for religion relative to the bias observed for other topics. Similarly, there have been suggestions that the DDC shows greater Western bias than does the LCC (Sultanik, 2022), but prior work has not provided comprehensive quantitative analyses to support this claim.

4.4 Study 2: Item Gender Bias

We now turn from Western bias to gender bias and use our definitions of item bias to quantify gender bias at the item level in both the LCC and the DDC.

4.4.1 Methods

To study item-level gender bias, we worked with all books in our dataset that were classified under both the LCC and the DCC. We only considered books with non-empty author fields in their MARC records.[7] Each of these books was tagged with the gender of its author.

[7]Author names are stored in the Main Entry-Personal Name field of a book's MARC record. This field records the person mainly responsible for the work (Library of Congress, 2022), whether they are the primary author in a multi-authored work or the editor of an anthology. For simplicity, we use the term "author" in all cases.

To determine an author's gender, we used data from the author-name-index and author-gender tables created and kindly shared by Ekstrand and Kluver (2021) as part of their book data integration pipeline, PIReT Book Data Tools. These tables store processed versions of author name and gender data from the Virtual International Authority File (VIAF). The VIAF stores author information, including the variants of an author's name and their gender. As discussed by Ekstrand and Kluver (2021), the VIAF, unfortunately, codes gender as binary and does not code for non-binary gender identities. Each author record is coded as either male, female, or unknown. For now, our analysis is thus limited to analyzing item-level biases between male and female authors. When more accurate author data is available, the same analyses can be performed including non-binary gender identities.

There is no linking identifier between a MARC record and its author's VIAF record. We followed the method for linking records used in the PIReT Book Data Tools (Ekstrand & Kluver, 2021). At a high level, string-matching was used to tag a book with its author's gender, and there are three main cases where an author's gender cannot be determined. The first is when a book's author matches to a VIAF record with the gender code "unknown." The second is when a book's author matches multiple VIAF records with conflicting known gender identities and thus has an ambiguous gender code. The third is when a book's author does not match any author record in the VIAF dataset. We discarded books in any of these three cases from our item-level analysis.

4.4.2 Results

2.54 million of the 3.31 million MARC records had a non-empty author field, and 1.95 million of these could be tagged with an author's gender. Table 4.4 contains a breakdown of the record-matching process. Less than 1% of the records were tagged as ambiguous. Only 5% of the MARC records could not be linked to any VIAF record. These results are similar to results achieved by the PIReT Book Data Tools. In all the datasets they were applied to, somewhere between 3.3% and 6.9% of book records could not be matched to a VIAF record (Ekstrand & Kluver, 2021).

Table 4.4: Breakdown of the results of the record-matching process. No author (MARC) refers to the percentage of MARC records with an empty or missing Main Entry-Personal Name field. No author (VIAF) refers to the percentage of MARC records with a Main Entry-Personal Name field that does not match a record in the Author-name-index table.

Subset	**% of records**
No author (MARC)	23.12
No author (VIAF)	5.00
Unknown gender	12.36
Ambiguous gender	0.55
Known gender	58.97

Count Bias

To measure item count bias, we compared the percentage of female authors to the percentage of male authors in the books classified by the Ohio academic libraries. We found that female authors are substantially underrepresented. In the set of items with both a DDC and LCC, 16.25% of the authors are coded as female, and 83.75% are coded as male. This differs from the percentages of men and women in the VIAF dataset, where 28.55% of authors with a known gender identity are coded as female and 71.45% as male. Because males and females account for roughly equal proportions of the general population, any departure from an even split provides evidence of external bias.

Table 4.5: 2007 circulation statistics for male versus female authors.

	F	M
% items in circulation	0.95	0.95
% items circulated	0.37	0.29
Rate of circulation	3.30	2.91

We also compared the circulation of books by men to the circulation of books by women. The results are shown in Table 4.5. The percentage of books by women in circulation is equal to the percentage of books by men. In 2007, 37% of circulating books by women were borrowed versus 29% of books by men. Similarly, borrowed books by women were taken out more times on average (3.30) than borrowed books by men (2.91). These findings reveal that demand alone cannot explain the under-representation of female authors and suggest that there are other forces that systematically reduce their representation.

Level Bias

We measured item level bias by determining the most specific (deepest) category each book in the Ohio academic libraries was assigned to. We then compared the depth at which books written by women tended to be classified to the depth at which books written by men tended to be classified. In the LCC, the mean depth of male and female authors is 4.55. The median for both is 4. In the DDC, the mean depth of female authors is 4.65, and the mean depth of male authors is 4.45. The medians for women and men are 5 and 4, respectively.

We plotted the distributions of books over the classification tree depths in Figure 4.7 and computed the difference in the mean depths of books by men and women. In the LCC, the difference is 0.004, and in the DDC it is 0.191. To test the significance of the results, we performed a permutation test. We shuffled the classification depths among books and recomputed the difference between the mean depth of books by men and the mean depth of books by women. This was repeated 1000 times, and the proportion of times the random

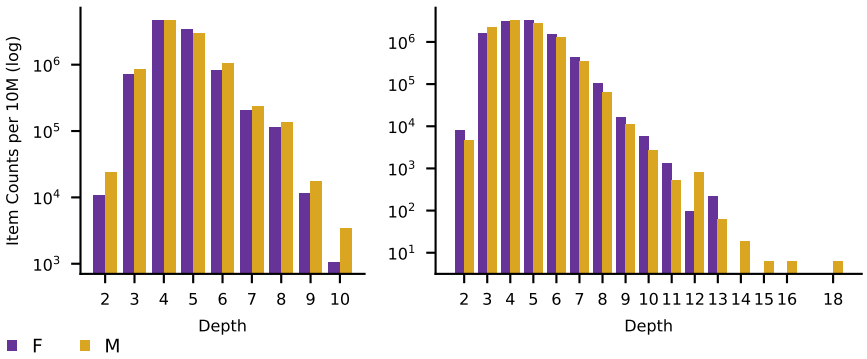


Figure 4.7: Item level gender bias in the LCC (left) and DDC (right). Data is scaled to counts per ten million.

difference in means was greater than or equal to the actual difference in means was used as the p-value. In the LCC $p = .61$, and in the DDC, $p < .001$. For the books in the Ohio academic libraries, LCC classifications did not yield a significant difference between the mean depth of classification for books written by men and the mean depth of books written by women. For the DDC, the difference is statistically significant, but small.

As an alternative statistic for item level bias, we calculated the probability that a randomly selected female-authored book is deeper in the tree than a randomly selected male-authored book. A male-authored and a female-authored book were randomly sampled 10,000 times. The depths of the two books were compared to determine the number of times the female-authored one was deeper in the tree (ties were discarded). The resulting statistic measured the probability that a female-authored book would be deeper in a tree than a male-authored book, given they were not of the same depth. In the LCC, this probability is 51%, and in the DDC it is 57%. To test the significance of these results, we performed a permutation test that measured the proportion of times the absolute value of the difference between the recalculated probability and 50% in the permuted samples is greater than the actual difference. In the LCC, $p = .29$, and in the DDC, $p < .001$. Again, the results do not indicate the presence of any item level bias for the Ohio library books when classified using the LCC and only minimal item level bias when using the DDC.

Distributional Bias

Distributional bias is evident when the distribution of male authors across children of a given node tends to be flatter than the distribution of female authors. To test for this bias, we collected every node that had at least 100 Ohio library books, at least 2 children, and both male and female authors. This approach yielded 822 LCC nodes and 2,832 DDC nodes that could be used to compare the distributions of books by male and female authors. See Appendix A.3 for a breakdown of the nodes that could not be used in this analysis.

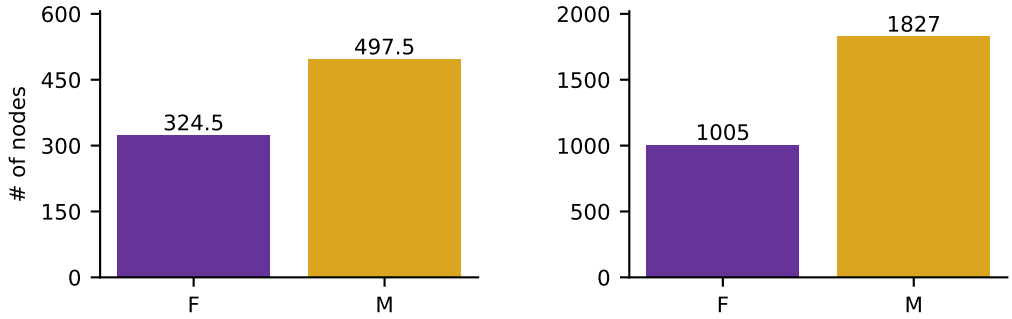


Figure 4.8: Distributional Bias in the LCC (left) and the DDC (right). The gold columns (M) show the number of nodes for which the distribution of male authors is flatter than the distribution of female authors, and the purple (F) columns show the number of nodes that yielded the opposite result. Ties were handled by adding a count of 0.5 to both columns.

To compute the distribution of male authors for each node, the number of male authors in each direct child node was divided by the total number of male authors across all child nodes. We did not use the total number of male authors in the parent node because not all items in a parent node are classified into one of its children. The same was done for the female author distribution. For example, in Figure 4.3biii there are ten purple books, and one is assigned to the first child node, two to the second, and seven to the third. Thus, the distribution of purple authors for that node is [0.1, 0.2, 0.7]. To determine which distribution was flatter at each node, we compared the Shannon entropy of female-authored books against male-authored books, using the Miller-Madow estimator to correct for small-sample bias (Büth et al., 2025; G. Miller, 1955).[8]

Figure 4.8 shows the number of nodes for which the male distribution was flatter than the female distribution. In both the LCC and the DDC, the distribution of male authors among a node's children tended to be flatter than the distribution of female authors. The effect is slightly stronger in the DDC, as the relative difference in size between the two counts is 1.82 as opposed to 1.53. To test the significance of the results reported above, a permutation test was performed. The entire set of author gender labels was shuffled among the classified items. For each node, the author gender distributions were recalculated, and the same flatness comparison was applied. In both the DDC and the LCC, the results were significant with $p < .001$.

4.4.3 Summary

The books in the Ohio academic libraries exhibit clear external item gender bias, as there are approximately five times more books written by men than by women. When these books are classified by either the LCC or the DDC, they also exhibit distributional gender

[8]In the published version (Warburton et al., 2024), we used plug-in Shannon entropy, which systematically underestimates entropy in smaller samples. Those results are reported in Appendix A.4 for transparency.

bias, as the books written by women tend to concentrate on a smaller subset of nodes than the books written by men. This bias is evident in both systems and somewhat stronger in the DDC. Finally, there is no strong evidence of level bias when the books are classified by either the LCC or the DDC.

4.5 Discussion

Category systems often contain inbuilt biases towards certain groups or concepts, and we developed methods that quantify these biases in hierarchical category systems. We tested our methods on two case studies of documented bias in the LCC and DDC. The first quantified Western bias in the internal node (category) structure of both systems, and the second quantified gender bias in the leaf nodes (items). The results confirm previous work documenting bias in the LCC and the DDC, allow for a detailed understanding of the different ways bias can manifest in the leaf nodes and internal nodes of these systems, and facilitate the comparison of bias between systems.

In the first case study, we found that language & literature in the DDC had significant category count and level bias, and that religion and history had significant count bias in both systems. These biases were in favour of Western nodes and confirm previous findings that the DDC is biased in its categorization of non-Western language and literature (Kua, 2004), and that non-Western religions and topics are under-represented in both the DDC and LCC (Westenberg, 2022; Zins & Santos, 2011). We also found that a category system that has count bias does not necessarily have level bias or descendant bias and vice versa, suggesting that the three proposed biases can be used to quantify different aspects of category bias and provide a relatively nuanced picture of how it manifests. Finally, we found that the DDC tends to show a higher degree of Western category bias than does the LCC. Specifically, there was evidence of strong category count and descendant bias in the DDC, whereas there was no evidence of descendant bias in the LCC.

In the second case study, we found that women are underrepresented in the set of books we considered and that there is a distributional bias in favour of men in both the LCC and DDC. Previous studies have documented that topics relating to women in the LCC and DDC tended to be restricted to specific categories (Olson & Ward, 1997; M. N. Rogers, 1993), and our analyses support a similar conclusion by suggesting that books by women tend to be restricted to relatively limited sections of the LCC and DDC. Despite strong evidence of item count bias and distributional bias, we do not find much evidence of item level bias in favour of men or women in either system. Like the three category biases we define, the three item biases can provide a detailed picture of the different ways in which item biases can be found in a system or the set of items it classifies.

4.5.1 Why Library Classification Systems are Biased

Our first case study found clear evidence of Western category bias, and to some extent this bias can be explained by literary warrant, or the idea that the system reflects and is justified by the literature it classifies. The system is biased because the literature is biased, and it is possible that when these systems were created, they were fully representative of the literature that they classified. Over time, however, library collections have changed, and although the LCC and DDC are both regularly updated, it seems likely that both embody stronger Western category bias than is justified by the holdings of a modern Western university library. Some initial evidence in this direction is provided by our analysis of data from the Ohio Academic Libraries. In 2007, Ohio academic libraries held more Western than non-Western books, but we found that the relative proportion of Western and non-Western categories was not justified by the relative proportion of Western and non-Western books. In particular, we found that there was a higher average number of non-Western books per non-Western category than Western books per Western category. Similarly, although in some cases there were more Western books per Western starting node than non-Western starting nodes, this was not consistently true. So, at least in the case of the Ohio Academic Libraries, the biases we find cannot solely be attributed to the books the system classifies.

A second possibility is that the biases reflect not the intrinsic content of the holdings, but the preferences of the users who access these holdings. For example, if library patrons were more likely to borrow Western books than non-Western books, it may make sense to have finer-grained categories for Western topics than for non-Western topics. Our analysis of circulation statistics, however, suggests that across the Ohio Academic Libraries in 2007, the average demand for each Western book was roughly equal to the average demand for each non-Western book. Differences in the preferences of modern library users can therefore not account for the finding that Western categories tend to be more fine-grained (i.e. include fewer books on average) than do non-Western categories.

Thirdly, the observed biases can be partially attributed to biased decisions made by the individuals who created these systems. Western bias has been widespread in Western culture over the past century, and has inevitably shaped the thinking of those who build and maintain library classification systems. The underlying psychological mechanisms that bias the decisions of librarians are likely to include mechanisms that drive biased categorization in general. One example is the out-group homogeneity effect, or the tendency to perceive out-group members as less diverse than in-group members. The descendant bias in the DDC seems to mirror this effect because finer-grained categories are used for Western (in-group) than for non-Western (out-group) topics. Like the out-group homogeneity effect, descendant bias in the DDC could be potentially attributed to greater familiarity with and exposure to Western literature over non-Western literature, or increased attention to and better memory for topics and features of literature that are relevant to the in-group (Das-

Smaal, 1990; Park & Rothbart, 1982). Category count bias, similarly, could also occur because features of in-group literature are easier to perceive and recall, and thus it is easier to differentiate this literature and create more categories.

In our second case study, our results for gender demonstrate that the three item biases in Figure 4.3 are sensitive to the preferential treatment of different groups of items. As mentioned earlier, however, these biases may be the result of external social pressures affecting the items classified by a system or may be imposed on the items by the classification system itself. The item count bias we find is clearly external to both classification systems, but it is unclear to what extent the distributional bias found is internal or external to either the LCC or DDC. Comparing the two systems provides some evidence that the item biases found in the DDC have an internal component. We found that the DDC has a somewhat stronger distributional gender bias than does the LCC and had a very slight item level bias where the LCC had none. These differences occurred even though the set of books considered was held constant across the two systems. Our results therefore suggest that some proportion of item bias is internal to the DDC, but do not allow us to tell whether the LCC is also subject to internal item biases.

4.5.2 Beyond Library Classification

Although our case studies focused on library classification, our methods are general and can potentially be applied to a broad range of hierarchical category systems. To illustrate, we apply our methods to WordNet (G. A. Miller, 1994). Both Western bias and gender bias are potentially relevant. For example, previous studies have documented Western biases in ImageNet (Liu et al., 2021; Luccioni & Rolnick, 2023), and these biases are likely inherited from WordNet, the source of the ImageNet hierarchy. However, to illustrate the range of our methods, we consider a third kind of bias. Using a procedure described in Appendix A.5, we identified synsets in WordNet that correspond to species of mammals, tagged these species as wild or domestic, and then used our methods to measure the extent to which WordNet prioritizes domestic species ahead of wild species. Although domestic mammals account for less than 1% of all mammal species (Mammal Diversity Database, 2023), Table 4.6 shows that English WordNet 3.0 displays a clear bias for domestic over wild mammals. Despite a larger number of starting nodes for wild than for domestic species, count bias is present because there are more categories (i.e. WordNet synsets) overall for domestic than wild mammals. Descendant bias is also present because domestic categories tend to have more subcategories (i.e. hyponyms) than do wild categories, leading to a more fine-grained representation.

WordNet lies somewhere between an institutional category system and a natural category system, but our approach can also be used to quantify cultural and individual differences in natural category systems. Names of plants (Berlin, 1992), animals, artifacts (Rosch

Table 4.6: Category bias analysis for domestic versus wild mammal species in English WordNet 3.0. The 3 measures of category bias reported are the total number of synsets (Count Bias), the mean depth of starting nodes (Level Bias), and the mean number of descendants per starting node (Descendant Bias). Starting nodes are WordNet synsets corresponding to different mammal species. See Appendix A.5 for full details of this analysis.

	Starting Nodes	**Count Bias**	**Level Bias**	**Descendant Bias**
Domestic	19	398	14.34	20.00
Wild	309	344	14.27	0.13

et al., 1976), body parts (Majid, 2010), and places (Basso, 1984; Burenhult & Levinson, 2008) are all organized into taxonomies or partonymies, and our methods could be applied to each of these cases. For example, consistent with our WordNet analysis, plant and animal names could be labelled as wild or domesticated, and future studies could measure the extent to which a folk taxonomy is biased towards domesticated species ahead of wild species. The degree of bias is likely to vary across cultures in line with existing findings that agricultural societies tend to have more names for plants than do hunter-gatherer societies (Balée, 1999; Berlin, 1992; C. H. Brown et al., 1985). Within cultures, the degree of bias is likely to correlate with factors such as expertise (Tanaka & Taylor, 1991). For example, Aguaruna Jivaro women have much more fine-grained categories for Manioc (a tropical root crop native to South America) than do men, and this difference aligns with the division of labour among men and women in Aguaruna Jivaro culture (Boster, 1985).

We defined bias as a preference for one group over another, and focused on two cases (gender and Western bias) where these preferences can cause harm, especially when systems that incorporate these preferences are perceived as objective. In the case of folk taxonomy, however, a preference for domestic over wild species may be beneficial in supporting communication about the species of most interest to a given culture. Preferences are not necessarily harmful, and can instead illuminate the different needs, values, or roles of the people and cultures who create and use category systems. Our approach therefore joins a set of existing quantitative techniques that can provide insight into conceptual variation both across and within cultures (Romney et al., 1986, 2000).

4.5.3 Limitations

Our work is limited in several important respects. Although our first case study revealed Western bias in both the LCC and the DDC, our analysis is limited to a relatively coarse distinction between Western and non-Western because we are unable to determine if the actual books assigned to each category are written from a Western or non-Western perspective. It is likely that, in a US-based library, books classified as non-Western still exhibit many kinds of Western bias.

Our case study of gender bias was limited by its reliance on incomplete author data. As mentioned earlier, the VIAF does not account for non-binary gender identities and only labels an author as "male," "female," or "unknown." In addition, for simplicity, our analyses were based on the gender of the first author or primary person associated with a book, which means that some nuance is lost because books with multiple authors of different genders are not considered.

A key limitation that applies to both our case studies is that we focused on two US-based, Western classification systems. Future work could aim to apply our methods to a more diverse set of library classification systems, including non-Western systems such as Russian and Chinese library classification systems (Zhang, 2003), and systems like the Universal Decimal Classification, which was designed to be more comprehensive than the Dewey Decimal Classification. It is also important to note that in both studies our book-level statistics are based on data from the Ohio Academic Libraries. The analyses of bias in the LCC and DDC based on these statistics are thus limited to how biased the LCC and DDC are with respect to this specific group of Western libraries. Despite their limitations, however, our analyses seem sufficient to demonstrate that our methods are capable of capturing biases in hierarchical category systems.

4.5.4 Future Work

Our work opens up several directions for future studies. We have focused on characterizing and measuring several kinds of bias, and the natural next step is to attempt to remediate these biases. For example, our methods demonstrate that the classification of non-Western languages and literatures encodes significant level bias. The LCC has no subcategory limits, so changes such as promoting the category for African languages and literatures to the subclass (second) level could help to remediate some Western category bias.

Although we focused on hierarchical category systems, future work could apply some of our methods to measure bias in flat category systems. One previous study in this area focused on gerrymandering and developed methods for quantifying bias in the United States' congressional districts (McCartan & Imai, 2023). Some of our methods for detecting category and item biases in hierarchical category systems can be directly applied to flat systems. For example, item count and category count bias can be applied without modification. Distributional bias could also be applied by considering the distribution of different groups across an entire flat system instead of considering differences in the distribution across the subcategories of each internal node.

Our work documents and quantifies biases in hierarchical classification systems, and future work could study the cognitive mechanisms that give rise to these biases. Perception, attention, and memory can all help to explain how biased collections of library books are created (B. Quinn, 2012), and the same three mechanisms are likely to contribute to biases

in hierarchical category systems. For example, differences between in-group members are often perceived as larger than differences between out-group members, and therefore more worthy of being recognized in a classification system (Park & Rothbart, 1982). These perceptual differences may arise as a consequence of selective attention to features that are more relevant to in-group members than to out-group members (Das-Smaal, 1990). Familiarity and exposure can also lead to bias because frequently encountered items (i.e. in-group members) are more likely to come to mind than items encountered rarely (out-group members). Laboratory experiments have previously considered all of these factors, but more work can be done to explore how these factors produce biases in hierarchical systems of categories.

Finally, our methods could be used to explore how biases in category systems change and develop with time. Category systems are rarely formed all at once and instead develop over time in response to a sequence of items. The sequence in which items are encountered can affect the categories that are created (Medin & Bettger, 1994), and future work can examine how bias is compounded or reduced as items are encountered over time. (Knowlton, 2005) studies historical change in the LCC by manually documenting all the ways in which the subject headings have and have not changed three decades after (Berman, 1971) proposed modifications to reduce bias. With access to historical versions of the LCC, DDC, or other category systems, our methods could allow us to explicitly quantify how these systems differ on measures of category and item bias over time. We expect that institutional category systems should become increasingly unbiased with time, but it is possible that some structural biases may compound and increase instead.

4.6 Conclusion

We developed a suite of methods that allow us to systematically quantify and compare biases across hierarchical category systems. Although we focused on Western and gender bias in library classification systems, our methods are broadly applicable and could be used to quantify cultural and individual differences in natural category systems. These methods can provide a deeper understanding of the different ways in which the needs, beliefs, and values of an individual or a group are encoded in their category systems. In cases where bias is harmful, measuring this bias can support efforts to reduce or control for this bias. *View of the World from 9th Avenue* would not score well on our definitions of category bias, but quantifying its bias could serve as a first step towards creating a more balanced categorization.

Chapter 5

The Evolution of Hierarchical Category Systems

In the previous chapters, we established how a hierarchical category system can encode bias and demonstrated how these biases manifest in institutional category systems like Wikipedia and the Dewey Decimal System. We also demonstrate how biases manifest in more cultural systems represented by the words and concepts in a language. Here we turn to the question of how category systems evolve over time. With a computational model and an experiment, we examine how the sequential incorporation of items into a category system can result in a system that is a less optimal fit to the underlying item space than one that might have evolved under different conditions. We provide one mechanism by which category systems that are internally biased can arise, as we demonstrate how mismatches between the structure of a category system and the environment it represents can develop.

5.1 Introduction

Humans categorize the world at multiple levels of abstraction (Berlin, 1992; Murphy & Lassaline, 1997; Rosch et al., 1976), but how these hierarchies change over time remains an open question. The history of real-world category systems suggests two key phenomena. First, the order in which items appear can influence how and where they are incorporated into a hierarchical category system. For example, in Tenejapa Tzeltal, a language spoken in Chiapas, Mexico, sheep were introduced during Spanish colonization and were labelled as a kind of deer (*tunim čih*, lit. 'cotton deer'; Witkowski and Brown, 1983). In contrast, sheep had been present in Britain long before Modern English emerged (Ryder, 1964), and were long treated as a distinct category.

Second, as hierarchical category systems evolve, they can become skewed representations of the items they organize. As illustrated in Figure 5.1, early versions of the Dewey Decimal System classified fungi under botany (e.g., Dewey, 1885), reflecting the then-dominant

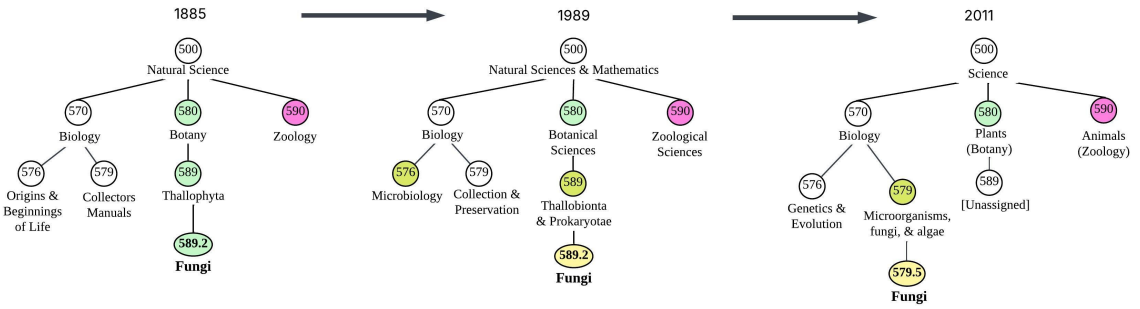


Figure 5.1: The evolution of the categorization of fungi by the Dewey Decimal Classification (DDC). Fungi were first classified under plants in the 2nd edition of the DDC (Dewey, 1885). It remained there in the 20th edition (Dewey, 1989) despite fungi being thought of as a distinct kingdom since at least the late 60s (Whittaker, 1969). By the 23rd edition (Dewey, 2011), the version that is still in use today, it had been moved out of plants.

view of fungi as a subcategory of plants. This classification persisted even after advances in molecular biology meant fungi were recognized as a separate kingdom (Whittaker, 1969). Although fungi were eventually reclassified, this example demonstrates how category decisions that were reasonable at the time can become suboptimal as our understanding of the set of items being classified changes.

Together, these examples suggest that the evolution of a hierarchical category system depends on the time at which items are encountered, with earlier category choices sometimes leading to suboptimal systems as the environment and how it is understood change. While prior work has largely focused on the emergence of flat category systems through cultural evolution (Carr et al., 2017; Carstensen et al., 2015; Ferdinand & Perfors, 2020), less is known about how individuals adapt hierarchical category systems to accommodate novel items they encounter in their environments over time. To address this gap, we introduce a hierarchical categorization model that sequentially incorporates novel stimuli into an existing hierarchy and then develop a task in which participants must do the same, allowing us to compare human and model behaviour. Starting from existing systems as opposed to from scratch allows us to observe how these systems are adapted to accommodate new items, and when and how the resulting hierarchies become skewed. This mirrors real-world scenarios where items must be explicitly categorized into an existing system without knowing the full space of items that could potentially be encountered.

5.1.1 Evolution & Development of Hierarchies

The hierarchical structure of human category systems is well-established (Murphy & Lassaline, 1997). People combine and divide items into more general and more specific groups, demonstrating that things like plants, animals, and emotions can be categorized at multiple levels of specificity (Boster, 1987; Hosoya et al., 2017; Medin et al., 1997, 2006). These levels of abstraction are typically grouped into broad superordinates (e.g., building), basic-level

categories (e.g., house), and more specific subordinate categories (e.g., bungalow). The basic level is considered privileged in that it reflects the level of abstraction most commonly used for naming everyday items, is maximally informative about the attributes of the items it categorizes, and is the most cognitively efficient such that it is the level of abstraction that is easiest for participants to learn and use (Atran et al., 1997; Mervis & Rosch, 1981; Murphy & Smith, 1982; Rosch et al., 1976).

Understanding why hierarchies have the structure they do requires considering how they form. From an evolutionary perspective, category systems evolve to optimally partition a space of meanings into sets of categories that preserve the most information while being simple enough that they are learnable (Carr et al., 2017; Carstensen et al., 2015). Under this view, hierarchies may emerge because different environmental demands drive the development of categories at different levels of specificity, with each level suited to different task demands (Hawkins et al., 2018). From a developmental perspective, different lexical categories (e.g., nouns) bias children toward learning particular levels of abstraction (Waxman, 1990; Waxman et al., 1991), and implicit cues in child-directed speech convey information about hierarchical structure, prompting children to learn hierarchical relationships between categories (Fourtassi et al., 2020). There is ongoing debate about which level of the hierarchy emerges first. Many argue that basic-level categories precede both superordinate and subordinate ones (R. Brown, 1958; Djalal et al., 2017; Mervis & Crisafi, 1982; Murphy, 2016; Waxman, 1990), while evidence from infant categorization has led others to argue that superordinate categories emerge first (Croteau et al., 2026; Mandler & McDonough, 1998; P. C. Quinn & Johnson, 2000), and still others question whether infants represent hierarchical relations at all (Rakison, 2000). While these studies help explain how hierarchies form, it still remains unclear how systems adapt as new items are incorporated and the extent to which their evolution is dependent on the order in which things are learned.

Research into conceptual change offers a starting point. People adapt their category systems to new information either by placing the new information into existing category structures or by more drastically restructuring the hierarchy so that it better accounts for the space of things it represents (Carey, 1985; Chi, 1992; Keil, 1989). However, these accounts still do not directly address the role of order highlighted by the real-world examples discussed previously. In particular, they do not explain how sequence affects the development of an individual's category system or how these systems can sometimes become suboptimal. The possibility of restructuring in conceptual development suggests that existing category systems do not always remain well-suited to the items they represent and may need to be revised as new information is encountered. More broadly, the processes through which category systems develop can lead to structures that are suboptimal representations of the environment. For example, communicative pressures can cause early conventions to become established and shape later category structure, even when those structures are not optimal partitions of the underlying space (Silvey et al., 2013). People encounter their environments

sequentially across time, and so the order in which information is learned may influence the systems that develop. If people begin by incorporating new items into existing categories, then early category decisions may affect later ones and influence how well a category system fits the environment it represents.

Prior work shows that presentation order can influence how easily categories are learned (Carvalho & Goldstone, 2017; Stewart et al., 2002), how stimuli are divided based on competing features (Anderson, 1991; Zaki & Salmi, 2019), and where category boundaries are drawn along continuous feature dimensions (Egré et al., 2013; Stöttinger et al., 2016). Most research on sequence effects, however, treats category systems as flat. When hierarchy is considered, the focus is typically on how sequence affects the level of abstraction at which items are categorized (Mack & Palmeri, 2015), rather than on how sequential item presentation shapes the development of the hierarchy itself. This leaves open the question of how the order in which items are encountered influences the structure of a category system over time. When new items are incorporated into an existing hierarchy, they must be placed within a structure shaped by previous categorization decisions, meaning that early decisions may influence how later items are categorized. The present work addresses this question by examining how the sequential presentation of stimuli affects where they are placed within a category system and whether different orders lead to different hierarchical structures.

5.1.2 Computational Models of Categorization

Computational models have been used to test the mechanisms underlying categorization and to better understand how categories are formed, learned, and adapted over time. These models encode different assumptions about what makes a category system "good" and use these principles to simulate how category structures are acquired and updated through experience. Among these approaches, Bayesian rational models of categorization have been particularly influential, providing a formal framework for understanding how category structures can be inferred from experience (Anderson, 1991; Canini & Griffiths, 2011; Griffiths et al., 2007; Sanborn et al., 2010).

Anderson's rational model of categorization (RMC) formalizes category acquisition as Bayesian inference over possible category structures (Anderson, 1991). Under this view, a good category system captures the statistical structure of the environment by grouping together items with similar patterns of features, such that category membership supports prediction about the features an item is likely to have. Because exact inference over all possible partitions becomes intractable as the number of items increases, the RMC has been implemented using approximate inference methods. Anderson's original model used an incremental maximum a posteriori (MAP) approximation, where items are incorporated sequentially into the most likely existing partition (Anderson, 1991). Later work introduced alternative approximations, including Gibbs sampling and particle filtering, with

particle filtering providing a better account of human categorization behaviour and order effects (Sanborn et al., 2010). Both the MAP RMC and the particle filtering model can be used to simulate incremental categorization and have been previously shown to capture order effects (Sanborn et al., 2010), but they represent category systems as flat partitions over an items space rather than hierarchical structures.

Hierarchical extensions to rational models represent categories as nested structures and infer the most probable hierarchical organization of observations (Canini & Griffiths, 2011). These approaches primarily focus on recovering the underlying organization of observations from their category labels, rather than modelling how hierarchical category systems develop through sequential presentation of new items. Hierarchical clustering, on the other hand, is concerned with learning the underlying organization of a set of items without any labels (Blundell et al., 2011; Cohen-Addad et al., 2019; Heller & Ghahramani, 2005). Objective-based hierarchical clustering in particular can provide a framework for defining and evaluating the quality of hierarchical category systems through explicit objective functions (Naumov et al., 2022). These objectives guide the construction of a hierarchy by determining how groups of items should be combined or divided to optimize the overall structure of the hierarchy. A common class of hierarchical objective functions evaluates category quality in terms of within-category similarity and between-category dissimilarity (Cohen-Addad et al., 2019; Naumov et al., 2022). This connects to notions of category goodness in the literature, such as compactness (Carr et al., 2020), and the idea that optimal category systems will maximize within-category similarity while minimizing similarity between categories (Mervis & Rosch, 1981).

Other approaches to hierarchical clustering focus on the incremental formation of hierarchies as new items are encountered (Ribert et al., 1999). In the present work, we combine these approaches by adapting a hierarchical clustering objective to an incremental setting. Rather than constructing a hierarchy after observing the complete set of items, our model incorporates items sequentially and uses the objective function to determine how each new item should be incorporated into the evolving hierarchy. Specifically, newly encountered items can either be incorporated into existing categories or motivate the creation of new categories, with these decisions based on which option most improves the fit of the hierarchy according to the objective. This provides a mechanism for studying how sequential item presentation shapes the evolution of category systems, and how early decisions can lead to structures that differ from those that would emerge if all items were available simultaneously.

5.1.3 The Present Work

In the present work, we investigate how people adapt existing hierarchical category systems as they incorporate novel items over time. Rather than examining how categories emerge

from an initially unstructured space, we begin with an established hierarchy and ask how sequentially encountering the space of items shapes the resulting system. This allows us to better understand how the history and sequence of categorization decisions influence the development of a hierarchy.

We introduce a computational model that formalizes category system development as a sequential process in which novel items tend to be placed in the category location that best improves the fit of the hierarchy to the item space. The model assumes that category systems are structured to efficiently represent the relationships among items. Specifically, category systems with more optimal fits to their item space group similar items into more specific categories deeper in the hierarchy, and dissimilar items only at higher, broader levels of categorization. By updating the hierarchy after each new item is encountered, the model captures how current categorization decisions can influence future decisions. By basing category assignment on the fit of the hierarchical structure to the item space, we introduce a model of category system development that is explicitly hierarchical.

We then develop an experimental task in which human participants perform the same sequential categorization process. Participants begin with an existing hierarchical category system and incorporate novel items one at a time. To better understand how hierarchical category systems evolve, our experiment manipulates not only the order of stimulus presentation but also the similarity of new stimuli to existing categories and the depth of the hierarchy. Varying similarity allows for cases where category systems can become suboptimal, while varying system depth allows us to examine how the structure of an existing hierarchy shapes its development. The ability of our model to account for participant behaviour is compared to two rational models of categorization, the original MAP-based RMC of Anderson (1991), and the single-particle filter RMC, which has been shown to account well for human categorization behaviour (Sanborn et al., 2010). Both these models simulate the growth of flat category systems, and thus comparison to these allows us to evaluate whether explicitly representing hierarchical structure improves our ability to capture how people adapt category systems over time.

The rest of this chapter is organized as follows. We first illustrate how sequence and hierarchy interact through a simple scenario. We then describe our computational model and the experiment used to test its predictions. Finally, we compare our model with rational models of categorization by fitting them to participant data. To preview our results, presentation order, inter-item similarity and the hierarchical structure of the initial system all influence the final category system participants create, sometimes leading to skewed representations of the items they categorize. Our model was better able to capture participant behaviour than existing rational models and tended to mirror participant behaviour with a few notable differences, suggesting a need for intrinsically hierarchical categorization models.

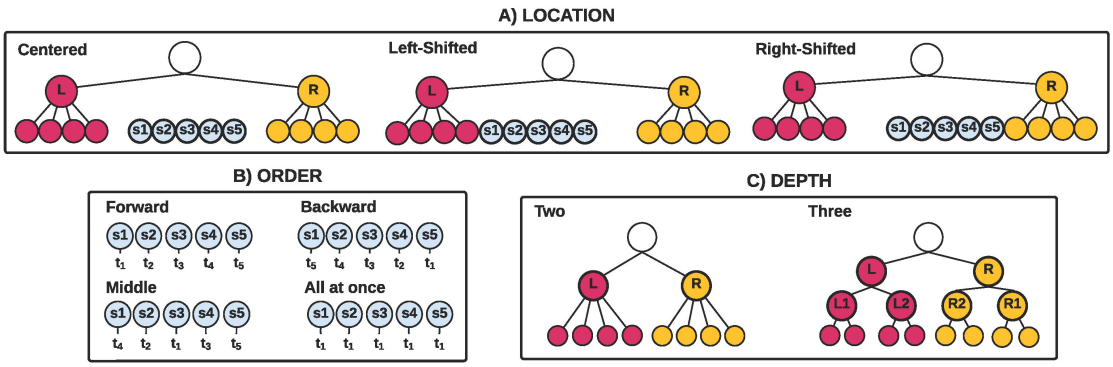


Figure 5.2: Illustration of the three factors affecting category evolution: location (3 levels), order (4 levels), and depth (2 levels). Location and order affect the novel stimuli presented to participants. Depth affects the existing category system that they must be incorporated into. Small circles at the bottom denote items, large labelled circles denote subcategories, and white circles denote the root category. Novel items (s_i) are blue; stimuli assigned to category L are pink, and to R are yellow. Order is indicated by time stamps (t_i).

5.2 Sequence Effects in Category Evolution

In this section, we introduce the factors that we propose shape the evolution of category systems when items are encountered sequentially. As illustrated in Figure 5.2, we focus on the location of novel items relative to existing categories in a one-dimensional similarity space, the order in which they are encountered, and the depth of the existing category hierarchy. We first motivate these factors with an example before presenting our computational model and hypotheses.

5.2.1 The Evolution of a Simple Category System for Shells

To motivate both our experiment and model design, consider a simple system of shells with two categories: L for small shells and R for big shells. Some of its properties are illustrated in Figure 5.2. Suppose there exists a set of shells, S, that has not yet been categorized. If all shells were arranged along a one-dimensional line representing size, S would be centred so that the middle shell in S is halfway between the largest shell in L and the smallest shell in R

Now imagine encountering the shells in S sequentially, incorporating each into the system before seeing the next. In a forward order (s_1 to s_5), s_1, being closer to L, will likely be placed there, which in turn makes it more likely that subsequent shells will be assigned to L. In a backward order, s_5 might be assigned to R, similarly influencing the category assignment of later shells. If the middle shell s_3 is seen first, it could appear equally dissimilar to L and R, prompting a new medium shell category, X, with the rest of S assigned there.

If S is shifted left so that its shells more closely resemble L, simultaneous presentation would likely lead to all shells being assigned to L. Forward sequential presentation might produce the same outcome, but a backward order could result in most shells being assigned to R, creating a system that differs from one formed with full knowledge of the feature space. Shifting S to the right would produce the opposite pattern, with the forward order potentially producing a system incongruent with the feature space.

Finally, consider a more fine-grained system in which L is subdivided into *small* and *very small* shells, and R into *big* and *very big* shells. With more lower-level category options, shells from S can be incorporated into existing subcategories or form new ones within L or R. For instance, when S is left-shifted, encountering the middle shell first may lead to its assignment as a new subcategory of L rather than creating a novel category X at the same depth as L and R. As the number of subcategory options increases, the formation of a new higher-level category becomes less likely, even when presentation order would favour it in a coarser system.

Together, these scenarios illustrate how both the order in which items are encountered and the hierarchical structure of the category system influence item assignments and the resulting system. We next introduce a model to simulate these effects.

5.2.2 Computational Model: Ordered CKMM

By some definitions, an optimal category system is "compact" such that it jointly maximizes within-category similarity and between-category dissimilarity (Carr et al., 2020; Mervis & Rosch, 1981). In the shell examples, we imagined a person encountering items one at a time, with each category decision affecting the next. We now describe a model that formalizes this sequential process. The model's goal is to construct a compact hierarchical category system, where similar items share narrow categories deep in the hierarchy and dissimilar items only share broad categories near the top.

To measure this, we use the Cohen-Addad, Kanade, Mallmann-Trenn, and Mathieu objective (CKMM; 2019), which assigns a hierarchical category system H a score

$$Q(H) = \sum_{i<j} d_{ij} \cdot |LCA_H(i,j)| \tag{5.1}$$

where, for all pairs of items (i, j), the size of the lowest common ancestor (LCA)[1] is multiplied by their dissimilarity (d_{ij}), measured as the distance between the items' vector representations. The higher the score, the more compact the system. Consistent with the intuition above, an optimal (or compact) category system will therefore maximize the CKMM objective by multiplying large distances with large LCAs and small distances with small LCAs.

[1] In other words, the deepest or smallest category shared by both items.

The CKMM objective was designed for binary category systems where every category splits into exactly two subcategories. Building on Naumov et al. (2022), we extend the CKMM objective to non-binary systems, where categories may contain more than two subcategories. We do this by averaging the score over ordered binarizations of non-binary categories. For example, if a category contains three subcategories $\{A, B, C\}$ which are ordered such that $A \prec B \prec C$, we average the CKMM score across the binary systems $\{A, \{B, C\}\}$ and $\{\{A, B\}, C\}$, but not $\{\{A, C\}, B\}$, since A and C are not adjacent. We therefore assume that sibling categories have an underlying linear order corresponding to the order in which they would be encountered during a left-to-right depth-first traversal of the hierarchy. To make this more concrete, consider the Dewey Decimal Classification, where categories are arranged along a hypothetical linear shelf such that 200 (Religion) comes before 300 (Social Sciences), which comes before 400 (Language). The ordered CKMM objective would only consider the binarizations {200, {300, 400}} or {{200, 300}, 400} to preserve this ordering. The objective is formally defined in Appendix B.1.1.

Given the ordered CKMM objective, a novel item s can be assigned to whichever category g maximizes the score of H, where candidate categories include both existing categories and potential new ones consisting solely of the new item. If an item is assigned to a subcategory, it is automatically assigned to all connected parent categories as well. Rather than always selecting the single best category, we assume assignment is probabilistic and define the probability that s is assigned to g as a Softmax over all possible category assignments:

$$P_{CKMM}(g \mid s, H) \propto \exp\left(\frac{Q_g}{\tau}\right) \tag{5.2}$$

where Q_g is the CKMM score of H after assigning s to g, and temperature τ controls how deterministic the model is. Specifically, lower τ values make the model more likely to always select the category that maximizes the CKMM score.

We further assume that some proportion of category assignments are random, giving the mixture model

$$P(g \mid s, H) = (1 - \alpha) \cdot P_{CKMM}(g \mid s, H) + \alpha \cdot \frac{1}{K} \tag{5.3}$$

where K is the number of candidate categories and $\alpha \in [0, 1]$ is a noise parameter that controls the contribution of a uniform random assignment of items to categories. Specifically, α represents the probability weight assigned to the noise component of the mixture model, where assignments occur according to the CKMM predictions with probability $(1 - \alpha)$ and according to a uniform distribution over candidate categories with probability α. In theory, an item can be added to a category system by being placed in any existing category or by creating a new subcategory of any existing category; for simplicity and tractability, when simulating our categorization task, we imposed constraints on the number of new categories

in our human experiments, and correspondingly limited the category options during simulation. See the experiment procedure for more details. This process is applied sequentially to each item $s \in S$, where the category system H is updated after each assignment and used to compute probabilities for the next item. Implementation details are provided in Appendix B.1.2.

5.2.3 Hypotheses

Having defined the model, we now make predictions about how a category system evolves in response to a novel set of items S with varying similarity to existing categories L and R. Figure 5.2 summarizes the three experimental factors manipulated in this study: the order in which items are encountered, the location of novel items relative to existing categories, and the depth of the category hierarchy. The following hypotheses were pre-registered on AsPredicted (#233, 910).

H1. **Sequence Effects.** Order affects assignment to categories L, R, and X. More specifically:

a. Centred items are more likely to be assigned to (i) L when presented in the forward compared to the backward order, (ii) R in backward versus forward order, and (iii) a new middle category X in the middle versus forward or backward orders.

b. Left-shifted items are more likely to be assigned to L in the forward compared to the backward order.

c. Right-shifted items are more likely to be assigned to R in the backward compared to the forward order.

H2. **Priming Effects.** Because items are encountered in roughly sequential order, items are more likely to be assigned to the same category as the initial item in that sequence.

H3. **Hierarchy Effects** Participants are more likely to create a new second-level category (X) in a two-level category system than in a three-level one.

H4. **Suboptimal Systems**. We assume that participants who view all items simultaneously produce more optimal category systems than those who view items sequentially, as they have complete knowledge of the stimulus space. We therefore treat the all-at-once condition as a baseline, and define divergence from this baseline as a measure of suboptimality. We predict that divergence will be greatest when presentation order conflicts with item location. Specifically:

a. In the left-shifted condition, items shown in the backward order will diverge more from the all-at-once baseline than those shown in the middle or forward order.

b. In the right-shifted condition, items in the forward order will diverge more from the all-at-once baseline than those shown in the middle or backward order.

5.3 Experiment

Here we describe an online experiment which asks human participants to incrementally extend a category system. This allows us to systematically study the effects of item order, item similarity, and system complexity on the evolution of a category system. The experiment design and analyses were pre-registered on AsPredicted (#233, 910). Frequentist and Bayesian analyses were pre-registered for Hypotheses 1–3 and a permutation test for Hypothesis 4. Frequentist analyses are reported in the main text, and Bayesian analyses are reported in Appendix B.2. Bayesian results are discussed in the main text only when they diverge from the frequentist conclusions. Stimuli, data, and code can be found on the Open Science Framework. This experiment was approved by the University of Melbourne Human Research Ethics Committee (ID: 29169)

5.3.1 Methods

Participants

We collected data from 419 participants across three pools: 80 university undergraduates who received course credit (pool 1); 159 Prolific participants paid £4.50 for 30 minutes (pool 2); and 180 additional Prolific participants from a pool screened for attention and use of large language models (pool 3). This pool was paid £3.20 for approximately 21 minutes, as task duration had been more precisely estimated. The total sample exceeded our pre-registered target ($n = 260$) because some information required for computational modelling (e.g., distractor order) was not recorded in the first two pools.

Data from 33 participants were excluded for incorrectly answering at least one comprehension question after three attempts ($n = 15$), miscategorizing at least 25% of distractors ($n = 14$), or technical issues ($n = 4$). The final sample consisted of 386 participants (230 women, 155 men) with ages ranging from 18 to 75 years (median: 33). All were fluent English speakers who provided informed consent prior to participation. Preliminary analyses using a first-order Rao-Scott adjusted chi-squared test of homogeneity (Rao & Scott, 1981) detected no significant difference in the category distribution of the three pools, $\chi^2(4) = 17.16$, adjusted $p = .55$, and the effect size was negligible (Cramér's $V = .02$), suggesting that the pools are equivalent (Cohen, 1988).[2] Results are reported for the combined sample.

Stimuli

Pictures of imaginary alien species were used as stimuli. Four sets were created, each containing 31 aliens presented on a square white background. See Figure 5.3 for examples.

[2]The first-order Rao-Scott adjustment was implemented to account for non-independence of category assignments within a trial.

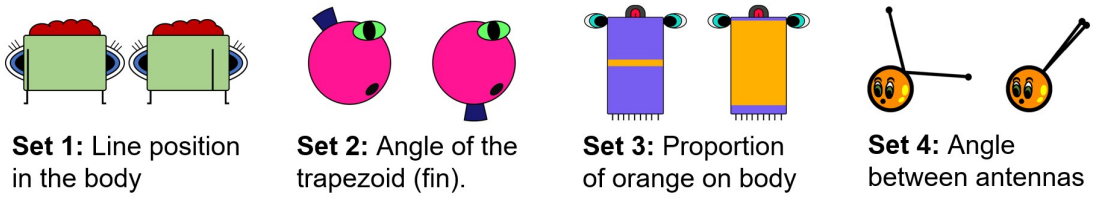


Figure 5.3: Left and right extremes of the four stimulus sets presented to participants. Stimuli are called aliens. In each set, all other stimuli fall between alien 1 (left) and alien 31 (right). Below each set of examples is a description of the feature that changes.

Within each set, aliens varied in equal intervals along a single continuous feature, with aliens 1 and 31 representing the two endpoints. The manipulated features were: the horizontal position of a line within the alien's rectangular body (Set 1); the position of a fin (a trapezoid) along the body's circumference (Set 2); the proportion of the body coloured orange (Set 3); and the angle between the two antennas (Set 4). All other features were held constant within a set.

We avoided placing perceptual attractors (e.g., right angles; Huttenlocher et al., 1991) at the midpoint stimulus (alien 16) in an effort to reduce the likelihood that participants would rely on pre-existing intuitions about spatial midpoints when categorizing the stimuli. As a result, in Set 1, the line of alien 16 is slightly left of centre; in Set 2, its fin is positioned below the horizontal midpoint; in Set 3, its body contains more purple than orange; and in Set 4, its antenna angle is less than 90°.

Design

To examine how hierarchical category systems evolve, we conducted an experiment that manipulated three factors illustrated in Figure 5.2. These were the **location** of novel stimuli in a feature space, the **order** in which novel stimuli were presented to a participant, and the **depth** of the existing category system.

The first two factors, location and order, concerned the stimuli that participants categorized. Location (3 levels, Figure 5.2A) determined which portion of the stimulus set participants saw, reflecting similarity to existing categories. Novel stimuli could be centred, meaning they were equally similar to categories L and R, left-shifted, meaning they tended to be more similar to stimuli in category L, or right-shifted, making them more similar to stimuli in category R.

Order (4 levels, Figure 5.2B) determined the sequence in which stimuli were presented. In the forward condition, stimuli were shown left-to-right along the continuous feature, and for stimuli in the backward condition this was right-to-left. Both sequences were slightly permuted to prevent participants from anticipating the pattern. In the middle condition, presentation started at the centre and alternated outward. Because there are two ways

to create the middle order (e.g., either s_3, s_2, s_4 or s_3, s_4, s_2), participants in the middle condition randomly saw one of the two. In these conditions, participants categorized stimuli one at a time. As a baseline, we also introduced an all-at-once condition where stimuli were displayed simultaneously rather than sequentially.

The final factor, depth (2 levels, Figure 5.2C), concerned the granularity of the existing categories in the hierarchical system. A system could have either two levels, with a root category divided into L and R, or three levels where L was further divided into subcategories $L1$ and $L2$, and R into $R1$ and $R2$. Depth affected where participants could add new categories.

Location, order, and depth were combined in a $3 \times 4 \times 2$ factorial design, yielding 24 distinct trial types. The trials were grouped into six blocks of four using an incomplete block design. Blocks were designed such that, for each set of four trials, there were two trials for each depth, one for each presentation order, and at least one for each location. Because there are three location conditions, one had to be repeated within a block. The repeated location was always assigned to the other depth condition (see Appendix B.3 for a list of the blocks). Within blocks, trial order was randomized, and the four stimulus sets were randomly assigned to the four trials without replacement. Participants were assigned to one block each using a queue that reassigned a block to the next participant if they dropped out before completion. This was done to ensure a balanced distribution of participants across blocks.

Procedure

Participants started the experiment by completing a series of practice tasks where they were asked to categorize images of animals into an animal taxonomy. These tasks were designed to be equivalent to those in the experiment and to familiarize participants with the task interface. Once completed, participants had to correctly answer six comprehension questions before proceeding. Participants then completed four experimental trials from their allocated block. In each trial, they were asked to categorize aliens into a partially completed alien taxonomy. Between each trial, participants were shown a screen stating how many sets of aliens out of four they had finished categorizing.[3]

Participants could add aliens to a category system in one of two ways: by placing them into an existing category at the bottom of the taxonomy, or by creating a new category using an available button. In the two-level condition, they could add one new category, and in the three-level condition, they could add up to three. During the practice rounds, participants were instructed to only create a new category if they felt the existing categories were not sufficient for the current item. This was done to reduce the likelihood that participants would feel the need to add a new category just because the button was there.

[3]For more details, a click-through experiment demo is hosted at katiewarburton.com/Alien-Taxonomist/.

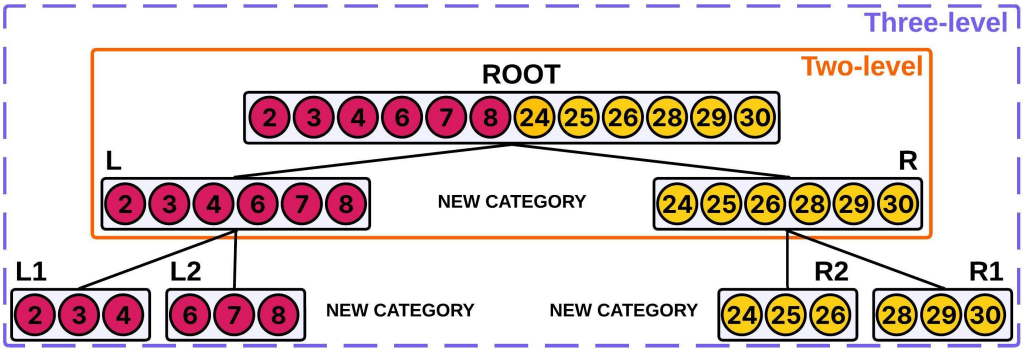


Figure 5.4: Visualization of the hierarchical category system setup at the start of a trial. Numbered circles represent stimuli and their surrounding boxes correspond to categories. "NEW CATEGORY" buttons were visible to participants; category labels (e.g., $L1$, R) were not and are only included here for explanatory purposes. Both the two-level (solid orange outline) and three-level (dashed purple outline) starting category systems are depicted.

At the start of each trial, participants were asked to familiarize themselves with the existing alien taxonomy before clicking "Begin Categorizing." This taxonomy was the starting category system and had a depth of either two or three, as depicted in Figure 5.4. For both category system depths, the top-level category contained 12 aliens, divided into two categories of six. Category L always appeared on the left, and category R on the right. In the three-level condition, each of these categories was further divided into subcategories of three stimuli. These were $L1$ and $L2$ under category L, and $R1$ and $R2$ under category R. Category L always initially included aliens 2, 3, 4 ($L1$) and 6, 7, 8 ($L2$), while R included aliens 24, 25, 26 ($R2$) and 28, 29, 30 ($R1$). Aliens always appeared in the category system in sorted order. Within the taxonomy display, there were either one or three "NEW CATEGORY" buttons. In both conditions, there was one button between categories L and R. In the three-level condition, there were two additional buttons. One was directly to the right of $L2$ and the other to the left of $R2$.

Once a participant clicked "Begin Categorizing," they were shown nine novel aliens (S) and four distractors (D). Distractors were always aliens 1, 3, 29, and 31 in a given stimulus set. They were designed to be clear examples of stimuli that should be placed into categories L and R. Specifically, 3 and 29 were duplicates of aliens that already appeared in the starting taxonomy, and 1 and 31 were extreme examples of categories L and R, respectively. Novel aliens were aliens 12-20 in the centred condition, aliens 9-17 in the left-shifted condition, and aliens 15-23 in the right-shifted condition.

Aliens were shown to participants either one at a time (forward, backward, middle) or all at once. When shown one at a time, aliens appeared in a square at the upper right corner of the screen, as depicted in the top left of Figure 5.5. Novel aliens and distractors were interleaved as follows: $[s, d, s, s, s, d, s, d, s, s, s, d, s]$. Distractor order was randomized, and novel stimuli followed a forward, backward, or middle order. Participants added aliens to

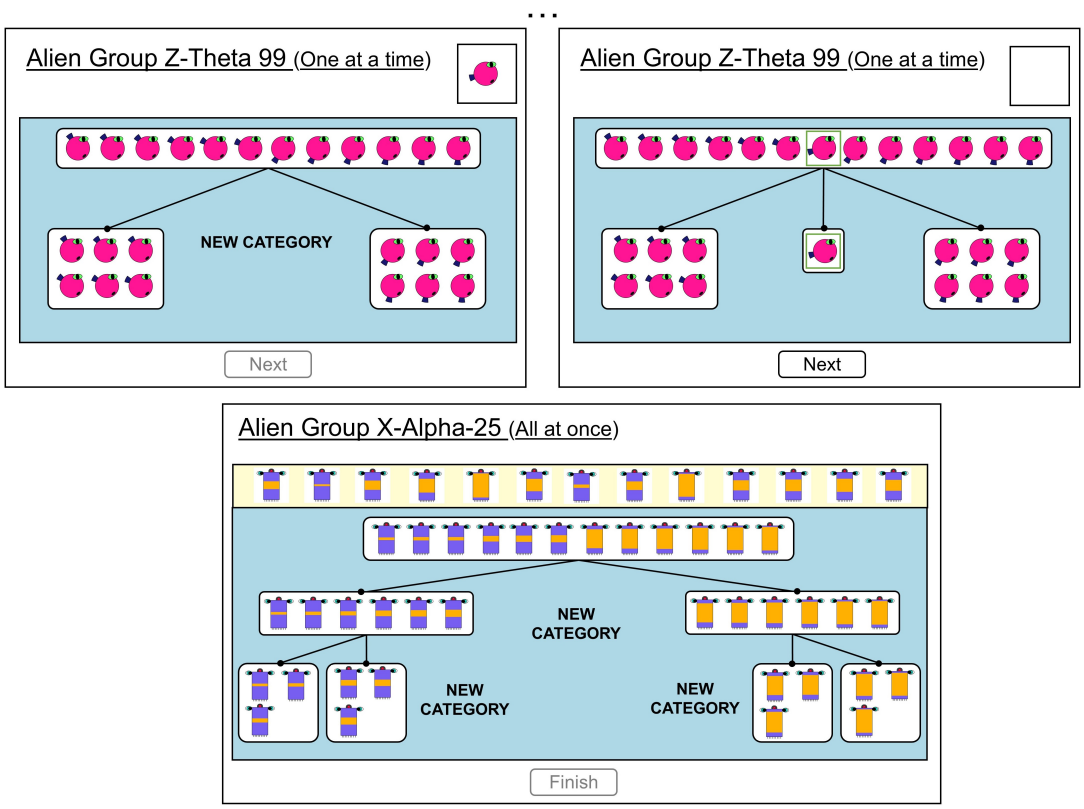


Figure 5.5: Example displays from "One-at-a-Time" and "All-at-Once" trials. The top depicts the trial setup when aliens are added one at a time and the bottom when aliens are added all at once. The top left is the trial before the first alien is added, and the right is after it has been incorporated. Newly added aliens are highlighted by a green box (depicted in the top right) which disappears when participants click "Next" (one-at-a-time) or when participants incorporate another alien (all-at-once).

the category system by clicking on the desired category. Each alien had to be incorporated before a participant moved on to the next, and only the current alien could be moved between categories. Once a participant added the final alien, they could click "Finish" to end the trial. When shown all at once, aliens appeared in a rectangular banner spanning the top of the taxonomy display. This is depicted at the bottom of Figure 5.5. Participants had to drag and drop aliens into the desired category and could not finish the trial until all aliens were categorized.

When an alien was placed into an existing category, it was inserted in sorted order based on item similarity.[4] When a new category was created, it appeared at the location of the selected "NEW CATEGORY" button and was linked to its parent category. In the two-level condition, the new category was labelled X. In the three-level condition, participants could create $L3$ as a child of L, and $R3$ as a child of R. They could also create a new category X like in the two-level condition; however, a single child, $X1$, was also created to ensure a

[4]E.g, if item 9 was added to category $L1$, it would be placed after item 4.

consistent category system depth. In all cases, the newly added alien was highlighted with a green border, allowing participants to easily track its placement (see Figure 5.5).

Participants' category assignments were recorded per trial and per alien. Their data were excluded from analysis if they failed to answer all six comprehension questions after 3 attempts or if they miscategorized 4 or more of 16 distractors across 4 trials. To account for variation in participants' abilities to perceive minor differences between stimuli, miscategorization was only analyzed at the second level (e.g., L, R), not the third level (e.g., $L1$, $R1$), regardless of category system depth.

Computational Model Fitting & Simulation

We fit our ordered CKMM model to participant data from pool 3 ($n = 165$) using 11-fold cross-validation ($n = 15$ per fold). For each fold, parameters were estimated by maximizing the log-likelihood of participant category choices on the training set over a discrete grid of parameter values. Final values for the temperature parameter τ and mixture parameter α were selected by majority voting across folds, yielding $\tau = 1.0$ (11/11 folds), $\alpha = 0.15$ (9/11 folds). See Table B.1 for the full range of parameter values considered.

To assess how well the fitted model captures participant behaviour, we simulate data using the fitted parameters and compare the simulated results to human results. Simulated trial counts per condition are matched to participant trial counts to ensure comparability. For simplicity, we do not simulate stimulus-level variation and instead represent stimuli as evenly spaced values between 0 and 1, where 0 corresponds to alien 1 and 1 to alien 31, regardless of the actual stimulus set used. For the forward, backward, and middle conditions, stimulus representations were presented to the model in the same order as participants. For the all-at-once condition, stimulus order was randomized. In this condition, although participants viewed all stimuli simultaneously, category assignments were made sequentially using drag and drop, making randomized presentation a reasonable approximation. However, the model cannot revise previous category assignments, whereas participants could move stimuli between categories in the all-at-once condition.

5.3.2 Results

Participants completed a total of $1,544$ trials across three pools, categorizing $13,896$ novel items and $6,176$ distractors. To keep analyses comparable between the two-level and three-level systems, results are reported for the second-level category assignment of items (L, R, or X) regardless of depth. Before turning to our hypotheses, we report some general trends in category assignment. Then, for each hypothesis, we report results from the participant data before assessing whether the simulated data captures the same pattern. After Hypothesis 4, we report a complementary exploratory analysis using the ordered CKMM objective as a direct measure of category system quality.

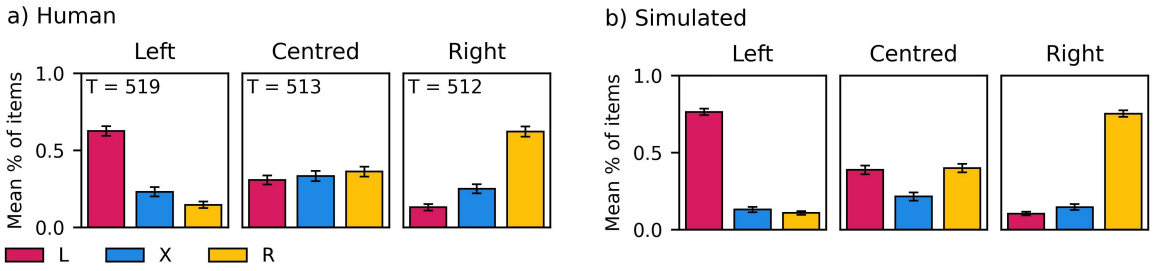


Figure 5.6: Mean proportion of items in categories L, X, R by location. Results are reported for both human and simulated data. The number of trials (T) are reported in the top right corner of each human panel, with simulated trial counts matched to human counts. Error bars represent 95% confidence intervals computed from 5000 bootstrap samples.

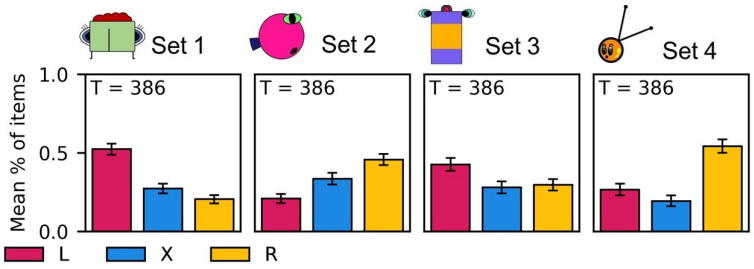


Figure 5.7: Mean proportion of items in categories L, X, R by stimulus set. Results are from human data. Each stimulus depicted at the top is the midpoint of its set (16 out of 31). The number of trials (T) are reported in the top right corner of each panel. Error bars represent 95% confidence intervals computed from 5000 bootstrap samples.

General Trends in Category Assignment

Participants assigned items to L, X, and R with probabilities of 0.355, 0.270, and 0.375, respectively, showing a slight preference for L and R over the novel category X. The model amplified this preference by underestimating assignment to X. Specifically, the simulated probabilities were 0.419, 0.163, and 0.418 for L, X, and R, respectively.

Figure 5.6 shows the mean proportion of items assigned to each category across location conditions. Participants were sensitive to item location, as a first-order Rao-Scott adjusted chi-squared test confirmed a significant association between location and category, $\chi^2(4) = 3152.9$, adjusted $p < .001$, with a medium effect (Cramér's $V = .34$). Participants tended to assign items to L when left-shifted and to R when right-shifted, with roughly equal assignments across categories for centred items. The model captured this pattern, also showing a significant association between location and category, $\chi^2\,(4) = 4858.8$, adjusted $p < .001$, though with a larger effect size (Cramér's $V = .42$), indicating the model overestimates the influence of location relative to participants.

Figure 5.7 shows the mean proportion of items assigned to each category across stimulus sets and suggests that participants were also sensitive to differences in the stimuli. To

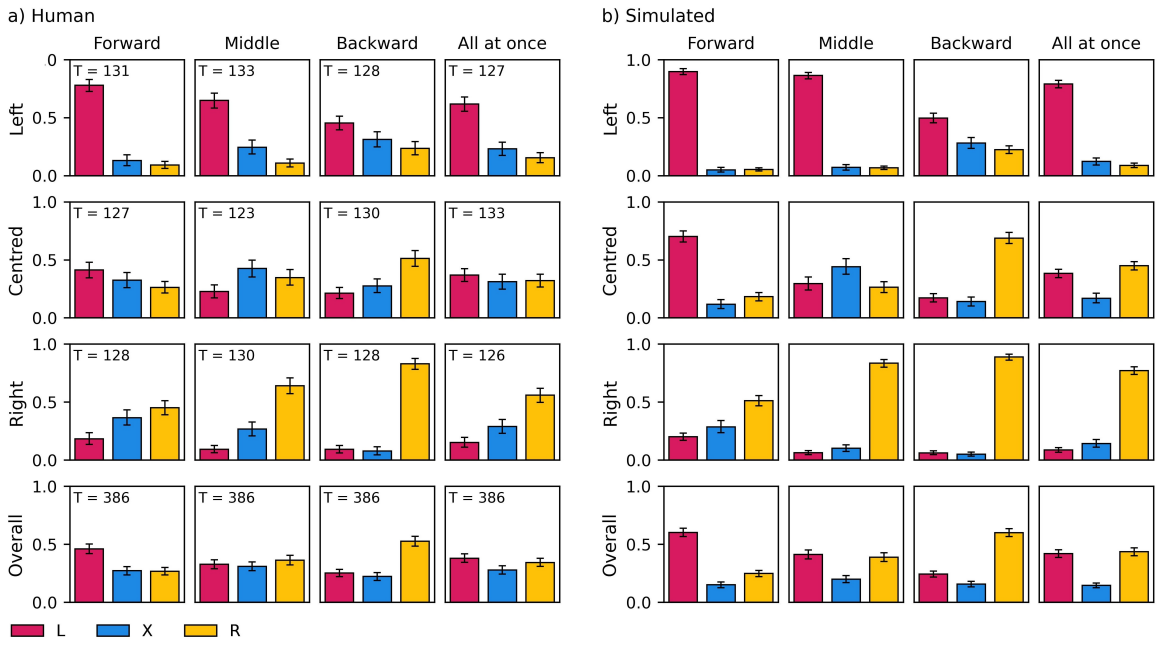


Figure 5.8: Mean proportion of items in categories L, X, R by location and order. Results are reported for both human and simulated data. The rows are the item locations and the columns the presentation orders. The number of trials (T) are reported in the top right corner of each participant panel. Simulated trial counts in each panel are matched to human trial counts. Error bars represent 95% confidence intervals computed from 5000 bootstrap samples.

support this, a first-order Rao-Scott adjusted chi-squared test found a significant difference in category assignment across sets, $\chi^2\,(6) = 1391.3$, adjusted $p < .001$, with a medium effect size (Cramér's $V = .22$). This is expected as the sets are not designed to be perceptually equivalent. Individual differences in categorization behaviour are also expected as people may vary in how they perceive and categorize stimuli. To account for these sources of variation in subsequent analyses, random intercepts for stimulus set and participant are included in all regression models of human data. As the CKMM model does not simulate stimulus- or participant-level variation, these are replaced by a random intercept for trial in item-level regression analyses to account for the non-independence of item assignments within a trial.

Sequence Effects

Hypothesis 1 (H1) predicted that presentation order would influence assignment of novel items to categories L, R, and X, with specific predictions for each item location condition (H1a–c). We first test the overall effect of order before examining each sub-hypothesis.

The mean proportion of novel items assigned to each category by order and location is shown in Figure 5.8. Examination of this figure suggests that, on average, the proportion

Table 5.1: Multinomial regressions predicting category assignment from order. Models are reported separately for each location condition and for the overall simulated and human datasets. The intercepts represent the log-odds of assigning L or R relative to the reference category (X) in the middle-order condition. Forward and backward coefficients represent deviations from these reference conditions.

Location	Category	Order	Human		Simulated	
			β	95% CI	β	95% CI
Left	*L vs. X*	Intercept	0.98	[0.84, 1.11]	2.51	[2.29, 2.73]
		Forward	0.81	[0.59, 1.03]	0.39	[0.05, 0.74]
		Backward	-0.61	[-0.80, -0.42]	-1.94	[-2.20, -1.68]
	R vs. X	Intercept	-0.82	[-1.02, -0.61]	-0.05	[-0.35, 0.26]
		Forward	0.45	[0.13, 0.77]	0.13	[-0.34, 0.60]
		Backward	0.53	[0.27, 0.79]	-0.18	[-0.53, 0.17]
Centred	*L vs. X*	Intercept	-0.63	[-0.78, -0.47]	-0.40	[-0.54, -0.26]
		Forward	0.87	[0.66, 1.07]	2.19	[1.96, 2.42]
		Backward	0.37	[0.15, 0.60]	0.61	[0.36, 0.86]
	R vs. X	Intercept	-0.20	[-0.34, -0.07]	-0.51	[-0.66, -0.37]
		Forward	-0.01	[-0.21, 0.20]	0.95	[0.69, 1.21]
		Backward	0.83	[0.64, 1.02]	2.10	[1.88, 2.33]
Right	*L vs. X*	Intercept	-1.05	[-1.27, -0.83]	-0.47	[-0.76, -0.18]
		Forward	0.36	[0.09, 0.63]	0.11	[-0.22, 0.45]
		Backward	1.21	[0.86, 1.57]	0.69	[0.23, 1.14]
	R vs. X	Intercept	0.88	[0.74, 1.01]	2.11	[1.92, 2.31]
		Forward	-0.66	[-0.85, -0.48]	-1.53	[-1.77, -1.30]
		Backward	1.49	[1.23, 1.74]	0.77	[0.45, 1.10]
Overall	*L vs. X*	Intercept	0.06	[-0.03, 0.14]	0.73	[0.64, 0.82]
		Forward	0.47	[0.35, 0.59]	0.66	[0.52, 0.79]
		Backward	0.07	[-0.06, 0.20]	-0.29	[-0.43, -0.15]
	R vs. X	Intercept	0.16	[0.08, 0.24]	0.67	[0.58, 0.76]
		Forward	-0.18	[-0.30, -0.06]	-0.17	[-0.31, -0.03]
		Backward	0.70	[0.58, 0.82]	0.67	[0.54, 0.80]

of times an item is assigned to each category varies with presentation order, both overall and within each location condition. This holds for both human and simulated data. To test whether order significantly influenced category assignment, we fit multinomial logistic regression models predicting category assignment as a function of order, compared to null models without order, using likelihood ratio tests. Models were fit separately for each location condition and did not include any random intercepts.

Human and simulated regression coefficients are summarized in Table 5.1. For both datasets, category assignment varied with order, although the pattern of variation differed across location conditions. Likelihood ratio tests comparing full models including order to null models containing only an intercept are summarized in Table 5.2. For both the human and simulated data, likelihood ratio tests preferred full models including order over

Table 5.2: Likelihood ratio tests for the effect of order on category assignment in human and simulated data. Full multinomial regression models including order are compared to null models without order. Tests are reported separately for each location condition and for the overall datasets.

	Human		Simulated	
	$\chi^2(4)$	p	$\chi^2(4)$	p
Left	287.68	$< .001$	599.56	$< .001$
Centred	232.14	$< .001$	1153.0	$< .001$
Right	418.42	$< .001$	353.57	$< .001$
Overall	576.75	$< .001$	1105.7	$< .001$

null models, supporting Hypothesis 1. To assess the overall effect of order across locations, additional models were fit to both datasets. Likelihood ratio tests again preferred models including order, although β coefficients were smaller than those obtained when each location was analyzed separately.

The preregistered analysis specified mixed-effects multinomial models to account for participant- and stimulus-level variation in the human data. However, the mixed-effects multinomial models implemented using the `R` package `mclogit` rely on quasi-likelihood estimation. Because quasi-likelihood-based deviances are not likelihood-based quantities, differences in deviance could not be used for likelihood ratio comparisons between full and null models. Therefore, the frequentist analyses reported here used multinomial models with fixed effects only. Trial-level random intercepts were also not included in the simulated data analyses to maintain consistency with the model structure used for the human data.

As a supplementary analysis, we estimated Bayesian mixed-effects multinomial models with random intercepts for participant and stimulus in the human data and trial-level random intercepts in the simulated data (see Appendix B.2.1). Models including order were compared with matched null models using leave-one-out cross-validation. Coefficient estimates differed from those of the fixed-effects regressions reported here because of differences in model structure; however, the Bayesian analyses led to the same overall conclusions. Models including order generally showed better predictive performance than the matched null models, providing additional support for Hypothesis 1. The one exception was the model fitted to simulated data from the centred condition (see Table B.3), which did not provide strong evidence that including order improved predictive performance relative to its null model.

To examine order effects in more detail, separate mixed-effects logistic regression models were fit to both the human and simulated data to predict whether an item was assigned to a specific category (vs. not assigned to that category) as a function of order. Order effects were examined through binary comparisons between conditions (e.g., forward vs. backward, or middle vs. forward/backward). For each comparison, order was coded as a

binary predictor (0/1), and the model tested whether the likelihood of assigning an item to the target category differed between the two conditions. Likelihood ratio tests were used to compare each model, including order as a predictor, against a corresponding null model without order.

In the centred condition, H1a(i) and H1a(ii) were supported. Participants were more likely to assign novel items to category L in the forward than in the backward order, as the log odds of category L increased by $\beta = 2.14$, 95% CI $[1.76, 2.53]$, and a likelihood ratio test preferred the full model to the null model, $\chi^2(1) = 161.64$, $p < .001$. The same pattern was observed in the simulated data, $\beta = 3.43$, 95% CI $[2.93, 3.93]$, $\chi^2(1) = 173.86$, $p < .001$, with a stronger estimated effect of order. Category R was more likely in the backward than in the forward order, $\beta = 1.30$, 95% CI $[1.01, 1.58]$, with the full model again preferred to the null model, $\chi^2(1) = 79.41$, $p < .001$. Again, the same pattern was observed in the simulated data, with a larger estimated effect, $\beta = 3.28$, 95% CI $[2.78, 3.78]$, $\chi^2(1) = 163.03$, $p < .001$.

H1a(iii) was not supported by the human data. We predicted that centred items would be more frequently assigned to category X in the middle order than in the forward or backward orders; however, we did not find evidence for this effect, as $\beta = -0.28$, 95% CI $[-0.58, 0.01]$ for the middle versus backward and forward orders, and a likelihood ratio test did not favour the full model with order over the null model without it, $\chi^2(1) = 3.34$, $p = .07$. The simulated data, by contrast, showed a strong effect in the predicted direction, with category X assignments more likely in the middle than in the forward and backward orders, $\beta = 3.53$, 95% CI $[2.68, 4.37]$, and order significantly improving model fit, $\chi^2(1) = 80.15$, $p < .001$.

In the left-shifted condition, participants' results supported H1b, as category L assignments were more likely in the forward than in the backward order, $\beta = 2.75$, 95% CI $[2.11, 3.40]$, and order significantly improved model fit, $\chi^2\,(1) = 73.15$, $p < .001$. The same pattern was observed in the simulated data, with category L assignments more likely in the forward than backward order, $\beta = 2.55$, 95% CI $[2.20, 2.91]$, and order significantly improving model fit, $\chi^2(1) = 183.08$, $p < .001$. For the right-shifted condition, both participant and simulated results supported H1c. Human data showed that category R assignments were more likely in the backward than forward order, $\beta = 3.77$, 95% CI $[3.00, 4.54]$, with order improving model fit, $\chi^2(1) = 105.31$, $p < .001$. Simulated data showed a weaker effect in the same direction, with $\beta = 2.35$, 95% CI $[2.00, 2.70]$, $\chi^2(1) = 158.44$, $p < .001$.

Altogether, presentation order influenced category assignments in the predicted directions for all hypotheses except H1a(iii). Across both human and simulated data, the same items could be assigned to different categories depending on their position within a sequence, demonstrating that category assignments were influenced by presentation order. The simulated data showed stronger order effects than participants in the centred condition, but similar or weaker effects in the left- and right-shifted conditions. This may be

because, as demonstrated in the previous section, the model is more sensitive to stimulus location than participants, which reduces the magnitude of order effects when category boundaries are shifted. Participant responses, however, are influenced by additional variability across stimulus sets and individual differences in perceptual discrimination ability, which may contribute to the variability observed in the centred condition.

Priming Effects

Hypothesis 2 (H2) predicted that novel items would be assigned to the same category as the first item in a sequence at above-chance rates. To test this, trials from the sequence conditions (forward, backward, middle) were combined ($T = 1,158$), and each subsequent item was coded as 1 if it was assigned to the same category as the first item and 0 otherwise.

Across trials, participants assigned 72.7% of subsequent novel items to the same category as the first item. To evaluate whether this tendency differed from chance, we fit a logistic mixed-effects regression predicting same-category assignment, with a fixed intercept and random intercepts for participants and stimuli. The intercept was $\beta = 1.25$, 95% CI $[0.78, 1.72]$. A likelihood ratio test comparing this model to a model with the intercept constrained to zero indicated that the observed tendency was significantly above chance, $\chi^2(1) = 8.35$, $p = .004$. The fitted CKMM model captured a similar pattern, with 66.7% of simulated items assigned to the same category as the first item. A logistic mixed-effects regression on the simulated data, with a random intercept for trial, found a significant intercept, $\beta = 1.21$, 95% CI $[1.08, 1.34]$. A likelihood ratio test comparing this model to a model with the intercept constrained to zero again indicated an above-chance tendency to assign novel items to the same category as the first item, $\chi^2(1) = 302.62$, $p < .001$.

The main difference between participants and the model is shown in Figure 5.9. Participants assigned all eight novel items to the same category as the first in 42.4% of trials, compared to only 27.5% for the model. The model was nearly as likely to assign seven items to the same category (21.7% of trials), whereas participants showed a stronger tendency toward categorizing all novel items within a single category.

One possible explanation for this stronger tendency is that participants were failing to switch categories or responding inattentively. However, evidence from their categorization of distractors suggests otherwise. Two of the distractors were designed to be clear examples of L and two of R. Attentive participants would therefore be expected to assign at least two distractors to a different category than the remaining items. When considering both novel items and distractors, only 0.3% of trials involved assigning all 12 items to the same category, whereas 30.1% involved assigning 10 of 12 items to the same category as the first. Furthermore, in nearly all of these 10-of-12 trials, that first category was L or R rather than X[5], suggesting participants were switching categories when appropriate. This

[5]Had participants assigned 10 out of 12 items to X, this would indicate errors on at least two distractors.

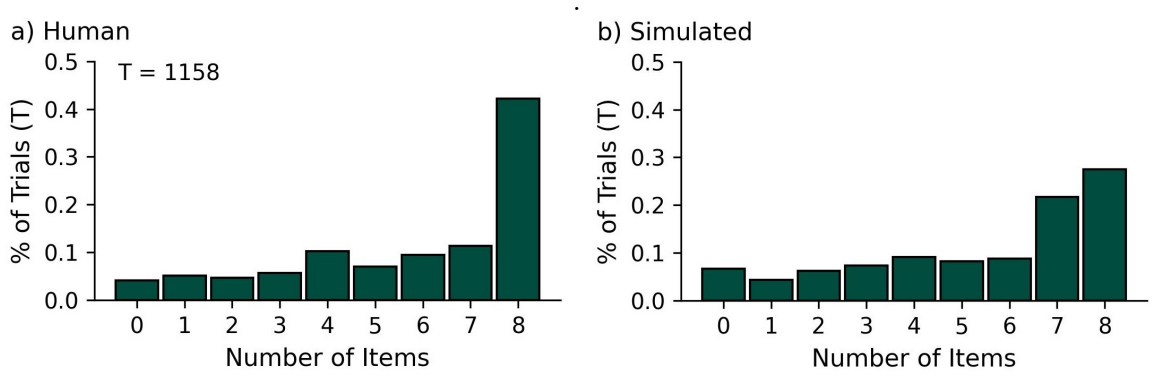


Figure 5.9: Distribution over number of same category assignments for novel items in a sequence trial. Sequence trials include forward, backward, and middle order trials. There are nine novel items total in a trial. Eight means that all subsequent items were assigned to the same category as the first; zero means that none were. Results are reported for both human and simulated data.

interpretation was supported by an exploratory logistic regression (See Appendix B.4.2 for more details).

Overall, H2 was supported. Both participants and the model assigned novel items to the same category as the first item at rates well above chance, although participants showed a stronger tendency to assign all novel items to a single category. This stronger tendency may reflect differences in perception, with participants potentially perceiving adjacent novel items as more similar than the model did with its purely numerical representations. However, the pattern is unlikely to reflect inattentiveness or a general failure to switch categories when warranted.

Hierarchy Effects

Hypothesis 3 (H3) predicted that participants would be more likely to create a new second-level category X in the two-level condition than in the three-level one. To assess this, all 1,544 trials were pooled and coded as 1 if X was created and 0 otherwise.

Participants created X in 53.9% of two-level trials compared to 32.9% of three-level trials. A logistic regression with random intercepts for participant and stimulus set predicting creation of X as a function of depth confirmed a significant effect, with the two-level condition significantly more likely to produce X relative to the three-level condition, $\beta = 1.13$, 95% CI $[0.89, 1.38]$. A likelihood ratio test preferred this model over a null model with only a fixed intercept, $\chi^2\,(1) = 90.80$, $p < .001$. The CKMM model showed a similar but slightly weaker pattern, creating X in 55.2% of two-level trials and 34.8% of three-level trials. A logistic regression without random effects confirmed a significant but weaker effect in the same direction, with the two-level condition more likely to produce X relative to the three-level condition, $\beta = 0.83$, 95% CI $[0.63, 1.04]$, and a likelihood ratio test again preferred

Table 5.3: Jensen–Shannon divergences (JSD) of the sequence conditions compared to the all-at-once baseline. JSDs are reported for human and simulated data. *P*-values are from a permutation test with 10,000 iterations. Within each row, bolded entries indicate the sequence condition associated with the greatest divergence in category assignments relative to the all-at-once baseline.

		Forward		Middle		Backward	
		JSD	p	JSD	p	JSD	p
Left	Human	**0.125**	$< .001$	0.048	.34	0.117	.002
	Simulated	0.109	$< .001$	0.071	.002	**0.219**	$< .001$
Right	Human	0.076	.07	0.071	.09	**0.222**	$< .001$
	Simulated	**0.193**	$< .001$	0.04	$< .001$	0.122	$< .001$

the model including depth over a null model with only a fixed intercept, $\chi^2(1) = 54.53$, $p < .001$.[6]

Together, these results support H3, as both participants and the model were more likely to create X in the two-level condition than the three-level one. This suggests that the system's current structure shapes its subsequent development. Systems with greater depth by nature have more subcategories and thus more ways to incorporate new items deeper in the hierarchy, reducing the need to create new categories near the top.

Creating Suboptimal Systems

Suboptimality & Divergence From Baseline Hypothesis 4 (H4) predicted that presentation orders conflicting with item location would produce the most divergent category assignments relative to an all-at-once baseline: the backward order in the left-shifted condition (H4a) and the forward order in the right-shifted condition (H4b). In the all-at-once condition, participants have complete knowledge of the stimulus space, and their responses therefore reveal which systems seem best to them given all relevant information. We computed the Jensen-Shannon divergence (JSD) between the average distribution of items across categories in each sequence condition and the all-at-once baseline, and compared these divergences across orders within each location condition. Permutation tests (10,000 iterations) were used to assess the significance of both the baseline divergences and the differences between them. Baseline divergences are reported in Figure 5.3.

In the left-shifted condition, participants' category systems differed significantly from the all-at-once baseline in the forward and backward orders, but not the middle order. To test H4a, we compared JSD values across orders: the backward order had a significantly higher JSD than the middle order (difference = 0.069, $p = .001$), but not the forward order

[6] A random effect for trial was not included, as analyses are performed at the trial level rather than at the item level.

(difference = −0.008, p = .70), providing partial support for H4a. In the right-shifted condition, only the backward order differed significantly from the all-at-once baseline. The forward order was not significantly more divergent than the middle order (difference = 0.005, p = .47), and although the difference between forward and backward orders was numerically large, it was in the opposite direction and therefore not significant in the direction of interest (difference = −0.146, $p > .99$), providing no support for H4b.

The simulated data showed a different overall pattern, with all presentation orders differing significantly from the all-at-once baseline and divergences generally larger than the corresponding human effects. In the left-shifted condition, the backward order had a significantly larger JSD than the middle order (difference = 0.111, p = .002) and the forward order (difference = 0.148, p = .001), unlike the human data, in which the backward order did not significantly differ from the forward order. In the right-shifted condition, divergences were significantly larger in the forward order than in the backward order (difference = 0.135, $p < .001$) and the middle order (difference = 0.070, p = .002), opposite to the pattern observed in the human data.

Overall, H4 was not reliably supported by the JSD analyses. The most divergent orders in the experimental data tended to align with item location rather than conflict with it. The simulated data showed the predicted pattern in terms of which order produced the largest divergence but did not mirror the human data. The CKMM score analyses in the following section offer a complementary perspective on these findings.

Suboptimality & CKMM Scores The JSD analyses treated the all-at-once condition as a benchmark for optimal performance, but did not reliably support H4. The ordered CKMM objective function (Equation B.2) provides a measure of category system quality, allowing us to directly evaluate the optimality of category systems across all conditions. We computed the ordered CKMM score for every category system produced by participants and the model, and report mean scores by order and location condition in Table 5.4. Scores are expressed relative to the lowest mean within each location condition, such that 0 indicates the least optimal system and larger values indicate more optimal systems.

Table 5.4: Mean CKMM Scores relative to the minimum score across order conditions. Scores are reported separately for human and simulated data in the left- and right-shifted conditions. For each row, the column with zero is the presentation order that received the lowest mean CKMM score and is therefore less optimal under the ordered CKMM objective. All other mean scores are relative to that minimum.

		Forward	Middle	Backward	All at once
Left	Human	32	27	**0**	25
	Simulated	34	34	**0**	24
Right	Human	**0**	16	22	12
	Simulated	**0**	22	24	22

Across both location conditions, the order conflicting with item location produced the lowest CKMM scores (Table 5.4). We fit a linear regression predicting CKMM score from item presentation order, with the all-at-once condition as reference and random intercepts for participant and stimulus set. For the left-shifted condition, order had a significant overall effect on CKMM scores, $\chi^2(3) = 40.91$, $p < .001$. The backward order produced lower CKMM scores than the all-at-once baseline ($\beta = -22.27$, 95% CI $[-36.73, -7.80]$), while neither the forward ($\beta = 10.77$, 95% CI $[-3.67, 25.21]$) nor the middle order ($\beta = 6.53$, 95% CI $[-7.87, 20.94]$) notably differed. For the right-shifted condition, order was again significant, $\chi^2(3) = 33.90$, $p < .001$, but the pattern reversed: the forward order produced lower CKMM scores than the all-at-once baseline ($\beta = -19.30$, 95% CI $[-31.36, -7.23]$), while neither the backward ($\beta = 4.48$, 95% CI $[-7.90, 16.86]$) nor the middle order ($\beta = 2.33$, 95% CI $[-9.59, 14.26]$) differed.

The CKMM model showed a similar pattern in both location conditions. A linear regression without random effects confirmed a significant overall effect of order on CKMM scores, $\chi^2(3) = 25.78$, $p < .001$ in the left-shifted condition. The backward order produced lower CKMM scores than the all-at-once baseline ($\beta = -24.45$, 95% CI $[-39.73, -9.17]$), while neither the forward ($\beta = 10.04$, 95% CI $[-5.16, 25.23]$) nor the middle order ($\beta = 9.49$, 95% CI $[-5.64, 24.63]$) strongly differed. For the right-shifted condition, order was also a significant predictor of CKMM scores, $\chi^2(3) = 15.28$, $p = .002$, and the forward order differed from the all-at-once baseline in optimality ($\beta = -22.45$, 95%CI $[-36.50, -8.41]$) whereas the backward ($\beta = 1.57$, 95% CI$[-12.48, 15.61]$) and middle ($\beta = -0.72$, 95%CI $[-14.71, 3.28]$) did not.

Overall, for participants, orders that conflicted with item location led to the creation of less optimal category systems in both the left- and right-shifted conditions, and the simulated data captured this pattern. However, in both the human and simulated data, orders that did not conflict with item location did not differ significantly from the all-at-once baseline. This suggests that the all-at-once baseline does not necessarily produce the most optimal category systems, and that divergence from this baseline does not necessarily reflect suboptimality.

5.4 Model Comparison

The fitted CKMM model captured the broad trends in behaviour across our hypotheses, reproducing sequence, priming, and hierarchy effects, though the picture was less clear for the JSD analyses. Here we compare our model to existing rational models of categorization (RMCs) to determine which best accounts for the data at the participant level. We only fit these models to data from the third participant pool, as information such as distractor order had not been recorded for the initial two pools, but was needed to fit the model to the data.

Typically, RMCs are fit to data from simultaneous presentation tasks where likelihood is computed over final category assignments. To fit the incremental nature of our task, the maximum a posteriori version of the rational model of categorization (MAP-RMC; Anderson, 1991) and the particle filter version (particle-RMC; Sanborn et al., 2010) were each adapted. To make the particle filter model incremental, we used a single-particle. Both models were modified to fit the constraints of the task. Specifically, the coupling constraint c was ignored once the maximum number of categories permitted by the experiment was reached, and probability mass was then spread across the existing categories. An α mixture parameter was included to allow the models to be fit to participant data. Without α, the MAP-RMC's deterministic assignment rule assigns zero probability to all but the optimal response at each step, making likelihood-based fitting impossible. We include it in the particle-RMC as well for consistency as a noise parameter.

As both RMCs assume flat category structures, model fitting was performed separately for each depth condition. In contrast, the CKMM model was fit to both depth conditions simultaneously, with the same parameters selected for both. Results are reported separately for each depth to allow for comparison. In addition to the full CKMM model, we evaluated two ablated versions: one without the α mixture parameter, and one replacing the Softmax choice rule with a greedy policy that assigned all probability mass to the highest-scoring category system while retaining α for model fitting. For the greedy CKMM model, $\alpha = 0$ for the same reason as in MAP-RMC. Without the α mixture parameter, the greedy assignment rule assigns zero probability to all but the highest-scoring category system.

Model performance was assessed using 11-fold cross-validation on participant data from pool 3 ($n = 165$). For each fold, a grid search was used to identify the parameters that maximized the log-likelihood of the training data, and the log-likelihood of the data in the held-out fold was then computed using the selected parameters. For the CKMM model and its ablations, the grid search was over τ and α; for the MAP-RMC and particle filter model, it was over α and the coupling constraint c. The results of the cross-validation are illustrated in Figure 5.10. Log-likelihoods are reported as gains over a random baseline, where each item has an equal probability of being assigned to any possible category.

Across both depth conditions, the ordered CKMM model and its ablations tended to outperform the MAP-RMC and particle-RMC models, achieving higher held-out log-likelihoods across cross-validation folds. Greedy CKMM better accounted for participant data than MAP-RMC in both depth conditions and did so with less weight placed on the uniform noise component (e.g., $\alpha = .27$ for Greedy CKMM vs. $\alpha = .39$ for MAP-RMC in the three-level condition). In both the two-level and three-level conditions, particle-RMC provided the best fit when $\alpha = 0$. Softmax CKMM outperformed particle-RMC in the two-level condition in both the ablated ($\alpha = 0$) and full ($\alpha = .15$) versions. In the three-level condition, however, Softmax CKMM without α did not outperform particle-RMC. Only the full Softmax CKMM model, which included a non-zero weight on the uniform noise com-

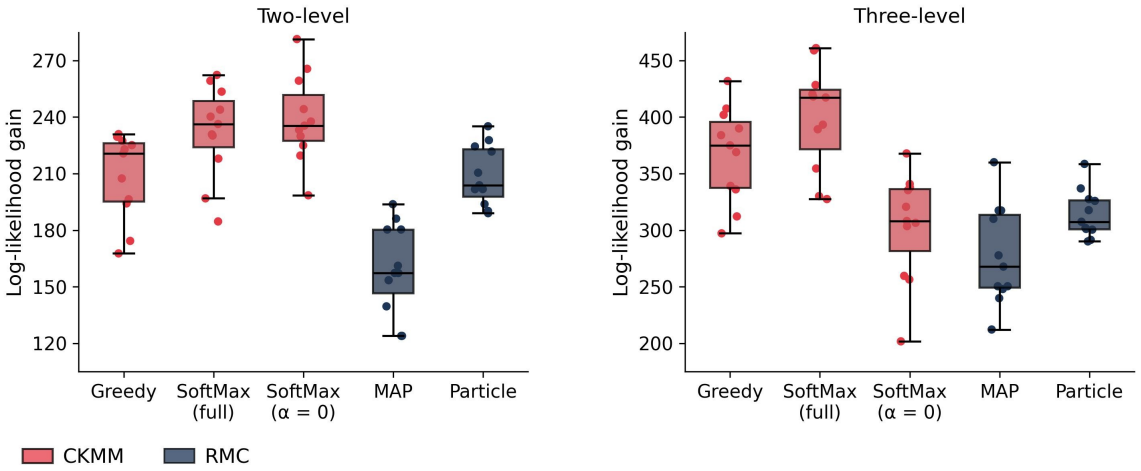


Figure 5.10: Log-likelihood gain over random baseline for CKMM and RMC models on held-out folds ($n = 15$ per fold). Results are reported separately for the two-level and three-level conditions. Log-likelihoods are reported as gain over the log-likelihood of the data under random assignment to categories (-428.5 for two-level systems and -758.9 for three-level systems). Results are from an 11-fold cross-validation.

ponent, achieved higher held-out log-likelihoods than particle-RMC. Full details, including all best-fitting parameter values, are provided in the Appendix B.5.

Overall, the ordered CKMM model provides a better account of participants' incremental categorization behaviour than the adapted RMC models across both depth conditions. This advantage was observed despite the CKMM model being fit with a single set of parameters across both depth conditions and the RMC models being fit separately for each condition. These results demonstrate that a model incorporating hierarchical structure can better capture behaviour in the present sequential categorization task than models based on flat category representations.

5.5 Discussion

Category systems do not form all at once but instead develop as items are encountered over time. We examined this process with a novel experiment and a model of categorization designed to simulate the development of a category system in response to sequentially encountered stimuli. In the experiment, participants were asked to incorporate novel stimuli into an existing taxonomy, and there were four key results. First, presentation order consistently shapes how a category system evolves. Second, participants are more likely than not to put all novel items into the same category as the first item they categorize. Third, the structure of the existing system matters: participants are more likely to introduce a new second-level category when the taxonomy has two levels rather than three. And fourth, although items presented all-at-once did not lead to significantly more optimal systems

when compared to all orders, certain item sequences can lead to the creation of less optimal category systems when optimality is defined in terms of within- and between-category similarity.

The ordered CKMM model formalized category system evolution as a sequential process in which items are probabilistically assigned to categories based on the relative improvement each assignment provides to the compactness of the hierarchy. Assignments that better preserve within-category similarity and between-category dissimilarity are more likely to be selected. This model was able to capture the broad patterns in participant behaviour. Specifically, the model exhibited similar sequence, priming, and hierarchy effects, and tended to create suboptimal systems under the same conditions as participants. The ordered CKMM model also achieved higher cross-validated log-likelihoods, indicating better predictive fit to participant data than the rational models we tested.

The results of the experiment have several implications. First, while sequence effects are well established, our study departs from previous findings in some respects. Previous work using "dynamic Sorites" tasks shows that when participants categorize colour patches that vary along a single dimension from blue to green, the boundary between these two categories is closer to blue when the patches are presented in order from blue to green than when the patches are presented from green to blue (Egré et al., 2013; Raffman, 2011). This "negative hysteresis" effect contrasts with the "hysteresis effect" that we found: left-to-right presentation leads to more items categorized as L, and right-to-left leads to more items categorized as R. To us, the hysteresis effect seems more intuitive in our setting, but warrants further exploration.

One possible explanation for the difference in hysteresis effects concerns the information available to participants during categorization and how the category system was represented. In Egré et al. (2013), negative hysteresis was observed in a linguistic task in which participants categorized stimuli using the labels "blue" and "green". In contrast, no significant hysteresis effect was found when participants judged colours relative to visual blue and green category endpoints. The authors suggested that these visible endpoints acted as anchors, allowing participants to estimate the boundary between categories and judge category membership relative to it. In our task, participants also had access to category endpoints, but additionally saw the full hierarchical structure of the system and all of their previous categorization decisions. This may have anchored participants' decisions to the developing category structure and encouraged consistency with previous choices, potentially contributing to the hysteresis effect we observed. Future work could investigate the effects of internal versus external category representations by asking participants to learn and construct category systems without a visual representation of the hierarchy. Participants would be required to maintain and update category structures internally, while remembering their past categorization decisions. Comparing this condition with one in which the hierarchy

remains visible could help establish how internal versus external representations of category systems influence the direction and magnitude of hysteresis effects.

The priming effect we find has implications for how institutional category systems develop, particularly when new information must be integrated into existing structures. Early associations can influence how emerging categories are organized even after those initial associations become less warranted. Returning to the example of fungi, early category decisions can continue to influence institutional organization even after scientific understanding changes. Fungi were historically studied within botany and classified alongside plants, and this relationship is reflected in the continued organization of fungal collections within botanical institutions. For example, the University of British Columbia Herbarium includes plants, fungi, algae, and lichens, potentially reflecting the historical organization of these fields within botanical research.[7]

Our findings also suggest that the structure of an existing category system influences where new categories are most likely to emerge. In fine-grained systems, modifying superordinate categories would likely require substantial revisions, making it more natural to introduce categories at lower levels. In domains where categorization is still coarse, it may be more natural to establish broader superordinate categories before going on to define more specific ones. This has implications for the order in which levels of abstraction emerge during category system evolution and development. While some work suggests that the basic level emerges before superordinate or subordinate levels (Mervis & Crisafi, 1982), our results align more closely with accounts in which superordinate categories emerge first (P. C. Quinn & Johnson, 2000). It is worth noting, however, that those studies concern the emergence of categories in early development, whereas our results presuppose an existing category system, which limits direct comparison. Regardless, the more subcategories that exist, the more opportunities there are to create lower-level categories or place items into existing ones rather than introducing new categories at the top. Thus, it seems plausible that higher-level divisions are more common early in the development of a category system, while more established, fine-grained systems see the introduction of comparatively fewer categories near the top.

Finally, the modelling results provide evidence that compactness captures an important aspect of category system optimality, while also revealing how systems can become less optimal with time. Ordered CKMM's ability to predict participant data and simulate overall behavioural patterns suggests that compactness is a good measure of hierarchy quality, as participants tended to make decisions consistent with maximizing it. The model also demonstrated that an agent making decisions to maximize compactness can ultimately produce a less compact and less optimal system due to the incremental nature of its construction. This suggests that category systems can become increasingly suboptimal as agents encounter new information and revise their understanding of the environment. At

[7] See https://beatymuseum.ubc.ca/research-2/collections/herbarium/

the same time, the assumptions that allow the model to capture this process also influence how robust it is when participants deviate from the predicted categories. When α was excluded from the ordered CKMM model, its performance fell more in line with that of the particle-RMC. This is likely due to the assumption of linear ordering in ordered CKMM. When items such as distractors were miscategorized, they had a large negative effect on the CKMM score, resulting in very little probability mass being assigned to any category other than the one intended for distractors. Although the RMC would also assign these category choices lower probability mass, the effects would be less extreme.

Although the assumption of an underlying linear order can limit the model in some cases, this assumption also allows the model to capture structure in the category systems being considered. Because our model is implicitly hierarchical, we expected larger performance gains in the three-level than in the two-level condition, which had a less hierarchical structure. The ordered CKMM model was better able to account for participant data in both conditions, even when fit jointly to two- and three-level representations using the same parameters. The RMC had to be fit separately to each level and could not jointly account for both. Some of these advantages likely stem from the ordering assumptions built into the CKMM objective. When computing the CKMM score, we consider only a subset of splits based on the ordering of items along a line and the corresponding ordering of categories. Because both the category system and stimulus set were designed to respect this ordering, it accounts for part of the performance advantage. In settings where such an ordering is less explicit, this advantage may be smaller – though knowledge of hierarchical structure could still lead our model to outperform models that assume flat partitions.

Overall, the results of both the experiment and the modelling suggest that the evolution of a hierarchical system is shaped by the history of its construction. Specifically, the existing structure, the order in which items are encountered, and early categorization decisions all influence the final structure and its optimality. The ability of our model to predict participant behaviour while maximizing compactness suggests that compactness captures some meaningful notion of category goodness. Its stronger predictive performance relative to flat RMC models also underscores the need for models that are intrinsically hierarchical.

5.5.1 When Does the Model Diverge from Human Behaviour?

Although the ordered CKMM model outperformed existing rational models in predicting participants' responses, its behaviour diverged from that of participants in three key respects. First, while the model supported hypothesis H1a(iii), we did not find evidence that suggested participants were more likely to create a new category X when items were presented in the middle order as opposed to the backward or forward order. One possible explanation is that system depth obscured the effect of order in participants' data. If two of a participant's four trials were in the centred condition, one was always in the two-level

condition and the other in the three-level condition. Because creation of category X is less likely in the three-level condition, if a participant's middle order trial is in the three-level condition and their forward or backward order trial is in the two-level condition, then the increase in likelihood of category X might be cancelled out by the decrease in likelihood attributed to a three-level system. Controlling for depth in a subsequent regression analysis (see Appendix B.4.1) yielded a positive effect ($\beta = 0.60$, 95% CI $[0.24, 0.95]$) of the middle order with respect to creation of category X that improved fit relative to a depth-only model ($\chi^2(1) = 10.29, p = .001$). However, the effect was weak in comparison to the other observed order effects and does not provide as strong evidence in favour of H1a(iii) for participants.

Second, the model showed a weaker tendency than participants to place all items into the same category, more frequently assigning one out of eight items to a separate category. However, as discussed previously, this difference could reflect perceptual limitations in humans that are not captured by the model. It may also reflect a slight bias toward assigning novel items to the category that participants have already been using when the item is similarly related to examples from both categories. Future work using stimuli that become increasingly divergent from the category participants have primarily relied on could further test this tendency toward same-category assignment. Alternatively, attempting to replicate the study with more distinct or complex stimuli that can still be represented along a continuous range of similarity (e.g., books) could help address potential perceptual limitations of the very similar-looking aliens we used.

Finally, the model supported hypothesis H4, whereas the experimental results did not. While the experimental results showed that item distributions across categories differed significantly between the forward and backward sequence conditions and the all-at-once baseline, they did not show that the largest divergences occurred when presentation order conflicted with item location (e.g., forward presentation in the right-shifted condition). This could reflect greater variability in the category systems participants constructed across conditions when compared to the model, though the fact that the CKMM-based analyses reveal the predicted effects despite also averaging across participants suggests it may instead reflect a limitation in the precision of the JSD as a measure of divergence.

The fact that a clearer pattern of optimality emerges with the CKMM objective, and that participants and the model tend to agree on optimality in this case, suggests that it may be a better measure of category goodness. The CKMM score is a more precise measure as it considers the entire system rather than averaging across individual category assignments. This is particularly relevant in the three-level condition, where the CKMM score accounts for third-level category assignments, whereas the JSD analysis only considers assignments at the second level.

5.5.2 Limitations & Future Work

There are several limitations and avenues for future work. In this chapter, we considered an incremental categorization model (ordered CKMM) based on a hierarchical clustering objective (Cohen-Addad et al., 2019; Naumov et al., 2022) and compared it to flat RMCs. This model better accounted for our participant data than the flat models and allowed us to explicitly formalize a definition of optimality that scored the entire hierarchical structure of the category system rather than separate levels in isolation. This objective aligned with psychological ideas of category goodness. Still, other models of categorization are explicitly hierarchical. Future work, for example, could adapt hierarchical Bayesian models of categorization (e.g., Canini & Griffiths, 2011) to compare them against our clustering objective.

Although our model captures human behaviour well, it is not intended as a cognitively plausible process model of category system evolution. The ordered CKMM model describes how incremental categorization decisions can produce hierarchical structures that optimize compactness, but it does not specify the mechanisms by which people actually form, maintain, and revise these structures internally. A process model would need to account for how people represent existing category systems, incorporate newly encountered items, and decide whether to extend existing categories or reorganize the hierarchy. In particular, to better explore how systems become suboptimal and how they are revised, there would need to be some mechanism that detects increasing inconsistencies between the category system and the environment. For example, people might accumulate prediction and communication errors when newly encountered items are poorly accommodated by the existing category structure, with sufficiently large or persistent errors eventually triggering revision. Alternatively, people might maintain some internal proxy for category coherence or compactness, with increasing dissimilarity among category members providing a signal that the current structure is no longer useful and should be revised.

Our experimental paradigm did not allow participants to revise previously made decisions, but future work could introduce greater flexibility to study when and how people revise existing hierarchies. This could include designing an experiment that examines the interaction between system complexity and the willingness of participants to perform major restructuring, particularly to top-level categories. An important question to consider is how the costs of reorganizing a system trade off against the potential gains in optimality that reorganization would bring. These gains could be operationalized in terms of improvements in compactness (e.g., the CKMM objective) or informativeness, while the costs of reorganization would likely vary across types of systems. For example, major (and even minor) changes to library classification systems require substantial effort, including relocating thousands of books, updating catalogues, and communicating these changes to the many libraries that use a given system. In cultural systems, revising conceptual hierarchies likely

incurs individual cognitive costs, as well as group-level communicative costs, as the changes must be learned and adopted by others.

Future work could operationalize restructuring costs in terms of factors such as system complexity (e.g., the depth and number of categories), the number of system users, and the scale of the change (e.g., the number of items affected). Future models predicting restructuring would therefore need to account for a trade-off between the benefits of improved optimality and the costs of reorganization, such that restructuring is more likely to occur when the expected gains exceed a threshold determined in part by these costs. This threshold may be higher for systems in which reorganization is more costly, allowing suboptimal structures to persist even when more optimal alternatives are available. Considering structural reorganization would also connect our framework to theories of conceptual change and acquisition (Chi, 1992; Keil, 1989), and enable the study of how these trade-offs shape the development and revision of category systems over time.

The version of the CKMM objective we used was extended both to non-binary hierarchies and to hierarchical clustering of high-dimensional embeddings (Naumov et al., 2022). This opens the possibility of combining our approach with high-dimensional representations of books (Schmidt, 2018) to simulate how a library classification system might be sequentially updated over time. By simulating the evolution of institutional systems like library classifications, we can ask whether they become gradually less optimal over time, and whether major reorganizations (e.g. the reclassification of Fungi) help correct this drift away from optimality. Work on Wikipedia has employed a measure of internal cluster coherence based on links between articles and found evidence consistent with this view, observing drops in these measures preceding periods of major reorganization (Suchecki et al., 2012). Beyond its theoretical interest, the CKMM objective could have practical applications. For example, it could be used to propose optimal updates to a classification system or suggest category placements for novel books.

Similarly, the ordered CKMM model could be applied to word or image embeddings of animals to examine sequence effects in incorporating new creatures into a folk taxonomy. For example, predictions of where animals such as sheep would be placed within an established taxonomy could be compared with simulations of early taxonomy formation, linking these results to the English and Tenejapa Tzeltal examples discussed in the introduction. Other examples of late introduction of animal species into a culture could be examined to see if ordered CKMM predicts where and how they are categorized.

5.6 Conclusion

Category systems are everywhere and shape how we make sense of the world around us. We have demonstrated how the evolution of a category system is dependent on early category decisions. In particular, we have shown that item presentation order can lead to the

evolution of category systems that are less optimal (or biased) representations of the items they categorize than those that would be created under different conditions. We have also shown that the structure of an existing system can constrain its future development by examining how hierarchical organization influences the creation of new top-level categories. Altogether, these results demonstrate that the structure of a category system reflects more than the items it categorizes; it also reflects the history of its construction. Although we have primarily focused on the evolution of a synthetic system in which items vary along a single dimension, we have introduced a model that can be extended to more complex stimuli and connected these ideas about category evolution to broader questions about the development of institutional systems. Studying how these systems develop and become skewed is important for understanding how we represent the world internally and may help explain why large-scale institutional systems, such as the Dewey Decimal Classification, take the forms that they do.

Chapter 6

Conclusion

The structure of a hierarchical category system is not just dependent on the items it categorizes. It also reflects the biases and perspectives of the people who created it and the history of its construction. This is true whether the category system is learned through development and shared within a culture, maintained as a large-scale institutional taxonomy, or constructed synthetically to investigate a specific psychological process in the lab. Because hierarchical category systems shape how information is stored, interpreted, and retrieved, understanding their structure and development is central to explaining how people represent and make sense of the world. Therefore, in this thesis, I presented a large-scale computational analysis of the structure and evolution of hierarchical category systems that examined both how their structures encode bias and how they evolve in response to incrementally encountered items.

Table 6.1: Space of current and future research on hierarchical category systems revisited.

	Cultural	**Institutional**	**Synthetic**
Structure	Chapter 4	Chapter 3 & 4	•
Evolution	•	Chapter 3	Chapter 5

Table 6.1 summarizes the space of current and future directions for researching the structure and evolution of cultural, institutional, and synthetic hierarchical category systems. In Chapters 3 and 4, I focused on the structure of institutional category systems, demonstrating how they encode biases in their categories, items, and labels. I also demonstrated how the bias framework developed in Chapter 4 could be extended to cultural category systems such as folk taxonomies by finding evidence of bias towards domestic mammals in English WordNet. Chapter 3 began to investigate the evolution of institutional category systems by demonstrating that while gender bias in Wikipedia's categories had reduced over time, it had not noticeably changed in the Library of Congress Subject Headings (LCSH). Finally, in Chapter 5, inspired by observed trends in the history of institutional category systems,

I designed an experiment and a computational model that demonstrate how hierarchical category systems evolve to accommodate sequentially encountered items. In particular, I demonstrated how some orders lead to the construction of suboptimal or skewed category systems.

Because hierarchical category systems are such a prominent form of representation, their structure and evolution are of interest to a broad array of disciplines beyond psychology. As I have discussed, two that are particularly relevant to the work in this thesis are library and information science and computer science. In the next sections, I describe the main findings and their implications for the three fields listed above. I then discuss the limitations of this work and future directions that can address both the limitations and some of the gaps that remain in Table 6.1.

6.1 Main Findings

In this thesis, I had three main research questions related to the structure and evolution of hierarchical category systems. The first question (RQ1) asked if biases in institutional category systems were consistent with psychological principles of categorization. In Chapter 3, I addressed this by demonstrating that, like with biases in cultural categories, some of the labelling biases could be attributed to people's perceived frequency of membership in the category. Previous work has suggested that frequency of instantiation might account for gender biases in humans (Bailey et al., 2019), which lines up better with perceived frequency than actual frequency, as frequency of instantiation also affects people's perceptions of who is frequent and therefore more typical or default. Though the focus of this chapter is not as explicitly about hierarchy as the rest, it demonstrates how bias can manifest in the structure of a hierarchical category system through the inclusion and explicit labelling of certain sub-categories (those relating to women) over others (those relating to men). Additionally, by exploring principles of typicality that are traditionally examined in cultural contexts, I show the utility of using institutional category systems to study human category systems and categorization. Moreover, the biases found in institutional category systems can both exaggerate and mitigate the biases found in cultural category systems, like the concepts we express through language.

The second research question (RQ2) asked how the structure of a hierarchical category system can encode bias. In Chapter 3, we demonstrated how it can encode bias through the unequal labelling of subcategories, and in Chapter 4, we extended this to the number and arrangement of the categories and items a system contains. In particular, we found that although both the Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC) contain bias, the DDC contained stronger bias than did the LCC, and in particular, this was because it had a more pronounced category count and descendant bias, and stronger distributional gender bias. We also found that our framework could be

extended to cultural category systems, as we used it to pick out category biases in the conceptual organization of mammals by English WordNet.

Finally, the third research question (RQ3) asked how a hierarchical category system evolves in response to sequentially presented stimuli. Our experiment demonstrated that participants' final category systems depended primarily on the first category choices they made, the order in which subsequent stimuli were encountered, and the existing structure (depth) of the system. Their decision to include a new category was influenced by the granularity of the systems, such that shallower systems increased the likelihood of new category creation. Overall, this demonstrated that category systems are very dependent on their history of construction.

The results of Chapter 5 also provide an explanation for why institutional category systems might become biased or suboptimal. As discussed throughout the thesis, biases can arise in the system because of external pressures. If the system is designed to accurately reflect the items it categorizes, and they themselves contain bias, then the system will also be biased. But biases can also arise internally because the system misrepresents or skews the space of things it is meant to categorize. With the computational model introduced in Chapter 5, we demonstrated one such mechanism by which systems can become internally biased. If we define bias as a mismatch between the underlying item-space and the structure of the system, then category systems can become biased because of the order in which items are encountered. Early category systems that were warranted at the time of their conception might become more biased as new things are discovered, and their existing structures limit how they are adapted to accommodate these new things.

6.2 Implications

The work presented in this thesis has implications for understanding how hierarchical structure develops in cultural category systems. It also has implications for library classification systems specifically, as well as for the broader categorical datasets that underpin many contemporary technologies.

6.2.1 Implications for the Basic-Level Advantage

These results have broader implications for how levels of abstraction emerge within a hierarchy. In particular, the finding in Chapter 5 that higher-level categories are more likely when a hierarchy is shallower suggests that superordinate categories may come first in the development of hierarchical category systems. These results suggest that people more readily make broad, high-level divisions when a representational space is still sparse and not yet finely structured.

This connects to debates about the basic-level advantage in categorization. It has long been suggested that basic categories emerge first in the development of category hierarchies (R. Brown, 1958; Djalal et al., 2017; Mervis & Crisafi, 1982). However, both connectionist models and information-theoretic models of category development have suggested that broad superordinate categories may precede more specific basic and subordinate ones (Khishigsuren et al., 2026; McClelland & Rogers, 2003) and research with infants has backed this up with evidence of early superordinate-level groupings (Croteau et al., 2026; Mandler & McDonough, 1998; P. C. Quinn & Johnson, 2000). There is debate about how to interpret these early representations: some accounts argue that infant categories are not yet adult-like in nature (Rakison, 2000), while others suggest that these categories may reflect perceptual groupings rather than superordinate-like conceptual structure (Murphy, 2016). However, they do not necessarily rule out the idea that category systems start with broad groupings of a conceptual space, just that they are not like the superordinate and taxonomic categories we traditionally discuss with respect to adults. The work in Chapter 5 is consistent with the idea that superordinate categories emerge first, but it only deals with very simple hierarchies adapted by adults. More research is needed to make more concrete claims.

6.2.2 Practical Implications for Library Classification

Concerns about bias in library classification systems are not new. Librarians and library scientists have been discussing these issues since at least the 1970s, when Berman (1971) published *Prejudices and Antipathies*, proposing numerous modifications to the LCSH to reduce embedded biases, particularly Western and gender biases. Certain gendered labels have been modified (e.g., changing "Women as Scientists" to "Women Scientists"), but this thesis suggests that broader structural patterns of asymmetric labelling persist, with little substantial change over the last thirty years. If gender bias remains prominent after decades of criticism and revision, it is likely that many other biases identified in Berman (1971) also persist. This conclusion is consistent with the assessments of both Knowlton (2005) and Mullender (2025), who note that significant work remains to be done despite 30 and 50 years of bias reform efforts, respectively. In contrast, our analysis of Wikipedia suggests that bias can sometimes be reduced through ongoing collective revision. Although far from unbiased, gendered nationality and ethnicity categories were substantially more balanced than equivalent terms in everyday language. This raises an important question: why has Wikipedia changed while the LCSH has changed comparatively little? I discussed this briefly in Chapter 3, but more consideration is warranted.

Part of the answer likely lies in the mechanisms through which change occurs. Revising subject headings in the LCSH is a lengthy process. Proposed headings must be carefully researched, checked against existing headings, appropriately formulated, and formally submitted before undergoing review. Similar procedures underlie the modification or removal of

existing headings. On the one hand, this is understandable as library classification systems organize large collections of material, and any modification can affect both current users and future cataloguing decisions. However, the same procedures that promote stability also slow change, allowing biases to persist long after they have been identified.

Some aspects of this process could potentially be automated. Natural language processing techniques could be used to identify candidate terms from the literature, while the methods developed in Chapters 3 and 4 could help identify structurally unbalanced regions of a classification system. With further development, these approaches may even help detect forms of bias in areas where researchers do not already expect them to occur. Regardless of the specific solution, mechanisms are needed to make large-scale classification systems more responsive because the biases they contain have real-world consequences.

Wikipedia offers a useful contrast. Although imperfect, it has undergone substantial change over time, in part because of its collaborative editing model. Weinberger (2007) argues that neutrality on Wikipedia emerges through a process of continual revision and contestation in which editors challenge, modify, and refine each other's contributions until a rough consensus is reached. Categories, like articles, are subject to this process of negotiation and revision. Consensus, of course, does not guarantee objectivity. Wikipedia's editor population remains disproportionately male, English-speaking, and concentrated in the United States, meaning that consensus often reflects the perspectives of dominant groups. While Wikipedia performed better than the LCSH on our measures of gender bias, it still exhibited asymmetries that largely mirror broader patterns of language use (excluding identities). Additionally, other work has demonstrated systematic differences in how women are represented across Wikipedia's articles and categories (Schmahl et al., 2020; Sun & Peng, 2021).

These concerns extend beyond the LCSH. The LCC and DDC are used throughout the world, with the LCC mainly being used in North American academic libraries, while the DDC, along with modified versions such as the Universal Decimal Classification (UDC), is used internationally. Like the LCSH, these systems have been criticized for Western bias, gender bias, and other forms of bias. The analyses presented in this thesis demonstrate how such biases can become embedded within the structure of a classification system itself. Although improvements have been made over time (the very biased version of Christianity that exists in the DDC today is already an improvement on previous versions), the processes required to implement large-scale structural changes are often even more difficult than those required to revise subject headings. For a category to be moved, the number it used to occupy must remain vacant for a period of time so that the libraries have time to reorganize their books and catch up. Inserting a new category between existing ones means books need to be re-shelved to make room.

Institutional category systems will likely always lag behind the resources they aim to represent. Perfectly accurate and unbiased classification is unattainable, and some delay

between identifying a problem and fixing it is inevitable. Nevertheless, the methods developed in this thesis provide tools for identifying structural biases at scale and may help institutions respond more quickly. This matters not only because library classification systems shape how knowledge is organized and accessed, but because they increasingly serve as foundations for digital taxonomies and other large-scale information systems (Godby & Stuler, 2013), allowing embedded biases to propagate beyond libraries themselves.

One practical improvement would be to establish more consistent principles for geographic and historical classification. For example, continents should generally be represented at comparable hierarchical depths rather than having large regions treated at the same level as individual countries. A system in which the entire continent of Africa is classified at the same hierarchical level as countries such as Greece or Britain (as it is in the LCC) creates an imbalance in the representation and granularity of geographic categories. Another possible improvement would be to reduce the extent to which gender categories are embedded within occupational or nationality categories in the LCSH. Rather than relying on combined headings such as "Women scientists", gender and occupation could be represented through separate headings for categories such as "Women" and "Scientists." Since catalogued works can receive multiple subject headings, this approach would allow relationships between concepts to be represented without creating asymmetric categories for particular groups. The ordering and display of headings could follow established cataloguing conventions, while the underlying structure would remain more flexible and consistent. These changes would be by no means easy to implement, but they illustrate how classification systems could move toward more balanced presentations of the world, without being re-created entirely from scratch.

Finally, work at the intersection of library and information science and computer science has attempted to automate library classification (Ávila-Argüelles et al., 2010; Frank & Paynter, 2004; Larson, 1992). Clustering techniques like those of Naumov et al. (2022), which are designed to handle the clustering of high-dimensional embeddings combined with effective vectorized representations of books like the HathiTrust book embeddings (Schmidt, 2018), might serve as one approach toward developing effective automatic library classification. This could be used to discover latent categorical structures in books that might not be immediately obvious and used to improve existing classification systems. However, the benefits of automated classification remain constrained by the collections on which it is based. If the collection itself is biased, then any classification designed to accurately represent the collection will be biased. Thus, improvements to categorization also require improving the collection.

6.2.3 Implications for NLP, Computational Technologies, & Beyond

As discussed previously, library classification systems often serve as a basis for the development of ontologies and taxonomies. The other systems I examine function in similar ways. WordNet underlies ImageNet, a system with well-documented bias that is widely used as the basis for machine classification and object recognition (Luccioni & Rolnick, 2023; Stock & Cisse, 2018). WordNet is also used in natural language processing more broadly, including in the construction of semantic embeddings and knowledge graphs. Similarly, Wikipedia is frequently scraped to build large-scale text corpora and support downstream NLP tasks, including ontology construction (Ben Aouicha et al., 2016; Hejazy & El-Beltagy, 2013; Zesch & Gurevych, 2007). Across all of these systems, hierarchical organization plays a central role, and this structure is also where many of these biases are introduced and reinforced, whether through asymmetric labelling of subcategories or more direct structuring of items and categories within a hierarchy. Because these systems frequently form the basis of technologies such as image search, the biases embedded in them carry through into downstream applications. As such, this work also points to broader concerns about how hierarchical category systems shape the technologies built on top of them. Additionally, many databases are built on top of hierarchical structured data for efficient retrieval (e.g., B-trees), but if these hierarchies can project biased structures onto the data, then they can have consequences for how information is retrieved.

6.3 Limitations

Individual limitations of each of the studies presented in this thesis are discussed in their respective chapters. However, two key limitations are worth highlighting across all three studies.

First, the work presented in this thesis is very Anglo-centric and focuses on Western, Educated, Industrialized, Rich (WEIRD) category systems and populations (Henrich et al., 2010). In Chapter 3, we consider labels from English versions of Wikipedia and the LCSH and compare them to labels in the English Ngram corpus. In Chapter 4, we investigate US-based classification systems, and in Chapter 5, our data come from English-speaking participants, the majority of whom were from the US and UK. This limits the generalizability of our results, as they do not reflect the full cognitive diversity of those who construct and use hierarchical category systems (Barrett, 2020). For example, although Berlin (1992) and López et al. (1997) find that folk taxonomies of natural kinds such as animals share cross-cultural commonalities, they also differ in what is grouped together, at what level this grouping occurs, and which aspects of animals the groupings are based on.

Future studies could extend our framework for category and item bias to systems such as the Classification for Chinese Libraries (Zhang, 2003) to determine whether our approach

is useful for identifying non-Western biases in non-Western contexts. Extending the work of Chapter 3 to other languages beyond English, especially those that include grammatical gender, could provide insight into how gender marking and typicality differ across languages. Falenska and Çetinoğlu (2021) already investigates labelling asymmetries in other language versions of Wikipedia, and these could additionally be compared to usage statistics and n-grams in other languages where data exists. Finally, the experiment in Chapter 5 could be translated and used to collect data from a broader range of countries and populations.

Second, although throughout our work we propose measures for quantifying bias, we do not propose methods for reducing it. Further work could apply these measures in library and information science to develop interventions that seek to reduce bias. This would be challenging, as there may be tensions between accurately representing biased holdings and actively reducing bias, and it would also likely be highly context-specific. However, it would offer avenues for change that may be more feasible to implement in large-scale systems. In a similar vein, we have proposed a measure of optimality and used it to predict and simulate human behaviour; however, it is not yet clear whether it can be used to assess the quality of real-world systems. This method was initially designed to be extended to high-dimensional vector representations such as books (Naumov et al., 2022), so it could potentially be used to measure the quality of existing systems and propose changes to increase optimality. Future work could also improve our understanding of what makes a system optimal by testing whether measures of optimality correspond to subjective judgments of quality, ease of browsing, or retrieval. Improving the fit of a system would likely involve balancing tensions between these criteria.

6.4 Future Directions

With this thesis, I also defined a space of research for the study of hierarchical category systems illustrated in Table 6.1. Although this thesis has addressed some of what was defined, gaps still remain, and here I discuss future work that could help to address this.

6.4.1 Biases Across Cultural, Institutional, and Synthetic Systems

Throughout this thesis, I have demonstrated that biases can emerge across a range of category systems. Chapters 3 and 4 established the presence of bias in four institutional category systems. In Chapter 4, I briefly demonstrated how the same framework could be extended to cultural category systems by examining the uneven representation of domestic and wild mammals in WordNet. Similarly, the asymmetries in gendered language observed in Google Books can be interpreted as reflecting asymmetric cultural categories for men and women, consistent with the people-as-men bias (Silveira, 1980). There are, however, opportunities for future work.

One possibility would be to more directly examine how bias develops within the structure of a category system over time. Although we proposed our order effects mechanisms as a way that bias can propagate, a more targeted study is warranted. An experiment where participants are primed to have certain needs, priorities, or biases that are reflected in the systems they create over time could be useful to better understand how biases are encoded into these systems. We could ask broader questions like, under what conditions is bias reduced, and under what conditions does it persist or become amplified? Can small asymmetries introduced early in the construction of a category system compound over time as early categorization decisions constrain later ones? The mechanism explored in Chapter 5 suggests that this may be possible. Following the trajectory of bias through the development of a category system could therefore provide insight into both the origins of structural bias and the conditions under which it is maintained or corrected. This could also be investigated through the application of the bias frameworks to versions of the LCC and DDC across time. One big challenge is finding historical book classifications so that historical versions of the system can be accurately represented as they were at the time. Comparing the evolution of bias in both cultural and institutional contexts may help identify common mechanisms through which biases emerge and become embedded within category structures.

6.4.2 Evolution of Institutional Systems

While Chapter 3 briefly examines changes in gendered category labels in the LCSH and Wikipedia, it does not address the broader mechanisms that drive change in institutional category systems. Future work could reconstruct historical versions of the Library of Congress Classification (LCC) and Dewey Decimal Classification (DDC) systems to identify patterns of category additions, splits, merges, and relocations over time. Such analyses could help explain how category systems develop from their initial conception to their current form and test broader hypotheses about the dynamics of category system growth. For example, they could examine whether densely populated regions of a hierarchy are less likely to see the addition of new categories, as the results of Chapter 5 might suggest.

Specific case studies, such as the incorporation of computer science into the DDC, could provide further insight into how new domains of knowledge are integrated into established classification systems. More broadly, comparing patterns of change across systems such as the DDC, LCC, and Wikipedia could reveal how category evolution differs between institutionally maintained, top-down systems and collaboratively constructed systems. Such comparisons may also help clarify the extent to which institutional category systems follow dynamics similar to those observed in cultural category systems. For example, although much of the work was inspired by perceived order effects in the evolution of institutional category systems, experiments could be used to determine the extent to which this is actually happening.

6.4.3 Emergence and Development of Cultural Systems

Our results demonstrate how category systems can become suboptimal, but a substantial body of work suggests that category systems will ultimately converge on efficient representations of the domains they describe (Carr et al., 2017; Kemp & Regier, 2012). One mechanism through which this may occur is cultural transmission (Silvey et al., 2019). Future work could therefore use an iterated learning paradigm to examine how hierarchical category systems emerge and change over successive generations of learners (Kirby et al., 2008). This paradigm could also test whether cultural transmission reduces suboptimalities that emerge within a generation as conceptual systems are adapted through individual learning and communication. Participants could inherit either full or partial hierarchical category systems, which they would use and modify while completing classification and naming tasks. The set of stimuli could also be altered partway through the transmission chain to examine how category systems adapt to new knowledge or changing environments, and whether such changes produce broader reorganization of the category structure. This would additionally allow researchers to investigate whether increasing suboptimality precedes major structural changes within a system. More generally, further work is needed to understand how hierarchical category systems are maintained, transformed, or lost through cultural transmission.

The work of Carey (1985), Chi (1992), and Keil (1989) also suggests that people make structural changes to their hierarchical category systems as they learn more about the world. New information is either absorbed into an existing system or forces a "radical restructuring" of the underlying categorical hierarchy itself. The mechanism examined in Chapter 5 primarily demonstrates how category systems become suboptimal through the assimilation of items into an existing category structure. Future work could explore the point at which systems become sufficiently suboptimal to necessitate more radical restructuring by allowing participants to reorganize or reclassify items within the hierarchies they are incrementally constructing. It is likely that, as systems become more complex, the threshold for such restructuring becomes higher because more work is required to modify the existing system. Future research could investigate this possibility by varying the complexity of the category systems participants create.

Beyond questions of how hierarchical systems change over time, there remain more fundamental questions about why hierarchical structure emerges in human category systems in the first place. Do hierarchies form because the items being categorized possess an underlying hierarchical structure (e.g., Kemp & Tenenbaum, 2008)? Or do humans impose hierarchical organization because different goals and tasks require representations at varying levels of specificity (e.g., Barsalou, 1991; Hawkins et al., 2018)? Although previous work has explored compositionality in the evolution of language and the formation of recursive substructures, this does not fully explain how hierarchies emerge, change, and persist over

time. While some accounts suggest that basic-level categories emerge first, followed by more general and more specific categories (Berlin, 1977), more recent work has argued that superordinate categories may possess an informational advantage and that a mixture of basic and superordinate categories emerges before other levels of abstraction (Khishigsuren et al., 2026). Future work could address these questions by analyzing the historical development of languages and naturally occurring category systems, and then use these analyses to inform the design of experiments and models simulating the emergence of hierarchies in individuals and groups. Such studies could examine which levels of granularity emerge first and under what conditions multiple levels of abstraction emerge and coexist.

6.5 Final Remarks

Category systems are inevitably subjective, reflecting the perspectives of the people who create them, the cultures and environments in which they are formed, and the processes through which they are constructed. Throughout this thesis, I have emphasized these points by demonstrating how the structure of a category system can encode bias through unequal labelling of subcategories, asymmetries in the arrangement and number of categories, and the ways items are distributed throughout a system. I have additionally proposed a model that demonstrates how participants can construct globally suboptimal systems through a series of locally optimal choices and how early category decisions can shape and constrain future ones.

No category system can serve every person and every purpose equally well at every point in time. Compromises must be made for a system to function effectively, whether they are made explicitly during the construction of an institutional system or emerge implicitly as a mind attempts to optimally represent a space. The structures we create are constrained by both our cognition and the environments in which we live. However, it remains important to understand the skewed and biased ways in which these systems can represent the world. This is especially true when considering institutional category systems and the categorical datasets that underpin many of the technologies through which we access and organize knowledge. While no system can ever be completely neutral, there is always room for improvement. The work presented in this thesis provides several avenues through which the biases embedded in category systems can be assessed and discussed, and offers a starting point for thinking about how such systems might better represent the world they seek to describe.

Appendix A

Supplementary Material for Chapter 4

A.1 LCC and DDC Category Examples

Subcategories of the category "Handicrafts. Arts and Crafts" in the LCC and its DDC equivalent "Handicrafts" (Table A.1).

Table A.1: "Handicrafts. Arts and Crafts" (TT1-999) subcategory names in the LCC and "Handicrafts" (745.5) subcategory names in the DDC.

LCC	DDC
TT161–170.7 Manual Training. School shops	**745.51** Wood crafts
TT174–176 Articles for children	**745.52** Fiber crafts
TT180–200 Woodworking. Furniture Making. Upholstering	**745.53** Leathercraft
TT201–203 Lathework. Turning	**745.54** Paper crafts
TT205–267 Metalworking	**745.55** Shellcraft
TT300–382.8 Painting. Wood Finishing	**745.56** Metal craft
TT387–410 Soft home furnishings	**745.57** Polymer clay work
TT490–695 Clothing manufacture. Dressmaking. Tailoring	**745.58** Beadwork
TT697–927 Home arts. Homecrafts	**745.59** Toy making
TT950–979 Hairdressing. Beauty culture. Barbers' work	
TT980–999 Laundry work	

A.2 Computational Details

To create a tree structure representation of the LCC, we parsed Excel files representing the 21 top-level classes of the LCC. The Excel files were created by extracting data from the 21 LCC outlines stored at https://www.loc.gov/catdir/cpso/lcco/. In the outlines, tabbing is used to indicate the hierarchical relationships between categories. In the Excel

spreadsheets, rows and columns were used to represent this relationship. For example, consider two categories where category A is stored in the first row of the spreadsheet and category B in the second. If category A is in the first column and category B in the second, then category B is a subcategory of A. If B is also in the first column, then A and B are at the same depth.

To create a tree structure representation of the DDC, we used two sources. The first was a set of Excel versions of the OCLC DDC 22 summaries[1] at https://www.oclc.org/research/activities/browser/terms.html. These summaries represented the first 3 levels of categorization in the DDC. For categories at depths greater than 3, we used the MDS table at https://www.librarything.com/mds/. Only categories associated with both a name and a classification number were used. In the DDC, the hierarchical structure of the system is represented by the position of the digit that differentiates a category. After the third level, digits occur following a decimal place. For example, 295.5 is a subcategory of 295.

Finally, to represent the books (or leaf nodes) in the LCC and DDC, we parsed MARC records using the `Pymarc` library. MARC records contain noisy data, so string processing was used to create a uniform format for LCC and DDC numbers, respectively. For the DDC, this meant that the number had to be at least 1 digit long, contain exactly one decimal if greater than 3 digits, and contain only digits otherwise. For the LCC, this meant that the number had to be at least two characters long, contain at least one alphabetic character, and contain at most one number that falls within the range of 1 to 9999. The number can be a natural number or a decimal number, and the alphabetic characters have to precede the numeric ones. Both DDC and LCC classification numbers can contain Cutter numbers. These are additional alphanumeric character sequences that follow another decimal (Cutter). These numbers provide information about a book such as author, publication date, and additional subject information. Cutter numbers were discarded as they were not related to the general category structure of the system. If a book's classification numbers could not be converted to the regularized format, the book was excluded from analysis. Next, for each record with both a DDC and LCC, we determined if the LCC and DDC numbers were valid classification numbers in their respective systems. We called a classification number valid if it could be placed into a category in our tree structure representations. To place books into categories, we found the deepest category that its classification number fell into, and then recursively added that book to the parents of the category.

[1] All copyright rights in the Dewey Decimal Classification system are owned by OCLC, Inc. Dewey, Dewey Decimal Classification, and OCLC are registered trademarks of OCLC, Inc.

Table A.2: Nodes in the LCC and DDC not used in the distributional bias analysis. We separate counts by the following exclusion criteria: nodes that had fewer than 2 children, nodes that had fewer than 100 books, nodes that had books written by men but not women assigned to child categories, and vice versa. The criteria are applied successively from left to right in the table. For example, the counts listed under 'No women' represent nodes with 2 or more children and 100 or more books that still have no books by women distributed among their child nodes. The column named 'Valid nodes' reports the counts for nodes that could be used in the analysis.

	< 2 children	< 100 books	No women	No men	Valid nodes	Total
LCC	5,858	187	3	0	822	6,870
DDC	27,668	3,149	97	5	2,832	33,751

A.3 Distributional Bias Node Data

Table A.2 describes the nodes that were excluded from the distributional bias analysis because they had fewer than 2 child nodes, fewer than 100 books classified under them, or did not have books written by both men and women distributed among their child nodes.

A.4 Distributional Bias: Plug-in Shannon Entropy

In this section, I report the distributional bias results from Warburton, K., Kemp, C., Xu, Y. & Frermann, L. (2024). Quantifying Bias in Hierarchical Category Systems. *Open Mind: Discoveries in Cognitive Science, 8,* 102–130. This version used plug-in Shannon entropy, which is underestimated for small sample sizes. Because the OhioLINK dataset contains fewer records for female-authored books than male-authored books, the plug-in entropy used in the original paper underestimated entropy in the smaller (female-authored) sample, leading to results that are more extreme than those reported in Section 4.4.2.

Figure A.1 shows that the Shannon entropy of the distribution of books by male authors among a node's children tended to be higher than that for female authors in both the LCC and DDC. The effect is stronger in the DDC, where the relative difference between the two counts is 2.34, compared to 1.80 in the LCC. To test the significance of these results, we performed a permutation test in which author gender labels were shuffled among the classified items and the gender distributions were recalculated. In both the DDC and the LCC, the results were significant ($p < .001$).

A.5 Domestic Bias in WordNet

To study bias in WordNet we first extracted a list of 5,587 unique Wikipedia page names, each corresponding to a page for a different mammal species. 28 of the names represented

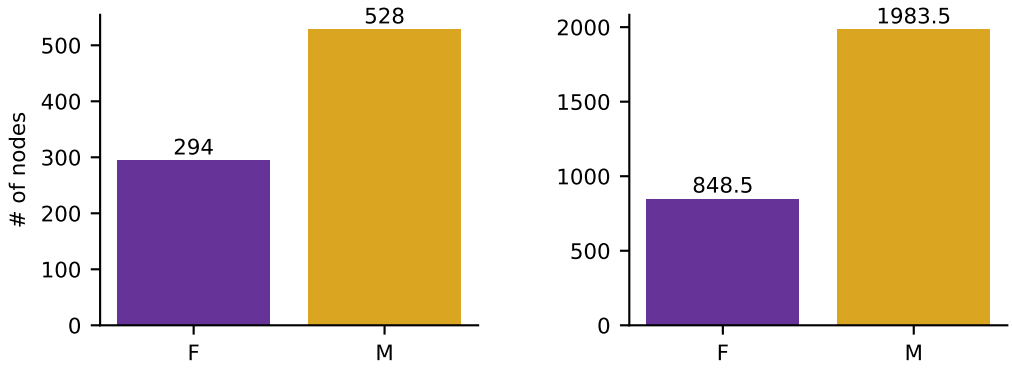


Figure A.1: Distributional bias in the LCC (left) and the DDC (right). The gold columns (M) show the number of nodes for which the distribution of male authors is flatter than the distribution of female authors, and the purple (F) columns show the number of nodes that yielded the opposite result. Ties were handled by adding a count of 0.5 to both columns.

pages for domesticated mammal species, and 5,559 were for wild species. The names were scraped from several Wikipedia pages (Table A.3) that contained lists of various mammal subgroups. The domestic mammal names were taken from a page that listed all domesticated animals, and we manually collected names from this page to ensure that we only had names for mammals. The domesticated animals list sometimes contained subspecies that were domesticated variants of a species that had both wild and domestic subspecies. To keep the analysis consistent, we used the mammal species as opposed to the subspecies and counted it as domesticated. As of November 2023, the Catalogue of Life checklist (Bánki et al., 2023) lists 5,952 extant species of mammals, so our collected list of mammal species seemed relatively comprehensive. Any mammal species that does not have a Wikipedia page is unlikely to have a WordNet synset. We therefore believe that the domestic and wild mammal lists collected from Wikipedia provide a sufficient basis for collecting synsets for these species in WordNet.

To link Wikipedia pages for mammal species with WordNet, we used BabelNet (Navigli & Ponzetto, 2012), a multilingual semantic network that links Wikipedia pages to WordNet synsets when the same concept exists in both datasets. Thus, using the lists of Wikipedia page names for wild and domestic mammal species, we queried BabelNet to see whether a BabelNet entry for each page existed, and if so, whether this entry contained a WordNet ID. If a WordNet ID existed, then we extracted the WordNet synset corresponding to the queried mammal species.[2] All of the Wikipedia pages for domesticated mammals linked to BabelNet entries, and 165 (2.97%) of the wild mammal Wikipedia pages were not found in BabelNet. Of the 28 domesticated BabelNet entries, 19 (67.86%) were linked to unique WordNet synsets. Of the 5,394 wild BabelNet entries, 309 (5.73%) appeared in WordNet.

[2]Sometimes a Wikipedia name directly linked to a BabelNet ID (e.g. "Bali_cattle") and other times it had to be converted to lowercase (e.g. "dog"). We therefore considered both possibilities before concluding that a BabelNet entry for a Wikipedia page could not be found.

Table A.3: Wikipedia pages scraped to obtain Wikipedia IDs for mammal species. URLs are shown relative to `https://en.wikipedia.org`. Pages indented under "List of placental mammals" were linked from that page and scraped separately. Species indicates the number of species extracted from each page.

Page	**URL**	**Species**
List of domesticated animals	/wiki/List_of_domesticated_animals	28
List of placental mammals	/wiki/List_of_placental_mammals	232
List of rodents	/wiki/List_of_rodents	2,495
List of primates	/wiki/List_of_primates	531
List of lagomorphs	/wiki/List_of_lagomorphs	89
List of bats	/wiki/List_of_bats	1,246
List of carnivorans	/wiki/List_of_carnivorans	293
List of artiodactyls	/wiki/List_of_artiodactyls	348
List of monotremes & marsupials	/wiki/List_of_monotremes_and_marsupials	349

We treated the synsets corresponding to mammal species as starting nodes and did not analyze any categories that were parents (hypernyms) of them. We therefore worked with 19 domestic starting nodes and 309 wild starting nodes.

WordNet is not a strict hierarchy, and sometimes synsets can be descendants of multiple nodes. To measure count bias, we therefore counted all unique synsets that either belonged to the list of starting nodes or were descendants of these starting nodes. There were more domestic nodes than wild nodes in WordNet (398 and 344 unique synsets for domestic and wild mammals, respectively). To measure level bias, we calculated the mean depth across all domestic and wild starting nodes individually. The depth was measured from the synset, "entity.n.01", at the top of the WordNet hierarchy.[3] Domestic starting nodes, on average, were slightly deeper in the WordNet hierarchy than wild starting nodes (average depths were 14.34 and 14.27 for domestic and wild nodes, respectively). Finally, to measure descendant bias, we calculated the mean number of descendants across domestic and wild starting nodes. We counted all direct and indirect hyponyms below a synset as descendants. Domestic starting nodes had, on average, 20 descendants, whereas wild starting nodes had 0.13. The low result for wild nodes indicates that most of these nodes had no children. The results for all 3 category biases are summarized in Table 4.6.

The WordNet analyses just described were carried out at the category level (Figure 4.2) rather than the item level (Figure 4.3). For any given person, the items in their semantic representation would be individual mammals they had encountered (e.g. their dog Rover). WordNet has instances of some categories (e.g. Barack Obama is an instance of President of the United States) but not for animals, and is therefore best suited for category-level rather than item-level analyses.

[3] Again, some synsets have multiple parents (hypernyms). If this was the case, we took the average depth between the parents and counted that as the starting node's depth.

Appendix B

Supplementary Material for Chapter 5

B.1 Model Details

In this section, we provide more details about the ordered CKMM model, including a formal definition of the objective (Appendix B.1.1), the implementation of the model (Appendix B.1.2), and how the model was used to simulate participant data (Appendix B.1.3).

B.1.1 Ordered CKMM Objective

We defined the base form of the CKMM objective (Cohen-Addad et al., 2019) in Section 5.2.2. This objective assigns a score to a **binary** hierarchical category system based on the distance between two items (i, j) in vector space and the size of their Lowest Common Ancestor (LCA).[1] We then described the **ordered**, **non-binary** version of the CKMM objective (ordered CKMM) without formally defining the equation. Here, we first introduce the non-binary extension proposed by Naumov et al. (2022). We then build on this extension to formally define the ordered CKMM objective used in our work.

To extend CKMM to non-binary hierarchical category systems, Naumov et al. (2022) modified the objective to account for cases where a triplet of items (i, j, k) share the same pairwise LCA. More specifically, $LCA(i, j) = LCA(j, k) = LCA(i, k)$. Such cases occur in non-binary hierarchies where a category can have more than two child categories. Since the original CKMM objective is defined for binary hierarchies, binary and non-binary structures are not directly comparable. The authors therefore assigned these triplets the expected CKMM score under a random binarization. Let S denote the set of triplets that share the

[1] The category they share that is at the greatest depth in the hierarchy. To make this more concrete, all pairs of items will share the root node of a category system (i.e., the top-level category), but the root node will only count as their LCA if they do not also share any subcategories of the root node.

same LCA. The resulting non-binary CKMM objective is:

$$Q_{NB}(H) = \sum_{(i<j<k)\notin S} \left(d_{\hat{i}\hat{k}} + d_{\hat{j}\hat{k}}\right) + \frac{2}{3} \sum_{(i<j<k)\in S} (d_{ij} + d_{ik} + d_{jk}) + 2\sum_{i<j} d_{ij} \tag{B.1}$$

where $(\hat{i}, \hat{j}, \hat{k})$ denotes a permutation of the indices of the triplet (i, j, k) such that $\hat{i}$ and $\hat{j}$ have the deepest LCA in the hierarchy.[2]

We further extend this objective by restricting the binarizations considered for non-binary categories. Specifically, we compute the expected CKMM score over binarizations that preserve the linear order imposed by the category hierarchy. We assume that the hierarchy defines a linear order over categories, such that binary splits may only occur between adjacent groups of ordered categories. This induces an ordering over items corresponding to a depth-first left-to-right traversal of the hierarchy. For triples of items that share the same LCA, let $(\tilde{i}, \tilde{j}, \tilde{k})$ denote the permutation of the indices of triplet (i, j, k) such that $\tilde{i} \prec \tilde{j} \prec \tilde{k}$ under the linear ordering. The ordered CKMM objective is as follows:

$$Q_{ORD}(H) = \sum_{(i<j<k)\notin S} \left(d_{\hat{i}\hat{k}} + d_{\hat{j}\hat{k}}\right) + \frac{1}{2} \sum_{(i<j<k)\in S} \left(d_{\tilde{i}\tilde{j}} + 2d_{\tilde{i}\tilde{k}} + d_{\tilde{j}\tilde{k}}\right) + 2\sum_{i<j} d_{ij}. \tag{B.2}$$

This version of the CKMM objective is the score function used by our computational model when evaluating candidate category assignments for novel items. The higher the CKMM score, the more optimal the category system is considered to be.

B.1.2 Model Implementation

In this section, we describe the implementation of the ordered CKMM score and the sequential procedure used to assign items to categories. We represent the two- and three-level category systems as non-binary trees in `Python`. For ease of implementation, all items are represented as individual leaf nodes below their deepest category assignment.

We implement ordered CKMM recursively over these tree structures, as defined in Algorithm 1. The recursive procedure computes the ordered CKMM score by decomposing the hierarchy into smaller subtrees. At each category node, the algorithm recursively evaluates the scores of child subtrees and adds those to the score introduced by the current category split. Because the CKMM objective depends on lowest common ancestors, the contribution of a category depends only on item pairs whose lowest common ancestor is that category. This allows the global objective to be computed by recursively combining scores from lower levels of the hierarchy.

[2] In other words, the lowest common ancestor with the smallest cardinality.

Algorithm 1 Ordered CKMM

Require: Category M and item distance matrix D

function orderedCKMM(M, D)
  $children \leftarrow$ ordered children of M
  $k \leftarrow |children|$
  if $k = 0$ **then**
    return 0
  else if $k = 1$ **then**
    return orderedCKMM($children[1]$, D)
  else if $k = 2$ **then**
    $A, B \leftarrow children$
    $score_A \leftarrow$ orderedCKMM(A, D)
    $score_B \leftarrow$ orderedCKMM(B, D)
    $d_{AB} \leftarrow \sum_{x \in \text{items}(A)} \sum_{y \in \text{items}(B)} D_{xy}$
    $n_A \leftarrow$ number of items in A
    $n_B \leftarrow$ number of items in B
    return $score_A + score_B + d_{AB}(n_A + n_B)$
  else
    $score \leftarrow 0$
    for $i = 1$ to $k - 1$ **do**
      $A \leftarrow children[1 : i]$
      $B \leftarrow children[i + 1 : k]$
      $score_A \leftarrow$ orderedCKMM(A, D)
      $score_B \leftarrow$ orderedCKMM(B, D)
      $d_{AB} \leftarrow \sum_{x \in A} \sum_{y \in B} D_{xy}$
      $n_A \leftarrow$ number of items in A
      $n_B \leftarrow$ number of items in B
      $score \leftarrow score + score_A + score_B + d_{AB}(n_A + n_B)$
    end for
    return $score/(k - 1)$
  end if
end function

For leaf categories, the recursive function returns zero because a leaf introduces no further partition of the hierarchy and therefore contributes no additional CKMM scores. For categories with a single child, the category does not introduce a new separation, so its score is inherited from the child subtree. When a category has exactly two children, the algorithm defaults to binary CKMM. The algorithm first recursively computes the scores of the two child subtrees. It then computes the cross-category distance contribution: $d_{AB} = \sum_{x \in A} \sum_{y \in B} D_{xy}$, where A and B are the two child categories. This contribution is weighted by the number of items separated by the split, resulting in the term $d_{AB}(n_A + n_B)$ from the CKMM objective. The total score is therefore the sum of the child subtree scores and the contribution introduced by the current category split.

Algorithm 2 Sequential Category Assignment

Require: Initial category system H_0, item sequence $S = \{s_1, s_2, \ldots, s_n\}$, temperature τ, mixing parameter α, item distance matrix D

$H \leftarrow H_0$
$assignments \leftarrow [\,]$
for each item s in S **do**
    $scores, candidates \leftarrow \{\}, \{\}$
    for each available category g in H **do**
        $H' \leftarrow$ copy of H with item s added to g
        $scores[g] \leftarrow \textsc{ordered}\text{CKMM}(\text{root}(H'),\ D)$
        $candidates[g] \leftarrow H'$
    end for
    $P_{CKMM} \leftarrow \text{Softmax}(scores, \tau)$
    $P \leftarrow (1-\alpha) \cdot P_{CKMM} + \frac{\alpha}{|candidates|}$
    $g^* \leftarrow$ sample from P
    $H \leftarrow candidates[g^*]$
    $assignments \leftarrow assignments + [(s, g^*)]$
end forreturn *assignments*

For non-binary categories, the algorithm follows the ordered CKMM formulation by considering all valid binary splits that preserve the ordering of child categories. Given k ordered children, there are $k-1$ possible split points. For each split, the children are divided into a left and right group, and the score of the resulting binary partition is computed recursively. The category score is then averaged over these ordered binarizations. This corresponds to computing the expected CKMM score over the restricted set of binary decompositions that preserve the hierarchy's linear ordering.

By recursively applying this procedure from the root category to the leaves, the algorithm computes the complete ordered CKMM score without explicitly enumerating all item triplets or category relationships. To further improve computational efficiency, we store the scores of previously evaluated subtrees and reuse these values whenever the same subtree is encountered again. This avoids repeatedly calculating identical intermediate scores during the evaluation of different ordered splits.

After defining the ordered CKMM score, we use this objective to guide the sequential assignment of novel items to categories as outlined in Algorithm 2. At each step, the model evaluates the consequences of assigning the current item to each available category and uses the resulting CKMM scores to determine the probability of each possible assignment.

For each item, candidate category systems were generated by temporarily assigning the item to each available category in the current hierarchy. The ordered CKMM score of each resulting category system was then calculated and converted into assignment probabilities using a Softmax transformation. The temperature parameter τ controlled the sensitivity of these probabilities to differences in CKMM scores, with lower values producing more deter-

Table B.1: Grid search parameter values for fitting ordered CKMM models to participant data. The range of values and step size (Δ) are used to define the search space. The value count is reported for each parameter.

Parameter	**Search space**	**Count**
Temperature	[1.0, 8.5], $\Delta = 0.5$	16
Noise (α)	[0.00, 0.50], $\Delta = 0.01$	51

ministic assignments and higher values flattening the probability mass. A noise parameter, $\alpha \in [0, 1]$, controlled the contribution of random assignment by mixing the CKMM-based probabilities with a uniform distribution over available categories. Once a category was selected, the category system was updated to reflect the choice, and the procedure continued with the next item in the sequence.

For the greedy version of the CKMM algorithm used for model comparison, the algorithm remained unchanged except for the calculation of P_{CKMM}. Instead of converting CKMM scores into probabilities using a Softmax transformation, the greedy version assigns probability as follows:

$$p_i = \begin{cases} \frac{1}{|C_{\max}|}, & \text{if } i \in C_{\max} \\ 0, & \text{otherwise} \end{cases}$$

where $C_{\max}$ is the set of category assignments with the maximum ordered CKMM score in a set of scores. When multiple categories share the highest score, the final assignment is selected randomly among the tied categories.

B.1.3 CKMM Simulations

To simulate participant data, we first estimated the parameters of the ordered CKMM model using participant-level category assignments from Pool 3. Model parameters were estimated using 11-fold cross-validation. Within each fold, the model was fit by maximizing the log-likelihood of observed category assignments under the CKMM-based assignment probabilities P_{CKMM} over a discrete grid of values for the temperature parameter τ and noise parameter α. The parameter combination that produced the highest likelihood was selected for each fold, and the final parameter values were determined by majority voting across the best-fitting parameter combinations identified in the folds. The full range of parameter values considered is summarized in Table B.1.

Although α can take any value between 0 and 1, here we only considered probabilities between 0 and 0.50 to reduce the search space. No cross-validation folds selected parameter values at the upper boundary of the search space, suggesting that the ranges considered were sufficiently broad. For temperature, we only considered values where $\tau \geq 1$ because the magnitude of CKMM score differences resulted in highly concentrated Softmax probabilities

at lower temperatures and already behaved quite deterministically. Because the likelihood values increased until reaching a maximum and then consistently decreased, we reduced computation time by stopping the search for a given parameter once likelihood values began to decline.

Once the best-fitting parameters were selected, trials were simulated following Algorithm 2. The temperature and noise parameters were set according to the estimated values. Simulations followed the same trial sequences as participants to allow direct comparison of trial-level behaviour. For the all-at-once condition, item sequences were randomized. For the sequential conditions, item sequences matched the order specified in the experimental trials.

B.2 Bayesian Analyses

For Hypotheses 1 to 3, we pre-registered Bayesian analyses alongside frequentist ones. In this section, we report the Bayesian analyses on the human and simulated data. All Bayesian models were estimated using the `brms` package in `R`.

B.2.1 Sequence Effects

For Hypothesis 1, we only pre-registered Bayesian analyses for 1a–1c, which look at binary comparisons between different orders on the likelihood of assignment to individual categories. For consistency with the analysis in the main text, we also include a Bayesian version of the multinomial analysis.

Hypothesis 1: Multinomial Regressions

For Hypothesis 1, we first fit a series of multinomial models predicting category assignment (L, X, R) from the order in which items were presented (forward, backward, middle). The models include random intercepts for participant and stimulus set in the human data and a random intercept for trial in the simulated data. These models involved complex hierarchical structures with separate regression parameters for each category comparison and a random intercept. This presented challenges for model estimation and convergence. To support estimation, all multinomial regression models were estimated with weakly informative priors on the log-odds scale. Regression coefficients for order effects were assigned $\mathcal{N}(0, 2)$ priors, and intercepts were assigned slightly wider $\mathcal{N}(0, 2.5)$ priors to allow greater uncertainty in baseline category probabilities. Standard deviations of participant, stimulus, and trial random intercepts were assigned Exponential(1). Models were estimated using 4 chains with 4,000 iterations per chain, including 2,000 warmup iterations. All models of human

Table B.2: Bayesian multinomial regressions predicting category assignment from order. All models include random intercepts for stimulus set and participant. The reported values are posterior estimates of regression coefficients (β) with 95% credible intervals (CrI) for each category comparison. The intercepts represent posterior estimates of the log-odds of assigning L or R relative to the reference category (X) in the middle-order condition. Forward and backward coefficients represent deviations from this reference condition.

Location	Category	Order	Human		Simulated	
			β	95% CrI	β	95% CrI
Left	*L vs. X*	Intercept	1.40	[0.27, 2.43]	2.53	[2.33, 2.75]
		Forward	1.05	[0.77, 1.33]	0.35	[0.03, 0.65]
		Backward	−0.60	[-0.85, -0.36]	−1.65	[-1.90, -1.40]
	R vs. X	Intercept	0.36	[-0.41, 1.11]	1.17	[0.95, 1.39]
		Forward	0.76	[0.46, 1.07]	0.29	[-0.04, 0.62]
		Backward	0.10	[-0.16, 0.37]	−0.76	[-1.02, -0.50]
Centred	*L vs. X*	Intercept	0.26	[-0.54, 1.05]	0.14	[-0.08, 0.36]
		Forward	0.46	[0.22, 0.69]	2.06	[1.74, 2.39]
		Backward	0.04	[-0.24, 0.32]	0.82	[0.49, 1.15]
	R vs. X	Intercept	0.45	[-0.19, 1.06]	0.08	[-0.13, 0.29]
		Forward	−0.10	[-0.33, 0.13]	1.15	[0.82, 1.48]
		Backward	0.46	[0.18, 0.73]	1.83	[1.53, 2.15]
Right	*L vs. X*	Intercept	0.29	[-0.23, 0.80]	0.87	[0.67,1.08]
		Forward	−0.20	[-0.52, 0.12]	−0.55	[-0.79, -0.30]
		Backward	1.05	[0.68, 1.42]	0.58	[0.26, 0.91]
	R vs. X	Intercept	1.33	[0.19, 2.38]	2.24	[2.05, 2.43]
		Forward	−0.85	[-1.20, -0.51]	−1.35	[-1.59, -1.12]
		Backward	1.23	[0.84, 1.62]	0.60	[0.29, 0.91]

and simulated data showed satisfactory convergence (all $\hat{R} \leq 1.01$[3]). Model parameters are reported in Table B.2.

Coefficient estimates (β) differed between the Bayesian and frequentist models (Table 5.1) because the Bayesian models included random intercepts, whereas the frequentist models did not explicitly account for participant and stimulus variation in the human data or trial-level variation in the simulated data. To determine whether order continued to improve model performance when this hierarchical variation was accounted for, we performed a leave-one-out (LOO) cross-validation comparison between the full Bayesian model and a null model estimated using the same priors and random-effects structure but without the fixed effects of order. Results are summarized in Table B.3.

LOO comparisons indicated that Bayesian models including order generally had higher expected predictive performance than matched null models. This pattern was observed for all human location conditions and for the left and right simulated conditions. In the centred

[3]All model parameters showed $\hat{R} = 1.00$ except for the standard deviation of the trial-level random intercept for category L in the model fitted to simulated data from the right-shifted condition ($\hat{R} = 1.01$)

Table B.3: Leave-one-out (LOO) cross-validation comparison of Bayesian multinomial regression models. Values show the difference in expected log predictive density (ΔELPD) between the full and null models, with associated standard errors (SE) of the differences. Positive values indicate higher expected predictive accuracy for the full model.

	Human		Simulated	
Location	ΔELPD	SE	ΔELPD	SE
Left	49.6	9.0	103.4	14.9
Centred	28.0	9.7	-0.3	9.3
Right	13.1	6.7	77.6	12.4

simulated condition, however, the full model did not improve predictive performance relative to the null model ($\Delta ELPD = -0.3, SE = 9.3$). Given that the estimated difference was small relative to its uncertainty, this provides little evidence either for or against including order in this condition. This differs from the corresponding frequentist analysis, which favoured the model including order, and from both the frequentist and Bayesian analyses of the human data, which favoured the full model as well.

Overall, despite some differences in the estimated coefficients and model comparisons, the Bayesian analyses largely agreed with the frequentist analyses. Models including order tended to show better predictive performance, with one exception for simulated data in the centred condition.

Hypothesis 1: Logistic Regressions

For Hypotheses 1a–c, we fit Bayesian logistic mixed-effects regressions using default priors. Models of human and simulated data were estimated using four chains with 2,000 iterations per chain, including 1,000 warmup iterations. All models showed successful convergence (all $\hat{R} = 1.00$). Models predicted whether an item was assigned to a specific category versus not assigned to that category from pairwise order comparisons. Human data models included random intercepts for participant and stimulus, whereas simulated data models included a random intercept for trial. Hypotheses were considered supported when the 95% credible interval for the order effect excluded zero and the posterior mean indicated a positive effect in the predicted direction.

For Hypothesis 1a, we tested whether pairwise order comparisons in the centred condition predicted the likelihood of assignment to categories L, R, and X. Specifically, Hypothesis 1a(i) tested whether presenting items in the forward (vs. backward) order increased the likelihood of assignment to category L. Hypothesis 1a(ii) tested whether presenting items in the backward (vs. forward) order increased the likelihood of assignment to category R. Hypothesis 1a(iii) tested whether presenting items in the middle (vs. non-middle) order increased the likelihood of assignment to category X.

For the human data, the posterior estimate for the forward (vs. backward) order contrast of Hypothesis 1a(i) was $\beta = 2.16$, with a 95% CrI of [1.78, 2.57], supporting the hypothesis. The posterior estimate for the backward (vs. forward) order contrast of Hypothesis 1a(ii) was $\beta = 1.30$, with a 95% CrI of [1.00, 1.59], also supporting the hypothesis. However, the posterior estimate for the middle (vs. non-middle) order contrast of Hypothesis 1a(iii) was $\beta = -0.30$, with a 95% CrI of [−0.60, 0.01], providing no evidence in support of Hypothesis 1a(iii).

For the simulated data, the posterior estimate for the forward (vs. backward) order contrast in Hypothesis 1a(i) was $\beta = 3.45$, with a 95% CrI of [2.95, 3.98]. The posterior estimate for the backward (vs. forward) order contrast in Hypothesis 1a(ii) was $\beta = 3.31$, with a 95% CrI of [2.81, 3.84]. And the posterior estimate for the middle (vs. non-middle) order contrast in Hypothesis 1a(iii) was $\beta = 3.24$, with a 95% CrI of [2.52, 4.00]. Consistent with the frequentist analyses, Hypotheses 1a(i) and 1a(ii) were supported in both the simulated and human data, whereas Hypothesis 1a(iii) was supported only in the simulated data.

For Hypothesis 1b, we tested whether item presentation in the forward (vs. backward) order increased assignment of items to category L in the left-shifted condition. For the human data, the posterior estimate for order was $\beta = 2.74$, with a 95% CrI of [2.09, 3.41], supporting the hypothesis. For the simulated data, the posterior estimate for order was $\beta = 2.57$, with a 95% CrI of [2.21, 2.95].

For Hypothesis 1c, we tested whether item presentation in the backward (vs. forward) order increased assignment to category R in the right-shifted condition. For the human data, the posterior estimate for order was $\beta = 3.71$, with a 95% CrI of [2.97, 4.53], supporting the hypothesis. For the simulated data, the posterior estimate for order was $\beta = 2.36$, with a 95% CrI of [2.01, 2.72].

Overall, the Bayesian analyses were consistent with the frequentist analyses, as they estimated similar effects of order on category assignment for both human and simulated data.

B.2.2 Priming Effects

For Hypothesis 2, we fit Bayesian logistic mixed-effects regressions using the default priors. Models of human and simulated data were estimated using four chains with 4,000 iterations per chain, including 2,000 warm-up iterations, for both human and simulated data. Both models converged successfully (all $\hat{R} = 1.00$).

For the human data, the model predicted same-category assignment with a fixed intercept and random intercepts for stimulus set and participant. The posterior mean of the intercept was positive ($\beta = 1.19$, 95% CrI $[0.12, 2.13]$), with the 95% credible interval excluding zero, indicating that same-category assignments were estimated to occur above

chance levels. For the simulated data, the model predicted same-category assignment with a fixed intercept and a random intercept for trial. The posterior mean of the intercept was also positive ($\beta = 1.22$, 95% CrI [1.08, 1.36]), with the 95% credible interval excluding zero. The Bayesian results were consistent with the frequentist analyses, as estimates of the intercept were comparable for both the human and simulated data.

B.2.3 Hierarchy Effects

For Hypothesis 3, we fit Bayesian logistic mixed-effects regressions using the default priors. Models for the human and simulated data were estimated using four chains with 2,000 iterations per chain, including 1,000 warmup iterations. Both models converged successfully (all $\hat{R} = 1.00$).

For the human data, the model predicted creation of category X with a fixed effect of depth and random intercepts for participant and stimulus set. The posterior mean of the depth contrast comparing the two-level condition to the three-level condition was positive ($\beta = 1.16$, 95% CrI [0.92, 1.40]), with the 95% credible interval excluding zero, indicating that category X was more likely to be created in the two-level than the three-level depth condition. For the simulated data, the model predicted creation of category X using the same fixed effect of depth but no random intercepts (see main text for reasoning). The posterior mean of the depth contrast was positive ($\beta = 0.84$, 95% CrI [0.63, 1.05]), with the 95% credible interval excluding zero. The Bayesian analyses were consistent with the frequentist analyses reported in the main text, with both supporting the same conclusions.

B.3 Experiment Block Design

Table B.4 defines the trial blocks used in the experiment described in Section 5.3.1. These blocks were assigned to participants using a queue of repeating numbers from one to six. When a participant dropped out of the experiment before completing it, that block would be reassigned to the next participant to ensure a more balanced distribution of participants across blocks.

B.4 Extended Experiment Results

B.4.1 Hypothesis 1a(iii): Controlling for Depth in Participant Data

Neither the frequentist nor Bayesian analyses of participant data supported Hypothesis 1a(iii), which predicted that, in the centred condition, creation of category X was more likely in the middle order than in the backward or forward orders. Because depth also affected creation of category X, we conducted an additional logistic mixed-effects regression

Table B.4: Experiment trial blocks. Rows within a block represent trials, and each block consists of four unique trials. Across all six blocks, the 24 total trial types are covered. Trial types are determined by a combination of the experimental factors: location, order, and depth.

Block	**Trial Types**		
	Location	Order	Depth
One	Left-shifted	Forward	Two-level
	Centred	Backward	Three-level
	Right-shifted	Middle	Three-level
	Right-shifted	All-at-once	Two-level
Two	Centred	Middle	Two-level
	Right-shifted	Forward	Two-level
	Left-shifted	Backward	Three-level
	Right-shifted	All-at-once	Three-level
Three	Centred	Backward	Two-level
	Right-shifted	Middle	Two-level
	Centred	Forward	Three-level
	Left-shifted	All-at-once	Three-level
Four	Centred	Forward	Two-level
	Centred	Middle	Three-level
	Right-shifted	Backward	Three-level
	Left-shifted	All-at-once	Two-level
Five	Left-shifted	Middle	Two-level
	Right-shifted	Backward	Two-level
	Left-shifted	Forward	Three-level
	Centred	All-at-once	Three-level
Six	Left-shifted	Backward	Two-level
	Left-shifted	Middle	Three-level
	Right-shifted	Forward	Three-level
	Centred	All-at-once	Two-level

including depth as a predictor alongside order (coded as middle vs. non-middle orders), using the same random-effects structure described above.

As expected, the model showed a significant effect of depth on creation of category X ($\beta = -1.05$, 95% CI $[-1.31, -0.78]$), indicating that category X was less likely to be created in the three-level than the two-level depth condition. However, the model also revealed a significant effect of order ($\beta = 0.60$, 95% CI $[0.24, 0.95]$), indicating that category X was more likely to be created in the middle than the non-middle orders. A likelihood ratio test comparing the full model (depth + order) to a model containing depth only showed that adding order significantly improved model fit, $\chi^2(1) = 10.29$, $p = .001$. Similarly, the full model provided a significantly better fit than the original model containing order as its only fixed effect, $\chi^2(1) = 56.19$, $p < .001$. Accounting for depth revealed an effect of order that was not detectable in the original model, providing support for Hypothesis 1a(iii). However,

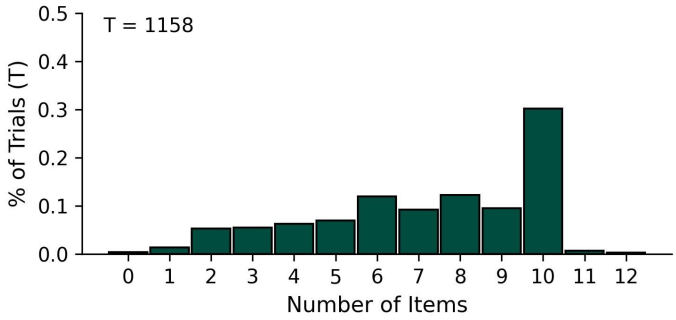


Figure B.1: Distribution over number of same category assignments for all items in a sequence trial. Sequence trials include forward, backward, and middle order trials. There are 13 items total in a trial (9 novel items, 4 distractors). 12 means that all subsequent items were assigned to the same category as the first; zero means that none were.

the effect of order was weaker in this analysis than in other analyses examining different order and location conditions.

B.4.2 Priming Effects: Including Distractors in Participant Data

In Section 5.3.2, we argued that the observed priming effect is unlikely to reflect a simple failure to switch categories or inattentive responding. To further assess this possibility, we examined participants' category assignments when the four distractor items were included.

Figure B.1 summarizes the distribution of the number of items assigned to the same category as the first item in each trial. If participants were simply failing to switch categories, we would expect a substantial number of trials in which all 12 subsequent items were assigned to the same category as the first. However, only 0.3% of trials showed this pattern. Similarly, only 0.7% of trials involved assigning 11 of 12 items to the same category. In contrast, 30.1% of trials involved assigning 10 of 12 items to the same category as the first.

For more detail, Table B.5 summarizes the distribution of first-item categories in trials with high levels of same-category assignment. Across these 10 out of 12 trials, 99.7% of first items were assigned to L or R, with only 0.3% assigned to X. This pattern is consistent with participants grouping novel items together while correctly categorizing distractors. When the first item is assigned to L or R, assigning subsequent novel items to the same category while correctly categorizing distractors produces 10 out of 12 same-category responses. Conversely, in trials where 8 out of 12 subsequent items matched the first-item category, the first item was most often assigned to X (57.0% of cases), consistent with participants grouping novel items into X while correctly classifying distractors, which should not be assigned to X.

These results suggest that the high levels of same-category assignment do not reflect participants simply failing to switch categories. To further support this interpretation, we conducted an exploratory mixed-effects logistic regression predicting same-category assign-

Table B.5: Distribution of first-item category assignments as a function of the number of subsequent same-category assignments. The maximum number of subsequent same-category assignments is 12. Trial counts are reported, and values for L, R, and X represent percentages of these trials where the first item is assigned to the referenced category.

Same as First	**Trials**	**First Item Category**		
		L	R	X
8 of 12	142	0.261	0.169	0.570
9 of 12	110	0.373	0.618	0.009
10 of 12	349	0.464	0.533	0.003
11 of 12	8	0.500	0.500	0.000
12 of 12	4	0.750	0.250	0.000

ment of subsequent items (1 = same as first, 0 = different) from item type (distractor vs. novel), first-item category assignment (X vs. L or R), and their interaction, with random intercepts for participants and stimuli. For the reference category (novel items when the first item in the sequence was assigned to L or R), participants were significantly more likely to choose a same-category assignment, $\beta = 1.10$, 95%CI $[0.81, 1.39]$ and distractors significantly reduced the likelihood of a same-category assignment, $\beta = -1.23$, 95%CI $[-1.32, -1.14]$, to about chance. There was also a significant interaction between item type and first-item category. While having X as the first item category did not significantly alter same-category rates for novel items $\beta = -0.10$, 95%CI $[-0.22, 0.02]$, for distractors it caused a significantly larger decrease in the log-odds of same-category assignment $\beta = -3.91$, 95%CI $[-4.36, -3.45]$ than did having the first category assignment be L or R. This means that a distractor was very unlikely to be assigned to category X if the first item in its sequence was assigned to X.

B.5 Cross-Validation Details

This section provides additional details on the cross-validation and parameter selection procedure used for model comparison. CKMM models were fit using the same parameter search space in Table B.1 and procedure in Appendix B.1.3. The MAP-RMC and single-particle RMC models were fit using a separate grid search over the coupling constraint parameter c and noise parameter α. For each model and depth condition, parameters were estimated using 11-fold cross-validation. Within each fold, the parameter combination that maximized the log-likelihood of the training data was selected. The final reported parameters were determined by majority vote across folds, and held-out log-likelihoods were calculated using these selected parameters. The parameter values considered for the MAP-RMC and single-particle RMC models are summarized in Table B.6.

Table B.6: Grid search parameter values for fitting rational models of categorization (RMC) to participant data. The range of values and step size (Δ) are used to define the search space. The value count is also reported for each parameter. The same search space is used for fitting both MAP and single-particle variants of the rational models.

Parameter	**Search space**	**Count**
Coupling constraint (c	[0.10, 0.90], $\Delta = 0.05$	17
Noise (α)	[0.00, 0.50], $\Delta = 0.01$	51

Table B.7: Model comparison using k-fold cross-validation. Validations were conducted separately for two-level and three-level conditions using 11 folds with 15 participants (30 trials) each. Likelihoods (log scale) are the mean across all held-out test sets. The best parameters were selected by majority vote across folds.

Model	**Two Level**		**Three Level**	
	Likelihood	Best Parameters	Likelihood	Best Parameters
MAP RCM	-268.7 ± 23.9	$\alpha = .34$, $c = .15$	-481.5 ± 43.8	$\alpha = .39$, $c = .20$
Particle RCM	-219.4 ± 16.0	$\alpha = .00$, $c = .35$	-444.5 ± 21.1	$\alpha = .00$, $c = .30$
Greedy CKMM	-219.7 ± 22.6	$\alpha = .27$	-391.3 ± 41.9	$\alpha = .27$
Softmax CKMM (no α)	-189.4 ± 23.0	$T = 5.0$	-455.4 ± 47.5	$T = 5.0$
Softmax CKMM (full)	-196.1 ± 24.6	$\alpha = .15$, $T = 1.0$	-359.0 ± 46.3	$\alpha = .15$, $T = 1.0$
Random baseline	-428.5		-758.9	

For the RMC models, parameter fitting and cross-validation were performed separately for each depth condition, whereas CKMM parameters were shared across depth conditions. CKMM held-out likelihoods are reported separately for each depth condition to allow for comparison. The resulting model fits are summarized in Table B.7. Reported values reflect mean held-out log-likelihoods across cross-validation folds. Best-fitting parameters were selected by majority vote across folds. The random baseline assumes that each item has an equal probability of being assigned to any available category at each step of the sequence. The baseline log-likelihood was calculated separately for each depth condition and used to express model performance as gains over random assignment.

Bibliography

Adler, M., Huber, J. T., & Nix, A. T. (2017). Stigmatizing disability: Library classifications and the marking and marginalization of books about people with disabilities. *The Library Quarterly*, *87*(2), 117–135. https://doi.org/10.1086/690734

American Psychological Association. (2018). *Bias*. APA Dictionary of Psychology. Retrieved October 9, 2026, from https://dictionary.apa.org/bias

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Angell, K., & Price, C. (2012). Fat bodies in thin books: Information bias and body image in academic libraries. *Fat Studies*, *1*(2), 153–165. https://doi.org/10.1080/21604851.2012.641895

Antin, J., Yee, R., Cheshire, C., & Nov, O. (2011). Gender differences in Wikipedia editing. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 11–14.

Ashby, S. R., & Zeithamova, D. (2022). Category-biased neural representations form spontaneously during learning that emphasizes memory for specific instances. *The Journal of Neuroscience*, *42*(5), 865–876. https://doi.org/10.1523/JNEUROSCI.1396-21.2021

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*(4), 547–569. https://doi.org/10.1017/S0140525X98001277

Atran, S., Estin, P., Coley, J., & Medin, D. (1997). Generic species and basic levels: Essence and appearance in folk biology. *Journal of Ethnobiology*, *17*(1), 17–43.

Ávila-Argüelles, R., Calvo, H., Gelbukh, A., & Godoy-Calderón, S. (2010). Assigning Library of Congress classification codes to books based only on their titles. *Informatica*, *34*(1).

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, *122*(8), e2416228122. https://doi.org/10.1073/pnas.2416228122

Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2019). Is man the measure of all things? A social cognitive account of androcentrism. *Personality and Social Psychology Review*, *23*(4), 307–331. https://doi.org/10.1177/1088868318782848

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, PEOPLE=MEN. *Science Advances*, *8*(13), eabm2463. https://doi.org/10.1126/sciadv.abm2463

Bailey, A. H., Williams, A., Poddar, A., & Cimpian, A. (2025). Intersectional male-centric and White-centric biases in collective concepts. *Personality & Social Psychology Bulletin*, *51*(10), 2085–2102. https://doi.org/10.1177/01461672241232114

Balée, W. (1999). Mode of production and ethnobotanical vocabulary: A controlled comparison of Guajá and Ka'apor. In T. L. Gragson & B. G. Blount (Eds.), *Ethnoecology: Knowledge, resources, and rights* (pp. 24–40). University of Georgia Press.

Bánki, O., Roskov, Y., Doring, M., Ower, G., Hernández Robles, D. R., Plata Corredor, C. A., Stjernegaard Jeppesen, T., Orn, A., Vandepitte, L., Hobern, D., Schalk, P., DeWalt, R. E., Ma, K., Miller, J., Orrell, T., Aalbu, R., Abbott, J., Adlard, R., & Adriaenssens, E. M. (2023). Catalogue of life checklist (version 2023-11-24). https://doi.org/10.48580/dfgnm

Barrett, H. C. (2020). Towards a cognitive science of the human: Cross-cultural approaches and their urgency. *Trends in Cognitive Sciences*, *24*(8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 1–64, Vol. 27). Academic Press. https://doi.org/10.1016/S0079-7421(08)60120-6

Basso, K. (1984). Western Apache place-name hierarchies. In E. Tooker (Ed.), *Naming systems* (pp. 78–94).

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*(3), 217–239. https://doi.org/10.1016/0010-0277(95)00669-X

Ben Aouicha, M., Hadj Taieb, M. A., & Ezzeddine, M. (2016). Derivation of "is a" taxonomy from Wikipedia Category Graph. *Engineering Applications of Artificial Intelligence*, *50*, 265–286. https://doi.org/10.1016/j.engappai.2016.01.033

Berlin, B. (1977). Speculations on the growth of ethnobotanical nomenclature. In *Sociocultural dimensions of language change* (pp. 63–101). Academic Press. https://doi.org/10.1016/B978-0-12-107450-0.50011-5

Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton University Press.

Berlin, B., Breedlove, D. E., & Raven, P. H. (1966). Folk taxonomies and biological classification. *Science*, *154*(3746), 273–275. https://doi.org/10.1126/science.154.3746.273

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. University of California Press.

Berman, S. (1971). *Prejudice and antipathies: A tract on the LC subject heads concerning people*. Scarecrow Press.

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, *45*(1), 106–125. https://doi.org/10.1177/0361684320977178

Bliss, H. E. (1933). *The organization of knowledge in libraries and the subject-approach to book* (2nd ed., rev. and partly rewritten). H. W. Wilson Company.

Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: From efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1124553

Blundell, C., Teh, Y. W., & Heller, K. A. (2011). Discovering nonbinary hierarchical structures with Bayesian rose trees. In K. L. Mengersen, C. P. Robert, & D. M. Titterington (Eds.), *Mixtures: Estimation and applications* (1st ed., pp. 161–187). Wiley. https://doi.org/10.1002/9781119995678.ch8

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364. https://dl.acm.org/doi/abs/10.5555/3157382.3157584

Boster, J. S. (1985). "Requiem for the omniscient informant": There's life in the old girl yet. In J. W. D. Dougherty (Ed.), *Directions in cognitive anthropology* (pp. 177–197). University of Illinois Press.

Boster, J. S. (1987). Agreement between biological classification systems is not dependent on cultural transmission. *American Anthropologist*, *89*(4), 914–920. https://doi.org/10.1525/aa.1987.89.4.02a00100

Boster, J. S., & D'Andrade, R. (1989). Natural and human sources of cross-cultural agreement in ornithological classification. *American Anthropologist*, *91*(1), 132–142. https://doi.org/10.1525/aa.1989.91.1.02a00080

Boster, J. S., & Johnson, J. (1989). Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist*, *91*(4), 866–889. https://doi.org/10.1525/aa.1989.91.4.02a00040

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001

Brekhus, W. (1998). A sociology of the unmarked: Redirecting our focus. *Sociological Theory*, *16*(1), 34–51. https://doi.org/10.1111/0735-2751.00041

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126

Brewer, M. B. (2007). The social psychology of intergroup relations: Social categorization, ingroup bias, and outgroup prejudice. In *Social psychology: Handbook of basic principles* (pp. 695–715). Guilford Press.

Brown, C. H. (1976). General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist*, *3*(3), 400–424. https://doi.org/10.1525/ae.1976.3.3.02a00020

Brown, C. H., Anderson, E. N., Bulmer, R., Drechsel, P., Ellen, R. F., Hays, T. E., Headland, T. N., Howe, L., Hyndman, D. C., Jensen, K.-E., Morris, B., & Reason, D. (1985). Mode of subsistence and folk biological taxonomy [and comments and reply]. *Current Anthropology*, *26*(1), 43–53. https://doi.org/10.1086/203224

Brown, R. (1958). How shall a thing be called? *Psychological Review*, *65*(1), 14–21. https://doi.org/10.1037/h0041727

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1986). *A study of thinking* (2nd ed.). Routledge. https://doi.org/10.4324/9781315083223

Bubeck, S., Meilă, M., & von Luxburg, U. (2012). How the initialization affects the stability of the K-means algorithm. *ESAIM: Probability and Statistics*, *16*, 436–452. https://doi.org/10.1051/ps/2012013

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81*, 77–91.

Burenhult, N., & Levinson, S. C. (2008). Language and landscape: A cross-linguistic perspective. *Language Sciences*, *30*(2-3), 135–150. https://doi.org/10.1016/j.langsci.2006.12.028

Büth, C. M., Acharya, K., & Zanin, M. (2025). Infomeasure: A comprehensive Python package for information theory measures and estimators. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-14053-5

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Canini, K., & Griffiths, T. (2011). A nonparametric Bayesian model of multi-level category learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *25*(1), 307–312. https://doi.org/10.1609/aaai.v25i1.7891

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, *1*(1), 37–41. https://doi.org/10.1207/S15327647JCD0101N_3

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*, *41*(4), 892–923. https://doi.org/10.1111/cogs.12371

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289. https://doi.org/10.1016/j.cognition.2020.104289

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*, 303–308.

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Cheryan, S., & Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, *127*(6), 1022–1052. https://doi.org/10.1037/rev0000209

Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (pp. 129–186). University of Minnesota Press.

Chi, M. T. H., Hutchinson, J. E., & Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly*, *35*(1), 27–62.

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, *8*(1), 15951. https://doi.org/10.1038/s41598-018-34203-2

Clapper, J. P. (2015). The impact of training sequence and between-category similarity on unsupervised induction. *Quarterly Journal of Experimental Psychology*, *68*(7), 1370–1390. https://doi.org/10.1080/17470218.2014.981553

Cohen, J. (1988). *Statisticsl power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen-Addad, V., Kanade, V., Mallmann-trenn, F., & Mathieu, C. (2019). Hierarchical clustering: Objective functions and algorithms. *Journal of the ACM*, *66*(4), 1–42. https://doi.org/10.1145/3321386

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Comer, D. (1979). The ubiquitous B-tree. *ACM Computing Surveys*, *11*(2), 121–137. https://doi.org/10.1145/356770.356776

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*(2), 291–303. https://doi.org/10.1037/0033-2909.111.2.291

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & Society*, *36*(4), 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

Croteau, J., Cheries, E., & Xu, F. (2026). The development of kind concepts: Insights from object individuation. *Psychological Review*, *133*(1), 134–155. https://doi.org/10.1037/rev0000527

Crovisier, R., & Intner, S. S. (1987). Classification for astronomy: The QB schedule of the Library of Congress classification. *Cataloging & Classification Quarterly*, *7*(3), 23–36. https://doi.org/10.1300/J104v07n03_04

Cutter, C. A. (1876). Rules for a printed dictionary catalogue. In U.S. Bureau of Education (Ed.), *Public libraries in the United States of America; Their history, condition, and management (Part II)*. U.S. Government Printing Office.

Das-Smaal, E. A. (1990). Biases in categorization. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Advances in psychology* (pp. 349–386, Vol. 68). North-Holland. https://doi.org/10.1016/S0166-4115(08)61332-1

De Vries, T., Misra, I., Wang, C., & Van der Maaten, L. (2019). Does object recognition work for everyone? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 52–59.

Dewey, M. (1885). *Dewey Decimal classification and relative index* (Second edition, revised and greatly enlarged). Library Bureau.

Dewey, M. (1989). *Dewey Decimal classification and relative index* (J. P. Comaromi, J. Beall, W. E. Matthews, & G. R. New, Eds.; 20th ed., Vol. 2). Forest Press.

Dewey, M. (2003). *Dewey Decimal classification and relative index* (J. S. Mitchell, J. Beall, G. Martin, W. E. Matthews, & G. R. New, Eds.; 22nd ed., Vol. 2). OCLC Online Computer Library Centre Inc.

Dewey, M. (2011). *Dewey Decimal classification and relative index* (J. S. Mitchell, J. Beall, R. Green, G. Martin, & M. Panzer, Eds.; 23rd ed., Vol. 2). OCLC Online Computer Library Center, Inc.

Dilevko, J., & Grewal, K. (1997). A new approach to collection bias in academic libraries: The extent of corporate control in journal holdings. *Library & Information Science Research*, *19*(4), 359–385. https://doi.org/10.1016/S0740-8188(97)90026-8

Djalal, F. M., Storms, G., Ameel, E., & Heyman, T. (2017). Feature taxonomy: What type of features do children associate with categories and how do they fare in predicting category judgments? *Acta Psychologica*, *178*, 114–123. https://doi.org/10.1016/j.actpsy.2017.04.009

Dubova, M., & Goldstone, R. L. (2021). The influences of category learning on perceptual reconstructions. *Cognitive Science*, *45*(5), e12981. https://doi.org/10.1111/cogs.12981

Duffy, S. A., & Keir, J. A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, *32*(4), 551–559. https://doi.org/10.3758/BF03195846

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, *62*(1), 617–629. https://doi.org/10.1111/bjso.12560

Egré, P., Gardelle, V. d., & Ripley, D. (2013). Vagueness and order effects in color categorization. *Journal of Logic, Language and Information*, *22*(4), 391–420. https://doi.org/10.1007/s10849-013-9183-7

Ekstrand, M. D., & Kluver, D. (2021). Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, *31*(3), 377–420. https://doi.org/10.1007/s11257-020-09284-2

Ellen, R. F. (1977). Anatomical classification and the semiotics of the body. In J. Blacking (Ed.), *The anthropology of the body* (pp. 343–373). Academic Press.

England, E. (2023). Perpetuating whiteness: Disrupting the dominant narrative in academic library collections. *Journal of Radical Librarianship*, *9*, 98–115.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2020). Categorical biases in human occipitoparietal cortex. *The Journal of Neuroscience*, *40*(4), 917–931. https://doi.org/10.1523/JNEUROSCI.2700-19.2019

Falenska, A., & Çetinoğlu, Ö. (2021). Assessing gender bias in Wikipedia: Inequalities in article titles. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 75–85. https://doi.org/10.18653/v1/2021.gebnlp-1.9

Ferdinand, V., & Perfors, A. (2020). The evolution of category systems within and between learners. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*, 648–654.

Fernandez, M. J., & Eastman, C. M. (1990). Basic taxonomic structures and levels of abstraction. *Proceedings of the 1st ASIS SIG/CR Classification Research Workshop*, 59–70.

Fourtassi, A., Wilson, K., & Frank, M. C. (2020). Discovering conceptual hierarchy through explicit and implicit cues in child-directed speech. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*, 2430–2436.

Fox, V. (2019). *Why are the 200s so heavily focused on Christianity? An explainer*. Retrieved November 21, 2022, from https://web.archive.org/web/20221205234752/https://ddc.typepad.com/025431/2019/11/why-are-the-200s-so-heavily-focused-on-christianity-an-explainer.html

Frank, E., & Paynter, G. W. (2004). Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology*, *55*(3), 214–227. https://doi.org/10.1002/asi.10360

Freeman, B. G., & Pennell, M. W. (2021). The latitudinal taxonomy gradient. *Trends in Ecology & Evolution*, *36*(9), 778–786. https://doi.org/10.1016/j.tree.2021.05.003

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gehlke, C. E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, *29*(185), 169–170. https://doi.org/10.2307/2277827

Gepperth, A., & Hammer, B. (2016). Incremental learning algorithms and applications. *European Symposium on Artificial Neural Networks*, 357–368.

Glushko, R. J., Hemerly, J., Petras, V., Manoochehri, M., Wang, L., Shedlock, J., & Griffin, D. (2016). Classification: Assigning resources to categories. In R. J. Glushko (Ed.), *The discipline of organizing: Professional edition* (pp. 417–466). O'Reilly Media.

Glushko, R. J., Maglio, P. P., Matlock, T., & Barsalou, L. W. (2008). Categorization in the wild. *Trends in Cognitive Sciences*, *12*(4), 129–135. https://doi.org/10.1016/j.tics.2008.01.007

Godby, C. J., & Stuler, J. (2013). The Library of Congress classification as a knowledge base for automatic subject categorization. In I. McIlwaine (Ed.), *Subject retrieval in a networked environment: Proceedings of the IFLA satellite meeting held in Dublin, OH* (pp. 163–169).

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037//0096-3445.123.2.178

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

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*(3), 548–568. https://doi.org/10.1111/tops.12212

Gould, S. J. (1990). Taxonomy as politics: The harm of false classification. *Dissent*, *37*, 73–78.

Govindarajan, V. S., Atwell, K., Sinno, B., Alikhani, M., Beaver, D. I., & Li, J. J. (2023). How people talk about each other: Modeling generalized intergroup bias and emotion. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2496–2506. https://doi.org/10.18653/v1/2023.eacl-main.183

Green, R. (2006). *Vocabulary alignment via basic level concepts* (OCLC/ALISE Research Grant Report). OCLC Research. https://www.webjunction.org/content/dam/research/grants/reports/green/rg2005.pdf

Green, R., & Panzer, M. (2014). The interplay of big data, WorldCat, and Dewey. *Advances in Classification Research Online*, *24*(1), 51–58. https://doi.org/10.7152/acro.v24i1.14677

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, *69*(7), 669–684. https://doi.org/10.1037/a0036056

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*, 323–328.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107. https://doi.org/10.1080/03640210701801974

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576

Hajibayova, L. (2013). Basic-level categories: A review. *Journal of Information Science*, *39*(5), 676–687. https://doi.org/10.1177/0165551513481443

Hall, S. (2018). The West and the rest: Discourse and power [1992]. In D. Morley (Ed.), *Essential essays, volume 2: Identity and diaspora* (pp. 141–184). Duke University Press. https://doi.org/10.1215/9781478002710-009

Hamilton, M. C. (1991). Masculine bias in the attribution of personhood: People = male, male = people. *Psychology of Women Quarterly*, *15*(3), 393–402. https://doi.org/10.1111/j.1471-6402.1991.tb00415.x

Hampton, J. A. (1982). A demonstration of intransitivity in natural categories. *Cognition*, *12*(2), 151–164. https://doi.org/10.1016/0010-0277(82)90010-5

Harrison, S., Gualdoni, E., & Boleda, G. (2023). Run like a girl! Sport-related gender bias in language and vision. *Findings of the Association for Computational Linguistics: ACL 2023*, 14093–14103. https://doi.org/10.18653/v1/2023.findings-acl.886

Hawkins, R. X. D., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*, 463–468.

Hejazy, K. A., & El-Beltagy, S. R. (2013). An approach for deriving semantically related category hierarchies from Wikipedia category graphs. In Á. Rocha, A. M. Correia, T. Wilson, & K. A. Stroetmann (Eds.), *Advances in information systems and technologies* (pp. 77–86). Springer. https://doi.org/10.1007/978-3-642-36981-0_8

Heller, K. A., & Ghahramani, Z. (2005). Bayesian hierarchical clustering. *Proceedings of the 22nd International Conference on Machine Learning*, 297–304. https://doi.org/10.1145/1102351.1102389

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Higgins, M. (2016). Totally invisible: Asian American representation in the Dewey Decimal classification, 1876-1996. *Knowledge Organization*, *43*(8), 609–621. https://doi.org/10.5771/0943-7444-2016-8-609

Hilles, S. (2023). The canon no longer applies: How the fine arts (N) range of the Library of Congress classification system perpetuates bias by privileging fine art over craft. *Art Libraries Journal*, *48*(4), 100–104. https://doi.org/10.1017/alj.2023.22

Hobart, E. F. M. (2020). Reflecting diversity in the Library of Congress Subject Headings. In *Social justice and activism in libraries: Essays on diversity and change* (pp. 110–117). McFarland.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Holman, E. W. (2005). Domain-specific and general properties of folk classifications. *Journal of Ethnobiology*, *25*(1), 71–91. https://doi.org/10.2993/0278-0771_2005_25_71_dagpof_2.0.co_2

Hosoya, G., Schindler, I., Beermann, U., Wagner, V., Menninghaus, W., Eid, M., & Scherer, K. R. (2017). Mapping the conceptual domain of aesthetic emotion terms: A pile-sort study. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(4), 457–473. https://doi.org/10.1037/aca0000123

Howard, S. A., & Knowlton, S. A. (2018). Browsing through bias: The Library of Congress classification and subject headings for African American studies and LGBTQIA studies. *Library Trends*, *67*(1), 74–88. https://doi.org/10.1353/lib.2018.0026

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101. https://doi.org/10.1109/JRPROC.1952.273898

Hupp, S. L. (1991). The Left and the Right: A preliminary study of bias in collection development in Ohio libraries. *Collection Management*, *14*(1-2), 139–154. https://doi.org/10.1300/J105v14n01_09

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352–376. https://doi.org/10.1037/0033-295X.98.3.352

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*(2), 220–241. https://doi.org/10.1037/0096-3445.129.2.220

Hyde, K. (2002). *From suffrage to postfeminism: An evolution of the Library of Congress subject treatment of women's issues* [Master's thesis, University of North Carolina at Chapel Hill]. https://doi.org/10.17615/06kn-c924

Intner, S. S., & Futas, E. (1996). The role and impact of Library of Congress classification on the assessment of women's studies collections. *Library Acquisitions: Practice & Theory*, *20*(3), 267–279. https://doi.org/10.1016/0364-6408(96)00067-1

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends*, *52*, 515–540.

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, *126*(3), 248–277. https://doi.org/10.1037/0096-3445.126.3.248

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788. https://doi.org/10.1037/0022-3514.54.5.778

Kam, V. (2007). Subject headings for Aboriginals: The power of naming. *Art Documentation: Journal of the Art Libraries Society of North America*, *26*(2), 18–22.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52. https://doi.org/10.1016/j.cognition.2017.05.001

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*, 71–76.

Kay, P. (1971). Taxonomy and semantic contrast. *Language*, *47*(4), 866–887. https://doi.org/10.2307/412161

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, *88*(3), 197–227. https://doi.org/10.1037/0033-295X.88.3.197

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press. https://doi.org/10.7551/mitpress/2065.001.0001

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. https://doi.org/10.1126/science.1218811

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128. https://doi.org/10.1146/annurev-linguistics-011817-045406

Khishigsuren, T., Mollica, F., & Kemp, C. (2026). Information theory predicts that superordinates emerge before basic-level categories. *PsyArXiv*. https://osf.io/preprints/psyarxiv/45ht9_v1

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/10.1016/j.conb.2014.07.014

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM*, *59*(10), 16–17. https://doi.org/10.1145/2983270

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Knowlton, S. A. (2005). Three decades since *Prejudices and Antipathies*: A study of changes in the Library of Congress Subject Headings. *Cataloging & Classification Quarterly*, *40*(2), 123–145. https://doi.org/10.1300/J104v40n02_08

Kua, E. (2004). Non-Western languages and literatures in the Dewey Decimal Classification scheme. *Libri*, *54*(4), 256–265. https://doi.org/10.1515/LIBR.2004.256

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kwasnik, B. (2000). The role of classification in knowledge representation and discovery. *Library Trends*, *48*(1), 22–47.

Lakoff, G. (1987). *Women, fire, and dangerous things: What catergories reveal about the mind*. The University of Chicago Press.

Lamo de Espinosa, E. (2017). *Is Latin America part of the West?* Elcano Royal Institute. Retrieved February 2, 2023, from https://www.realinstitutoelcano.org/en/work-document/is-latin-america-part-of-the-west/

Lancaster, F. (1970). Vocabulary control in information retrieval systems. *Advances in Librarianship*, *7*, 1–40.

Larson, R. R. (1992). Experiments in automatic Library of Congress classification. *Journal of the American Society for Information Science*, *43*(2), 130–148. https://doi.org/10.1002/(SICI)1097-4571(199203)43:2⟨130::AID-ASI3⟩3.0.CO;2-S

Lerman, K., & Hogg, T. (2014). Leveraging position bias to improve peer recommendation. *PLoS ONE*, *9*(6), e98914. https://doi.org/10.1371/journal.pone.0098914

Lerner, J., & Lomi, A. (2018). Knowledge categorization affects popularity and quality of Wikipedia articles. *PLoS ONE*, *13*(1), e0190674. https://doi.org/10.5281/zenodo.1005175

Levinson, S., & Meira, S. (2003). 'Natural concepts' in the spatial topological domain – Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*(3), 485–516. https://doi.org/10.1353/lan.2003.0174

Li, J., Luong, T., Jurafsky, D., & Hovy, E. (2015). When are tree structures necessary for deep learning of representations? *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2304–2314. https://doi.org/10.18653/v1/D15-1278

Library of Congress. (n.d.). Process for adding and revising Library of Congress Subject Headings. Retrieved February 2, 2025, from https://www.loc.gov/aba/cataloging/subject/lcsh-process.html

Library of Congress. (2022). *100 - Main Entry-Personal Name (NR)*. Retrieved October 21, 2022, from https://www.loc.gov/marc/bibliographic/bd100.html

Library of Congress, Policy and Standards Division. (2016). *Module 1.5: Introduction to LCSH*. Retrieved February 2, 2025, from https://www.loc.gov/catworkshop/lcsh/PDF%20scripts/1-5%20Intro%20To%20LCSH.pdf

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. *Proceedings of the Asso-*

ciation for Computational Linguistics 2012 System Demonstrations, 169–174. https://dl.acm.org/doi/10.5555/2390470.2390499

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., & Elliott, D. (2021). Visually grounded reasoning across languages and cultures. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10467–10485. https://doi.org/10.18653/v1/2021.emnlp-main.818

Lo, G. (2019). "Aliens" vs. catalogers: Bias in the library of congress subject heading. *Legal Reference Services Quarterly*, *38*(4), 170–196. https://doi.org/10.1080/0270319X.2019.1696069

Loehrlein, A. J. (2012). *Priming effects associated with the hierarchical levels of classification* [Doctoral dissertation, Indiana University].

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, *32*(3), 251–295. https://doi.org/10.1006/cogp.1997.0651

Losee, R. M. (1992). A Gray code based ordering for documents on shelves: Classification for browsing and retrieval. *Journal of the American Society for Information Science*, *43*(4), 312–322. https://doi.org/10.1002/(SICI)1097-4571(199205)43:4%3C312::AID-ASI7%3E3.0.CO;2-Z

Losee, R. M. (1993). The relative shelf location of circulated books: A study of classification, users, and browsing. *Library Resources and Technical Services*, *37*(2), 197–209.

Lu, K. (2020). *Wikipedia category tree* (Version 2). Kaggle. Retrieved November 19, 2024, from https://www.kaggle.com/datasets/kevinlu1248/wikipedia-category-tree

Luccioni, A. S., & Rolnick, D. (2023). Bugs in the data: How ImageNet misrepresents biodiversity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14382–14390. https://doi.org/10.1609/aaai.v37i12.26682

Lund, B., & Agbaji, D. (2018a). Use of Dewey Decimal classification by academic libraries in the United States. *Cataloging & Classification Quarterly*, *56*(7), 653–661. https://doi.org/10.1080/01639374.2018.1517851

Lund, B., & Agbaji, D. (2018b). What scheme do we prefer? An examination of preference between Library of Congress and Dewey Decimal classification among US-based academic library employees. *Knowledge Organization*, *45*(5), 380–392. https://doi.org/10.5771/0943-7444-2018-5-380

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, Article 54. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS ONE*, *5*(7), e11452. https://doi.org/10.1371/journal.pone.0011452

Mack, M. L., & Palmeri, T. J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General*, *144*(3), 551–569. https://doi.org/10.1037/a0039184

Mackie, D. M., & Worth, L. T. (1989). Differential recall of subcategory information about in-group and out-group members. *Personality and Social Psychology Bulletin*, *15*(3), 401–413. https://doi.org/10.1177/0146167289153010

Mai, J.-E. (2010). Classification in a social world: Bias and trust. *Journal of Documentation*, *66*(5), 627–642. https://doi.org/10.1108/00220411011066763

Majid, A. (2010). Words for parts of the body. In B. Malt & P. Wolff (Eds.), *Words and the mind: How words capture human experience* (pp. 58–71). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311129.003.0004

Mammal Diversity Database. (2023). Mammal diversity database (version 1.11). https://doi.org/10.5281/zenodo.7830771

Mandler, J. M., & McDonough, L. (1998). On developing a knowledge base in infancy. *Developmental Psychology*, *34*(6), 1274–1288. https://doi.org/10.1037/0012-1649.34.6.1274

Manzini, T., Yao Chong, L., Black, A. W., & Tsvetkov, Y. (2019). Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. https://doi.org/10.18653/v1/N19-1062

Markman, E. M. (1981). Two different principles of conceptual organization. In M. E. Lamb & A. L. Brown (Eds.), *Advances in Developmental Psychology* (pp. 199–236, Vol. 1). Psychology Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Markman, E. M., & Callanan, M. A. (1984). An analysis of hierarchical classification. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 325–365, Vol. 2). Lawrence Erlbaum Associates.

Martínez-Ávila, D., & Budd, J. M. (2017). Epistemic warrant for categorizational activities and the development of controlled vocabularies. *Journal of Documentation*, *73*(4), 700–715. https://doi.org/10.1108/JD-10-2016-0129

McCartan, C., & Imai, K. (2023). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *The Annals of Applied Statistics*, *17*(4), 3300–3323. https://doi.org/10.1214/23-AOAS1763

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McClure, E. F. (1975). Ethno-anatomy: The structure of the domain. *Anthropological Linguistics*, *17*(2), 78–88.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, *1*(2), 250–254. https://doi.org/10.3758/BF03200776

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to rome? *Cognitive Psychology*, *32*(1), 49–96. https://doi.org/10.1006/cogp.1997.0645

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., & Blok, S. (2006). Folkbiology of freshwater fish. *Cognition*, *99*(3), 237–273. https://doi.org/10.1016/j.cognition.2003.12.005

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Merritt, R. D., & Kok, C. J. (1995). Attribution of gender to a gender-unspecified individual: An evaluation of the people = male hypothesis. *Sex Roles*, *33*(3/4), 145–157. https://doi.org/10.1007/BF01544608

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115. https://doi.org/10.1146/annurev.ps.32.020181.000513

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201–233). Cambridge University Press.

Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, *53*(1), 258–266. https://doi.org/10.2307/1129660

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team and Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Miller, G. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 95–100.

Miller, G. A. (1994). WordNet: A lexical database for English. *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. https://aclanthology.org/H94-1111

Misch, A., Dunham, Y., & Paulus, M. (2022). The developmental trajectories of racial and gender intergroup bias in 5- to 10-year-old children: The impact of general psycholog-

ical tendencies, contextual factors, and individual propensities. *Acta Psychologica*, *229*. https://doi.org/10.1016/j.actpsy.2022.103709

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., Öttl, A., et al. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*, *46*(3), 841–871. https://doi.org/10.3758/s13428-013-0409-z

Moy, L. (2021). Facing injustice: How face recognition technology may increase the incidence of misidentifications and wrongful convictions. *William & Mary Bill of Rights Journal*, *30*, 337–372. https://doi.org/10.2139/ssrn.4101826

Mullender, A. (2025). Five decades since prejudices and antipathies. *Canadian Journal of Academic Librarianship*, *11*, 1–13. https://doi.org/10.33137/cjal-rcbu.v11.45357

Murphy, G. L. (2002). *The big book of concepts*. MIT Press. https://doi.org/10.7551/mitpress/1602.001.0001

Murphy, G. L. (2016). Explaining the basic-level concept advantage in infants…or is it the superordinate-level advantage? In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 57–92, Vol. 64). Academic Press. https://doi.org/10.1016/bs.plm.2015.09.002

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In *Knowledge concepts and categories*. Psychology Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*(1), 1–20. https://doi.org/10.1016/S0022-5371(82)90412-1

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, *2*(1), 86–97. https://doi.org/10.1002/widm.53

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: An overview, II. *WIREs Data Mining and Knowledge Discovery*, *7*(6). https://doi.org/10.1002/widm.1219

Naumov, S., Yaroslavtsev, G., & Avdiukhin, D. (2022). Objective-based hierarchical clustering of deep embedding vectors. *arXiv*. https://arxiv.org/abs/2012.08466v2

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

OhioLINK Collection Building Task Force, Gammon, J., & O'Neill, E. T. (2011). OhioLINK–OCLC collection and circulation analysis project 2011. https://doi.org/10.25333/C30341

Olson, H. A. (2001). Sameness and difference. *Library Resources & Technical Services*, *45*(3), 115–122. https://doi.org/10.5860/lrts.45n3.115

Olson, H. A. (2002). *The power to name: Locating the limits of subject representation in libraries*. Springer Netherlands. https://doi.org/10.1007/978-94-017-3435-6

Olson, H. A. (2004). The ubiquitous hierarchy: An army to overcome the threat of a mob. *Library Trends*, *52*(3), 604–616.

Olson, H. A., & Schlegl, R. (1999). Bias in subject access standards: A content analysis of the critical literature. *Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI*, 236–247. https://doi.org/10.29173/cais460

Olson, H. A., & Ward, D. B. (1997). Ghettoes and diaspora in classification: Communicating across limits. *Proceedings of the Annual Conference of CAIS/Actes Du congrès Annuel De l'ACSI.*, 19–31. https://doi.org/10.29173/cais45

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58. https://doi.org/10.1016/0010-0277(81)90013-5

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*(3), 495–503. https://doi.org/10.3758/BF03210840

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 446–457. https://doi.org/10.1145/3351095.3372843

Parducci, A. (1983). Category ratings and the relational character of judgment. *Advances in Psychology*, *11*, 262–282. https://doi.org/10.1016/S0166-4115(08)62067-1

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068. https://doi.org//10.1037/0022-3514.42.6.1051

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Quinn, B. (2012). Collection development and the psychology of bias. *The Library Quarterly: Information, Community, Policy*, *82*(3), 277–304. https://doi.org/10.1086/665933

Quinn, P. C., & Johnson, M. H. (2000). Global-before-basic object categorization in connectionist networks and 2-month-old infants. *Infancy*, *1*(1), 31–46. https://doi.org/10.1207/S15327078IN0101_04

Raffman, D. (2011). Vagueness and observationality. In G. Ronzitti (Ed.), *Vagueness: A guide* (pp. 107–121). Springer Netherlands. https://doi.org/10.1007/978-94-007-0375-9_5

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163. https://doi.org/10.2307/271063

Rakison, D. H. (2000). When a rose is just a rose: The illusion of taxonomies in infant categorization. *Infancy*, *1*(1), 77–90. https://doi.org/10.1207/S15327078IN0101_07

Rao, J. N., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, *76*(374), 221–230. https://doi.org/10.1080/01621459.1981.10477633

Reagle, J., & Rhue, L. (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication*, *5*, 1138–1158.

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

Ribert, A., Ennaji, A., & Lecourtier, Y. (1999). An incremental hierarchical clustering. *Proceedings of the Vision Interface Conference*, 586–591.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20. https://doi.org/10.1016/S0022-5371(73)80056-8

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*(4), 378–411. https://doi.org/10.1016/j.cogpsych.2004.10.001

Rodriguez, R. D. (1984). Hulme's concept of literary warrant. *Cataloging & Classification Quarterly*, *5*(1), 17–26. https://doi.org/10.1300/J104v05n01_02

Rogers, M. N. (1993). Are we on equal terms yet? Subject headings concerning women in LCSH, 1975-1991. *Library Resources and Technical Services*, *37*(2), 181–196.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Romney, A. K., Moore, C. C., Batchelder, W. H., & Hsia, T.-L. (2000). Statistical methods for characterizing similarities and differences between semantic structures. *Proceedings of the National Academy of Sciences*, *97*(1), 518–523. https://doi.org/10.1073/pnas.97.1.518

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338. https://doi.org/10.1525/aa.1986.88.2.02a00020

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192–233. https://doi.org/10.1037/0096-3445.104.3.192

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). Lawrence Elbaum Associates.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Rosenberg, A. L. (1979). Encoding data structures in trees. *Journal of the ACM*, *26*(4), 668–689. https://doi.org/10.1145/322154.322160

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, *35*(2), 345–366. https://doi.org/10.1007/s00357-018-9259-9

Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLoS ONE*, *15*(4), e0231189. https://doi.org/10.1371/journal.pone.0231189

Rubin, M., & Badea, C. (2012). They're all the same!. . . but for several different reasons: A review of the multicausal nature of perceived group variability. *Current Directions in Psychological Science*, *21*(6), 367–372. https://doi.org/10.1177/0963721412457363

Ryder, M. L. (1964). The history of sheep breeds in britain. *The Agricultural History Review*, *12*(1), 1–12.

Sahoo, N., Callan, J., Krishnan, R., Duncan, G., & Padman, R. (2006). Incremental hierarchical clustering of text documents. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 357–366. https://doi.org/10.1145/1183614.1183667

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. https://doi.org/10.1037/a0020511

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Schmahl, K. G., Viering, T. J., Makrodimitris, S., Naseri Jahfari, A., Tax, D., & Loog, M. (2020). Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 94–103. https://doi.org/10.18653/v1/2020.nlpcss-1.11

Schmidt, B. (2018). Stable random projection: Lightweight, general-purpose dimensionality reduction for digitized libraries. *Journal of Cultural Analytics*, *3*(1). https://doi.org/10.22148/16.025

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv*. https://doi.org/10.48550/arXiv.1711.08536

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

Silveira, J. (1980). Generic masculine words and thinking. *Women's Studies International Quarterly*, *3*(2-3), 165–178. https://doi.org/10.1016/S0148-0685(80)92113-2

Silvey, C., Kirbey, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 1312–1317.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. *Journal of Memory and Language*, *109*, 104051. https://doi.org/10.1016/j.jml.2019.104051

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467–482.

Sleator, D. D., & Tarjan, R. E. (1985). Self-adjusting binary search trees. *Journal of the ACM*, *32*(3), 652–686. https://doi.org/10.1145/3828.3835

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33. https://doi.org/10.1006/cogp.1997.0672

Smith, C. L. (1979). Children's understanding of natural language hierarchies. *Journal of Experimental Child Psychology*, *27*(3), 437–458. https://doi.org/10.1016/0022-0965(79)90034-1

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386. https://doi.org/10.1162/106454603322694825

Sokal, R. R. (1963). The principles and practice of numerical taxonomy. *Taxon*, *12*(5), 190–199. https://doi.org/10.2307/1217562

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, *38*(6), 1409–1438.

Sokal, R. R., & Rohlf, F. J. (1980). An experiment in taxonomic judgment. *Systematic Botany*, *5*(4), 341–365. https://doi.org/10.2307/2418517

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3–11. https://doi.org/10.1037/0278-7393.28.1.3

Stock, P., & Cisse, M. (2018). ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision – ECCV 2018* (pp. 504–519, Vol. 11210). Springer International Publishing. https://doi.org/10.1007/978-3-030-01231-1_31

Stöttinger, E., Sepahvand, N. M., Danckert, J., & Anderson, B. (2016). Assessing perceptual change with an ambiguous figures task: Normative data for 40 standard picture sets.

Behavior Research Methods, *48*(1), 201–222. https://doi.org/10.3758/s13428-015-0564-5

Suchecki, K., Salah, A. A. A., Gao, C., & Scharnhorst, A. (2012). Evolution of Wikipedia's category structure. *Advances in Complex Systems*, *15*(supp01), 1250068. https://doi.org/10.1142/S0219525912500683

Sultanik, A. (2022). Lessons learned about converting from Dewey Decimal classification to Library of Congress classification. *The Christian Librarian*, *65*(1). https://doi.org/10.55221/2572-7478.2322

Sun, J., & Peng, N. (2021). Men are elected, women are married: Events gender bias on Wikipedia. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 350–360. https://doi.org/10.18653/v1/2021.acl-short.45

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

Svenonius, E. (2000). *The intellectual foundation of information organization*. MIT Press. https://doi.org/10.7551/mitpress/3828.001.0001

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019). What are the biases in my word embedding? *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 305–311. https://doi.org/10.1145/3306618.3314270

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. https://doi.org/10.1016/0010-0285(91)90016-H

Thomson, S. A., Thiele, K., Conix, S., Christidis, L., Costello, M. J., Hobern, D., Nikolaeva, S., Pyle, R. L., Van Dijk, P. P., Weaver, H., Zachos, F. E., Zhang, Z.-Q., & Garnett, S. T. (2021). Towards a global list of accepted species II. Consequences of inadequate taxonomic list governance. *Organisms Diversity & Evolution*, *21*(4), 623–630. https://doi.org/10.1007/s13127-021-00518-8

Timeo, S., Farroni, T., & Maass, A. (2017). Race and color: Two sides of one story? Development of biases in categorical perception. *Child Development*, *88*(1), 83–102. https://doi.org/10.1111/cdev.12564

Tomer, E. (2019). Avian taxonomies compared. Retrieved June 1, 2026, from https://smbasblog.wordpress.com/wp-content/uploads/2019/03/avian-taxonomies-compared2.pdf

Trubetskoy, S. (2017). *List of Western countries*. Retrieved February 9, 2022, from https://sashamaps.net/docs/maps/list-of-western-countries/

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, *25*(6), 983–995. https://doi.org/10.1037/0012-1649.25.6.983

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169–193. https://doi.org/10.1037//0096-3445.113.2.169

United Nations, DESA, Population Division. (2022). *World population prospects 2022: Summary of results* (UN DESA/POP/2022/TR/NO. 3). United Nations, Department of Economic and Social Affairs, Population Division. https://www.un.org/development/desa/pd/content/world-population-prospects-2022

Van Berkel, L., Molina, L. E., & Mukherjee, S. (2017). Gender asymmetry in the construction of American national identity. *Psychology of Women Quarterly*, *41*(3), 352–367. https://doi.org/10.1177/0361684317707710

Verheyen, S., Ameel, E., Rogers, T. T., & Storms, G. (2008). Learning a hierarchical organization of categories. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*, 751–756.

Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 454–463.

Wang, Q., Wang, X., Chen, Z., & Wang, R. (2013). The category structure in Wikipedia: To analyze and know how it grows. In P. Liu & Q. Su (Eds.), *Chinese Lexical Semantics* (pp. 538–545). Springer. https://doi.org/10.1007/978-3-642-45185-0_56

Watson, J., Beekhuizen, B., & Stevenson, S. (2023). What social attitudes about gender does BERT encode? Leveraging insights from psycholinguistics. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6790–6809. https://doi.org/10.18653/v1/2023.acl-long.375

Waxman, S. R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, *5*(2), 123–150. https://doi.org/10.1016/0885-2014(90)90023-M

Waxman, S. R. (2021). Racial awareness and bias begin early: Developmental entry points, challenges, and a call to action. *Perspectives on Psychological Science*, *16*(5), 893–902. https://doi.org/10.1177/17456916211026968

Waxman, S. R., & Kosowski, T. D. (1990). Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. *Child Development*, *61*(5), 1461–1473. https://doi.org/10.2307/1130756

Waxman, S. R., Shipley, E. F., & Shepperson, B. (1991). Establishing new subcategories: The role of category labels and existing knowledge. *Child Development*, *62*(1), 127–138. https://doi.org/10.2307/1130709

Weinberger, D. (2007). *Everything is miscellaneous: The power of the new digital disorder*. Times Books.

Westenberg, J. (2022). Where 'Dewey' go from here? Perpetuation of difference in public library knowledge organization systems. *Aletheia: The Arts and Science Academic Journal*, *2*(2). https://doi.org/10.15173/a.v2i2.3002

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex roles*, *55*, 259–266. https://doi.org/10.1007/s11199-006-9078-z

Whittaker, R. H. (1969). New concepts of kingdoms of organisms: Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science*, *163*(3863), 150–160. https://doi.org/10.1126/science.163.3863.150

Wikipedia:Categorizing articles about people. (2025, January 23). https://en.wikipedia.org/w/index.php?title=Wikipedia:Categorizing_articles_about_people&oldid=1271312953

Wikipedia:Verifiability. (2025, January 26). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:Verifiability&oldid=1272031077

Wikipedia:WikiProject Women in Red. (2025, January 28). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Women_in_Red&oldid=1272297538

Witkowski, S. R., & Brown, C. H. (1983). Marking-reversals and cultural importance. *Language*, *59*(3), 569–582. https://doi.org/10.2307/413904

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2024). Word reuse and combination support efficient communication of emerging concepts. *Proceedings of the National Academy of Sciences*, *121*(46), e2406971121. https://doi.org/10.1073/pnas.2406971121

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, *11*(1), 8–21. https://doi.org/10.20982/tqmp.11.1.p008

Zaki, S. R., & Salmi, I. L. (2019). Sequence as context in category learning: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1942–1954. https://doi.org/10.1037/xlm0000693

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. https://doi.org/10.1073/pnas.1800521115

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *41*, 1254–1260.

Zebrowitz, L. A., White, B., & Wieneke, K. (2008). Mere exposure and racial prejudice: Exposure to other-race faces increases liking for strangers of that race. *Social Cognition*, *26*(3), 259–275. https://doi.org/10.1521/soco.2008.26.3.259

Zesch, T., & Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, 1–8. https://aclanthology.org/W07-0201/

Zhang, W. (2003). Classification for Chinese Libraries (CCL): Histories, accomplishments, problems and its comparisons. *Journal of Educational Media and Library Sciences*, *41*(1), 1–22.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. https://doi.org/10.18653/v1/N19-1064

Zins, C., & Santos, P. L. (2011). Mapping the knowledge covered by library classification systems. *Journal of the American Society for Information Science and Technology*, *62*(5), 877–901. https://doi.org/10.1002/asi.21481