An Experimental Study of the Evolution of Hierarchical Category Systems

Katie Warburton (kwarburton@cs.toronto.edu)[1,2], **Lea Frermann**[3], **Yang Xu**[2,4] **& Charles Kemp**[1]
[1]School of Psychological Science, University of Melbourne
[2]Department of Computer Science, University of Toronto
[3]School of Computing & Information Systems, University of Melbourne
[4]Cognitive Science Program, University of Toronto

Abstract

Hierarchical category systems evolve in response to a changing environment encountered over time. To better understand how such systems develop, we designed a task in which participants sequentially incorporated novel stimuli into an existing taxonomy. We manipulated presentation order, similarity to existing categories, and category system depth. All three factors shaped the hierarchical systems that emerged, and the likelihood of creating a new high-level category decreased as system depth increased. Our findings show that hierarchical category systems depend on the sequence in which items are encountered, and that sequence effects can result in hierarchical category systems that are skewed representations of the world.

Keywords: categorization; hierarchy; sequence effects

Introduction

Humans categorize the world at multiple levels of abstraction (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. In early versions of the Dewey Decimal System, fungi were classified under botany (e.g., Dewey, 1989), reflecting the then-dominant 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 the set of items being classified or our understanding of them 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. 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 as they encounter changing environments over time. To address this gap, we developed a task in which participants must sequentially incorporate novel stimuli into an existing hierarchy. Working with pre-existing systems allows us to observe how people adapt them 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.

Prior work on sequence effects shows that the order of item presentation can influence categorization (Clapper, 2015; Medin & Bettger, 1994), affecting 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 of this work treats category systems as flat, but when hierarchy is considered, the focus is on how item order can bias the level of abstraction at which participants most readily categorize items (Mack & Palmeri, 2015), without addressing how sequence affects the development of the system as a whole.

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. We first illustrate these ideas with a simple scenario demonstrating how sequence and hierarchy might interact, and then test them empirically with our experiment. 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.

Category Evolution in a Changing Environment

To introduce the study and our hypotheses, consider a category system with two categories of shells: *L* for small shells and *R* for big shells. Some of its properties are illustrated in Figure 1A. Suppose there exists a set of shells, *S*, that has not yet been categorized. If all shells were arranged along a

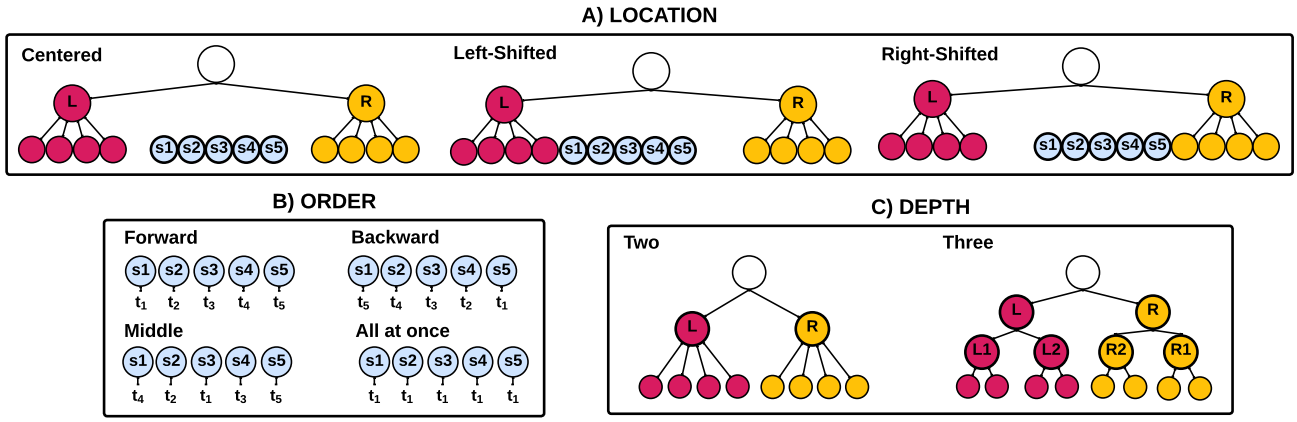

Figure 1: Illustration of the properties of novel items (A, B) and initial category systems (C). Small circles at the bottom denote items, large labelled circles denote subcategories, and white circles denote the root category. Novel items (s_i) are blue; stimuli assigned to category L are pink and to R are yellow. Order is indicated by time stamps (t_i).

one-dimensional axis 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. Based on these ideas, we make the following predictions about the development of a system with categories L and R in response to a novel set of items S with varying similarity to the existing categories.

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 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.

To test these predictions we introduce an experiment, described in the next section. Our hypotheses and their subsequent analyses are pre-registered on AsPredicted (#233,910).[1]

Experiment

This experiment was approved by the University of Melbourne Human Research Ethics Committee (ID: 29169). Stimuli, data, and code can be found on Github.

Methods

Participants We collected data from 419[2] participants across three pools: 80 university undergraduates who received course credit; 159 Prolific participants paid £4.50 for 30 minutes; and 180 additional Prolific participants from a

[1] We pre-registered two additional hypotheses that are not within the scope of this paper but will be included in a longer version.

[2] We exceeded our preregistered target ($n = 260$) because information needed for modelling was not initially recorded. No modelling is reported here, but we retain the full sample for transparency.

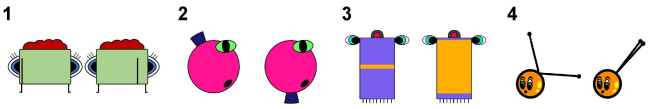

Figure 2: Example aliens from the four stimulus sets. In each set, all stimuli fall between alien 1 (left) and alien 31 (right).

pool screened for attention and use of large language models. This pool was paid £3.20 for approximately 21 minutes as task duration had been more precisely estimated.

Data from 33 participants were excluded for failing comprehension checks 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), aged 18–75 years (median: 33). All were fluent English speakers who provided informed consent. Preliminary analyses indicated a negligible effect size from a chi-squared test of homogeneity comparing category assignment across pools,[3] so 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. Within each set, aliens varied in equal intervals along a single continuous feature, with aliens 1 and 31 representing the two endpoints (Figure 2). 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) 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 manipulated three factors illustrated in Figure 1: the **location** of novel stimuli, the **order** of presentation, and the **depth** of the existing category system.

The first two factors, location and order, concerned the stimuli participants categorized. Location (Figure 1A) determined which portion of the stimulus set participants saw, reflecting similarity to existing categories. Stimuli were centred (equally similar to L and R), left-shifted (more similar to L), or right-shifted (more similar to R).

Order (Figure 1B) determined the sequence in which stimuli were presented. In the forward condition, stimuli were shown left-to-right; in the backwards condition, right-to-left. Both

[3] $\chi^2\,(4) = 17.16$, $p = .002$, Cramér's $V = .02$

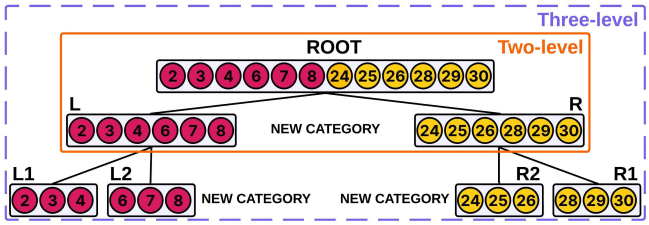

Figure 3: Representation of the two-level and three-level hierarchical category system at trial start. Numbered circles represent stimuli, and their surrounding boxes denote categories. Participants could see "NEW CATEGORY" buttons, but not category labels (e.g., $L1$).

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 (Figure 1C), 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 $R2$.

These factors were combined to create 24 unique trial types and we used an incomplete block design with six blocks to assign participants to four trials each. Every participant completed two trials at each depth, one trial for each presentation order, and at least one trial for each location. Within blocks, trial order was randomized, and stimulus sets were assigned randomly so that each set was seen only once per participant.

Procedure Participants started the experiment by completing practice tasks where they were asked to categorize images of animals into a taxonomy. These tasks mirrored the experimental trials and familiarized participants with the interface. They then had to correctly answer six comprehension questions before proceeding. Next, participants completed four experimental trials in their assigned block, categorizing aliens into a partially completed taxonomy. Between trials, a screen indicated how many of the four alien sets they had finished.

Participants could add aliens to the taxonomy either by placing them in an existing category or by creating a new one. In the two-level condition participants could add one new category, and in the three-level condition they could add up to three, one subcategory for each existing category plus an additional new category X.

At the start of each trial, participants familiarized themselves with the existing alien taxonomy (Figure 3). In both the two-level and three-level conditions, the top category contained 12 aliens split evenly between L and R. In the three-

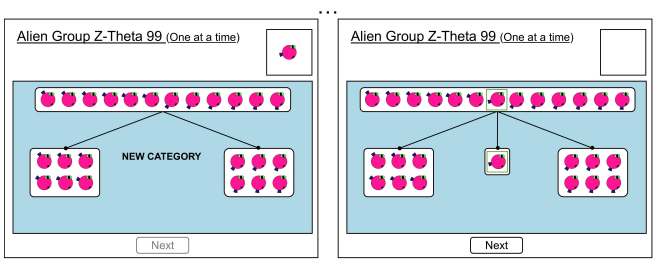

Figure 4: Example screens from a sequential trial before (left) and after (right) an alien is added to a new category. Newly added aliens are highlighted with a green box.

level condition, each of the four subcategories contained three aliens. Aliens were always displayed in sorted order. There were also buttons to create new categories and subcategories. In both conditions there was one button between L and R. In the three-level condition there were two additional buttons, one to the right of $L2$ and one to the left of $R2$.

Once participants started the trial they were shown nine novel aliens (S) and four distractors (D). Distractors were aliens 1, 3, 29, and 31, and served as clear examples of categories L and R. The novel aliens varied by condition, with aliens 12 to 20 shown in the centred condition, 9 to 17 shown when left-shifted, and 15 to 23 when right-shifted.

Aliens were presented either in sequence (forward, backward, middle), or all at once. When shown in sequence, aliens appeared in the upper right corner of the screen, as shown in Figure 4. Novel aliens and distractors were interleaved as $[s,d,s,s,s,d,s,d,s,s,s,d,s]$, with distractor order randomized. Each alien had to be categorized before moving on, and only the current alien could be moved between categories. When shown all at once, aliens appeared in a box spanning the top of the display. Unlike the sequence conditions, participants could move all presented aliens between categories.

Aliens placed into an existing category were inserted in sorted order. New categories appeared at the location of the selected “NEW CATEGORY" button and were linked to their parent category. In the two-level condition, the new category was labelled X. In the three-level condition, participants could create $L3$ under L, $R3$ under R, and X. For X, a subcategory, $X1$, was automatically created to maintain a consistent depth.

Category assignments were recorded per trial and per alien. Data were excluded if participants failed to answer all comprehension questions after three attempts or miscategorized at least four distractors across all trials. A distractor was miscategorized if an L distractor was not assigned to L or an R distractor was not assigned to R.

Results

Participants completed a total of 1544 trials across three pools, categorizing 13,896 novel items. 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 general trends in category assignment.

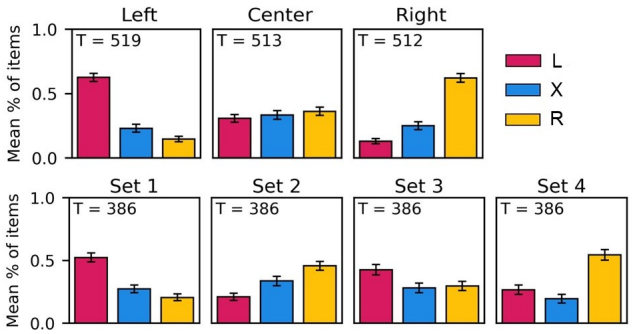

Figure 5: General trends in the second level category assignment of items by location (top) and stimulus set (bottom). The number of trials (T) are reported in the top left of each panel and error bars represent 95% confidence intervals computed from 5000 bootstrap samples.

We varied the location of the novel items and the stimulus set. Figure 5 shows the mean proportion of items assigned to each category across locations (top) and sets (bottom). Participants were sensitive to item location and a chi-squared test confirmed a significant association between location and category, $\chi^2(4) = 3152.9$, $p < .001$, with a medium effect (Cramér's $V = .34$). Participants tended to assign items to L when they were left-shifted, and to R when right-shifted. When centred, assignments to L, X, and R were roughly equal. We thus analyze locations separately when appropriate.

Participants were also sensitive to stimulus set as a chi-squared test found a significant difference in category assignment across sets, $\chi^2(6) = 1391.3$, $p < .001$, with a medium effect size (Cramér's $V = .22$). This is expected as the sets are not perceptually equivalent, but to account for this variation, all regressions include a random intercept for stimulus set. We expected individual differences in categorization and include a random intercepts for participant as well.

Sequence Effects We tested the overall effect of order on category assignment by examining the mean proportion of novel items placed in categories L, R, and X. These proportions are displayed in Figure 6, overall and by location.

Figure 6 suggests that, on average, the proportion of times an item is assigned to each category varies with presentation order, regardless of location. To confirm this, we fit multinomial logistic regression models predicting category assignment as a function of order and compared these to null models without order using likelihood ratio tests (see Table 1). Models were applied separately to each location and included random intercepts for stimulus set and participant.

Order was a significant predictor of category assignment for all location conditions, with likelihood ratio tests preferring models that included order over the null models. To assess the overall effect of order across locations, we fit an additional model including a random intercept for location. The like-

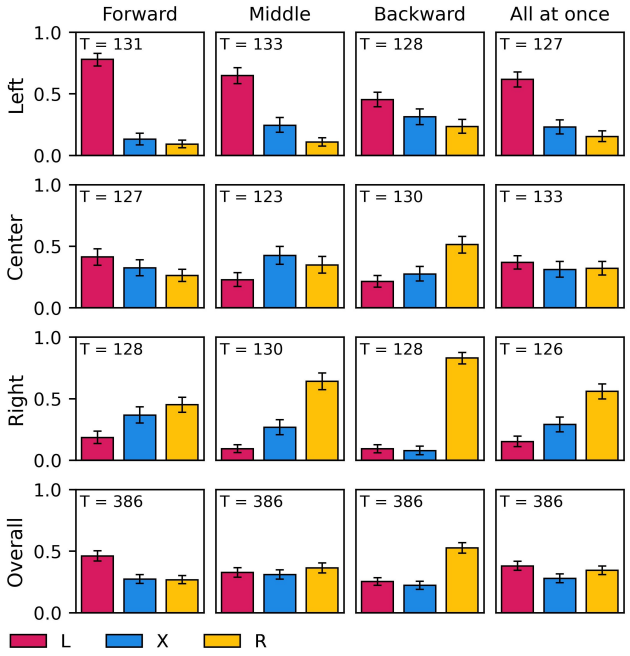

Figure 6: Mean Proportion of Items in Categories L, X, and R by Location and Order. The rows are the item locations and the columns the presentation orders. The number of trials (T) are reported in each panel and error bars represent 95% confidence intervals computed from 5000 bootstrap samples.

lihood ratio test again preferred the model including order, although the β coefficients were smaller than those obtained when each location was analyzed separately. Together these results are consistent with our prediction that presentation order influences category assignment.

Looking in more detail at the sequence effects, we tested specific predictions for each location condition (H1*a–c*). Logistic regressions were fitted to predict whether an item was assigned to a specific category or not, again with random intercepts for stimulus set and participant. Likelihood ratio tests compared each model including order to a null model without it.

In the centred condition, results supported Hypotheses *1a(i)* and *1a(ii)*. Participants were more likely to assign novel items to category L in the forwards than in the backwards 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, $\chi^2(1) = 161.64$, $p < .001$. Category R was more likely in the backwards than in the forwards order, $\beta = 1.30$, 95% CI $[1.01, 1.58]$, with the full model again preferred to the null model, $\chi^2(1) = 79.41$, $p < .001$.

Hypothesis *1a(iii)* was not supported. We predicted that centred items would be more frequently assigned to category X in the middle order than in the forward or backward orders, but found that items were less likely to be assigned to X in the middle order ($\beta = -0.28$, 95% CI $[-0.58, 0.01]$) when including random intercepts. This effect was not statistically

Table 1: Multinomial regressions predicting category assignment from order. All models include a fixed intercept, and random intercepts for stimulus set and participant. The overall model also includes a random intercept for location. All order coefficients are with reference to the middle order. χ^2 and p-values are from a likelihood ratio test between the full model and an equivalent null model without order.

Location	Category	Order	β	95% CI	$\chi^2(4)$	p
Left	*L vs. X*	Intercept	0.99	[0.34, 1.64]		
		Forward	0.97	[0.74, 1.20]		
		Backward	−0.64	[−0.84, −0.44]		
	R vs. X	Intercept	−1.10	[−1.79, −0.41]		
		Forward	0.58	[0.25, 0.90]		
		Backward	0.76	[0.49, 1.04]	355.41	< .001
Centred	*L vs. X*	Intercept	−0.78	[−1.52, −0.04]		
		Forward	1.00	[0.78, 1.22]		
		Backward	0.30	[0.06, 0.53]		
	R vs. X	Intercept	−0.31	[−0.80, 0.17]		
		Forward	−0.01	[−0.20, 0.22]		
		Backward	0.88	[0.68, 1.07]	256.38	< .001
Right	*L vs. X*	Intercept	−1.18	[−1.53, −0.84]		
		Forward	0.45	[0.17, 0.73]		
		Backward	1.21	[0.85, 1.56]		
	R vs. X	Intercept	0.96	[0.14, 1.78]		
		Forward	−0.98	[−1.19, −0.78]		
		Backward	1.49	[1.23, 1.76]	501.56	< .001
Overall	*L vs. X*	Intercept	−0.30	[−1.48, 0.88]		
		Forward	0.73	[0.60, 0.86]		
		Backward	−0.08	[−0.23, 0.06]		
	R vs. X	Intercept	−0.15	[−1.24, 0.94]		
		Forward	−0.33	[−0.46, −0.20]		
		Backward	0.88	[0.75, 1.00]	**865.38**	**< .001**

significant, and a likelihood ratio test did not favour the full model over the null model, $\chi^2(1) = 3.34$, $p = .068$.

In the left-shifted condition, results supported Hypothesis *1b* 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, $\chi^2(1) = 73.15$, $p < .001$. For the right-shifted condition, results supported Hypothesis *1c*. Category R assignments were more likely in the backward than in the forward order, $\beta = 3.77$, 95% CI $[3.00, 4.54]$, and a likelihood ratio test preferred the full model to the null model, $\chi^2(1) = 105.31$, $p < .001$.

All together, these results show that presentation order influenced category assignments in the predicted directions for all hypotheses except *1a(iii)*. The same items can end up in different categories depending on when they are encountered.

Hierarchy Effects We predicted that participants would be more likely to create new category X in the two-level condition than in the three-level one. This prediction was confirmed, and across all trials, participants in the two-level condition created new category X 53.9% of the time whereas participants in the three-level condition created X only 32.9% of the time.

To test for statistical significance, we pooled all 1544 trials and coded each trial as 1 if X was created and 0 otherwise. We fit a logistic regression predicting creation of X as a function of

Table 2: Jensen–Shannon divergence (JSD) between category distributions in the all-at-once and sequence conditions. *P*-values are from a permutation test with 10,000 iterations.

	Forward		Middle		Backward	
	JSD	p	JSD	p	JSD	p
Left	0.125	<.001	0.048	.342	0.117	.002
Right	0.076	.062	0.071	.088	0.222	<.001

depth, with a fixed intercept and random effects for participant and stimulus. The coefficient for depth two relative to depth three was $\beta = 1.13$, 95% CI $\left[0.89, 1.38\right]$, and a likelihood ratio test preferred the full model over a null model with only a fixed intercept, $\chi^2\left(1\right) = 90.80$, $p < .001$.

These results are consistent with our prediction that depth influences whether a new second-level category is created, and show that a system's current structure influences its subsequent development. Greater depth introduces more subcategories and thus more ways to incorporate an item into existing categories, reducing creation of new categories near the top.

Creating Suboptimal Category Systems 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 used Jensen-Shannon divergence between category assignments to test whether responses in the left- and right-shifted sequence conditions were significantly different from responses in the all-at-once condition. Table 2 shows that in the left-shifted condition, participants' systems differ significantly from the all-at-once baseline in the forward and backward orders, but not the middle order. In the right-shifted condition, only the backward order shows a significant difference. These findings suggest that some, but not all, orders lead to less optimal category systems when compared to the all-at-once benchmark.

Discussion

Category systems do not form all at once but instead develop as items are encountered over time. We examined this process with an experiment where participants sequentially incorporated novel stimuli into an existing taxonomy and found three main results. First, presentation order consistently shapes how a category system evolves. Second, 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. Third, sequential presentation can lead to systems that are different from those produced given full knowledge of the stimulus set, and therefore likely suboptimal.

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 more work is needed to understand the conditions that lead to positive as opposed to negative hysteresis effects.

With respect to hierarchy, our findings suggest that broader, high-level categories are more likely to emerge when the existing system is relatively coarse. Future work could connect this pattern to the levels of abstraction in category hierarchies (e.g., basic vs. superordinate; Rosch et al. 1976) and examine how the granularity of a hierarchy shapes where new categories are introduced. 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.

An important direction for future work is to apply computational models to our data. Models such as the rational model of categorization (Anderson, 1991; Sanborn et al., 2010) have been used to account for sequence effects, but most work only with flat category systems. We therefore believe that existing models are unlikely to explain all of our results, and in particular our finding that hierarchy depth affects the creation of new categories. Our results may therefore highlight the need for new models that are intrinsically hierarchical.

Another future direction is to build on our current analyses using a formal measure of optimality. We treated responses in the all-at-once condition as a benchmark which suggests that sequence effects lead to suboptimal systems, but a formal measure of optimality would allow stronger conclusions about the detrimental effect of certain orders and the positive effect of others. A formal measure could also potentially be applied to real-world hierarchies such as library classification systems to test the hypothesis that these systems become gradually less optimal over time, and that major reorganizations of these systems (e.g. the reclassification of Fungi) help to address this drift away from optimality.

Finally, our experimental paradigm did not allow participants to revise decisions they had previously made, but future experimental work could allow more flexibility in order to study when and how participants revise their existing hierarchies. Allowing structural reorganization connects with theories of conceptual change and acquisition (Chi, 1992; Keil, 1992), and would allow participants to make changes analogous to those that occur in library classifications and other institutional systems as they are periodically updated.

Category systems are everywhere and shape how we make sense of the world around us. Studying how these systems evolve and become skewed is important for understanding how we internally represent the world, and can help explain why everyday systems like the Dewey Decimal System, are organized the way that they are.

Acknowledgements

We would first like to thank Ben Stone for all of his help getting the experiment online. We would also like to thank Viola Pucci, Merrick Giles, Uyen Doan, Chunhua Liu, Matteo Guida, Damian Curran, Bryan Chen, Yilin Geng, Bel Moore, Jasmin Stariolo, and Zoë Wilson for piloting and providing feedback on various versions of the experiment. Data collection was funded by the Melbourne School of Psychological Sciences and the Complex Human Data Hub. KW is funded by the U of T–UoM IRTG program.

References

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

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

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*.

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

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

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

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.

Egré, P., De Gardelle, V., & 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

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*.

Keil, F. C. (1992). *Concepts, kinds, and cognitive development*. MIT Press.

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

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

Raffman, D. (2011). Vagueness and observationality. In *Vagueness: A guide* (pp. 107–121). Springer. https://doi.org/10.1007/978-94-007-0375-9_5

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

Ryder, M. L. (1964). The history of sheep breeds in Britain. *The Agricultural History Review*, *12*(1), 1–12.

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

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

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

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

Witkowski, S. R., & Brown, C. H. (1983). Marking-reversals and cultural importance. *Language*, *59*(3), 569–582. https://doi.org/10.2307/413904

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