Uncovering children's concepts and conceptual change

Pablo León-Villagrá (pablo_leon_villagra@brown.edu) [1] **Isaac Ehrlich**[2]
Christopher G. Lucas[3] Daphna Buchsbaum[1]
[1]Brown University, USA [2]University of Toronto, Canada [3]University of Edinburgh, UK

Abstract

Capturing the structure of human conceptual knowledge is a challenging but fundamental task. The most prominent approach, Multidimensional Scaling (MDS), usually requires many similarity judgments, which leads to long experiments, and only provides a representation of a fixed set of stimuli. In contrast, we present a more flexible method that can generalize to novel stimuli. This method uses a child-friendly task that allows researchers to uncover the development of categories with fewer participant judgments. We evaluate this approach on simulated data and find that it can accurately reveal representations even when trained on data generated by groups that categorize differently. We then analyze data from the World Color Survey and find that we can recover language-specific color organization. Finally, we use the method in a novel developmental experiment and find age-dependent differences in how fruit categories are structured. These results suggest that our method is widely applicable in developmental tasks.

Keywords: Categorization, psychological spaces, child development

Introduction

The ability to form categories develops in early infancy (Quinn, Eimas, & Rosenkrantz, 1993) and allows us to extract useful and generalizable features from individual exemplars across a variety of everyday tasks. For example, when people go grocery shopping, they rarely search for "green things", or "sweet things". Instead, people shop based on categories, like "apples" or "candy".

Adults tend to have similar expectations about many categories, having had numerous and broadly similar encounters with category members and their features, e.g., different fruits and their variations in shape, color, texture, and so forth. In contrast, in a domain where people must rely on sparse evidence – as many domains are for young children – it is plausible that different people will come to different conclusions about categories, based on the idiosyncrasies of their own experiences. For example, if one has only encountered a small set of fruits, say pineapples and bananas, one might deduce that all fruits are yellow. In contrast, if this early experience includes oranges, one might infer that fruit colors can be orange or yellow. As children go from knowing very little to having adult-like categories, we might expect their beliefs about categories to change systematically. Since children can differ in their experiences, and thus in the types of expectations they develop, experiments need to be able to uncover detailed, group-specific category structures that are robust to individual differences.

However, directly accessing these structures in experiments is challenging. As a result, it is more common to focus on the judgments and choices people make that are guided by their category structures, e.g., judgments of similarities between pairs of items. Then, one can infer the category structures that are most consistent with participants' similarity judgments. One of the most prominent approaches to infer these category structures is multidimensional scaling (MDS; Shepard, 1980). MDS is a method that assumes that the observed data, usually similarities between items, results from the items' distances in a geometric psychological space.

For example, if a participant deems oranges and limes similar but apples dissimilar from both, MDS would attempt to position oranges and limes close together in psychological space while keeping apples further away. This example highlights two important properties of the category representations that MDS finds: First, the dimensionality of the geometric space directly affects how closely the distances can mimic the similarity data. For some data, like the fictitious participant who deems oranges and limes similar, but not apples, even 1-dimensional spaces (a line) can be sufficient. On the other hand, high-dimensional spaces might be required to faithfully capture similarity judgments for complex data.

Second, the geometric spaces that explain participants' judgments do not have to match the perceptual features of the stimuli. In our example, limes share colors with some apples, but this similarity may be incidental to a person who knows about typical textures and shapes of these fruits. Since MDS is often used to uncover these psychological phenomena, the recovered spaces are called psychological spaces. MDS has been a crucial tool in uncovering psychological spaces ranging from the perception of colors to facial expressions (for an introduction and overview, see Borg & Groenen, 2005). Moreover, data obtained via MDS have been vital in developing models of human cognition, such as the universal law of generalization (Shepard, 1987).

MDS and Developmental Studies

While MDS offers a convenient way to chart psychological spaces, it relies on obtaining reliable human similarity judgments between items. Experimentally, obtaining these judgments poses several challenges. First, many experimental paradigms require participants to apply some implicit understanding of what kind of similarity is being measured – in

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

which respect are the items similar (Medin, Goldstone, & Gentner, 1993)? Furthermore, the participant must have an explicit understanding of graded similarity, with some items being more similar than others, which might not be warranted for young children (Medin et al., 1993). Second, since making a similarity judgment requires comparing two (or more) items, and traditional MDS methods require measurements of similarity for all item pairs, even small numbers of stimuli can lead to prohibitively taxing experimental setups.

One common simplification to address these shortcomings is to assume that the psychological space is 2D, and participants can express similarities in spatial distances (Goldstone, 1994). With this assumption, many similarity judgments can be derived at once by asking people to organize items spatially, putting similar items closer together and dissimilar items farther apart, a task that is easily understood even by preschoolers. Moreover, recent extensions of this technique have suggested that while the task imposes 2D organizations, higher-dimensional spaces can be learned when aggregating participants (Richie, White, Bhatia, & Hout, 2020) and the method can be used with young children (Unger, Fisher, Nugent, Ventura, & MacLellan, 2016). However, questions remain about how reliable spatial ordering tasks are and how strongly spatial biases influence participants' perceived similarity (Verheyen, White, & Storms, 2020).

Here, we propose a computational and experimental method that can recover *generalizable* psychological spaces. Instead of merely mapping observed stimuli to locations in a psychological space, like most previous MDS methods, this method learns a function from features of stimuli to a latent representation that explains participants' judgments. This learned function can be applied to arbitrary new stimuli, leading to psychological spaces that are generalizable: predictions can be made about the similarity of stimuli the participant has not seen or rated and their location in the psychological space can be obtained. This flexibility has important methodological implications and offers exciting prospects that alternative methods cannot provide. However, we do not argue that this method is universally preferable over metric or non-metric MDS, or that it recovers better spaces than alternative methods when abundant similarity judgments are available. Instead, we think that this approach is competitive with MDS solutions and has the unique feature of providing a generalizable solution.

Learning Similarities Implicitly

The method we are proposing as a computational and experimental paradigm is based on an approach from computer science called deep metric learning, a family of deep neural network architectures that learn similarities from groupings–items that have been sorted into groups taken to be the "same" type (with items across groups being "different"). The goal of training this network is to uncover the psychological spaces by learning to predict same-different judgments between experimental stimuli. Importantly, the classification into same or different is based on the distance of the two stimuli in a geometric space, which is encoded in the output layer of the network, see Figure 1.

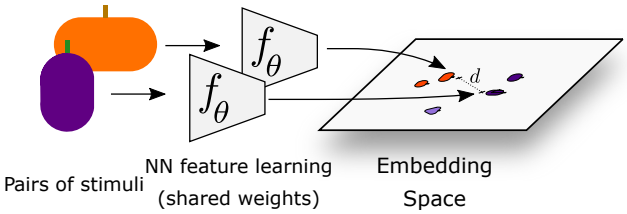

Figure 1: DrLIM is a twin (or "Siamese") deep metric learning network that uses two identical networks f with shared parameters θ to project pairs of stimuli into an embedding space. The network is then trained to minimize the error of predicting same-different judgments for all pairs, given the distance d in the embedding space.

These techniques produce a function that maps stimuli in the domain to points in psychological space. One such method, and the method we are using here, is dimensionality reduction by learning an invariant mapping (DrLIM; Hadsell, Chopra, & LeCun, 2006). In cognitive science, DrLIM has previously been used to elicit adult psychological spaces from same-different judgments (Sanborn, Griffiths, & Shiffrin, 2010)[1]. However, as in the approaches discussed above, the experiment in Sanborn et al. (2010) required each participant to produce many same-different judgments (90 judgments each), making the task too demanding for many developmental settings.

Here, we evaluate DrLIM as a method for learning psychological spaces in developmental studies. This goal raises two important requirements. First, while adult studies can sometimes afford to present each participant with the full set of possible category members, or ask them to make comparisons between all pairs of items, these high cognitive and attentional demands are not practicable in most developmental settings. Therefore, DrLIM must be applicable in experimental designs in which each child only encounters a subset of all stimuli, and psychological spaces are aggregated over those subsets. Second, if our aim is to understand how category representations change over development, we must be able to aggregate data by age, and recover commonalities that are robust to individual differences between children. In particular, DrLIM should be able to accommodate differences in categorization or judgment strategies, e.g., one child might decide all orange-like fruits are the same, and another might separate mandarins and navel oranges.

We first used simulated data to validate that DrLIM is robust to aggregated responses of heterogeneous data, as this allows us to compare the solutions obtained with DrLIM against the ground-truth data. Then, we tested DrLIM in the psychologically relevant domain of color categorization, with

[1] A similar method was also used in work by Lee (1997). In contrast to our work, this early approach was trained on pre-computed features of the stimuli to capture adult categorization results.

data from the World Color Survey(WCS; Kay, Berlin, Maffi, Merrifield, & Cook, 2009), and found that it could recover language-dependent color representations. Finally, we conducted an experiment with children and adults, showing that our approach is effective and can be used to recover categorical organization with children as young as four.

Simulations

To establish that DrLIM can recover meaningful spaces for aggregated agents, we first validated it on simulated data. This evaluation is critical since previous work by Sanborn et al. (2010) only analyzed individual participant data. However, when aggregating data, participants can differ in the categories they use, for example, categorizing at different levels of specificity, or, in developmental studies, using incongruent or idiosyncratic categories.

Simulating Aggregate Groupings

We constructed a synthetic category structure from four feature distributions a, b, c, d. The feature distributions were normal, bi-variate distributions, shifted in x and y according to two variables x_s and y_s, $\mu = [[-1,1],[-1,y_s],[x_s,1],[x_s,ys]]$. The two variables controlled how far apart in x and y the four distributions were on the 2D plane. We generated 30 features samples from each distribution. Samples were clearly separable in x, $x_s = 5$ and less separable in y, $y_s = 2.5$. For the resulting samples and spaces, see Figure 2.

We generated a set of 20 simulated agents, with each agent having an implicit categorization rule that grouped the synthetic features either at a high level or at the feature level. High-level classification merged pairs of features into two categories, for example $[a,b],[c,d]$. In contrast, agents classifying at the feature level produced four categories. Agents probabilistically classified each stimulus according to the softmax over the ground-truth class-membership likelihoods (the choice axiom, Luce, 1959). To simulate the behavior of a heterogeneous group of people categorizing items according to different criteria, we created two types of agents:

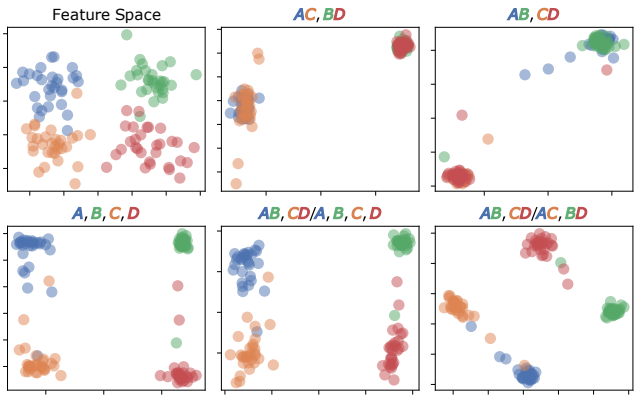

Figure 2: The true feature space that generated the category structures and 2D solutions for the five agent simulations.

high-level categorizers (AB,CD, or AC,BD) and feature-level categorizers (A,B,C,D). We then combined these agents in four types of agent-aggregations: all high-level categorizers, all feature-level categorizers, $50/50$ mixtures of feature-level and high-level categorizers ($AB,CD/A,B,C,D$), and a $50/50$ mixture of both types of high-level categorizers ($AB,CD/AC,BD$) .

We used a 3-layer network with interspersed dropout layers (10% dropout). The output layer varied from 1-4 units, corresponding to the dimensionality of the solutions. All other layers had 30 units and rectified linear activation functions. We optimized contrastive loss (Hadsell et al., 2006) using stochastic gradient descent for 800 epochs. The loss margin was set to 0.5, and we repeated the procedure 25 times.

Results

Overall, we could recover the simulated spaces, both in terms of the dimensionality of the best-fitting solution and the arrangement of points within those spaces. When all agents grouped items based on the low level, we recovered the full feature space. In contrast, if all agents grouped based on high-level features, our solution collapsed to those two features. Crucially, even when the data consisted of conflicting aggregations, we recovered features of the ground-truth space.

The difference in loss for dimensions 1-2 ranged from $M = 2 \times 10^{-4}$ for AC,BD to $M = 1.3 \times 10^{-2}$ for $AB,CD/AC,BD$. Subsequent dimensions did not reduce loss considerably, indicating that 2D solutions achieved acceptable results (all loss reductions $< 1 \times 10^{-4}$). For all 2D results, see Figure 2. Training loss reflected the complexity of the simulations: Homogeneous agent populations resulted in lower loss than mixtures of classification schemes, and overlapping feature boundaries (AB,CD) increased loss. Finally, heterogeneous simulations were the most difficult to train, with the mixture of inconsistent populations of high-level classification schemes $AB,CD/AC,BD$ producing the highest overall loss.

The spaces correctly collapsed onto two feature-groups for homogeneous high-level simulations, whereas the feature-level simulation, A,B,C,D, maintained the original structure. Crucially, mixtures of categorization schemes reflected the global structure of the feature space.

To quantify the accuracy of our solutions–how well the category space recovered by DrLIM represented the true underlying category distributions–we clustered the items according to their location in the 2D space learned by DrLIM, using k-means, with k set to the true number of item groups. The clustering recovered the true clusters with a high accuracy (k-means accuracy $AB,CD = 100\%$, $AC,BD = 98\%$, $A,B,C,D = 97\%$, $AB,CD/A,B,C,D = 86\%$, $AB,CD/AC,BD = 75\%$).

In addition, we compared our approach with non-metric MDS solutions trained on co-occurrence, using the widely-used SMACOF package (De Leeuw & Mair, 2009). Our solutions aligned well with the solutions obtained by non-metric MDS, both in terms of the dimensionality of the solution (2-dimensional solutions across all simulations were acceptable)

and the 2-dimensional spaces recovered[2]. Across simulations, our results strongly resembled the results obtained by MDS (all correlations between pairwise distances $R^2 > .93$, all $p < 0.001$).

Our simulation results confirmed that our method could recover the complexity and structure from aggregated data. Moreover, when simulations consisted of heterogeneous agents, we still recovered important features of the underlying feature space. Next, we test our methods in a more psychologically relevant domain - language-specific variability of color spaces.

Uncovering Language-specific Color Spaces

Color categorization is a central testing ground for theories of language, perception, and the origins of human cognition (Skelton, Catchpole, Abbott, Bosten, & Franklin, 2017; Watson, Beekhuizen, & Stevenson, 2019). While color perception is rooted in the biological transformation of light's physical properties, color categorization exhibits cross-cultural and individual variability. As such, color naming data provide an interesting test case for our approach. Here, we evaluate the applicability of DrLIM to heterogeneous datasets by training on data from the World Color Survey[3] (WCS). The WCS dataset contains color terms for 330 color chips from speakers of 110 languages in non-industrialized societies.

In order to evaluate DrLIM on color spaces of varying complexity, we evaluated six languages: two three-term languages (the smallest number of basic color terms in the WCS), two seven-term languages (the most frequent number of basic color terms in the WCS), and two 11-term languages (the number of basic color terms in English). To determine the number of color terms in each language, we assigned each chip to the majority color term used to label it. We then categorized all pairs of chips for each speaker in each language, categorizing them as the same if the speaker assigned the same term for both colors. We used the same network as for the agent simulations but increased the maximum output dimension to 6 to account for the data's higher complexity.

Results

Overall, DrLIM was able to uncover language-specific psychological spaces underlying color terms. Consistent with previous work, our results find good fit with 2D spaces. Furthermore, colors deemed highly salient, or focal, aligned well with clusters in the uncovered spaces. Similar to the simulation results and consistent with the structure of perceptual color spaces, 2D solutions achieved acceptable loss, with subsequent dimensions offering only minor reductions. Loss improved on average 0.14 for dimensions 1-2 (all subsequent improvements < 0.04). The resulting 2D spaces corresponded well to the number of terms within a language.

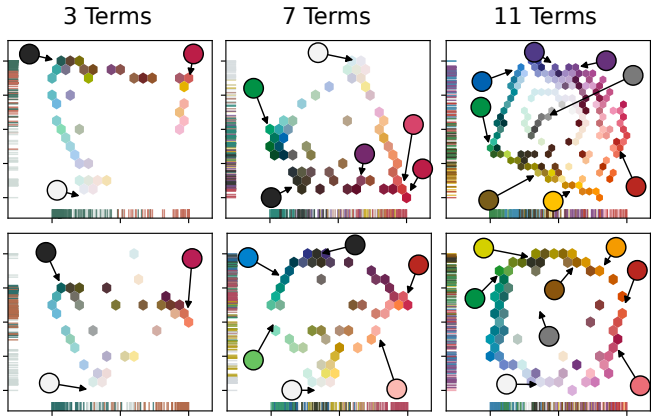

Figure 3: Six WCS languages that differ in their number of terms (columns). To display the high-dimensional data, we bin colors in a 25×25 hexagonal grid and only display bins with ≥ 2 colors. We then calculate the most representative color for each bin. In addition, rug plots display the distribution of terms within the 2D space. We overlay the most frequent focal colors, as collected in Kay et al. (2009).

Focal colors[4], which speakers deemed the most representative of each color term were positioned on separated clusters within the learned space. In contrast, colors for which speakers disagreed in their color terms often interpolated between the clusters, see Figure 3. To assess if the solutions captured the color terms used in each language, we applied k-means clustering to the 2D solutions, with k set to the number of majority color terms in the language. The clustering recovered the colors the speakers judged the same or different accurately (across all languages $\geq .72\%$).

Overlapping Subsets The color dataset corresponds to a balanced design with a large number of judgments per participant, with each speaker receiving and labeling all 330 color chips. However, often experiments require participants to rate subsets of items, with little overlap between participants. To evaluate if DrLIM is robust to such designs, we fitted the model to subsets of the WCS. We selected a set of 75 color chips at random and varied the overlap of colors between speakers. The sets were created by sampling the overlapping chips without replacement from the subset (25%, 50%, 75% out of 75) and selecting the remaining colors from the subset's complement. To evaluate the resulting models, we correlated the solutions provided by the overlapping data with the full dataset. All overlapping models produced highly correlated solutions (all $R^2 > .86$).

Interim Discussion

DrLIM produced encouraging results for both the simulated and color data, recovering domain structure and relevant clusters. These results are novel contributions to the approach

[2]For the aligned 2D solutions obtained via SMACOF, see https://osf.io/c85vd

[3]https://www1.icsi.berkeley.edu/wcs/

[4]These were colors that were collected from language informants in Kay et al. (2009) and that were considered the most representative for the color terms in each language.

in Sanborn et al. (2010), as they establish that DrLIM can be used with aggregated data even for heterogeneous simulations or when language speakers do not receive the same stimuli. This suggests that the method can produce meaningful psychological spaces in developmental experiments, even if children exhibit some individual differences in their sorting behavior or in experiments in which it is unfeasible to present all stimuli exhaustively.

Uncovering Age-dependent Fruit Spaces

We evaluated the item groupings of young children and adults for a set of stylized fruits, a stimulus set previously used to uncover psychological spaces and latent category distributions (Sanborn et al., 2010; León-Villagrá, Otsubo, Lucas, & Buchsbaum, 2020). To make the task more accessible to children, we collected same-different pairs via a grouping task, in which children were asked to place similar fruits in boxes.

Participants

Participants were split into three age groups of N = 30 each: 4-5 years old, 6-7 years old, and adults (18 years or older). Adults were recruited from the Greater Toronto Metro Area (M_{age} = 21.13, SD = 5.25, 25 female). Children were recruited from a local Toronto museum. An additional 15 children were excluded according to our preregistered criteria: 7 for picking an incorrect fruit in the familiarization task, 5 for placing all stimuli into one box, and 3 for not completing the experiment[5]. For 4-5-year-old children, M_{age} = 4.57, SD = 0.5, 16 female, and for 6-7-year-old children, M_{age} = 6.4, SD = 0.5, 13 female. Adult participants received $10 or course credit, and child participants received a small toy.

Materials and Procedure

The experiment consisted of a familiarization task to verify that the participant understood the task and a grouping task. In the familiarization task, the participant was told to imagine going to a grocery store and was presented with 16 cards (9×6 cm) displaying the fruits. The cards were presented in random order in a grid of 4×4 cards in front of the participant. The participant then was shown one of four possible fruit cards (the target fruit) and asked, "Can you help me find one more of this kind of fruit?". Once the participant selected a card, the chosen card was removed, and the question was repeated (five questions in total).

Fruits on the 16 choice cards were one of four colors (red, orange, green, or purple) and one of four shapes. Three cards matched a particular target exactly in color, and three matched exactly in shape. The remaining cards did not match in color or shape. Fruits were programmatically generated using the method described in Sanborn et al. (2010). Each fruit was a single-colored convex hull around three equally sized circles; see Figure 4. Six parameters determined the fruit's appearance: three determined its shape (radii, horizontal distance, and vertical distance), and three determined its color (hue, saturation, lightness). Each fruit was topped with a brown stem to orient the participants to the top of the fruit.

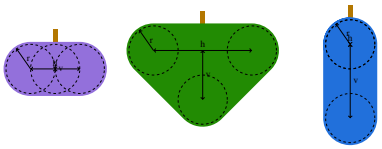

Figure 4: Examples of programmatically generated fruits

In the main task, participants were told they were "in charge of the grocery store" and presented with a random set of 36 cardboard tiles (5×5 cm) out of 180. To evenly cover the stimulus set, a 6-dimensional Sobol sequence determined the fruit's shapes and colors in the main task. In addition to the tiles, participants also received 36 paper boxes made by folding a single sheet of letter-size paper. Participants were asked to place all fruits into the boxes, with fruits of the same kind going into the same box. After completing the main task, participants were asked to explain why they grouped cards together for three boxes ("You put all of these fruits in the same box. What makes these the same kind of fruit?"). The three boxes were selected at random from all boxes that contained at least two fruits.

Results and Discussion

We used the same model and training procedure as for the WCS data, training on the six parameters defining the fruits. Again, across all age groups, loss was acceptable for 2D solutions allowing us to focus on these for comparison and visualization purposes, see Figure 5. The spaces of 4-5-year-olds exhibited gradual changes in the saturation and color of fruits but less differentiation in terms of shape, potentially reflecting a preference for color-based groupings. In contrast, adults exhibited less organization according to color. Instead, fruits were organized broadly around shape differences. We quantified how similar the three resulting 2D spaces were by calculating the distances between all points within an age group and correlating these distances. The youngest children exhibited the lowest mean distances ($M = 0.03$, $SD = .02$), while the spaces of 6-7 year-olds ($M = 0.16$, $SD = .08$) and adults ($M = 0.31$, $SD = 16$) where more dispersed. We then correlated the item-wise distances across age groups. These correlations matched our qualitative descriptions of the 2D spaces: 4-5-year-olds did not correlate strongly with 6-7-year-olds ($R = 0.11$) or adults ($R = 0.03$). In contrast, 6-7-year-olds exhibited similar distances to adults ($R = 0.33$).

To obtain an additional measure of which features participants deemed relevant, we coded the choices in the familiarization phase as matching in color or shape. Consistent with the 2D spatial organization, we found an effect of age on the number of shape matches, $F(2,87) = 8.48, p < .0001$. Post-hoc comparisons using a Tukey test indicated that adults selected significantly more shape matches than 4-5-year-olds (M_{adults} = 3.33, $M_{\text{4-5}}$ = 2.3, $p = .001$) and 6-7-year-olds ($M_{\text{6-7}} = 2.5, p = .005$). In contrast, we did not find an ef-

[5]The preregistration is accessible at https://osf.io/c85vd

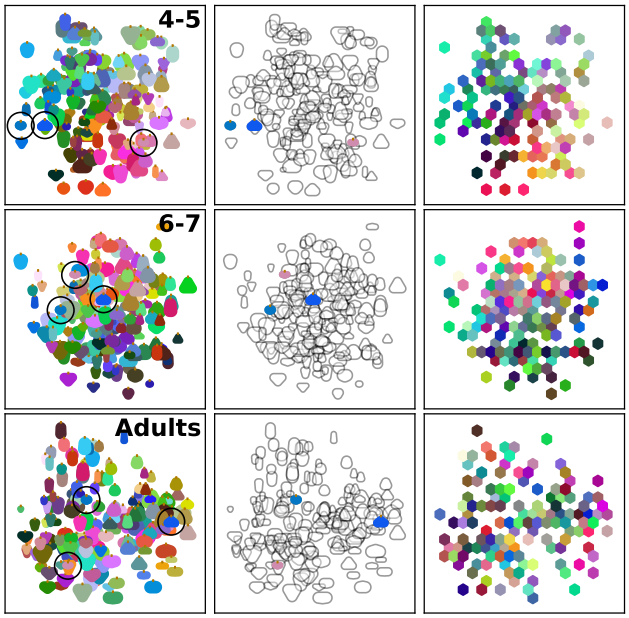

Figure 5: The 2D fruit spaces uncovered by our method (left column). To contrast the placement of fruits in the three spaces, we selected a set of three fruits, matching in color or shape, and highlighted their location. For 4-5-year-olds, fruits matching in color are placed nearby, whereas those matching in shape are on opposite sides. In contrast, for adults, the fruits matching in color are far apart. To facilitate inspection of the resulting spaces, we also display the shapes of the fruits (central column) and binned the space hexagonally, showing the most representative color in each bin (right column, following the procedure outlined in Figure 3).

fect of age on color matches, $F(2,87) = 2.75, p = .069$. Both adults ($M = 2.5$) and children produced similar numbers of color matches ($M_{4\text{-}5} = 3.0, M_{6\text{-}7} = 3.1$).

Explanations To examine whether the spaces were consistent with the participants' explanations, the two first authors transcribed the responses. Consistent with the spaces uncovered by our method, adults predominantly named shape (59 out of 115) and color (34) as the grouping feature. 6-7-year-olds gave color (33 out of 91) and shape (30) at comparable rates. Finally, 4-5-year-olds preferred color (36 out of 93) over shape (15), see Figure 6.

General Discussion

Our work develops DrLIM into a widely applicable computational and experimental paradigm, showing that the method can reconstruct meaningful psychological spaces in short experiments in which participants do not receive the full set of materials. In simulations, we found that the model could reconstruct the agents' categorization schemes, even for heterogeneous agent populations. Furthermore, we showed that we could uncover color representations for languages in the

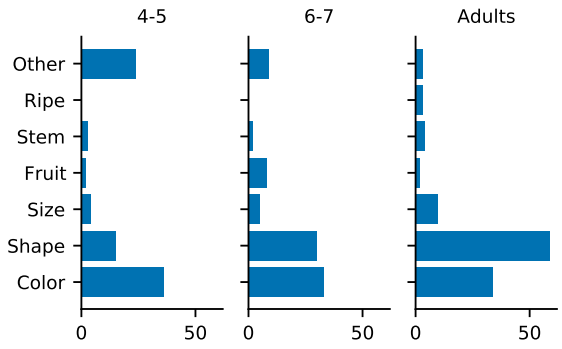

Figure 6: Adults mostly named shape followed by color when asked why they grouped the fruits. Instead, 6-7-year-olds named color and shape, and 4-5-year-olds preferred color.

WCS, even when we trained on datasets without full overlap.

We then performed the first application of DrLIM to a developmental setting, successfully recovering psychological spaces across multiple age groups, including children as young as four years old. We found that the recovered spaces exhibited age-dependent biases for how fruits are represented. These results were consistent with secondary measures, such as the sequence of choices in the familiarization task and participants' explanations. Our approach is a promising experimental paradigm for developmental studies of category representation, and our method can uncover similarity judgments in short and straightforward experiments.

Outlook and Future Studies

Future work should examine if deep metric learning can be generalized to uncover individual differences within age groups, for example extending DrLIM to weigh dimensions in the psychological spaces for each participant, similar to recent extensions to MDS approaches (Okada & Lee, 2016).

More generally, we see our method as a way to bridge a plurality of experimental results. Since we learn an implicit similarity *function*, stimuli that were not used to construct the function can be projected into a shared space, for example allowing results from different experiments to be projected into a shared psychological space. Results obtained through similarity judgments, spatial sorting, or category generalizations could be contrasted in a shared space, allowing broad comparisons between experimental paradigms. This aspect of the method offers the prospect of developing general *cognitive* embeddings, much like recent *universal* language representations (Cer et al., 2018).

Acknowledgements

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada [funding reference number 2016-05552]. We thank Kay Otsubo, Madeline Pelgrim, and Matthew Tatur for their help in developing and running the experiment. We also would like to thank the Royal Ontario Museum for providing their testing spaces. Finally, we would like to thank all members of the CoCoDevLab for helpful feedback and discussion.

References

Borg, I., & Groenen, P. J. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Science.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... others (2018). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 169–174).

De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of statistical software*, *31*, 1–30.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 1735–1742).

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The world color survey*. CSLI Publications Stanford, CA.

Lee, M. D. (1997). The connectionist construction of psychological spaces. *Connection Science*, *9*(4), 323–352.

León-Villagrá, P., Otsubo, K., Lucas, C. G., & Buchsbaum, D. (2020). Uncovering Category Representations with Linked MCMC with people. In *Proceedings of 42nd Annual Meeting of the Cognitive Science Society.*

Luce, R. D. (1959). *Individual choice behavior.* John Wiley.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35–44.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, *22*(4), 463–475.

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2020). The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior research methods*, 1–23.

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

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

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

Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, *114*(21), 5545–5550.

Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., & MacLellan, C. J. (2016). Developmental changes in semantic knowledge organization. *Journal of experimental child psychology*, *146*, 202–222.

Verheyen, S., White, A., & Storms, G. (2020). A comparison of the Spatial Arrangement Method and the Total-Set Pairwise Rating Method for obtaining similarity data in the conceptual domain. *Multivariate Behavioral Research*, 1–28.

Watson, J., Beekhuizen, B., & Stevenson, S. (2019). Identifying the Evolutionary Progression of Color from Crosslinguistic Data. In *Proceedings of 41st Annual Meeting of the Cognitive Science Society* (pp. 3071–3077).