

The Effects of Feature Verbalizability on Category Learning

Bailey N. Brashears (bbrashea@uwo.ca)

Department of Psychology &
The Brain and Mind Institute
Western University
London, ON N6A 3K7, Canada

John Paul Minda (jpminda@uwo.ca)

Department of Psychology &
The Brain and Mind Institute
Western University
London, ON N6A 3K7, Canada

Abstract

This study intended to investigate the effects of varying factors on the use of verbal and implicit classification systems when learning novel categories in an interactive video game environment by measuring the effects of feature type (easy vs difficult to describe verbally). Verbal and implicit classification were operationalized by measuring rule-based and family resemblance strategy use respectively. This experiment found that participants presented with stimuli that were easy to describe verbally were more likely to use rule-based classification, while participants presented with stimuli that were difficult to describe verbally showed no preference for one form of classification. The results of this study open up a novel field of research within category learning, further exploring the effects of feature verbalizability.

Keywords: Category learning, COVIS theory, feature verbalizability

Category learning is a part of everyday life. People form, update, and use various categories to make classification decisions about a variety of things in daily life, from animals to food to vehicles. Most category learning and use takes place without an individual's conscious awareness and - in the average adult - with a high level of accuracy. However, there is a wide variety of factors that can influence how we determine which things belong to what category, and how our brains process the categories themselves.

Review of Literature

The COVIS (Competition between Verbal and Implicit Systems) theory of category learning posits that category learning is accomplished by two separate but competing systems: the verbal system, which deals with learning explicit category rules, and the implicit system, which involves learning more complex categories through the procedural learning of multiple exemplars (Ashby & Maddox, 2011). The verbal system operates using active hypothesis testing,

with individuals making, testing, and revising rules on a conscious level. This approach is effective when learning categories with a feature-based rule or series of rules determining category membership. However, individuals are capable of learning more complex categories through the use of the implicit system. The implicit system can operate outside of conscious awareness and uses dopamine-mediated learning in order to gradually acquire categories based on covariation of features, family resemblance, and other complex and subtle distinctions that can broadly be described as non-rule-described categories (Minda & Miles, 2010). Generally, the verbal category learning system can be considered to learn categories with rules that are easy to verbalize, while the implicit category learning system can be considered to learn categories with rules that are difficult to verbalize (Ashby & Maddox, 2011). This dual-systems theory has a basis in neurobiological findings as well as behavioral findings (Ashby & Casale, 2003; Ashby & Maddox, 2011).

According to COVIS, both systems learn categories simultaneously, but there is a bias towards the verbal system that results in implicit learning not being initially expressed, as the implicit process is slower. In general, an individual will attempt to consciously find a rule-based solution during categorization tasks, and implicit categorization will occur if conscious verbal categorization is unsuccessful (Ashby et al., 1998; Maddox & Ashby, 2004). When classifying ambiguous stimuli that could be classified according to either rule-based (verbal) or family resemblance (implicit) strategies, both European Americans and Asian Americans as well as East Asians showed significant preference for using rule-based categorizations strategies (Norenzayan et al., 2002). However, there has been recent work calling this bias into question; a study modelled on the work of Norenzayan et al failed to replicate the results and instead found a preference for family resemblance categorization strategies in their

sample of European American participants (Murphy, Bosch, & Kim, 2017).

While the effects of language are easiest to see in verbal category learning, language can be an important factor for both category learning routes. Research on the role of language in category learning has found that merely introducing novel words as labels can highlight the item categories (Balaban & Waxman, 1997). The inclusion of existing category labels in a categorization task can also affect the perception of individual category member features (Lupyan, 2008). One study found that embedding a simple two-dimensional stimulus - a Gabor patch with varying orientation and spacing of lines - within category-irrelevant features and labelling the stimuli holistically as a “flower” decreased reliance on individual features of the stimulus (Perry & Lupyan, 2014). Perry and Lupyan concluded that placing a stimulus in a richer visual context distracted participants from the normally visually salient feature of orientation.

One specific factor that can affect which category learning strategies are favored is the verbalizability of the item being categorized or of the item’s features. Feature verbalizability refers to whether the individual features of a stimulus are able to be described verbally. While the effect on the verbalizability of stimulus features specifically has not been explored in the current literature, the findings of previous research on language and category learning as well as the theory of COVIS allow us to formulate predictions for this study. As verbal category descriptions facilitate category learning and verbal route categorization relies on easy to verbalize rules, we hypothesize that when presented with stimuli with features that are difficult to describe verbally, individuals will be less likely to use rule-based categorization. Lupyan’s work was largely concerned with assigning verbal labels to categories to increase holistic classification, while we expect the emphasis on verbalizability of the individual features to include rule-based classification.

The purpose of this experiment was to investigate the effects of feature type (easy to describe verbally vs difficult to describe verbally) on category learning preference. Feature verbalizability is a possible factor in the selection of a category learning strategy that has not been fully addressed by the current literature. The only other work on feature verbalizability was conducted by Zettersten & Lupyan (2020) using colours and shapes that were easy or difficult to assign a verbal label; this study found that category learning rate was faster when the features were easy to describe verbally. However, this study did not address categorization strategy.

Our hypothesis was that individuals who learn categories with features that are easy to describe verbally will show a preference for rule-based categorization, while individuals who learn categories with features that are difficult to describe verbally may show a preference for family resemblance categorization. Our reasoning for this is that when individuals are learning stimuli with features that are easy to describe verbally, it may be easier to identify a single

feature rule for classification, since verbal descriptions are important for this kind of category learning. Furthermore, when individuals cannot easily describe a feature in simple verbal terms, this may bias them towards using an implicit category learning strategy instead of trying to find a verbally described rule.

Methods

Participants

Seventy undergraduate students at Western University participated in this experiment. Participants were compensated with course credit in an undergraduate psychology course. The 70 participants in Experiment 1 were drawn randomly from this subject pool. Two of the 70 participants (one in the easily-verbalizable condition, one in the not-easily verbalizable condition) were not included in the final sample as they failed to learn the training stimulus within our criterion.

Materials

Stimuli. The stimuli were images of fictional monsters constructed of five binary features. The first set consisted of features selected to be easy to describe verbally: spots (two vs. four), eyes (two vs. three), ear colour (teal vs. orange), tail shape (square vs. triangle), and back colour (red vs. green). The second set consisted of features selected to be difficult to describe verbally: spots (uneven stripes vs. uniform polka dots), eyes (narrow and vertical vs. wide and horizontal), ear shape (short and pointed vs. long and floppy), nose shape (long and pointed vs. short and blunt), and back shape (short bumpy “spines” vs. tall “sail”). Prototype items are shown in Table 1.

Table 1: Stimulus Prototypes.

	Group A	Group B
Easily-Verbalizable		
Not-Easily Verbalizable		

In order to verify that the novel stimuli intended to have feature sets that were easily and not-easily verbalizable were verifiably such, a norming study was conducted. The norming study was designed to collect descriptions of the features from one group of individuals and determine if another group of individuals could then identify the same feature based on those descriptions. This study consisted of two phases. In the first phase, 63 individuals participated in a Qualtrics study through the University of Western Ontario’s SONA system. In this study, participants were shown one of the four stimulus prototypes followed by one of the features

in isolation. They were then asked to describe the individual features in the provided text box. Then, for each feature, the two most common descriptors from the 63 responses were selected to be used in Phase Two.

In the second phase, 30 individuals participated in a Qualtrics study through Amazon's MTurk system. They were shown the top two descriptors from Phase One and asked, out of the two paired features, which was best fit by the descriptors. The accuracy from Phase Two was then calculated compared to the results from Phase One. A Paired Sample Student's t-test found that accuracy was significantly higher for the easily-verbalizable items ($M = 0.95$, $SD = 0.09$) than for the not-easily verbalizable items ($M = 0.63$, $SD = 0.14$), $t(29) = 11.55$, $p < .001$.

The participants in the second phase were better able to differentiate the feature pairs from descriptions generated in the first phase for the feature pairs in the easily-verbalizable category. This means that when describing features in the easily-verbalizable category, the generated descriptions were more distinct from each other when compared to the description of the paired feature. To illustrate this, the description for the spot feature pair in the easily-verbalizable category set included "two yellow hearts" and "four yellow hearts" with "yellow spots" included in each description, a clear and easy to describe distinction. However, for the spot feature pair in the not-easily verbalizable category set, the descriptions collected included "green spot pattern" and "beige markings" with "green spots" included in each description; in this case, any of the descriptions used could apply to either of the members of the feature pair, so accuracy was lower in matching the description to the correct stimulus. From these results we can conclude that the easily verbalizable stimulus features are objectively easier to verbally describe than the not-easily verbalizable stimulus features.

Table 2: Training Stimulus.

Category A					Category B					
1	2	3	4	5	1	2	3	4	5	
A1	0	0	0	0	0	B1	1	1	1	1
A2	0	1	0	0	0	B2	1	0	1	1
A3	0	0	1	0	0	B3	1	1	0	1
A4	0	0	0	1	0	B4	1	1	1	0
A5	0	0	0	0	1	B5	1	1	1	1

The binary notation for the category sets was taken from Minda and Ross (2004) and is shown in Table 2. Like the original study, each training category contained fifteen stimuli with five exemplars presented in three sizes. Small stimuli were presented at 75 x 51 pixels, medium stimuli were presented at 150 x 102 pixels, and large stimuli were presented at 225 x 153 pixels. The size of the stimulus was

not relevant for categorization; this was meant only to add more variation to the stimulus set. Each category set generated for the training phase included the prototype and four stimuli that varied by a single feature; one random feature would always be present on all stimuli in the category. This was designed such that the category could be learned either by a family resemblance (FR) or a criterial attribute (CA) strategy. A family resemblance strategy could be used due to the similar appearance of each stimulus to the prototype, while a criterial attribute strategy could be used due to the single feature which perfectly predicts category membership.

Table 3: Testing Stimulus.

1	2	3	4	5	1	2	3	4	5
Exception Items					Single Features				
T1					T11				
T2					T12				
T3					T13				
T4					T14				
T5					T15				
T6					T16				
T7					T17				
T8					T18				
T9					T19				
T10					T20				

In addition to the training stimuli, a group of test stimuli was also generated to be used in the study and is shown in Table 3. The test stimulus consisted of three subsets of ten items each; all test stimuli were presented at the medium size. The first subset - items TA1-5 and TB1-5 - were repetitions of the training stimulus used to gage the participant's accuracy. The second subset - items T1-10 - were exception items taken from the original Minda and Ross (2004) study; these items had conflicting category membership based on whether the participant used a FR or CA strategy and were used to differentiate these strategies. The final subset - items T11-20 - were each of the single features presented in isolation used to gage the participant's attention towards the individual features that comprised the two category prototypes.

Procedure

The experiment was conducted using a video game programmed in GameMaker Studio 2. All responses were recorded by the program and saved to a plain text document on the testing machine. Randomization of condition and CA feature was also handled by the program.

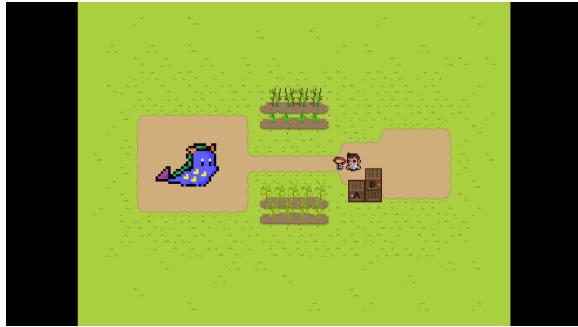


Figure 1: Screenshot of Gameplay (Easily-Verbalizable Condition).

Participants were randomly assigned to one of two stimulus sets: easily-verbalizable features and not-easily-verbalizable features. This affected only the visual appearance of the stimuli used and the experiment proceeded identically regardless of condition. Participants in both conditions were introduced to the game “Monster Farm” where they would play as a farmer and their job was to figure out which group each of the “monsters” living on a farm belonged to. They were also told that some features or aspects of the monsters might help them determine the correct group, either group A or group B. On each trial, they would need to select a collar for the monster with the group letter using either the A or B key. If the classification was correct, they were shown a check mark; if the classification was incorrect, they were shown an X. This feedback was displayed with the stimulus still visible until the next trial began automatically after a few seconds.

The training phase would continue until a subject had completed at least four blocks (60 trials) and completed at least twelve trials in a row correctly. If a subject did not complete this criterion within 400 trials or one hour, the game would end without the testing phase and the participant’s data would not be further analyzed. After reaching the criterion, both groups entered the testing phase. In this phase, subjects were shown each of the test stimuli in Table 2 and sorted the monster, as in the training phase, into the group they thought it belonged to. In this phase, there was no feedback on their decisions and each stimulus was shown once in a randomized order.

Results

Learning Rate Analysis

The first analysis focused on the number of trials it took a participant to reach criterion in each condition. This was defined as making a correct categorization decision on 12 trials in a row, after a minimum of 60 trials. A Welch’s Two Sample t-test found that the learning rate for the easily verbalizable condition ($M = 82.68, SD = 50.23$) and the not-easily verbalizable condition ($M = 89.50, SD = 43.2$) did not differ significantly from one another, $t(64.56) = -0.60, p = .550$. These results are visualized in Figure 2. Note that the minimum number of trials a participant could complete the training phase in was 60.

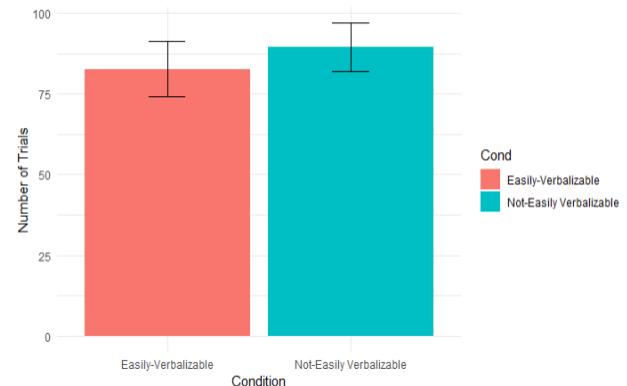


Figure 2: Mean Number of Trials to Criterion by Condition.

Accuracy Analysis

The second analysis focused on participants’ accuracy on the training items as presented during the testing phase in each condition. All ten training items were presented in a randomized order with the other testing stimuli and presented only in the medium size. A Welch’s Two Sample t-test found that accuracy was significantly higher in the easily verbalizable condition ($M = 0.93, SD = 0.12$) than in the not-easily verbalizable condition ($M = 0.80, SD = 0.18$), $t(58.47) = 3.50, p = .001$. These results are visualized in Figure 3.

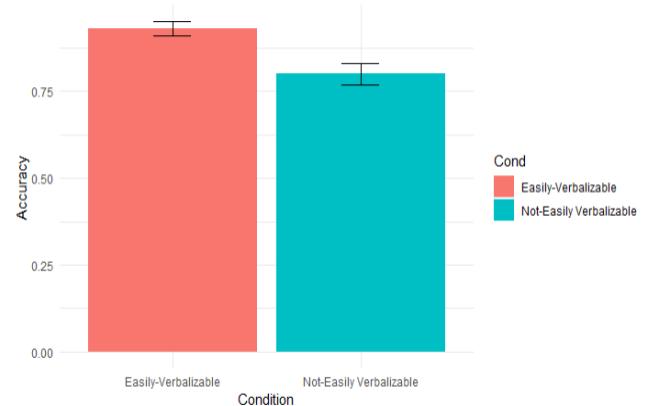


Figure 3: Mean Accuracy in the Testing Phase by Condition.

Categorization Strategy Modelling Analysis

The final analysis of this data set focused on the categorization strategies used by the participants. Modelling was conducted using participants’ responses to the ten transfer items presented in the testing phase, randomized with the other testing stimuli and presented only in the medium

size. The exception items were designed such that a participant using a FR rule and a participant using a CA rule would categorize the stimulus in opposite categories. Therefore, based on the participants' responses, we can determine what strategy each participant was most likely relying on.

These responses were then compared to the results predicted by two model responses and each participant was given a score out of one indicating how well their responses matched the predicted results of each model. For each matching item response, the individual was given a 1. For each response that didn't match, the individual was given a 0. These scores were added together and divided by the number of items (10) with a possible score of 1 indicating a perfect fit with the model. If a given strategy model correctly predicted a participant's classification for all ten stimuli, the model fit with a score of 10. If the model did not predict all the classifications, the score would be less than 1.0.

For the criterial attribute (CA) strategy model, individuals' responses were compared to the expected responses if one was only attending to the criterial attribute. This would entail responding with A when the criteria attribute matched group A but three or four of the remaining features were consistent with group B. This indicates that the individual was ignoring most or all of the other features of the stimulus and only using the criterial feature to make their category decisions. The family resemblance strategy model compared the responses to the opposite pattern, one that was expected if one was attending to overall family resemblance regardless of the criterial attribute. This would entail responding B when three or four of the features were consistent with group B but the criterial attribute matched group A. This would indicate that the individual is only looking at the overall family resemblance to make their category decisions and is ignoring the criterial feature.

A mixed 2x2 ANOVA was conducted with model type as a within-subjects independent variable (CA vs. FR), stimulus set as a between-subjects independent variable (easily-verbalizable vs. not-easily verbalizable), and model fit as the dependent variable. There was no main effect of condition, $F(1, 66) < .01, p > .999$. There was a significant main effect of model type, indicating that the CA model ($M = 0.67, SD = 0.35$) was overall a better fit for participants than the FR model ($M = 0.33, SD = 0.35$) regardless of condition, $F(1, 66) = 19.12, p < .001$. There was also a significant interaction effect, $F(1, 66) = 14.78, p < .001$. These results are visualized in Figure 4. Due to the binary nature of the responses and the design of the exception items, the FR and CA model fits are inverse of one another as these model fits are mutually exclusive.

Discussion

We found in this experiment that the easily-verbalizable stimulus set and not-easily verbalizable stimulus set were both equally easy to learn, as seen in the learning rate analysis, but participants were better at retaining the category

membership of the easily-verbalizable stimulus set into the training phase than the not-easily verbalizable stimulus set, as seen in the accuracy analysis. The similar learning rate is not consistent with the results of Zettersten & Lupyan (2020); however, this may be due to a floor effect since there was a minimum of 60 trials in the learning phase.

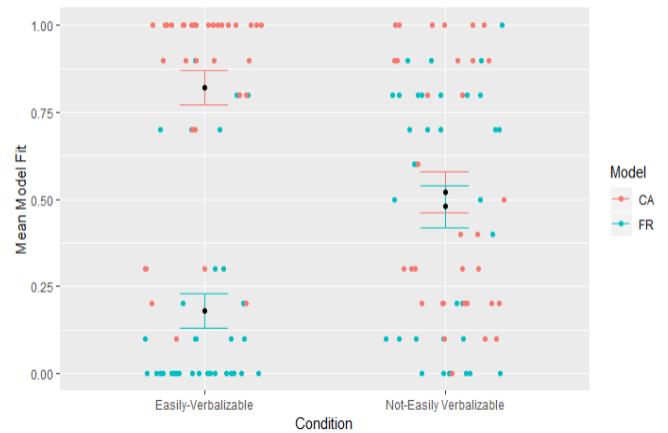


Figure 4: Model Fit (CA and FR) by Condition.

Additionally, our modelling analysis found that individuals in the easily-verbalizable condition were more likely to use CA strategies, while individuals in the not-easily verbalizable condition did not show a strong preference for either strategy. This result is consistent with our hypothesis that a stimulus with not-easily verbalizable features would facilitate implicit category learning strategies; individuals in the not-easily verbalizable condition were more likely to use family resemblance strategies than those in the easily-verbalizable condition even if there was not a strong preference for them.

This preference for rule-based learning is further seen by the finding that individuals were more likely to retain the category membership knowledge into the testing phase in the easily-verbalizable condition, as individuals were able to use the previously discovered rules to easily sort the stimulus again without having to rely on knowledge of the prototype.

Consistent with the findings of Perry and Lupyan (2014), we expected labelling the stimulus as "monsters" and placing them in a rich visual context to cause participants to view the stimuli more holistically, rather than attending only to the discrete features they contained. Similar to Perry and Lupyan's (2014) flowers, our stimuli had a variety of visual features irrelevant to their category membership. Furthermore, we attempted to extend this concept by varying the verbalizability of the individual features. From these results, we can conclude that feature verbalizability does play a role in category learning strategy selection and features that are easy to describe verbally are conducive to rule-based strategy use and verbal route category learning.

This experiment demonstrated an interesting and novel effect of feature verbalizability on preference for category learning strategies. Participants showed a strong and significant preference for rule-based categorization when classifying the stimulus set with easily-verbalizable features compared to the stimulus set with not-easily verbalizable features. There was no strong preference for either rule-based or family resemblance strategies when classifying stimulus with not-easily verbalizable features. While the role of verbal labels and names has been of interest in categorization research, the effects of the verbalizability of the individual features themselves is unexplored in the existing literature and this study demonstrates a need for further work.

While our findings are not definitive as to the cause, there are a few possible explanations. The first is that the easily-verbalizable features facilitated rule-based learning by encouraging the use of verbal route categorization. When individuals are presented with features that they can easily describe it is easier for them to formulate and test a hypothesis (e.g., two spots = group A, four spots = group B). If an individual is presented with features that are more difficult to describe verbally it is more difficult to formulate a straightforward hypothesis; two different spot patterns that are difficult to describe distinctly can be harder to group into verbally described rules, since they are harder to describe. Another possibility is that the easily-verbalizable features appeared more distinct from one another compared to the not-easily verbalizable features. The easily-verbalizable features were designed with simple colour and number contrasts in order to be easy to describe verbally. However, this may have caused a larger visual contrast in this group compared to the not-easily verbalizable group, which had consistent colours and numbers, which means the results may have been due to the features appearing more distinct, so individuals attend to the individual features more for this stimulus group. Furthermore, the not-easily verbalizable stimulus set might have been viewed more holistically due to this decreased visual contrast, leading to more use of family resemblance categorization.

In the future, we would like to conduct additional research to further investigate these findings and identify the underlying cause. One possibility is to examine the role of feature verbalizability or visual distinctiveness. A study like this could replicate the current experiment with two new stimulus sets. These stimulus sets would be balanced to control for their visual distinctiveness, possibly by using grey-scale stimuli or by adding colour contrasts to the not-easily verbalizable stimulus, with varying hues of the same colour to create a contrast that is visually distinctive but relatively difficult to describe verbally.

Overall, this study had novel findings that open a new possible line of category learning research and demonstrated that a previous category learning study has results that may not be replicable. Additionally, we successfully created a new platform for conducting classification experiments that can be developed and adapted for future studies as we continue to

better understand the effects of this video game environment on category learning. While our work on indirect feedback and diverted attention was inconclusive, we demonstrated that language is an important factor in category learning strategy use that should be explored further.

References

- Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). "A Neuropsychological Theory of Multiple Systems in Category Learning." *Psychological Review* 105 (3): 442–81.
- Ashby, F. G., & Casale, M. B. (2003). "The Cognitive Neuroscience of Implicit Category Learning." *Advances in Consciousness Research* 48: 109–42.
- Ashby, F. G., & Maddox, W. T. (2011). "Human Category Learning 2.0." *Annals of the New York Academy of Sciences* 1224 (April): 147–61.
- Balaban, M. T., & Waxman, S. R. (1997). "Do Words Facilitate Object Categorization in 9-Month-Old Infants?" *Journal of Experimental Child Psychology* 64 (1): 3–26.
- Lupyan, G. (2008). "From Chair to 'Chair': A Representational Shift Account of Object Labeling Effects on Memory." *Journal of Experimental Psychology. General* 137 (2): 348.
- Maddox, W. T., & Ashby, F. G. (2004). "Dissociating Explicit and Procedural-Learning Based Systems of Perceptual Category Learning." *Behavioural Processes* 66 (3): 309–32.
- Minda, J. P., & Miles, S. J. (2010). "Chapter 3 - The Influence of Verbal and Nonverbal Processing on Category Learning." In *Psychology of Learning and Motivation*, 52:117–62. Academic Press.
- Minda, J. P., & Ross, B. H. (2004). "Learning Categories by Making Predictions: An Investigation of Indirect Category Learning." *Memory & Cognition* 32 (8): 1355–68.
- Murphy, G. L., Bosch, D. A., & Kim, S. (2017). "Do Americans Have a Preference for Rule-Based Classification?" *Cognitive Science* 41 (8): 2026–52.
- Norenzayan, A., Smith, E. E., Kim, B. J. & Nisbett, R. A. (2002). "Cultural Preferences for Formal versus Intuitive Reasoning." *Cognitive Science* 26 (5): 653–84.
- Perry, L. K., & Lupyan, G. (2014). "The Role of Language in Multi-Dimensional Categorization: Evidence from Transcranial Direct Current Stimulation and Exposure to Verbal Labels." *Brain and Language* 135 (August): 66–72.
- Smith, E. E. (2008). "The Case for Implicit Category Learning." *Cognitive, Affective & Behavioral Neuroscience* 8 (1): 3–16.
- Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, 196, 104