A Priming Model of Category-based Feature Inference

Laura M. Hiatt (laura.hiatt@nrl.navy.mil)
US Naval Research Laboratory
Washington, DC 20375 USA

Abstract

Categorization has a large impact on how people perceive the world, especially when used to make inferences about uncertain features of new objects. While making these inferences, people tend to draw information from only one possible categorization of a new object; in addition, people are sensitive to pre-existing correlations between features. Here, we explain these trends of feature inference using a priming-based cognitive process model, and show that our model is distinguished in that it can explain not only these two main trends, but also cases where people seem to reverse the first trend and base inferences on information from multiple categories.

Keywords: categorization; priming; spreading activation; inductive inference; cognitive models

Introduction

Categorization is a fundamental tool in human cognition. One of its main functions is to allow people to more easily understand the world by making inferences about new objects based on existing knowledge that they already have. If one sees a furry animal coming towards it and categorizes it a loose dog, then it would be natural to further infer that the animal is probably friendly.

Systematic research into how these inferences are made has shown two major trends in performance (Nosofsky, 2015; Murphy & Ross, 1994, 2007; Griffiths, Hayes, & Newell, 2012). First, people seem to base inferences on a single identified category for an object, even if the object's categorization is uncertain (called the *single-category view*). So, for example, people would typically infer that the dog is friendly without considering that it might be a fox, which should be avoided. Second, people are sensitive to correlations between features, and are more likely to infer features that are strongly associated with the observed features of the new object. For example, people would be further biased towards inferring the dog is friendly if it were wagging its tail.

While there is a large body of research that supports these two trends, here we consider a series of experiments performed by Murphy and Ross (1994) that comprehensively considered several variants and extensions of the basic inference paradigm. The authors, however, admit that their overall results challenge many of the formal models of categorization and inference (Murphy & Ross, 1994, 2007), with none fully explaining the results. Recently, Nosofsky (2015) developed a exemplar model of feature inference that does qualitatively capture their results. Notably, however, Nosofsky's (2015) analysis does not discuss an important caveat of the first trend: that responses seem to shift towards a *multi-category view*, where more than one possible category is considered when making the inference, if participants do not explicitly identify the category before making the feature inference (Murphy & Ross, 1994; Griffiths et al., 2012).

We present here a priming-based process model of inductive feature inference that explains these two main results, including this caveat. Situated in the cognitive architecture ACT-R/E (Trafton et al., 2013), a critical aspect of our model is that its inferences are based not only on what stimuli have been seen, but also on what the model is currently thinking about (i.e., what is in its working memory). We show our model's ability to account for feature inferences in four main experiments that are particularly indicative of the trends of feature inference: Experiments 1, 5, 6, and 8 from Murphy and Ross (1994).

Experiments

In the four experiments we consider from Murphy and Ross (1994), participants were shown category structures with differently shaded geometric objects, grouped together and labeled with the category they represent (e.g., Figure 1). Participants were told that the categories represented different children who drew the objects, and that the objects were illustrative of a larger set of drawings by each child. Then, the experimenter told participants about a new drawing, but only shared one feature of it, such as a triangle; this feature, the *query feature*, was typically chosen to be ambiguous in which child drew it. Participants were then asked what they thought the other feature of the new drawing was (such as the triangle's color). Additionally, in some experiments, participants were asked to categorize the drawing (i.e., say which child drew it) before they inferred the second feature. The most likely category for each query is called the *target category*.

Experiment 1 focused on whether inferences are made using information from single, or multiple, categories. The categories are shown in Figure 1[1]. This experiment had two conditions. In the *increasing condition*, the query feature was a triangle. The target category for a triangle is Bob, since Bob has the most triangles. The *target-category feature*, or the feature that would be selected by primarily considering the target category using a single-category view, is black. This condition is called increasing because there is additional evidence outside the target category that the triangle would be black, since Sam and John also sometimes draw triangles, and they also sometimes draw black objects.

Now, consider a new drawing that is a square. Here, the target category is John, and the target-category feature is white. In this condition, the *neutral condition*, there is no evidence outside of the target category that the square would be white;

[1]While other variations of this category structure were used to counterbalance features and category locations, they preserved this same main category structure and so we discuss the experiment in terms of this one. We do this for the other experiments, as well.

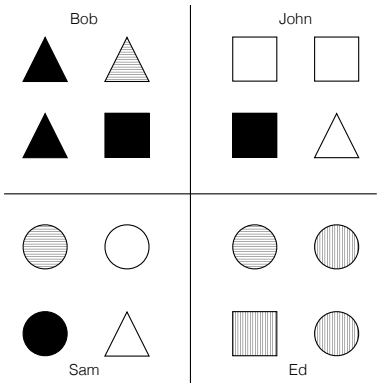

Figure 1: Category structure for Experiment 1. In the increasing condition, the query feature is a triangle, the target category is Bob, and the target-category feature is black. For the neutral condition, the query feature is a square, the target category is John and the target-category feature is white. Adapted from Figure 1 of (Murphy & Ross, 1994).

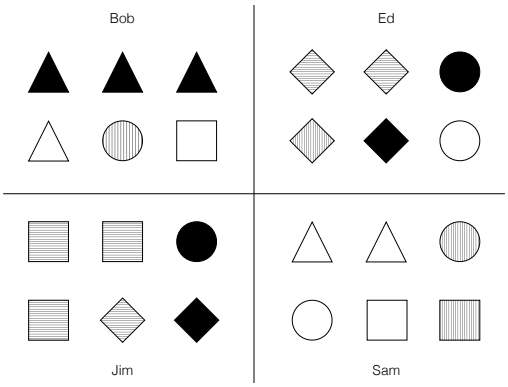

Figure 2: Category structure for Experiments 5 and 6. Here, the query feature is a triangle, the target category is Bob, and the target-category feature is black. Adapted from Figure 3 of (Murphy & Ross, 1994).

no other child draws both squares and white objects.

Almost all of the 29 participants selected the target category and target-category feature for both the increasing and neutral conditions, and so ceiling effects prevented them from being statistically compared. Participants also, however, provided a probability estimate of their certainty in their response. These probability judgments did not have a ceiling effect, yet provide no evidence of a difference between the two conditions: the average certainty for each condition was 53%. This parity supports the single-category view of feature inference by suggesting that, despite the additional evidence for the target-category feature present in the increasing condition, participants only took the target category into account when making their inference.

Experiments 5 and 6 used a category structure in which the single-category view and multiple-category view suggest different patterns of feature inferences (Figure 2). Further, they considered how the initial step of identifying the target category may affect participants' use of single vs. multiple categories in their inference. Here, the query feature is a triangle, and the target category is Bob, since he drew more triangles than the other children. The single-category view suggests black as the inferred feature; black is thus considered the target-category feature. A multiple-category view, however, suggests that black and white are equally likely.

The results of the experiments support both these views, depending on whether participants were asked to make the initial categorization step. In Experiment 5, where participants did not initially categorize the drawings, 58% of the 32 participants chose the target-category feature, black, with the majority of remaining responses as white. This difference was not significant, supporting the multiple-category view. In Experiment 6, however, where participants did categorize the drawings before predicting the other feature, 82% of the 36 participants responded with the target-category feature. Additionally, 88% of the participants that categorized the drawing into the target category responded with the target-category feature. This supports the single-category view and suggests that participants were biased by the target category when they identified it before making their inference.

Experiment 8 focused on exploring how feature correlations may affect predictions. Here, the query features in two conditions were explicitly controlled to have different degrees of correlation with the target-category features. All participants were asked to assign the drawing to a category before responding to the feature queries. Figure 3 shows an example category structure. In the *correlated condition*, the query and target-category features were perfectly correlated: the query feature was a circle, the target category was "D" and the target-category feature was "vertically striped." In the *uncorrelated condition*, the features are only weakly correlated, with a query feature of triangle, a target category of "C" and a target-category feature of white.

The results show that 95% of the 26 participants selected the target category across both conditions. More importantly, more participants selected the target-category feature for the correlated condition (94%) than for the uncorrelated condition (90%). This suggests that people are biased towards correlated features when they make inferences.

Model

We developed a priming-based process model of feature inference given uncertain categorizations, situated within a computational cognitive architecture, ACT-R/E, that allows us to model the processes people undergo as they perform tasks. In this architecture, concepts that are thought about at the same time become associated in memory, and then can prime one another; by using ACT-R/E, we are able to develop a priming-based account of feature inference that is supported by the underlying principles of this existing, well-studied theory of cognition. Here, we first describe the general principles

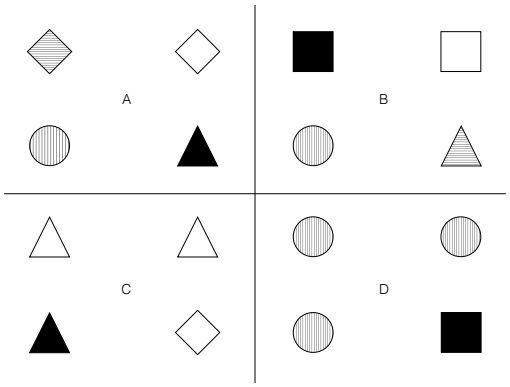

Figure 3: Category structure for Experiment 8. For the correlated condition, the query feature is a circle, the target category is D and the target-category feature is vertically-striped. For the uncorrelated condition, the query feature is a triangle, the target category is C and the target-category feature is white. Adapted from Figure 5 of (Murphy & Ross, 1994).

of our model. Then, we give further details of ACT-R/E, and discuss how our model's principles interact with the architecture to make specific predictions about feature inference.

The process model has two phases corresponding to the two phases of the experiment: an initial phase where the model looks at, encodes, and stores the categories and objects in memory; and an inference phase where the model makes the category and feature predictions. During the initial phase, each of the objects becomes associated with its underlying features; both the features and objects, in turn, also become associated with their corresponding category. These associations mean that the concepts prime one another when the model is thinking of them.

Then, during the inference phase, to predict the category of a new object, the model selects the category with the most priming, including priming from the query feature. Consequently, the model's category response is heavily influenced by the presence of the query feature in the category. To perform the feature prediction, the model selects the object in memory that has the most priming, including priming from the query feature and, when applicable, the selected category. The second feature of that object is then considered to be the inferred feature. This means that the predicted feature is heavily influenced by both the correlation between the two features, and the prevalence of that feature within the identified category (when the category is identified).

Model Architecture

The model was developed within the cognitive architecture ACT-R/E (Trafton et al., 2013), an embodied version of the ACT-R cognitive architecture (Anderson, 2007). At a high level, ACT-R/E is an integrated, production-based system, and models in ACT-R/E capture the core cognitive processes that people go through as they undergo tasks. At its core are the contents of its working memory; working memory indicates, for example, what the model is looking at, what it is thinking, and its current goal. At any given time, there is a set of *productions* (if-then rules) that may fire because their preconditions are satisfied by the current contents of working memory. From this set, the production with the highest predicted usefulness is selected to fire. The fired production can either change the model's internal state (e.g., by adding something to working memory) or its physical one (e.g., by pressing a key on a keyboard). In our discussion, we abstract over these productions and instead describe processes at a higher level (i.e., we say that we look at an object, instead of discussing the 3-4 productions that must fire to achieve that).

Working memory is represented as a set of limited-capacity buffers that can contain thoughts or memories. In addition to the symbolic information (i.e., factual information) represented as part of these memories, memories have activation values that represent their relevance to the current situation, and guide what memories are retrieved from long-term memory and added to working memory at any given time. Activation has three components, activation strengthening, spreading activation, and activation noise, that together have shown to be an excellent predictor of human declarative memory (Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson, 1983; Schneider & Anderson, 2011; Thomson, Harrison, Trafton, & Hiatt, 2017). Noise is a random component that models the noise of the human brain; since its presence would not affect our results, we ignore noise in the rest of this paper. Activation strengthening is learned over time and is a function of the frequency and recency with which the memory has been in working memory in the past. The predominant role of activation strengthening in this experiment relates to ordering effects, which the experimental stimuli's counterbalancing averages out. Therefore, we primarily focus the rest of our discussion of activation on its third component: spreading activation, or priming.

Priming is a short-term activation that sources from working memory, distributing activation along associations between the contents of working memory and other memories. Memories become associated when they are in working memory at the same time. Once established, an association from memory j to memory i has a strength value that affects the degree to which j primes i, and intuitively reflects the probability that memory i is relevant while thinking of memory j. This allows spreading activation to capture correspondences between memories that typically co-occur, as well as memories that are semantically related (such as an object and its color and shape). Association strengths are calculated in a Bayesian-like way, and are a non-standard adaptation of ACT-R's Bayesian-based priming mechanisms. We use this adaptation to account for the large numbers of associations and objects needed by the experiments we consider here, which ACT-R's original formulation is unable to do, as well as to capitalize upon its theory that priming stems from working memory; see Hiatt and Trafton (2016) for more information on our priming mechanisms.

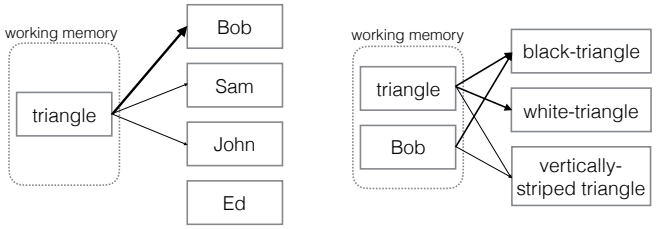

Figure 4: Priming for Experiment 1, increasing condition, for the query feature's categorization (left) and the inferred feature (right). Thicker arrows indicate more priming; thinner indicate less. The cumulative priming means that Bob will be selected as the category and black will be the inferred feature.

ACT-R/E models interact with the world using ACT-R/E's built-in functionality. Models can view visual items on a simulated monitor, and can act on the world by pushing keys on a simulated keyboard and clicking a simulated mouse. ACT-R/E models are also inherently tied to physical embodiment (i.e., executing models on a robot), but we do not use that functionality in this paper.

Model Details

Our model for feature inference starts out with only the task knowledge and productions necessary to complete the tasks. It also assumes prior exposure to the category names, since they are names participants would have encountered frequently in their daily lives (i.e., "A", "John", etc.). There are no initial associations; are all learned during the experiment. The model "looks at" the stimuli as the participants did via its simulated monitor.

During the initial experiment phase when the model is looking at the categories and objects, it first finds a category to look at, encodes it and adds it to working memory. While continuing to think of the category, the model then looks at, encodes and adds to working memory each of the objects in that category, while making note of their color and shape. Consequently, as it looks at each object: the object (i.e., "black-triangle") becomes associated with the category (i.e., "Bob"); the object's features (i.e., "black" and "triangle") become associated with both the object and the category; and the features of the object become associated with each other. When it has finished looking at all objects of a category, it repeats this process with the other categories until it has looked at all of the categories and objects on the screen.

During the inference phase, the model first adds the query feature to working memory as part of the process of interpreting the query. When asked to infer the category of an object, the model retrieves the category from memory with the highest activation, including both activation strengthening and spreading activation (i.e., priming), responds with the retrieved category, and leaves the category in working memory. For example, Figure 4, left side, shows the priming when selecting the category for Experiment 1's increasing condition. Then, when asked to infer the object's missing feature, the model retrieves an object while both the retrieved category (when applicable) and the query feature are in working memory. Again, the object with the highest activation, both activation strengthening and priming, is retrieved; Figure 4, right side, shows this for Experiment 1's increasing condition. The second feature of the retrieved object is given as the response to the query.

Model Results

In the original experiments, several versions of the basic category structure were created to counterbalance features and category locations. We varied our category structures accordingly, then used our model to simulate data from 500 participants per experiment to allow our results to better converge on the model's true predictions; our reported results are the proportion of the 500 model runs that responded with the target category, target-category feature, etc., for each query.

The model had the same parameters for each experiment. The activation strengthening decay parameter was 0.45 instead of its default of 0.5. The associative learning rate was 4.8, representing a moderate rate of learning. There is no real default value for this parameter. All other parameters were set to their default values.

The main experiment and model results are shown in Table 1. For Experiment 1, the model exhibited perfect performance, always selecting the target category, and always selecting the target-category feature for both the increasing and neutral conditions. This is comparable to the experimental results, where almost all participants also selected the target category and target-category features.

In this experiment, however, despite almost all participants selecting the target-category feature, participants' probability judgments of their responses were not as certain, with an average judgment for each condition of 53% for both the increasing and neutral conditions. While we have no a priori way of extracting probability judgments from the modeling framework we utilize, our model does informally support these results. This is because, from our model's point of view, these conditions' structures are the same. Both include two objects with the query feature and target-category feature in the target category; one object with just the query feature in the target category; and two objects with just the query feature outside of the target category. Thus, in both conditions, while the black-triangle object (or white-square object) is the highest activated object, it only receives about half of the total priming, suggesting a probability judgment of 50%.

For Experiment 5, the model selected the target-category feature 50% of the time, which moderately reflects the experiment's results. In Experiment 6, the model very strongly matched the experimental data, selecting the target-category feature 80% of the time, as compared to the experiment's 82%. Additionally, 89% of the model runs that categorized the drawing into the target category responded with the target-category feature, compared to the experiment's 88%.

For Experiment 8, 95% of model runs selected the target

Table 1: Model Results

Experiment	Condition/Participant Group	Measurement	Exp. Data	Model
Exp. 1	increasing	probability judgments	53%	50%
	neutral	probability judgments	53%	50%
Exp. 5	all participants	target-cat. feature	58%	50%
Exp. 6	all participants	target-cat. feature	82%	80%
	target cat. correct only	target-cat. feature	88%	89%
Exp. 8	all participants	target category	95%	95%
	correlated	target-cat. feature	94%	100%
	uncorrelated	target-cat. feature	90%	91%

category, the same as in the experiment. All of the model runs selected the target-category feature for the correlated condition, and 91% selected the target-category feature for the uncorrelated condition. Again, this strongly corresponds to the experimental results, where there was a significant difference between the two conditions, with 90% of participants selecting the target-category feature in the uncorrelated condition vs. 94% for the correlated condition.

Model Discussion

Recall the two main trends in research on feature inference for uncertain categorizations that are illustrated by the four experiments we consider here. First, people are biased towards the single-category view when making feature inferences; the bias seems to be modulated, however, when they do not categorize the object first. And second, people's inferences are also sensitive to correlations between features, selecting correlated features more often than non-correlated.

The model explains both of these trends via priming between the features, objects and categories. It explains the first trend, and its caveat, because its predictions are based on the sources of priming in working memory, and as such are not inherently based on the consideration of single- or multiple-categories. When making a feature prediction, the model always has the query feature in memory, which primes objects that are associated with it. This serves to provide suggestions compatible with the multiple-category view of what the predicted feature should be. For example, in Experiment 5, where triangle is the query feature and there is no categorization step, triangle equally primes black-triangle and white-triangle, because there are equal numbers of them. This leads to a roughly an equal likelihood (50%) of the predicted feature being black or white. While this underestimates the 58% response rate of the experimental data, given the lack of statistical significance in this experiment, we are comfortable concluding that our model explains this trend.

In conditions where participants categorize the feature before making their prediction, priming stems not only from the query feature but also from the category, which provides suggestions compatible with the single-category view of what the inferred feature should be. In Experiment 6, identical to Experiment 5 but with an added categorization step, when shown a triangle, the model generally selects Bob as the category (i.e., Figure 4). Bob then strongly primes black-triangle, since it has three of them, and weakly primes white-triangle, since it has only one of them. Combining this category priming with the priming from the query feature, black-triangle overall receives more. Again, this matches the data, where 82% of participants overall selected black as the inferred feature, and 88% of participants who identified Bob as the target category selected black as the inferred feature. Overall, then, the model's use of priming in memory allows the model to capture conditions both where participants seem to be biased towards the single-category view, and where they do not – a major contribution of the model.

The model also explains the second main trend of feature prediction, where participants are sensitive to correlations between features. There are two reasons for this. The first is that correlated objects, on average, have slightly higher activation strengthening, since they will be more familiar to participants than objects with less common feature pairings. The second reason is that correlated objects will receive much higher levels of priming from their underlying features because that priming is, in a sense, undiluted by other options. For example, in Experiment 8, where the correlated query feature is circle, the only object primed by circle is vertically-striped-circle. The target category, D, also spreads a high amount of activation to vertically-striped-circle, since there are three of them in that category, further underscoring the correlated feature as the answer. In contrast, for the uncorrelated query, both sources of priming (the query feature triangle, and the target category C) prime white-triangle in addition to strongly priming the target black-triangle. Thus, the model suggests that for the correlated condition, the target-category feature should almost exclusively be selected, whereas in the uncorrelated condition, the target-category feature should just mostly be selected. These explanations match the data, where the target-category feature was selected for 94% of correlated, but only 90% of uncorrelated, structures.

General Discussion

The authors of the experiments that we model here were ultimately interested in characterizing people's inference behaviors across different manipulations of categories and fea-

tures (Murphy & Ross, 1994). Recently, as we mentioned, Nosofsky (2015) proposed an exemplar model that qualitatively accounts for the majority of the results. The model is based on an equation that calculates the similarity between feature/category pairs using two parameters: the salience of the feature, and the salience of the category. The probability of inferring the target-category feature is then found by summing the similarity of the query feature/category pair to all displayed feature/category pairs with the target-category feature, and dividing by the summed similarity of the query feature/category pair to all displayed feature/category pairs (irrespective of the target-category feature).

Our view of this promising work, however, is that it does not consider an important result of the experiments: that of the difference in results between experiments where participants explicitly identified the target category, and where they did not (e.g., Experiment 5 vs. Experiment 6). Recall that when participants were asked to identify the target category before making their inference, a large and significant majority responded according to the single-category view; when participants were not asked to identify a target category before making their inference, however, participants' responses greatly shifted towards the multiple-category view. Nosofsky (2015) do not discuss this difference, and considers the results of Experiment 5, instead, as weakly supportive of the single-category view that is more strongly suggested by Experiment 6. Although dynamically adjusting the parameter settings depending on the specific queries of the experiment may lead to this difference in predictions, there is no intuition for how this parameter setting change may occur.

Our priming-based process model of feature inference, however, naturally answers that question as part of its core theory. Our model indicates that the difference in results is due to an underlying difference in the way that the experiments are processed by the human mind. It accounts for this difference because it includes the sources of priming in working memory to be a key part of its predictions. It suggests that when a person has explicitly thought about a category, the category is included as part of the inference process, biasing the model towards the single-category view; when a person has not, the model relies only on priming from the query feature, biasing the model towards the multiple-category view. Our model thus explains the same qualitative trends as Nosofsky (2015) while also accounting for this additional aspect of feature inference, and quantitatively matching the data.

Another model that has been proposed for explaining feature inference is the rational model and its associated variants (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2010). This model, while also rooted in Bayesian-based reasoning, has been shown to have trouble accounting for the breadth of the results we model here (Nosofsky, 2015). A recent promising version of this model was developed by Konovalova and Le Mens (2016), whose rational model is sensitive to uncertainty in categorization; our belief, however, is that it also would have trouble accounting for differences stemming from the presence or lack of an initial categorization step.

Acknowledgments

This work was supported by the Office of the Secretary of Defense and the Office of Naval Research. The views and conclusions contained in this paper do not represent the official policies of the U.S. Navy. We thank Greg Trafton, Sunny Khemlani and Tony Harrison for their advice and comments.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 576.

Hiatt, L. M., & Trafton, J. G. (2016). Familiarity, priming and perception in similarity judgments. *Cognitive Science*. (doi: 10.1111/cogs.12418)

Konovalova, E., & Le Mens, G. (2016). Predictions with uncertain categorization: A rational model. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148-193.

Murphy, G. L., & Ross, B. H. (2007). Use of single or multiple categories in category-based induction. *Inductive reasoning: Experimental, developmental, and computational approaches*, 205–225.

Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1929.

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.

Schneider, D. W., & Anderson, J. R. (2011). A memory-based model of hick's law. *Cognitive Psychology*, *62*(3), 193–222.

Thomson, R., Harrison, A. M., Trafton, J. G., & Hiatt, L. M. (2017). An account of interference in associative memory: Learning the fan effect. *Topics in Cognitive Science*, *9*(1), 69-82.

Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, II, F., Khemlani, S. S., & Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, *2*(1), 30-55.