HUMAN LEARNING AND GENERALIZATION OF CONCEPTS DEFINED BY FEATURAL RELATIONS

BY

MATT WETZEL

AA, AS, Lakeland Community College, 2012
BA, Cleveland State University, 2014
MS, Binghamton University, 2020

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Cognitive & Brain Sciences
in the Graduate School of
Binghamton University
State University of New York
2022

Accepted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Cognitive & Brain Sciences
in the Graduate School of
Binghamton University
State University of New York
2022

May 10, 2022

Kenneth J Kurtz, Chair
Department of Psychology, Binghamton University

Peter Gerhardstein, Member
Department of Psychology, Binghamton University

Ralph R Miller, Member
Department of Psychology, Binghamton University

Daniel Corral, Outside Examiner
Department of Psychology, Syracuse University

Abstract

Category and concept learning research has broadly focused on the mechanisms and principles that explain how humans learn functional mappings from a set of object features to a set of object categories. However, not all concepts can be predicted by feature values independently (e.g., the concept of *wealthier* does not depend on a particular range of values, but instead a comparison of multiple values from any numeric range). The present work provides 3 sets of novel experiments designed to test whether phenomena indicative of feature-based category learning emerge in a domain where subjects can leverage *featural relations* in addition to *features* themselves. The behavioral results partially support this hypothesis, but diverge from feature-based predictions in important ways. Further, the behavioral data was used as an empirical benchmark to test 2 different computational models of feature-based category learning and 3 different representational formats. The findings from the present work expand the explanatory scope of theoretical models from the feature-based category learning literature, and provide an empirical dataset to test theories spanning the less-explored domain in-between *featural* and *relational* inductive learning.

Acknowledgements

Acknowledgements and credit for this work go towards a number of parties. First, thanks go to Dr. Kenneth J. Kurtz for advising this project and my academic and scientific education for the past 6 years, especially for the very engaging scientific discussions and for holding your lab to a high standard. Major thanks also go to the dissertation committee for taking the time to read this document and for criticizing and evaluating its scientific merit, and also for providing insight on the theoretical relevance of the results. More thanks and acknowledgements go to the past and present grad students in the LaRC Lab – Dan, Sean, JD, Garrett, Nolan, Mercury, Alexus, and Josh – for many hours of helpful discussion, feedback, and advice. Further thanks go to the very competent undergraduate research assistants for providing insightful comments during lab discussions and for conducting nearly all of the in-person data collection. Thanks and acknowledgements also go towards the many coders in the open-source community (who I'll never meet) for creating and maintaining the software needed to conduct research in the modern world. Lastly, very heartfelt thanks and appreciation to the faculty and staff in the psychology department at Binghamton University for generously providing constant academic mentoring, sharing fascinating lectures & research talks, and for creating an intellectually challenging and enriching environment.

Contents

List of Tables

List of Figures

Introduction

The survival and fitness of an animal seems to depend in part on its ability to predict and infer the state of its environment. In some instances, this inference might come from a memory from a specific experience. For example, an animal might encounter predators at a particular location; it should ideally be wary each time it encounters that same location. In other instances, the animal might need to *generalize* to new experiences it hasn't encountered. If an animal encounters a new location that's *similar to* a previously experienced location with predators, it would behoove the animal to recognize that similarity. The set of all predator-ridden locations could be thought of as a *concept*. A *concept* can be described as a set of distinct objects or events that can be linked together in some way (Kemp, 2012) – often by some principled rule or metric of *similarity*.

Researchers often further distinguish between *featural* and *relational* concepts. Featural concepts are defined as a flat list of cues (or features), while relational concepts are thought to be defined by systems of relations (Corral et al., 2018; Gentner, 1983; Markman & Ross, 2003). How humans and nonhuman animals learn and generalize both featural and relational concepts has been a crucial question in the cognitive science literature (Ghirlanda & Enquist, 2003; Guttman & Kalish, 1956; Honig & Urcuioli, 1981; Murphy, 2004; Shepard, 1987).

While there are many ways to describe the problem of concept learning, this dissertation will follow the *inductive learning* framework described by Holland et al. (1989) and further expanded on by Kemp & Jern (2014). First, *induction* can be defined as the process through which human and nonhuman learners infer the state of the environment using whatever predictive variables are available to them (Holland et al., 1989). The predictive elements of the environment are often defined as *features*, *objects*, *categories*, and *relations*. Many inductive learning problems can be operationalized as a process where subjects learn either statistical or logical relationships between these elements (Kemp & Jern, 2014). The question of how perceptual data is mapped onto these elements is often treated as a *perceptual* problem that is outside the scope of concept learning (Chalmers et al., 1992). In the long run, this framing may turn out to be an oversimplification of the learning experience. However, it does a good job of encapsulating the assumptions and framing from which many theorists and modelers have come to understand the problem of induction and concept learning.

Category learning is a particular type of inductive learning domain where learners are tasked with inferring the initially unknown category of an object from a known set of features (Kurtz, 2015). Importantly, a set of features (and memory of prior categorizations) is all knowledge learners are provided with[1]. Consequently, many theoretical models of category learning use only feature values and prior knowledge to predict categorization decisions (Markman & Ross, 2003). Some argue that this lacks a key predictive element present in many natural domains: relations. A *relation* is some concept that binds multiple elements together (Gentner, 1983),

[1]Of course, they can always leverage whatever prior knowledge they bring prior to the experiment.

and is not necessarily dependent on the absolute values of features or objects themselves. For instance, the concept of *wealthier* cannot be deduced from a particular amount of money a person has; rather, it requires a comparison to the feature value (wealth) of another person. Theorists have argued that many key human cognitive capacities explicitly rely on relational reasoning (Gentner & Kurtz, 2005; Penn et al., 2008), and the lack of relations in models of category learning has been seen as a limitation (Kurtz & Boukrina, 2004; Markman & Ross, 2003; Medin et al., 1993).

The present work aims to make a step towards addressing the question of how relational information is integrated into the learning and formation of categories, and whether that integration is adequately described by the same mechanistic principles that have become central in the category learning literature. *Relational* information will be defined in a very narrow sense: any information that reflects the comparison of features in an inductive learning task. The goal of this exploration is to both (1) expand the scope of category learning theories, as well as (2) make progress towards building theories that span across the space of conceptual learning domains (Kemp, 2012). The next sections will briefly cover key findings and theoretical issues in the category and relational learning literature. Then, the present work describes 3 novel sets of experiments designed to falsify the hypothesis that relational cues are leveraged like featural cues in a categorization task.

Category Induction

Categories can be described as groupings of features, objects, and relations – often based on some critical principle (such as perceptual or statistical similarity,

or some formal rule). How learners *acquire*, *generalize*, and *generate* categorical knowledge is a fundamental question in cognitive science (Kemp, 2012; Kurtz, 2015). A typical category learning experiment (Bruner et al., 1956) is roughly analogous to an associative learning preparation: sets of featural cues are co-presented with some set of category labels (or, outcomes). In many instances, corrective feedback is provided after subjects predict the category label they think a set of features belongs to[2] (supervised learning). In other instances, subjects are never explicitly given the correct category labels to learn from (unsupervised learning). Researchers are often interested in why different category *structures* are easier or more appealing to learn than others, as well as how category knowledge is generalized to unseen examples. Empirical phenomena from this learning task have been leveraged as explanatory benchmarks for theoretical models of human and animal category learning (Kurtz, 2015; Murphy, 2004; Wills & Pothos, 2012).

The category learning literature has been fairly productive in building empirically predictive computational and mechanistic accounts of learning and generalization (Wills & Pothos, 2012). However, many theoretical accounts converge on their behavioral predictions despite utilizing fundamentally different learning frameworks (Kurtz, 2015). For example, ALCOVE (Kruschke, 1992), DIVA (Kurtz, 2007), SUSTAIN (Love et al., 2004), and RULEX (Nosofsky, Palmeri, et al., 1994) are all models of human categorization with different underlying frameworks; however, they all perform competitively when predicting human learning of the classic category structures from Shepard et al. (1961). This makes it difficult to validate which theoretical accounts are most valid. Despite the variability in frameworks, category

[2]Though learning can also occur via having subjects predict features from category labels (Chin-Parker & Ross, 2004; Yamauchi & Markman, 1998), or via passive observation of co-presented features and category labels (Levering & Kurtz, 2015).

learning theorists often invoke a consistent order of operations through which category knowledge is learned and generalized. First, data from the environment is represented as a set or structure of latent variables in either a continuous or featural *psychological space*. Then, some internal mechanism maps information from the psychological space onto some set of concept labels or behavioral responses. Finally, the learner makes an optimal or principled response based on whatever environmental challenge is presented.

Category learning frameworks tend to differ in terms of (a) how information is represented, (b) how information is processed during learning, and (c) what the optimization goals and environmental constraints of the learner actually are. Key phenomena from the category learning literature most relevant to the present work that will be discussed further in the next section.

Information Representation & Similarity

Many models of categorization (Anderson, 1991; Kruschke, 1992; Kurtz, 2007; Love et al., 2004; Nosofsky, 1986) begin with a set of features, objects, and category labels pre-discovered by an often unspecified perceptual process. While there has been relatively little discussion over how the perceptual system identifies the elements of an inductive learning task, there has been historical debate over how these elements are represented after they've been discovered (Jäkel et al., 2008; Medin et al., 1993; Tversky & Gati, 1982, 1982) – particularly in regards to the representation of stimuli. Many models of categorization rely on a *geometric* approach to stimulus representation, where predictive features of a stimulus are coded as numeric values indicating presence or magnitude; the collection of these

values are treated as coordinates in some psychological space.

In some instances, this psychological space can be explicitly defined by the raw feature values of stimuli (Nosofsky, Gluck, et al., 1994). However, this assumes that the perceptual variability manipulated by the experimenter matches the featural representations of a human observer, which may be an inaccurate assumption. So in other instances, the psychological space – that is, the hypothetical latent variables humans leverage in a particular learning task – are inferred using an external behavioral metric like *similarity judgements* or *confusability measures* (Nosofsky, 1992; Shepard, 1980). *Multidimensional scaling* (Shepard, 1962) is one such example of this approach. Inferring the psychological space using behavioral measures of *similarity* is a key process in many computational models of categorization (Kruschke, 1992; Nosofsky, 1986), and has the advantage of being arguably closer to a subject's perceptual representation of the stimuli.

The geometric approach has a number of advantages. First, it allows the modeled learner access to information about the magnitude of stimulus features; this seems useful given many real world learning problems involve stimuli or objects with features that vary in magnitude, and that magnitude can be a source for statistical inference. In addition, geometric representations make constrained predictions about the downstream effect of representing stimuli as spatial coordinates. For instance, *similarity* judgements between stimuli can often be accurately predicted by some function of featural *distance* (e.g., a *gaussian* function over the distance in height between two basketball players). One well-replicated finding in the concept learning literature is that the likelihood of generalizing a learned response to a new example follows a nonlinear, monotonic function of distance between that new

example and some anchored reference point in psychological space (Nosofsky, 1985; Shepard, 1987; see Figure 1).

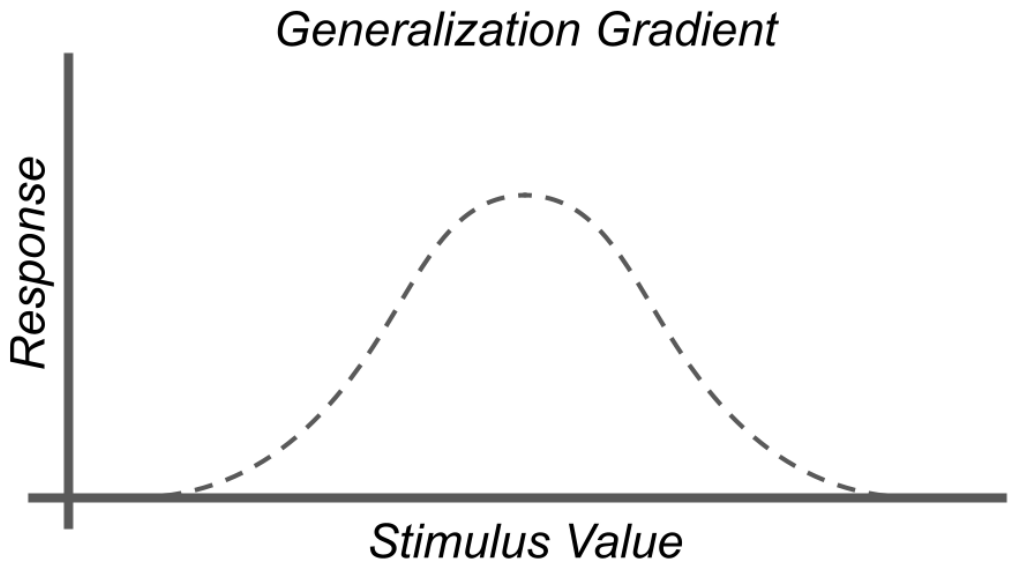

Figure 1: Example of a monotonic decreasing generalization function from the exponential family. *Monotonic* is in reference to distance from a reference point (in this case, the center).

Geometric representations of stimulus coordinates have been heavily adopted by theoretical models of categorization (Kruschke, 1992; Kurtz, 2007; Love et al., 2004; Nosofsky, 1986; Rosseel, 2002), and make a number implicit assumptions about any behavioral process that leverages a spatial representation of stimuli, particularly as a metric of *similarity*. One such assumption is bidirectionality: the similarity between object A and B should be equivalent to the similarity between object B and A. Another implicit assumption is the triangle inequality: the distance between A and B plus the distance between B and C must be greater than the distance between A and C (Jäkel et al., 2008; Tversky & Gati, 1982).

Importantly, Tversky & Gati (1982) demonstrated a number of empirical scenarios where human similarity judgements violate both of these implicit assumptions of geometric space. Tversky (1977) introduces an alternative approach to stimulus

representation where stimuli are represented purely by the presence of their features. Similarity judgements can then be predicted via a weighted combination of (a) the number of features that two stimuli share, and (b) the number of features that two stimuli do not share. Tversky's (1977) approach doesn't share the same mathematical constraints (bidirectionality, triangle inequality) of geometric stimulus representations, which makes it a candidate for explaining similarity and stimulus representation in cases where the geometric approach fails. However, it consequently is not applicable to the many scenarios where the geometric approach succeeds (Cheng, 2000; Kruschke, 1992; Nosofsky, 1989; Nosofsky, Gluck, et al., 1994; Shepard, 1987) – like in the prediction of human similarity judgements and category generalization (Nosofsky, 1986). Thus, the tension between these two theoretical approaches remains unresolved[3].

The implicit assumptions of both *geometric* and *featural* stimulus representation have been questioned as being restrictive and narrow in their explanatory scope (Medin et al., 1993; Murphy, 2004; Murphy & Medin, 1985). On the one hand the explicit assumptions of *geometric* and *featural* frameworks allow researchers to know exactly what subjects might be leveraging during inductive learning and generalization. On the other hand there are a number of instances where human judgements seem to reflect some type of knowledge that goes beyond unstructured lists of feature values (Murphy, 2004). For instance, in a relatively straightforward demonstration, Rips (1989) told subjects about a bird whose appearance was transformed to look like an insect as a result of living near toxic waste. Adult subjects still believed the animal was a bird despite its *featural* dissimilarity.

[3]Though see Jäkel et al. (2008) for examples of alternative spaces that could address the explanatory limitations of the geometric approach without sacrificing the concept of a *space*.

Subjects' response in this demonstration isn't surprising; what is surprising, is that such a simple scenario would be outside the explanatory scope of leading models of concept representation in their current form (Murphy, 2004).

Demonstrations like the one described have been referred to as *knowledge effects* (Murphy, 2004), and are a crucial motivating factor of the *theory view* of concepts (Murphy & Medin, 1985). The theory view posits that concepts reflect a broader, interconnected system of knowledge dependent on context and experience. While the theory view lacks an exact specification of how this knowledge is represented, there is another theory of stimulus representation that goes beyond treating concepts as collections of independent features. *Structural* stimulus representations (also called *propositional* or *symbolic* representations) have been foundational to many theories of concept learning (Gentner, 1983; Griffiths et al., 2010; Kemp, 2012; Pylyshyn, 1973). In the structural framework, stimuli are represented as elements in some representational language (Kemp, 2012) or network (Gentner, 1983; Pylyshyn, 1973) that integrates other elements of an inductive learning task. Like the historical geometric and featural approaches, the process through which the latent elements are derived from unstructured perceptual information is often unspecified and left to some theoretical perceptual process (Chalmers et al., 1992).

The structural approach to stimulus representation has one notable advantage (among others) over both geometric and featural stimulus representations: relations are explicitly represented and used for inference. This is important given various empirical demonstrations where humans appear to leverage relational knowledge (Goldstone et al., 1991). For example, consider the stimulus triad in Figure 6, where subjects might be asked to choose one of the *options* (bottom) that is most

similar to the *target* (top). If each stimulus were represented in terms of its features, the option most similar to the target would be option A. However, Goldstone et al. (1991) found that human adults[4] will reliably choose option B as more similar to the target – hypothetically because option B shared the same *relational structure* (ABA) as the target. *Structural* stimulus representations are also useful for explaining scenarios where people perceive similarity between systems of objects that have no obvious featural similarities – for example, the solar system and the Rutherford-Bohr model of the atom (Gentner, 1983). The structural approach to stimulus representation has been widely adopted in theoretical models of higher-order cognition (Doumas et al., 2008; Falkenhainer et al., 1986; Holyoak & Hummel, 2001; Larkey & Love, 2003), as well as some areas of standard concept learning (Kemp, 2012; Nosofsky, Palmeri, et al., 1994).

Each of these stimulus representation schemes has been argued as central to the problem of induction (Jäkel et al., 2008; Markman & Gentner, 1993; Medin et al., 1993), and are sometimes at odds when predicting specific behavioral phenomena. However, the domains where each of these approaches have predictive value are distinct in some ways. For example, the *geometric* approach often accurately accounts for similarity ratings and classification probabilities when stimulus features share the same latent dimension and are in *correspondence*[5]. When stimulus features are not in *correspondence* or directly comparable, subjects might instead leverage presence/absence or relational structure within a stimulus set. Providing a unified framework that integrates *geometric*, *featural*, and *structural* mechanisms is beyond the scope of the present work. However, the present work

[4]as well as children as they age (Gentner, 1988; Rattermann & Gentner, 1998)

[5]Features are described as *in correspondence* if they embody a common element between two objects (Goldstone & Medin, 1994)

does investigate whether relations[6] are leveraged in learning domains traditionally best described by the *geometric* framework.

Selective Attention

The invocation of similarity as a construct is almost ubiquitous – [though controversial (Medin et al., 1993) – in models of concept learning (Goldstone & Son, 2012). This includes relational concept learning (Goldstone et al., 1991). Typically, the learner's sense of *similarity* is described as some process of comparing and contrasting the *features* or *relational structure* between some set of stimuli or category abstractions. However, not all features (or relations) of a stimulus are equally useful towards the learner's objectives. It seems ideal to have some mechanism that focuses the learner's attention towards features that are task-relevant. In support of such a mechanism, there are many demonstrations where learners seem to *focus* only on highly predictive cues while simultaneously ignoring weakly predictive cues (Kruschke, 1993, 2003; Shepard et al., 1961). This phenomenon is aptly referred to as: selective attention[7].

Evidence for selective attention comes from the well-replicated finding that humans learn categories quicker when they can be distinguished on the basis of a few relevant features (Kruschke, 1993; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961). More evidence comes from the observance that humans seem *more* sensitive to diagnostic features than prototypical features (Chin-Parker & Ross,

[6]specifically, feature comparisons

[7]Interestingly, *if* a learner's goal is prediction, and *if* a learner is able to learn the statistical relationship between a large set of strongly and weakly predictive cues, *then* a statistically optimal learner should integrate the predictive value of as many cues as possible. Since human learners don't seem to do this, they may instead be prioritizing features in a way that serves some currently unknown or later-stage purpose.

2004) when engaging in a *classification* task. A similar phenomenon has been observed in the *multiple cue probability learning* literature, where humans are tasked with mapping noisy, nondeterministic cues (i.e., features) to outcomes (which are sometimes category labels). In this particular learning domain, humans seem to leverage a single, strongly predictive cue when making decisions – while ignoring the predictive utility of other, weakly predictive cues (a phenomenon referred to as *cue competition* effects). Importantly, *cue competition* effects have been found in probabilistic category learning (Kruschke & Johansen, 1999), function learning (Busemeyer et al., 1993), and causal learning (Baker et al., 1993) experiments.

Further, when asked to sort a group of simplified, artificial objects into categories, humans tend to sort objects on the basis of a single, perfectly predictive feature (Ahn & Medin, 1992; Medin et al., 1987; Milton et al., 2009). Consequently, selective attention mechanisms are integral to many leading accounts of human category learning (Kruschke, 1992; Nosofsky, 1986). Interestingly, this phenomena may not be restricted to humans[8] (Kruschke, 2003; Wasserman & Castro, 2021).

Rule-Based Classification Strategies

Leading theoretical models of categorization (Kruschke, 1992; Minda & Smith, 2011; Nosofsky, 1986) incorporate the various mechanisms that have been discussed thus far (stimulus-specific memory, similarity computation, selective attention). However, there is a very particular behavioral phenomena that can't be sufficiently captured by these mechanisms alone. A good illustrative example comes from Lee

[8]Kruschke (2003) has argued that the types of attentional processes in human category learning experiments may map onto the phenomena of *blocking* in the animal learning literature. In addition, recent work by Wasserman & Castro (2021) found preliminary evidence that pigeons will selectively attend to diagnostic features over nondiagnostic features in a category learning preparation.

et al. (2018), who trained subjects to discriminate between two objects that varied in color along a blue-green continuum. A *similarity-based* generalization response would predict that the strength or frequency of a response should *decrease* as a function of distance from a point in psychological space (i.e., the color wavelength that was reinforced during training) – a behavior consistent with how non-human animals typically generalize in a conditioning experiment. Importantly, Lee et al. (2018) found a proportion of learners whose generalization profile was strongest towards the most extreme regions in the feature space (Figure 2). This behavior is consistent with theoretical accounts that classify based on learned rules or thresholds (Ashby & Gott, 1988; Erickson & Kruschke, 1998), and inconsistent with accounts that can only leverage stimulus-memory and similarity computations alone (Medin & Schaffer, 1978; Nosofsky, 1986).

Typically, 'rules' in the category learning literature refer to binary rules (Nosofsky, Palmeri, et al., 1994) or dimensional boundaries that divide the psychological space in half (Ashby & Gott, 1988; Erickson & Kruschke, 1998). However, the rules humans can leverage for classification can be more complex. Evidence comes from Kurtz & Wetzel (2021), who trained subjects on a category structure where membership alternated along regions of feature space. They found that subjects could not only learn the rule that defined the training stimuli, but also extrapolate the alternating pattern to regions of feature space unobserved during training (Figure 3).

In addition to expanding the phenomenological boundaries of rule-based classification behavior in human learners, this demonstration also provides an additional case where *similarity-based* theories of generalization fail. Consider the visual

representation of the category structure in 3. During training, subjects observe stimuli whose single feature is represented by the labels A and B; items labeled t are those observed *after* training. Similarity-based accounts would predict that the t stimuli would be classified as members of the B category – since they are the *most similar* stimuli observed during training. The majority of learners who instead classified the t stimuli as members of the A category is a direct violation of *similarity-based* theories of generalization.

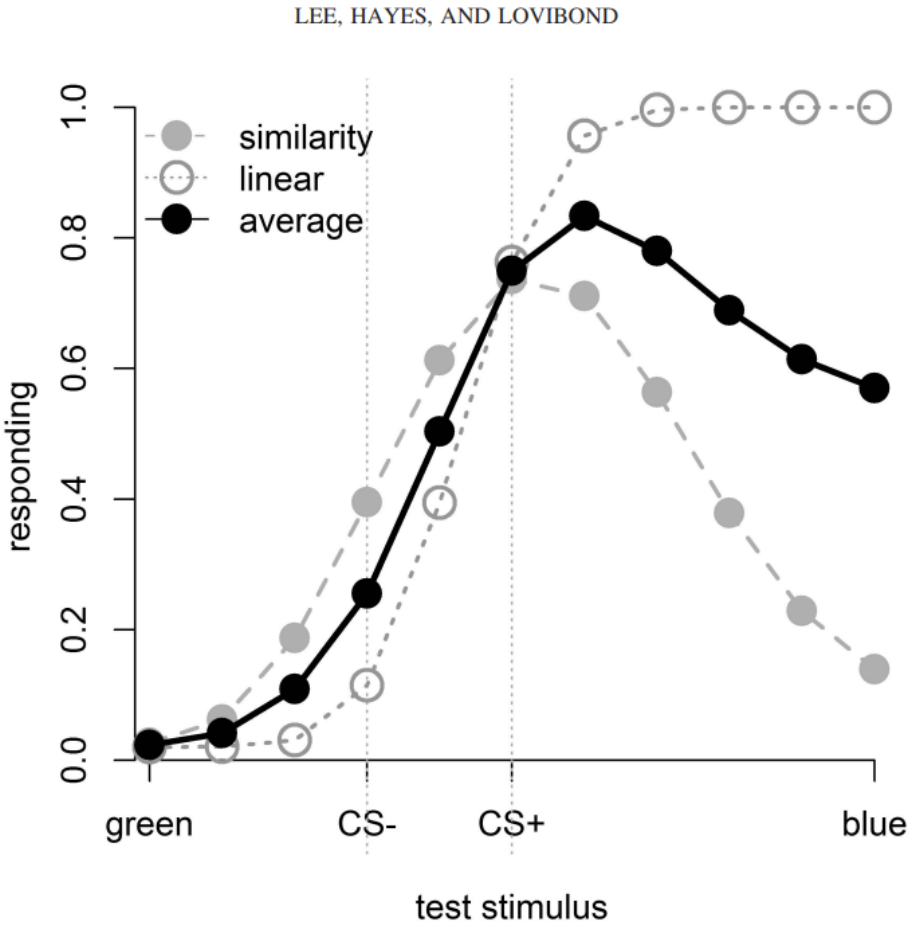

Figure 2: Generalization data from Lee et al. (2018), exemplifying a mix of rule-like and similarity-like generalization profiles. The y-axis shows aggregate probability of generalization; the x-axis is the range of feature values each stimulus could fall in (in ascending order). Taken directly from Lee et al. (2018).

As discussed, there is compelling evidence that human subjects may learn and leverage rules during concept induction. In fact, many of the earliest investigations

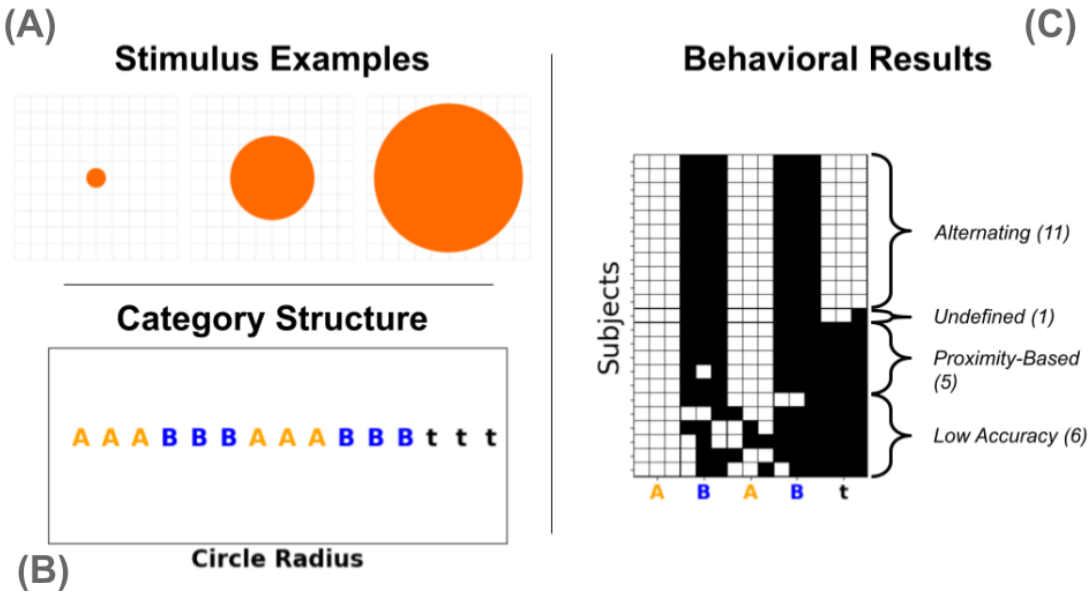

Figure 3: Visualization of the stimuli (A), category structure (B), and generalization profiles of each individual subject (C) from Kurtz & Wetzel (2021). In (B), each row is a subject, and each column is a stimulus with a particular feature value (in ascending order). Generalization profiles showed a mix of rule-like and similarity-like classification strategies. Taken directly from Kurtz & Wetzel (2021).

of concept induction (Bruner et al., 1956) invoked rule-based or hypothesis-testing explanations of behavioral phenomena. Additionally, many modern, leading accounts of human categorization leverage both rule-based and similarity-based mechanisms during learning (Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky, Palmeri, et al., 1994; Schlegelmilch et al., 2021). Whether instantiated as logical propositions (Nosofsky, Palmeri, et al., 1994) or geometric boundaries(Ashby & Gott, 1988; Erickson & Kruschke, 1998), existing frameworks primarily treat features as an independent, flat list of cues. One question of the present work is to explore whether rules can be learned based on *featural relations* instead of *features* independently.

Information Complexity

Acquisition difficulty on different concept learning problems has been utilized as a way to test the psychological plausibility of different theoretical frameworks. For example, a number of demonstrations show that humans are much faster at learning categories that can be classified on the basis of a single feature (Kruschke, 1993; Shepard et al., 1961). A key empirical result comes from Shepard et al. (1961), who trained subjects on 6 different category learning problems that produced an ordering of learning difficulty with human subjects. Each of the 6 category structures span over the domain of all possible stimuli with 3 binary features (8 in total); the category structures themselves reflect qualitatively different ways of grouping all 8 possible stimuli into 2 equally sized groups (see Figure 4). Models that produce the same *learning difficulty ordering* are typically regarded as more psychologically plausible than those that don't (Kurtz, Levering, et al., 2013; Nosofsky, Gluck, et al., 1994). In addition, the difficulty of learning the 6 problems can often be predicted using a number of interesting complexity metrics, such as: boolean complexity (Feldman, 2000), description length (Kemp, 2012), or information entropy (Pape et al., 2015)[9]. This is important because learning difficulty could reflect either (a) constraints of human learners, or (b) constraints inherent to the task.

[9]See Vigo (2013) for an alternative account.

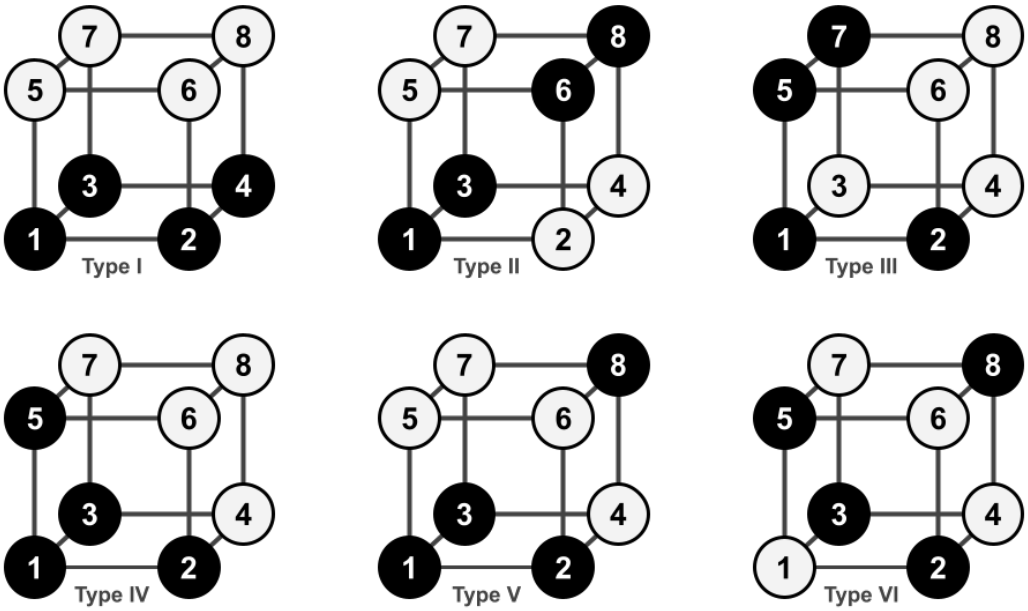

Figure 4: Visualization of the 6 category structures from Shepard et al. (1961), each represented as a cube. Each dimension of the cube represents the value of a binary feature; each point on the cube represents a different stimulus (of 8 possible).

Relational Inference

In a typical category (and associative) learning preparation, category labels (and outcomes) are defined over particular regions of the stimulus space – which is itself defined by the cues or features used during training[10]. In most cases, the features present during learning exist on perceptually distinct dimensions, and have no basis for direct comparison. In contrast, there are also inductive learning scenarios where features are in *correspondence* (Goldstone & Medin, 1994); these features typically exist on the same latent perceptual dimension, and can be directly compared. A primary question of interest for the present work is whether animals can leverage comparisons between *corresponding* features as a basis for inductive

[10]These may be the physical features of the stimuli, or the latent features in some psychological space (Shepard, 1987).

learning; this type of induction is often called *relational*[11] Many researchers have argued that relational induction is a (if not *the*) primary facet of human cognition (Gentner, 1983; Kemp, 2012; Medin et al., 1993), and there are many behavioral phenomena that are difficult to explain without assuming some mechanism for relational computation (Blaisdell & Cook, 2005; Goldstone et al., 1991; Köhler, 1938; Premack, 1983).

Abstract Concept Learning

Abstract concepts (or arguably, *relational concepts*) are defined as concepts that are not predictable from features of stimuli independently (Bodily et al., 2008), but instead from some inter- or intra-stimulus comparison (Zentall et al., 2008). In contrast to *concrete* concepts, *abstract* concepts aren't constrained to a particular feature space. Abstract concept learning has been well-studied in both human and nonhuman animals – some have suggested that it represents a meaningful form of relational reasoning (Lazareva, 2012; Zentall et al., 2008). While the simplest learning problems in the *abstract* concept learning literature may not capture the ecological breadth of larger literature on relational cognition, they nevertheless highlight scenarios where purely featural accounts fail to predict behavioral phenomena. A few of the key phenomena will be discussed.

The Transposition Effect

It has been well-demonstrated that animals can be trained to associate some region of stimulus space with some reward or outcome (Domjan, 1993; Guttman & Kalish, 1956). For example, a pigeon might be trained to produce a pecking

[11]Though there is controversy on what *relational* precisely means (Penn et al., 2008).

response each time it is presented with a circle with a given size. Intuitively, whatever is learned by the pigeon is likely based on the absolute magnitude of the cue (in this case, size). However, cues in the environment are seldom presented in isolation. Researchers have also demonstrated that animals in an inductive learning task may leverage the *difference between 2 cue values* rather than the absolute magnitude of any cue in particular (Köhler, 1938).

A notable example occurs in a discrimination learning task where animals are reinforced for selecting one target stimulus from a set. For example, suppose an animal is reinforced to select the larger of two circles with varying sizes (Figure 5) – and then shown a new set of circles with different sizes. Interestingly, nonhuman animals in some preparations will reliably choose the larger of the two circles (in a new set) *even though* that larger circle is *less featurally similar to* the circle that was originally reinforced. This phenomenon is referred to as the *transposition effect*, and has been central to the question of relational learning in animals[12].

Same/Difference Learning

The relational match-to-sample (RMS) task has been another popular way of investigating same/difference concept learning (Goldstone et al., 1991; Premack, 1983). In a typical RMS task, a subject is presented with a *target* stimulus (usually instantiated as a collection of shapes), along with 2 alternative *options*. The subject's task is to select the *option* that 'goes with' the *target*. Importantly, the target stimulus has a particular *relational* characteristic; for example (see Figure 6),

[12]Notably, there are alternative accounts of the transposition effect that do not require any explicit relational computation (Spence, 1937). However, an alternative version of the transposition effect (Gonzalez et al., 1954) requires animals to select the *intermediately* sized object (within a set of 3 objects of varying sizes), which provides an extra challenge for non-relational accounts.

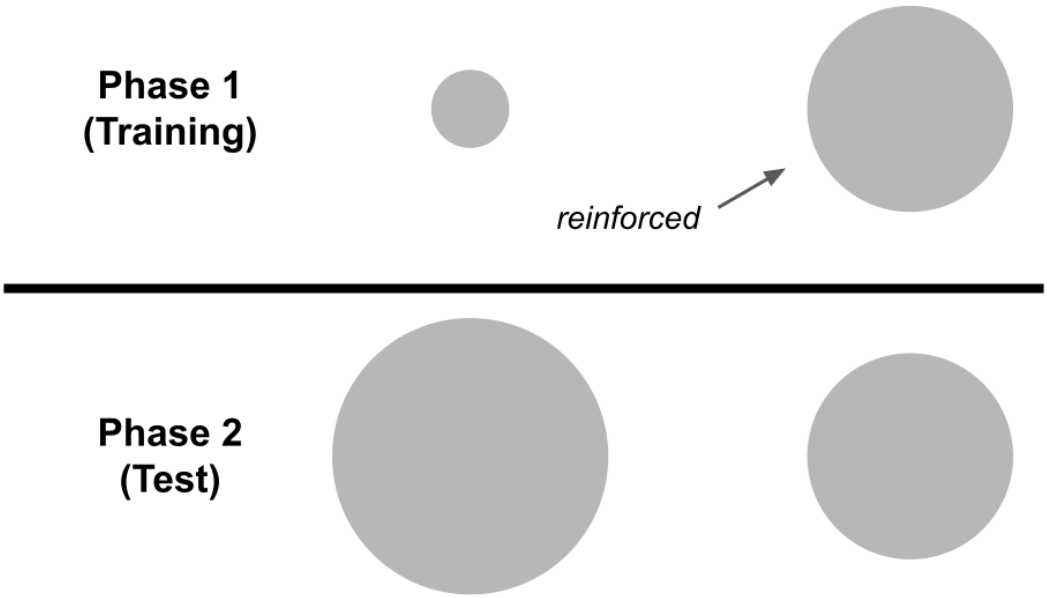

Figure 5: Example of the typical discrimination learning preparation for demonstrating the transposition effect. In the training phase, animals are reinforced to select the larger of two circles. In the test phase, the critical test is whether animals select the smaller circle (which is featurally similar to the training stimulus), or the larger circle (which matches the *relational* pattern an animal may have learned during training).

the target might consist of two objects that are the *same* shape. Typically, one of the *options* embodies the same relational characteristic (e.g., _same shape___) as the *target*; the opposing option will often be featurally similar or unrelated. Under certain conditions, humans[13] will reliably select the *relationally* similar option (Goldstone et al., 1991; Kotovsky & Gentner, 1996)

There has also been considerable investigation towards the question of whether nonhuman animals recognize the concept of *sameness* – that is, the awareness that two features, objects, or concepts are identical (Blaisdell & Cook, 2005; Penn et al., 2008; Zentall & Hogan, 1974). In classification learning tasks (Wasserman et al., 1995, 2001), there are conditions where animals can behaviorally differentiate between (a) object arrays that instantiate *sameness* across objects, and (b) object

[13]and even language trained chimps (Premack, 1983)

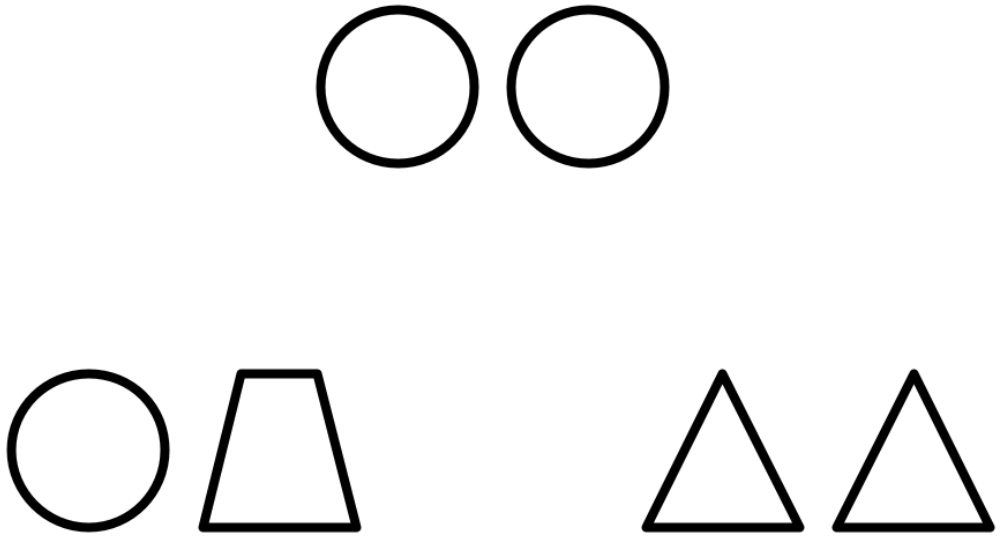

Figure 6: Example of the relational match-to-sample (RMS) task. Subjects are asked to match the target (top) with either of the two options (bottom). One option is *featurally* similar to the target (bottom left), the other option is *relationally* similar to the target (bottom right).

arrays that instantiate *difference* (Figure 7). The concepts of *same* and *different* are intuitively *relational* in nature, since they involve a comparison of objects (Zentall et al., 2008). Interestingly, Wasserman et al. (2004) highlight that *entropy* within a stimulus array is highly predictive of generalization of same/different concepts. It might be that some sort of entropy calculation provides the crucial cue nonhuman animals (and even humans; see Young & Wasserman, 2001) leverage when making *same* / *different* judgements.

Relational Category Learning & Theory Learning

The demonstrations of *abstract* (or relational) concept learning discussed thus far typically leverage a very particular experimental preparation. Often, only one feature of the stimulus is manipulated – explicitly highlighting what cues an animal has access to in an experiment. In addition, the animal is trained in a

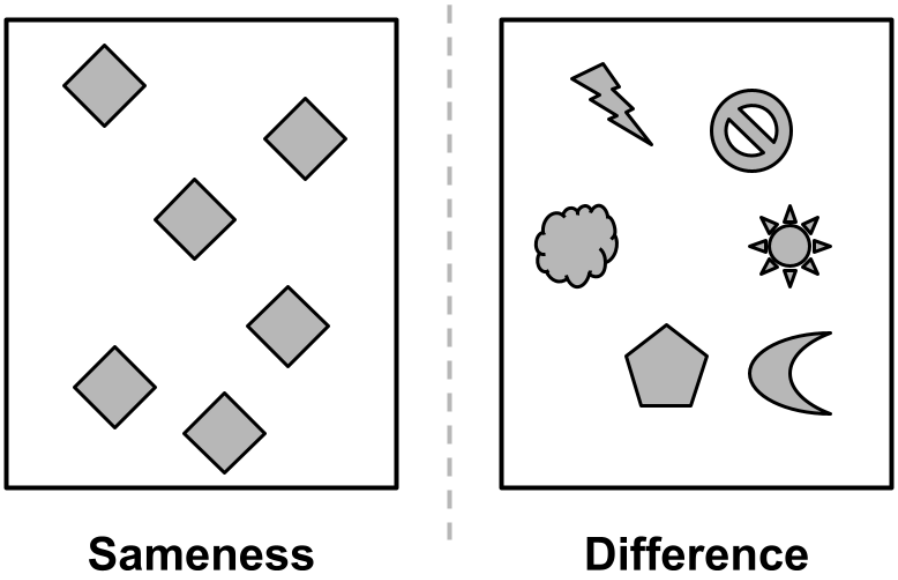

Figure 7: A representation of the type of object arrays that could be used in an abstract classification learning task; arrays vary in terms of the number and frequency of unique icons. Note: these are not exactly identical but similar to the stimuli used by prior investigations (Wasserman et al., 1995; Young & Wasserman, 2001).

discrimination learning task where it chooses a target stimulus among a set of alternative options. However, abstract concept learning has also been demonstrated within the context of a category learning preparation, where the task is to use the present stimulus configuration as the basis for selecting among a set of *category labels*. *Relational* categories are categories made up of stimuli that embody systems of relations between *features*, where membership is based on the structure of those relations *rather* than their particular feature values (Gentner & Kurtz, 2005).

In an important demonstration, Kurtz, Boukrina, et al. (2013) trained subjects to learn categories based on relational properties – such as monotonicity, identicality, and support (see Figure 8). Exemplars across categories were featurally similar single-component stimuli, each consisted of an arrangement of rocks of varying shape, size, and color. Importantly, exemplars within each category could be identified based on the particular relational property that they instanti-

ated. The greater-than-chance performance in a relational category learning task can be treated as evidence that subjects both recognize and generalize relational information during a standard inductive learning task.

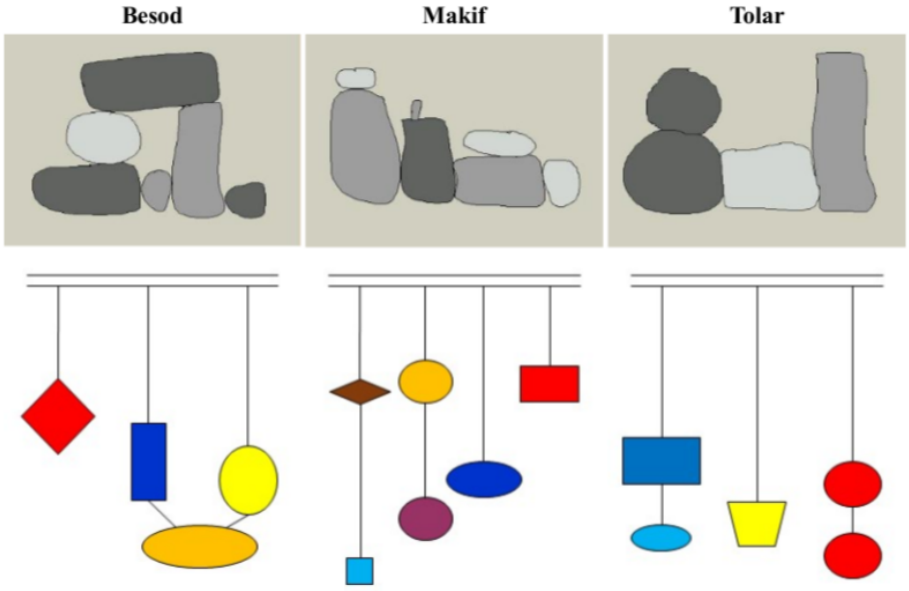

Figure 8: Example of the stimuli used in Kurtz & Boukrina (2004). The top row is the stimuli used during training; the bottom row is the stimuli used during a transfer test phase. While the training and transfer items are *featurally* very distinct, they share a potential relational 'structure' that subjects might be leveraging during learning. Taken directly from Kurtz & Boukrina (2004).

The relational category learning literature has explored how people learn to categorize systems of relations (which are typically instantiated as a single stimulus). A system of relations is referred to in some literatures as a *theory* (Carey, 1985; Davis, 2014; Kemp et al., 2010), and a considerable amount of work has addressed how the structure of relations in a system influence how relational knowledge is learned and generalized (Corral & Jones, 2014; Kemp et al., 2007). For example, Kemp et al. (2007) presented subjects with a set of English letter pairs; the structure that defined which pairs of letters went together was manipulated (Figure 9). Kemp et al. (2007) found that subjects' subjective ratings of complexity – as well as their accuracy in identifying appropriate letter pairings in a generalization

phase – was well predicted by the *representational length*[14] given an appropriate representational language.

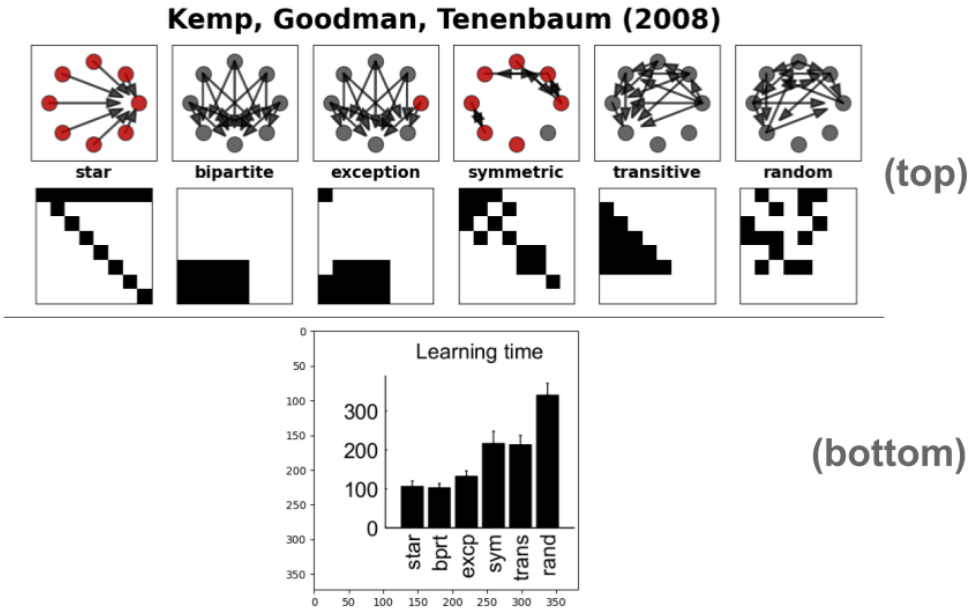

Figure 9: Top: Learning structures used by Kemp et al. (2007). Each node in the graph represents a unique stimulus; each connection represents whether those two stimuli had a relation. Bottom: Average time required for human subjects in each condition (taken directly from Kemp et al., 2007).

In similar work, Kemp et al. (2010) allowed subjects to interact with two sets of objects[15] (each uniquely identified with a letter of the English alphabet). Each object could be in one of two categories; the only way to identify category membership was to observe a perceptual event occurring when 2 objects made contact. The *structure* of each learning problem was defined by the rules defining how objects from each category interact. One finding was that probabilistic relationships between categories were harder to learn[16] than fully deterministic relationships. Another finding was that a directed relationship between the two categories (i.e., only one category can 'activate' another) was easier to learn than

[14] i.e., the number of required predicates and symbols used in describing the problem solution – which is related to but distinct from the concept of *description length* (Kemp, 2012)

[15] via dragging around on a screen

[16] 'Harder to learn' was operationalized by the number of individuals who successfully learned the relationship.

an undirected relationship (i.e., either category can activate the other, but not itself) – particularly in the probabilistic learning problems. All of this work focuses fundamentally on how the structure of a problem impacts learning and generalization within a domain of relational systems.

The work of Kemp et al. (2007) and Kemp et al. (2010) provide a tightly controlled experimental demonstration of relational learning. However, it is difficult to know what role *featural* information would have played since features didn't have diagnostic value. In contrast, work by Corral & Jones (2014) has explored how featural relations might be leveraged in a classification task. Here, subjects were trained to classify a stimulus based on the featural relations *within* its subcomponents. Though importantly, the featural relations themselves were not intended to be the source of induction; rather, subjects were found to be sensitive to the *topological* structure that described the presence of featural relations in a single stimulus. Interestingly, topological structure may correlate with *description length*; if so, Corral & Jones (2014) and Kemp et al. (2007) would provide converging evidence for how structural complexity impacts which relational systems humans can most easily learn[17].

Present Work

The category learning literature has been particularly productive at both discovering key inductive learning phenomena and evaluating theories explaining how those phenomena arise. Among these phenomena, (1) similarity-based generalization, (2) selective attention towards diagnostic features, and (3) information

[17]which may also align with work using network-based descriptions as well (Karuza et al., 2017; Sloman et al., 1998)

complexity as a predictor of learning difficulty have been well-replicated and very influential in ruling-out competing theories of category learning (Nosofsky, Gluck, et al., 1994). However, existing work in the category learning literature has typically used learning problems where categories are defined over single-component stimuli (i.e., stimuli embodying a single, perceptually identifiable object). Researchers from other literatures have pointed out that this fails to account for how *relational* information can be integrated into category learning and generalization (Goldstone & Medin, 1994; Markman & Ross, 2003; Medin et al., 1993).

There have been some investigation of how *relational* information is utilized in a standard category learning preparation. Some of these investigations have leveraged complex stimuli with a multitude of present relations where the exact nature of the information presented to subjects is less-specified (Kurtz, Boukrina, et al., 2013; Kurtz & Boukrina, 2004; Patterson & Kurtz, 2014). This lack of specification makes it difficult to test mechanistic theories of relational induction. Investigations with well-defined stimuli have utilized discrimination or relational match-to-sample tasks (Fagot et al., 2001; Köhler, 1938; Zentall & Hogan, 1974), or have utilized categorization tasks using stimuli defined by discrete measures of feature variability (Wasserman et al., 1995, 2001; Young & Wasserman, 2001). There is existing work that has leveraged relational stimuli with tightly controlled structure (Corral et al., 2018; Corral & Jones, 2014, 2017). In this particular work, stimuli are treated as a *single-components*, and the focus of investigation was how humans leverage relational structure in an inductive learning task. This leaves room for exploration on how featural relationships are leveraged to understand *multi-component*, compound stimulus pairs.

The present work had 3 goals. First, the present work explored how humans might integrate *featural relations* into classification decisions. This could help paint a more complete picture of the scope of information human subjects leverage in an inductive learning task. Second, the present work asked if key phenomena of category learning emerge in a relational learning domain[18]. This could provide insight into whether or not subjects treat featural comparisons as predictive cues when classifying compound stimulus pairs. Third, the present work tested whether existing computational models of category learning can explain human behavior when provided relational information in addition to featural information during learning. Here, we defined *relational* information as any predictive cue that emerges from the comparison of multiple features values (i.e., relative magnitude differences). The present work treated *relations* as feature value comparisons exclusively, though this is only one small aspect of what human *relational knowledge* may encompass (Corral & Jones, 2017; Gentner, 1983; Gentner & Kurtz, 2005).

[18]following in the theoretical direction of Corral et al. (2018)

Behavioral Investigation

In a standard classification learning task, subjects are presented with a single object along with a set of potential category labels the object could belong to. The standard elements of the learning task are (1) the object, (2) the object's features, and (3) the category labels the subject is supposed to guess[19]. This preparation in particular seems to discourage consideration of relational aspects of the learning domain, for a number of reasons:

1. comparisons between features *within* an object are ill-defined if they are not represented by the same perceptual dimension (like length, or color)[20],
2. comparisons of features *across* objects would require an accurate memory of objects previously encountered (which is difficult for human subjects), and
3. any relational information that could be inferred is typically task-irrelevant.

The present work made a very minor change to the standard category learning preparation: rather than showing subjects one object at a time, subjects were shown a pair of objects with *corresponding features*[21]. We call a collection of

[19]While various investigations have manipulated *which* of these elements subjects are required to guess (Chin-Parker & Ross, 2004; Yamauchi & Markman, 1998, 2000), these 3 elements themselves are typically the only immediate information subjects have to work with.

[20]Features in a category learning experiment are typically not represented by the same perceptual dimension (Kurtz, 2015; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961).

[21]Note that this is distinct from the discrimination learning preparation that tasks animals with selecting 1 of two stimuli; here, subjects must map a pair of stimuli to a distinct, categorical response.

corresponding objects a *compound stimulus*. In addition, the category subjects learned to classify was designed around *feature comparisons* rather than features independently. That is, an object pair could have belonged to a category even if it is featurally dissimilar from previously learned exemplar pairs *as long as* the feature comparison across objects was category consistent. This design followed the procedure of previous demonstrations of relational concept induction (Corral et al., 2018; Corral & Jones, 2014; Kemp et al., 2007).

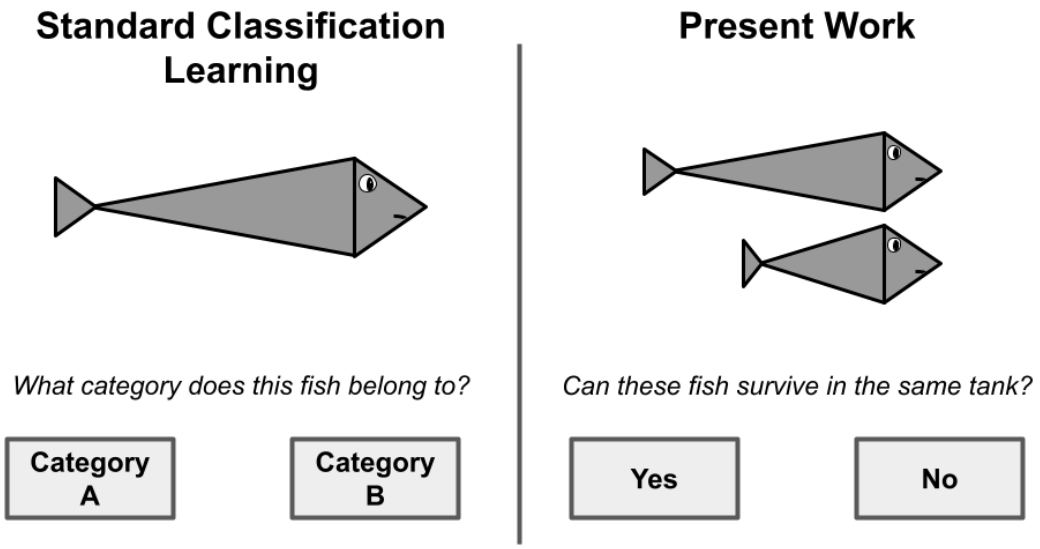

Figure 10: Comparison of a normal category learning task versus the task used in present work. Left: a single item is presented along with a set of category labels subjects could select; Right: the same procedure, but with a compound stimulus pair instead of one, single-component stimulus.

The next sections describe 3 behavioral experiments designed to expand our phenomenological understanding of inductive learning over compound stimulus pairs. In each experiment, subjects were required to learn a mapping *from* pairs of corresponding stimuli *to* an arbitrary set of category labels. The goal was to shed light on (a) how featural comparisons are leveraged during learning (if at all), (b) how this interacts with the predictive value of the features independently, and

(c) whether theories of category learning can account for the behavioral data we observe. In addition, the behavioral data served as the empirical benchmark for extending existing computational accounts.

All of the materials for the behavioral experiments can be found at https://osf.io/nkp46/. All experiments were coded using HTML and Javascript (using the oCanvas library; Koggdal, n.d.).

Experiment 1: Relational Generalization Gradients

Nonlinear monotonic generalization gradients have been a hallmark phenomenon of human and animal generalization (Guttman & Kalish, 1956; Shepard, 1987). Experiment 1 aimed to investigate how humans generalize on the basis of relations between features that vary in a continuous space, and whether that generalization behavior reflected what is typically observed when animals generalize on the basis of independent sets of features. Subjects learned to categorize pairs of objects that varied on the basis of 1 quantitatively-valued feature. Importantly, the category structures were designed based on the difference in magnitude between each stimulus feature (see Figure 11). After training, the rest of the stimulus space was sampled to estimate an average generalization gradient across subjects. In all cases, it was predicted that the aggregate generalization gradients will resemble a monotonic function of distance in a *relational* similarity space.

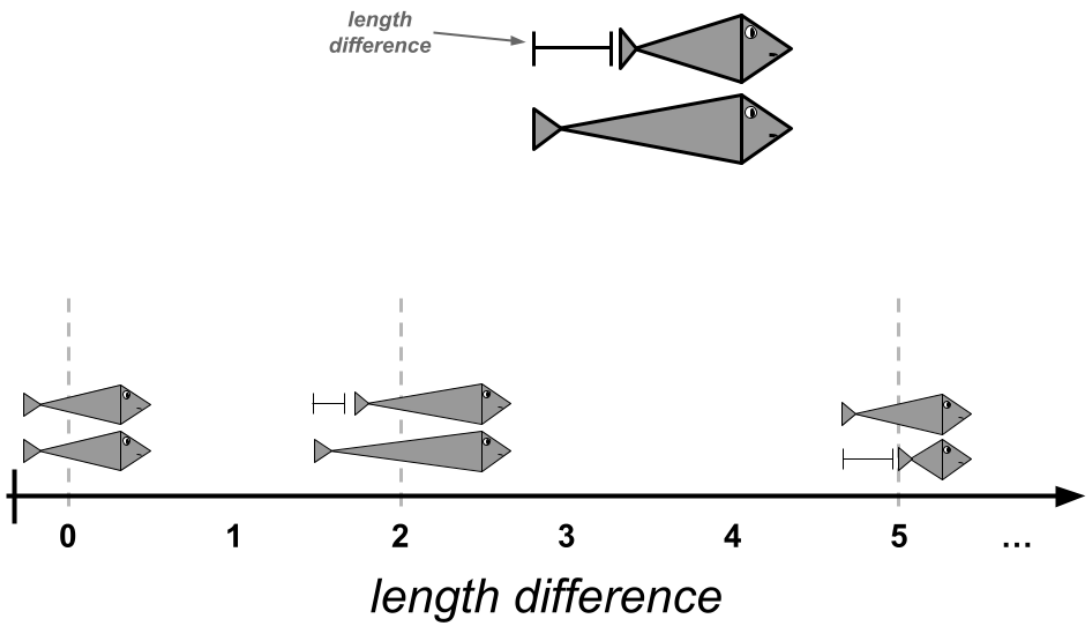

Figure 11: Visualization of the *stimulus space* in experiment 1. Each stimulus pair (shown together in each trial of learning) can be represented by their *featural* magnitude difference (in this case, *body length*).

Experiment 1a

The purpose of experiment 1a was to establish a baseline estimation of subjects' aggregate generalization behavior. Subjects first underwent an observational learning task, where they were shown a series of 1-featured stimulus pairs described as positive instances of a category; the subject was not required to make a classification decision. After this training phase, subjects were tested on a set of stimulus pairs that span the full range of possible magnitude differences between stimuli; in this phase, subjects were required to make a classification decision. Three different category structures were utilized: (1) a category structure where all stimuli differed by **8 units**, (2) a category structure where all stimuli differed by **1 unit**, and (3) a category structure where all stimuli differed by **0 units** (i.e., the stimuli in a pair are identical). The category structures were referred to as *diff-8*, *diff -1*, and *diff-0*, respectively. This choice in category structures was made in order to examine

whether generalization behavior differed depending on the region of space where positive category membership was assigned.

Participants

Subjects consisted of 203 undergraduate students from Binghamton University; students were compensated with partial credit towards a course requirement.

Materials & Design

Subjects completed the experiment online using their own computers at any location of their choosing[22]. The experiment display consisted of visual objects (buttons, shapes, text boxes) in a web browser, and could be navigated through button clicking or keyboard presses. The stimuli for the experiment consisted of fish that could vary on the basis of one feature: length. The length could vary between any integer value between 1 and 16. The *true size* was dependent on the subjects' computer screen size. On a laptop with a ~ 13 inch (~ 33cm) display[23], the length varied from 5-21 centimeters. Stimuli were designed using Google Slides, and the experiment was presented via HTML and oCanvas (Koggdal, n.d.). Subjects were told that their job was to learn about which pairs of fish could "go together" in the same water tank.

Training phase: The training phase consisted of 2 blocks of training; each block consisted of 1 exposure to all possible combinations of the 16 stimuli *that instantiated the target length difference*. The number of unique pairs differed depending on the target length difference. In the *diff-8*, *diff -1*, and *diff-0* conditions, the number of

[22]This undoubtedly leads to variation in screen size and screen coloration.
[23]with 1280 x 800 pixel resolution

unique pairs are 8, 15, and 16 respectively. To keep the training times relatively equal across conditions, stimuli from the *diff-8* condition were sampled twice per block. Each subject was only trained on one category structure, making this a between-subjects design.

Test phase: The test phase consisted of 2 exposures to *each possible length difference*, including the target length difference from training. The two exposures were a each random pair of fish sampled from a particular length difference. The test phase was nearly identical to the training phase, except subjects were instructed to make a classification response among a set of 2 possible options ('Yes' or 'No', instantiated on the screen as buttons).

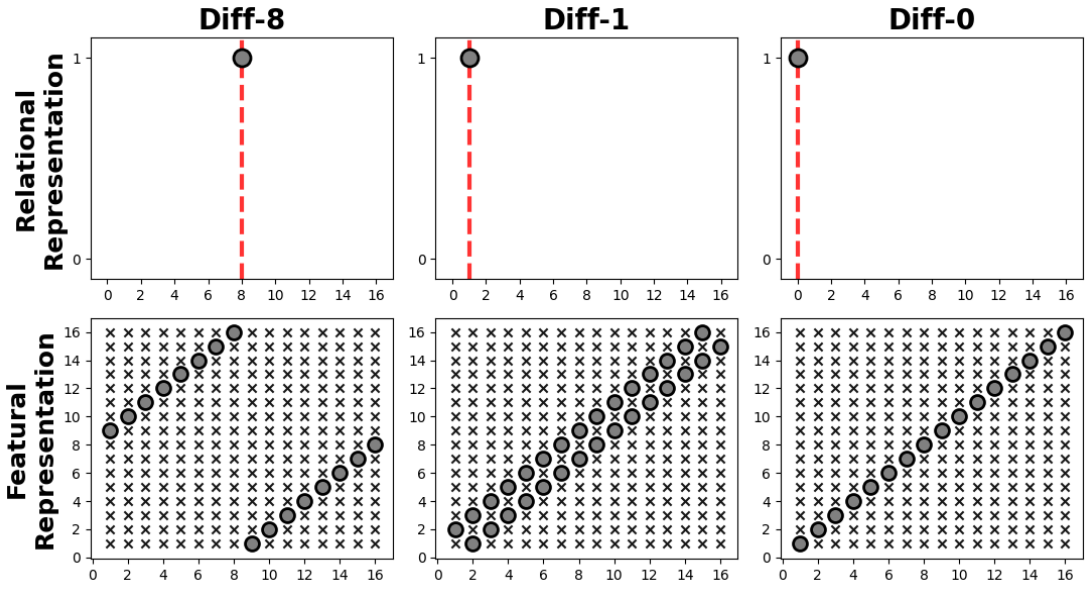

Figure 12: Category structures from experiment 1A. Top: a relational representation of the training stimuli plotted in a 1 dimensional space, defined by the *difference* in body length. Bottom: a purely featural representation of the stimulus pairs in a 2 dimensional space – with training items depicted as circles and generalization pairs depicted as x's. The y-axis represents the body length of the first fish in the pair, the x-axis represents the body length of the second.

Procedure

After approving an online consent form, subjects engaged in a series of unrelated experiments – one of which was the present study[24]. When subjects were assigned the present experiment, they were instructed that they'd be learning about which fish "go together" in the same tank. They were told there would be a test phase after training. They then proceeded to the training phase, which consisted of 2 blocks of observational learning. After training, they were told that in the next phase they would have to guess whether a presented fish pair could go together in the same tank. They then underwent 2 blocks of classification, with one pair of stimuli presented at a time, with no corrective feedback. This concluded the experiment.

Results & Discussion

In experiment 1a, the goal was to provide a description of generalization behavior post observational learning. Because there was no task during training, there is no direct measure of learning; the only metric we have for inferring whether subjects learned the correct category structure comes from their generalization decisions. Subjects' generalization decisions for new pairs that matched the *critical length difference* of the training items were used to assess learning accuracy. When looking at generalization decisions for test items matching the critical length difference at training, there appears to be little difference in accuracy between the *diff-8* (M = 93.056%, SD = 19.33%), *diff -1* (M = 91.912%, SD = 23.831%), and *diff-0* (M = 95.37%, SD = 14.629%) conditions, $F(2, 202) = .461$, $p = .631$. While a Bayesian

[24]This is true across all experiments.

analysis might be more appropriate given the shortcomings in using Frequentist statistics to support a null hypothesis (Kruschke, 2010), the learning accuracy is not necessarily of interest in the present study and will not be explored further. The key analysis of interest is the *shape* of the generalization gradient.

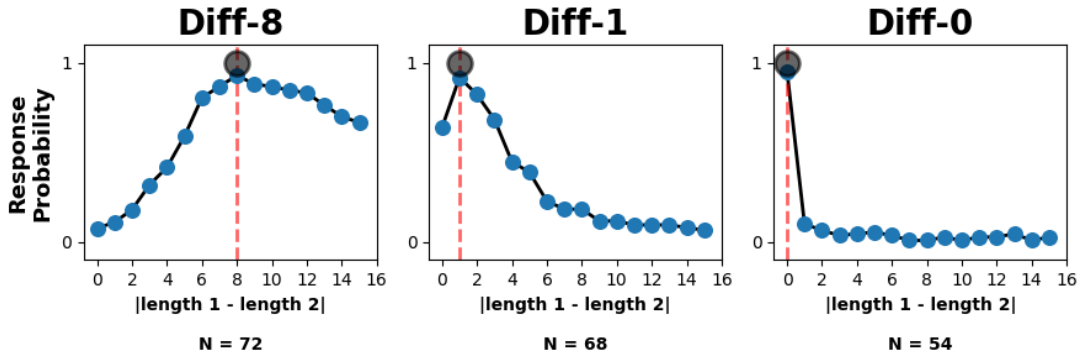

Figure 13: Generalization profiles for each of the conditions in experiment 1A, averaged across all subjects at each block.

Generalization Behavior To assess the overall shape of the generalization gradient across the possible *length differences* of stimulus pairs, 3 candidate models were chosen: a *radial basis* function, a *piecewise linear function*, and a *rectangular* function. Previous literature suggests that categorical generalization is best described by a function from the exponential family (e.g., an exponential function, a radial basis function, or a gaussian function) – this is why the *radial basis function* was chosen. The other two candidate models were chosen as theoretical alternatives that are close to the *radial basis function* in form but distinct enough to be valid alternatives – which would help rule out whether the present generalization data resemble expected behavior in a typical *feature-based* learning experiment (Nosofsky, 1986; Shepard, 1987).

Each candidate model has 2 parameters (corresponding to the 'center' of the function and the 'sharpness' of the function). Model fits were conducted via a

gridsearch that tests all possible parameter values within a certain range (with 1000 evenly spaced values within each parameter); the same range was used for each model (see Table 1). The adequacy of the model fit was operationalized as the *sum squared error* (SSE) between model predictions and average human behavior at each *length difference* humans observed during the test phase.

Table 1: Parameter ranges used for the search over *center* and *sharpness* of each function.

parameter	range
center	[-25, 25] (1000 steps)
sharpness	[-10, 10] (1000 steps)

For the *diff-8* condition, the *radial basis* function provided the closest fit to the average behavioral generalizations, while both the *square* and *piecewise-linear* functions provide relatively poor fits. For the *diff-1* condition, the *square* function produces the worst fit, followed by the *piecewise-linear* function. While the *radial basis* function provides the best fit to the data, it is not much better than the *piecewise-linear*; this is likely due to the asymmetry between points *0* and *2*. All 3 functions provide good fits for the *diff-0* condition, with the *piecewise-linear* and *radial basis* function providing the strongest. The *sharpness* parameter was much higher for both the best fitting *piecewise-linear* and *radial basis* functions, which highlights the much sharper generalization behavior exhibited by human subjects in the *diff-0* condition relative to the other conditions in experiment 1a. The *sharpening* of the generalization gradient in the *diff-0* condition may relate to

the observation that many subjects in a abstract concept learning task generalize *sameness* concepts *categorically* [a@young2001entropy] – discussed further in the general discussion.

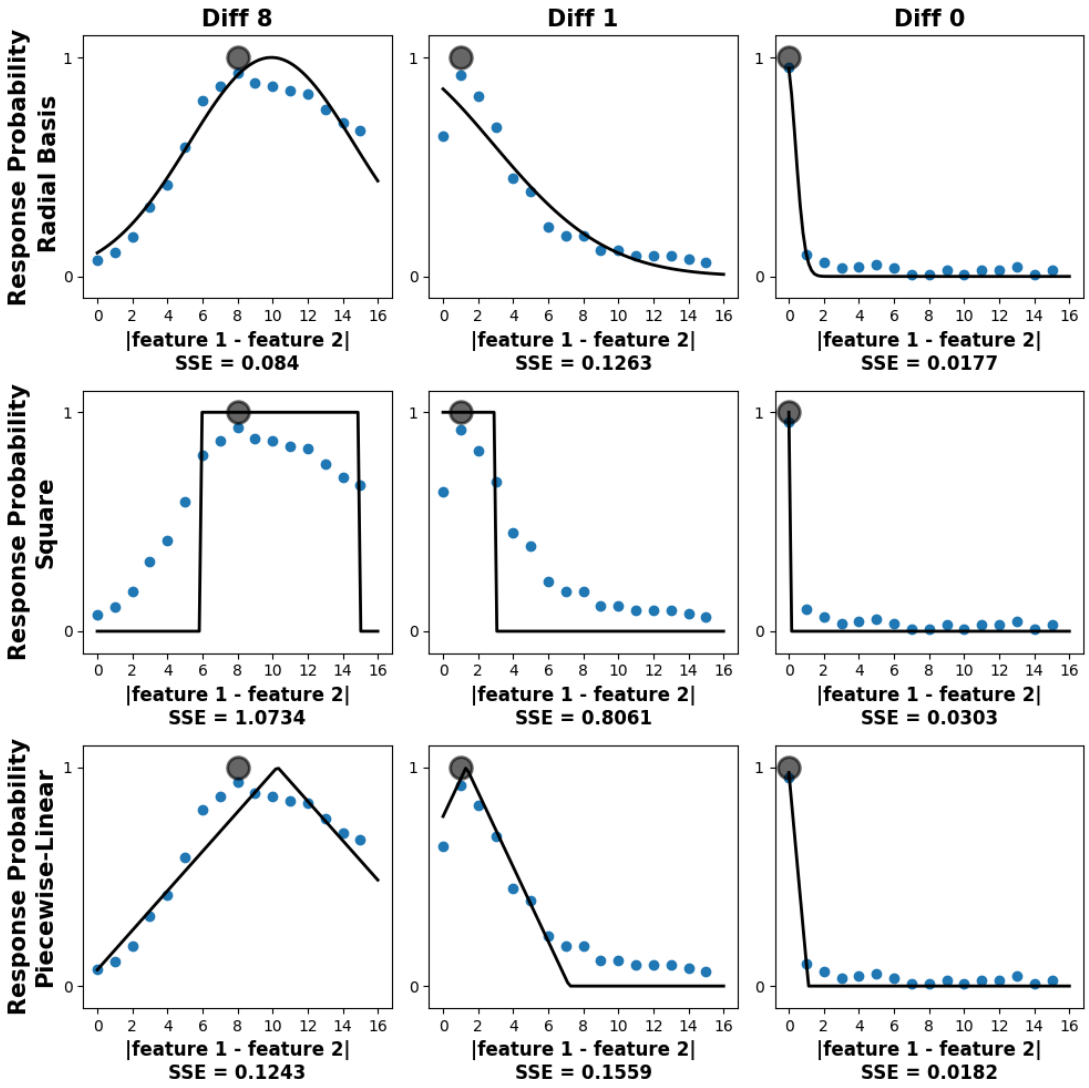

Figure 14: Visualization of model fits for each condition from experiment 1a.

Individual Profiles There is a particularly interesting asymmetry in the aggregate generalization data for the *diff-8* condition (see Figure 13). One possibility is that this asymmetry is a typical behavior of the *average* subject; however, further inspection (after the data were collected) suggested two particular generalization profiles: *rule-like* learners (who showed a sharp asymmetry) and *similarity-like*

learners (whose generalization behavior was roughly symmetrical). This trend mirrors what was observed in Lee et al. (2018), who trained subjects on similar category structures using a reinforcement learning paradigm (albeit with standard features instead of *relationally defined* features). Of particular interest was the quantity of rule-like learners in the present work. Learners were classified as *rule-like* if 6 out the 7 largest length-difference pairs were generalized as the target category learned during training.

This classification scheme resulted in 50 (out of 72) learners being classified as *rule-like* in the present work. Importantly, rule-like learning mechanisms have been widely prevalent in models of category learning (Ashby & Gott, 1988; Nosofsky, Palmeri, et al., 1994) – sometimes in conjunction with similarity-based mechanisms (Erickson & Kruschke, 1998). This may tentatively suggest that either (a) the learning mechanisms active in the present relational learning task are the same as those active in a traditional category learning study, or (b) the learning mechanisms are distinct – but leverage the same computational strategies. This rule-like behavior may also mirror the classification strategies used by subjects in the *diff-0* condition – though why this behavior was so pronounced in the *diff-0* condition is unclear[25].

Experiment 1b

In experiment 1a, subjects' generalization behavior *partially* resembled what's observed in category learning experiments (Nosofsky, 1989). One hypothetical explanation for this is that subjects are directly leveraging featural differences

[25]One possibility is that *exact sameness* may carry unique and valuable information useful for an animal's ecological fitness – though interestingly, there is still debate over whether nonhuman animals recognize the concept of *sameness* (Penn et al., 2008; Wasserman et al., 2004).

between stimulus pairs as the cue for classification decisions – and, that process follows the same phenomenological principles observed in standard feature learning experiments. This explanation predicts the generalization gradients observed in experiment 1a. However, it is important to note that existing theoretical models in the concept learning literature would predict a similar generalization profile even if they never explicitly computed the relation between features in the compound stimulus pair. To differentiate between these two explanations, experiment 1b used training stimuli sampled from a specific region of *feature space*. Then, the generalization stimuli were sampled from the *opposite* region of the *feature space* used in training. A featural account would predict generalization only in the region of feature spaced observed during training. If subjects were to continue to generalize the learned category beyond the feature space observed during training, this could be viewed as evidence against a purely featural account.

Participants

Subjects consisted of 135 undergraduate students from Binghamton University; students were compensated with partial credit towards a course requirement.

Materials & Design

Experiment 1b uses the same stimuli, materials, and preparation from experiment 1a. Again, subjects completed the experiment online using their own computers at any location of their choosing. The key difference is the restricted feature space used in training and test. In the training phase, stimulus pairs were only selected when their total length magnitude was less than 16 – this corresponds to the lower left diagonal in a 2D feature space (see Figure 15). The opposite

region was chosen for stimuli in the generalization phase. The same *diff-8*, *diff -1*, and *diff-0* category structures were used.

Training phase: Subjects were assigned to 1 of the 3 category structures and underwent 4 blocks of observational training (4 blocks were used to compensate for the reduced number of stimulus items relative to experiment 1a).

Test phase: The test phase was identical to the test phase of experiment 1a, except that stimulus pairs were sampled from the upper right diagonal of feature space.

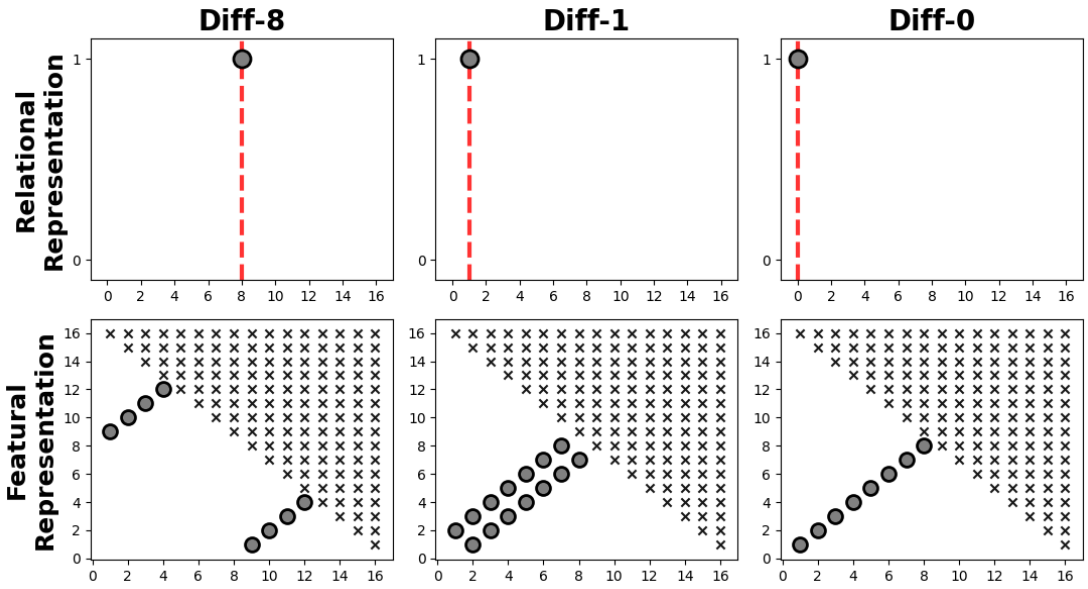

Figure 15: Category structures from experiment 1B. Top: a relational representation of the training stimuli plotted in a 1 dimensional *relational* space, defined by the *difference* in body length. Bottom: a purely featural representation of the stimulus pairs in a 2 dimensional space – with training items depicted as circles and generalization pairs depicted as x's. Generalization pairs were only sampled from the upper diagonal of the feature space to test against the possibility that subjects were representing the stimuli *featurally* instead of *relationally*.

Procedure

The procedure was the same as experiment 1a.

Results & Discussion

The purpose of experiment 1b was to rule out the hypothesis that featural information played a role in subjects' generalization responses. If it did, then we wouldn't expect to see differences in the generalization profiles of subjects in experiments 1a and 1b. The first analysis to consider is whether or not the reduced training space impacted subjects' accuracy towards generalizing the critical *pair length difference* observed during training. When looking at *test phase* generalization towards the critical training pairs, subjects' responses appeared accurate in all but the *diff-8* category structure. An analysis of variance testing the generalization accuracy across category structures was significant, $F(2, 134) = 46.429$, $p < .0001$. Post-hoc tests revealed that accuracy in the *diff-8* ($M = 40.5\%$, $SD = 33.9\%$) condition was significantly lower than both the *diff -1* ($M = 84.1\%$, $SD = 26.7\%$) condition, $t(\text{df}) = -6.759$, $p < .0001$, and *diff-0* ($M = 89.7\%$, $SD = 16.3\%$) condition, $t(\text{df}) = -8.807$, $p < .0001$. It was also apparent that the generalization profile for the *diff-8* condition differed between experiments 1a and 1b. Specifically, generalization for the critical training *relation* in the *diff-8* condition was less accurate in experiment 1a ($M = 93.1\%$, $SD = 25.5\%$) than in experiment 1b ($M = 39.3\%$, $SD = 48.9\%$), $t(107) = 10.73$, $p < .0001$.

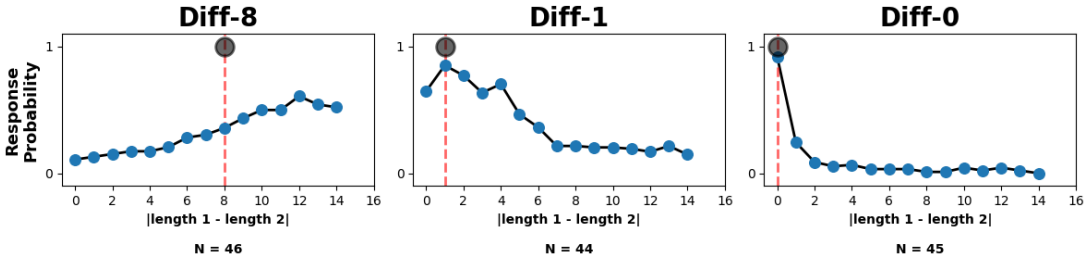

Figure 16: Generalization profiles for each of the conditions in experiment 1B, averaged across all subjects at each block.

There were no *a priori* predictions about how the manipulation in experiment 1b would affect the ratio of similarity- and rule-like generalization profiles in the *diff-8* condition. Nevertheless, an exploratory analysis (using the same classification scheme from experiment 1a) found 9 learners with *rule-like* generalization profiles (compared to 36 in experiment 1a). This might suggest that the rule-like learners in particular were using primarily *relational* cues during learning (rather than *featural* ones) – since they extrapolated the learned rule outside of the feature space observed during training. It was also apparent that the generalization profile for the *diff-0* condition was less sharp than in experiment 1b. When looking specifically at the responses for stimulus pairs with a length difference of *1* (closest to the target difference *0*), we see that the mean probability of generalizing the stimulus pair as a positive instance of the relation observed during training was significantly higher in experiment 1b ($M = 10.2\%$, $SD = 30.4\%$) than in experiment 1a ($M = 20.7\%$, $SD = 40.6\%$), $t(111) = 2.215$, $p = .029$ – though this was an entirely post-hoc observation.

Experiment 1c

Experiments 1a and 1b were useful for (1) establishing a baseline of generalization in the present domain, and (2) helping delineate the competing predictions of purely featural versus purely relational mechanistic explanations. Leading models from the concept learning literature make additional phenomenological predictions about generalization – like the prediction of rule-like generalization behavior (Ashby & Gott, 1988; Erickson & Kruschke, 1998; Nosofsky, Palmeri, et al., 1994). However, models in the concept learning literature *typically* learn using a classification

learning scheme (*not* using the observational learning mode used in experiments 1a and 1b). Experiment 1c used a training phase where subjects learned via supervised classification. By more closely matching the intended learning task of models in classification learning literature (Kruschke, 1992; Kurtz, 2007), we could assess how well theories of featural learning explain human learning and generalization of *featural relations*.

Participants

Subjects consisted of 152 undergraduate students from Binghamton University; students were compensated with partial credit towards a course research-participation requirement.

Materials & Design

Experiment 1c uses the same stimuli, materials, and online preparation from experiments 1a and 1b. A crucial difference was that subjects learned the categories via trial-and-error classification, where corrective feedback about true category membership was provided after a guess was made on each trial. Four category structures were selected. In the *1-8 structure*, one category of stimuli varied by 1 length unit and the other varied by 8 length units. This means that during training, subjects viewed fish pairs whose length difference varied by either 1 units or by 8 units; the correct classification was based on these length differences. In the *2-8* structure, one category of stimuli varied by 2 length units and the other varied by 8 length units. In the *7-8* structure, one category of stimuli varied by 7 length units and the other varied by 8 length units. Note that stimulus pairs from both categories in the 7-8 condition were highly similar (i.e., confusable). Due to

the very low learning performance observed in the first 15 subjects, data collection for the 7-8 condition was halted and instead a 4-8 condition was tested instead[26] (where the categories are less confusable). The same restricted stimulus sampling procedure from experiment 1b was used: training stimuli were sampled from the lower left diagonal of feature space and test stimuli were sampled from the upper right of feature space.

Training phase: Subjects were assigned to 1 of the 4 category structures and underwent 4 blocks of supervised classification training. On each trial, subjects are presented with a stimulus pair, and asked to guess whether or not the 2 fish could exist in the same tank (given two options labeled "Yes" and "No"). After making a guess, they are told whether or not their guess was correct. After subjects completed a minimum of 20 trials, the experiment was ended early *if* they correctly categorized 19 of their last 20 trials, where it was assumed they had or would have reached the solution[27].

Test phase: The test phase was the same as in experiments 1a and 1b, consisting of classification without corrective feedback.

Procedure

The procedure was the same as experiments 1a and 1b, except that the training phase consisted of classification learning.

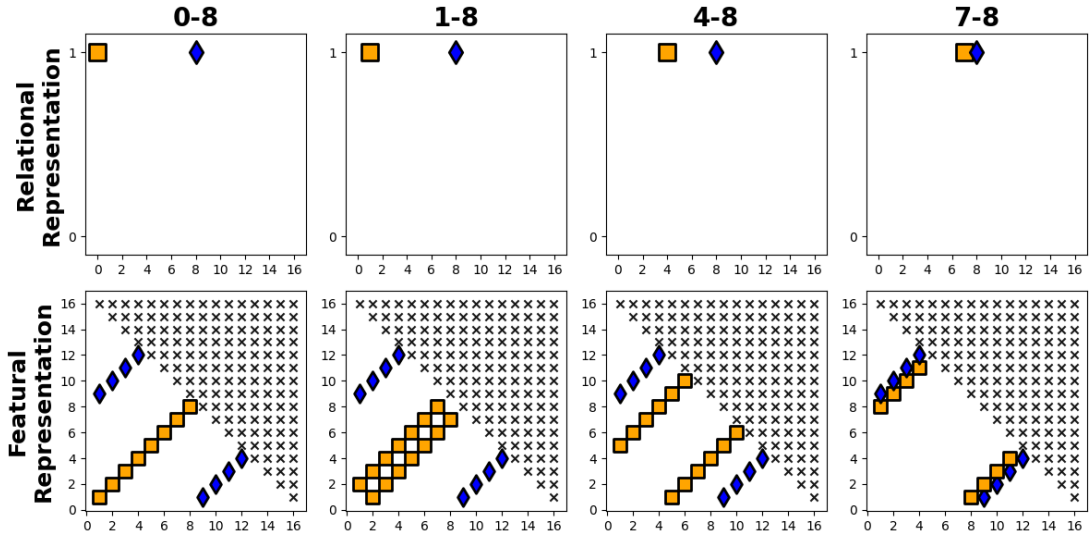

Figure 17: Category structures from experiment 1C. Top: a relational representation of the training stimuli plotted in a 1 dimensional space, defined by the *difference* in body length – orange squares represent one category while blue diamonds represent the other. Bottom: a purely featural representation of the stimulus pairs in a 2 dimensional space – with generalization pairs depicted as gray x's.

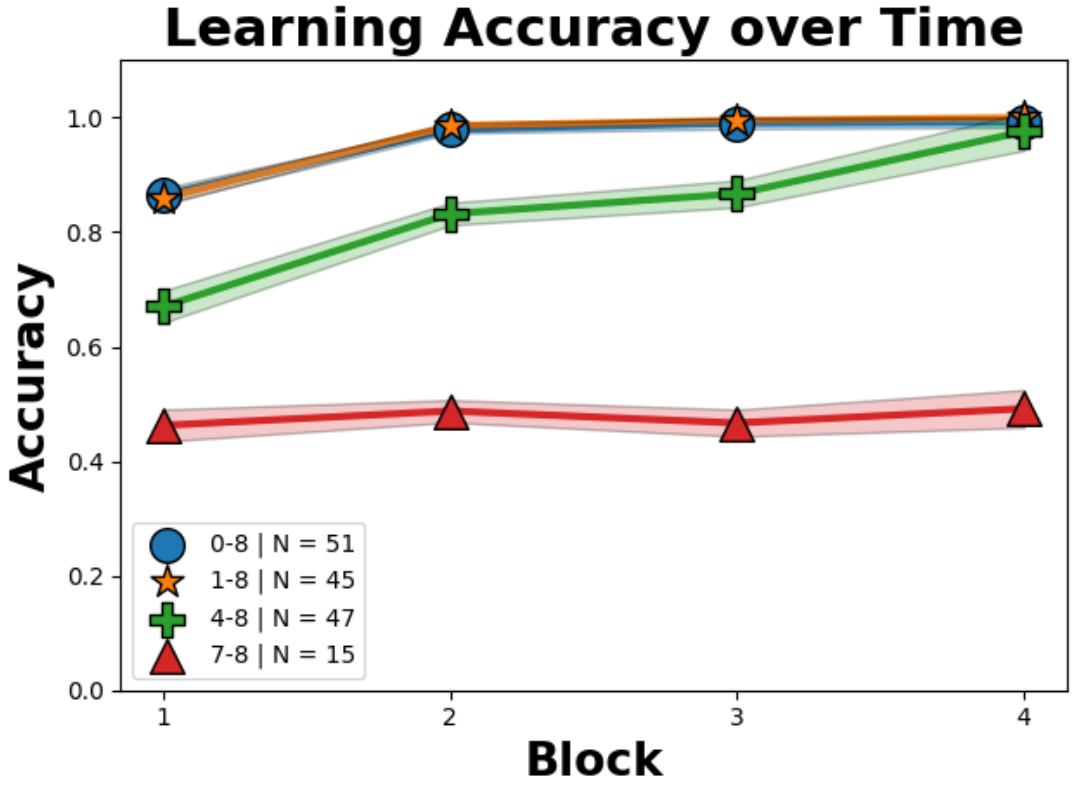

Figure 18: Plot of the learning accuracy for each block of training in each condition of experiment 1c. Width of the error bands represent -/+ the *standard error of the mean*.

Results & Discussion

The goal of experiment 1c was to see how generalization behavior shifts when subjects learn via a *classification* task (rather than the *observational* task utilized in experiments 1b and 1c). A few observations immediately stand out. First, it appears that the *0-8* and *1-8* conditions are dominated by *rule-like* generalization. Using the same criterion for selecting rule-learners from experiments 1a and 1b, we observe 47 (out of 51) rule-learners in the *0-8* condition, and 43 (out of 45) rule-learners in the *1-8* condition. One interpretation of this result is that *classification* learning drives learners towards *more discriminative* learning strategies[28] – a phenomenon that echoes what's been discovered in the feature-based category learning literature (Chin-Parker & Ross, 2004; Levering & Kurtz, 2015). In fact, this was an *a priori* motivation for moving to a classification learning phase in experiment 1c. When comparing the number of rule-like generalizers in the *diff-8* condition from experiment 1a and in the *0-8* condition from experiment 1c, a between-experiment chi-square test of independence was significant, $\chi^2\ (1) = 7.926$, $p = .005$[29].

Interestingly, there were only 34 (out of 47) rule-like learners in the *4-8* condition, and only 1 (out of 15) in the *7-8* condition. Because the *7-8* condition was cut short and does not have an adequate number of subjects to be meaningfully analyzed, it

[26]Note: while the raw data was observed throughout the study, no inferential tests were made at any point until data from all conditions in experiment 1c were collected.

[27]What would have been the remaining trials after ending early were coded as being perfectly accurate during the later analysis.

[28]In this case, we would define rule-like generalization as *more discriminative*, since it requires diagnostic but not necessarily prototypical category knowledge (Chin-Parker & Ross, 2004).

[29]The *diff-8* (experiment 1a) and *0-8* (experiment 1c) conditions wqere chosen *post-hoc* for this comparison because they were highly similar except in regards to the learning task and presence of a 2nd category.

will not be discussed further[30]. The reduced number of rule-like learners in the *4-8* and *7-8* conditions may reflect increased difficulty in understanding what the category structure was – further evidenced by the low accuracy when generalizing the critical *relation* observed during training.

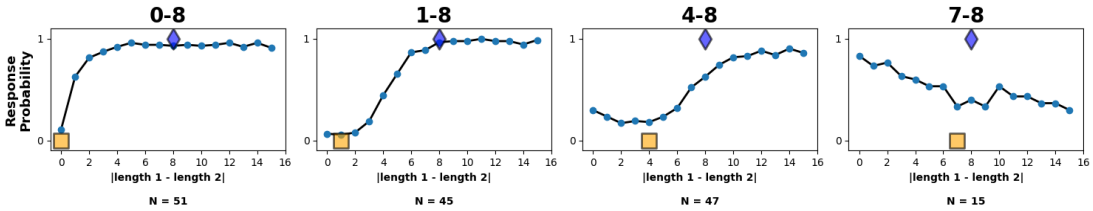

Figure 19: Generalization profiles for each of the conditions in experiment 1C, averaged across all subjects at each block.

Experiment 2: Relational Extension of Kruschke (1993)

Experimental set 1 provided insight into the baseline generalization behavior for both observational and classification training schemes for categories defined by featural differences between stimuli with corresponding latent dimensions. Importantly, because the stimuli only varied on 1 feature (length), the *relational space* that described the stimulus domain was 1 dimensional. Many key models from the feature-based concept learning literature (Anderson, 1991; Kruschke, 1992; Kurtz, 2007; Love et al., 2004) have been empirically tested using stimuli with multiple features, and make either explicit or implicit predictions about how multiple features are integrated in a classification decision. A specific prediction is

[30]Though interestingly, the aggregate generalization profile in *7-8* seemed to resemble a linear generalization anchored at 0; this might be indicative of a prior response behavior that's leveraged when the learning task is too difficult or noisy. Though again, the limited sample size warrants caution with this interpretation.

that subjects tend to restrict their attention towards a small set of highly predictive features; this has been referred to as *selective attention*. This would explain a well-replicated phenomenon where category structures are learned quicker when they can be differentiated on the basis of a small set of highly predictive features (Kruschke, 1993; Shepard et al., 1961).

The purpose of experiment 2 was to extend a classic demonstration of selective attention from Kruschke (1993) , where subjects learned to discriminate between 2 feature-based categories defined over a 2-dimensional stimulus space. In one condition, the objects could be categorized on the basis of a single feature, while in the other, both features were required for learning (see Figure 21). Experiment 2 uses the same category structures and training scheme as Kruschke (1993), except the learning materials were stimulus pairs instantiated in a 2D *relational space* instead of a 2D *feature space*. Similar to Kruschke (1993), one of the category structures in this space could be solved by comparing stimuli on the basis of *1 feature*; the other could not. The question here was whether performance would be stronger for the category structure with an *attentional* solution, which would support the notion that *selective attention* pressures are present in a relational learning domain.

Participants

Subjects consisted of 85 undergraduate students from Binghamton University; students were compensated with partial credit towards a course research-participation requirement.

Materials, Design, & Procedure

All experiments were run in-person on a computer display with a 1440 x 900 pixel resolution. The experiment display consisted of visual objects (buttons, shapes, text boxes) in a web browser, and could be navigated through button clicking or through keyboard presses. Stimuli were generated using the oCanvas library (Koggdal, n.d.). Subjects were told that they were playing the role of an "interior designer", and that their task was to determine whether a lamp pair belongs in (a) a museum, or (b) an office (Figure 20). After training, subjects engaged in an unsupervised test phase where they were not given corrective feedback. The stimuli for experiment 2 consisted of lamps that could vary on the basis of 2 features: height and shading. The height and shading could vary between any integer value from 0-5. On the computer display, the true height ranged from 3 - 9.5 cm, while the pixel shading varied from 5 - 225 grayscale.

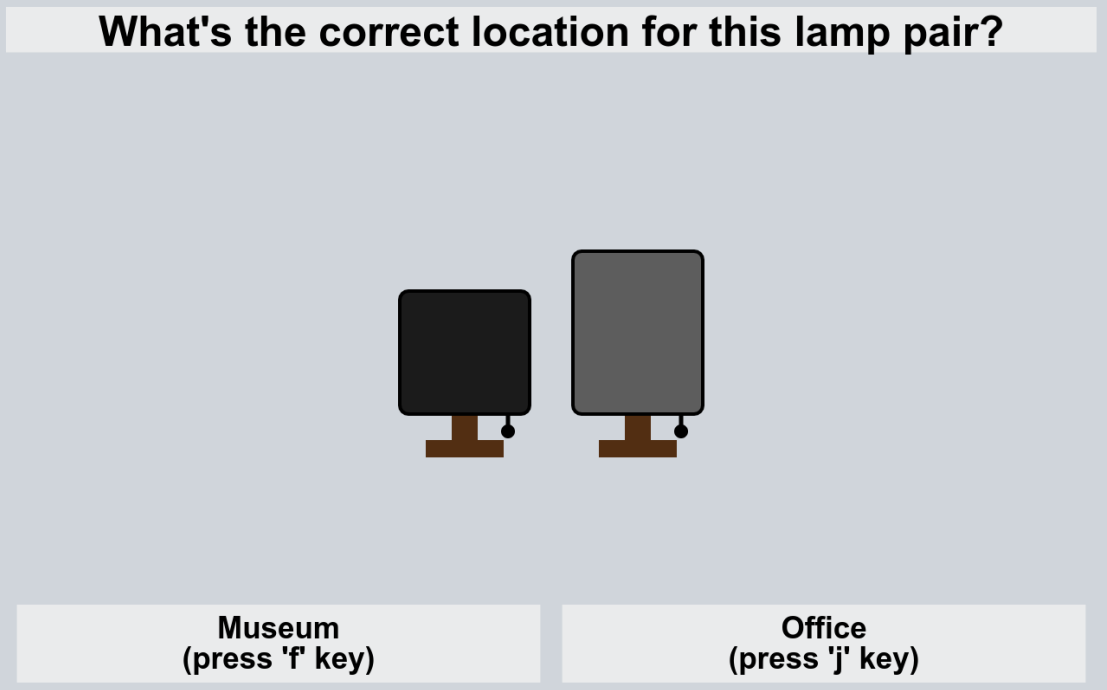

Figure 20: Screenshot of a training trial in experiment 2; taken on a screen with ~ 13 inch (~ 33cm) display.

Subjects were assigned to 1 of 3 categories. In the *filtration* condition, each category was on opposite sides of the 2D *relational space*; that is, it could be solved on the basis of 1 feature comparison. For example, all of the pairs in one category might have the property 'similar shading', while all pairs in the other category might have the property 'different shading'[31]. In the *condensation* condition, one category was on the lower left diagonal of the 2D space, while the other was in the upper right diagonal; consequently, it could not be solved on the basis of 1 feature comparison. In the *condensation-flipped* condition, one of the coordinates for the two categories were flipped. While this category structure doesn't have an attentional solution, it does have the particular property that objects in one category are *featurally similar* to each other, while objects in the other category are *featurally dissimilar* to each other. That is, pairs from one category had the property 'similar height and similar shading', while pairs from the other category had the property 'different height and different shading'. Interestingly, success in this condition would suggest that subjects consider aggregate similarity between feature comparisons as a basis for classification.

Training phase: The training phase consisted of 6 blocks of supervised classification training. Each block consisted of 1 exposure to a stimulus pair that varied in featural magnitude according to the coordinates of that stimulus's position in Kruschke (1993)'s original category structure (see Figure 21). Because multiple unique stimulus pairs can instantiate the same *featural relation*, only 1 appropriate stimulus pair was sampled in each block[32]. During each trial, subjects were pre-

[31]Note: the 2 features (height and shading) were counterbalanced randomly for each subject to mitigate the effect of any particular feature being more or less salient than another.

[32]This also resulted in a particular issue that occurred during the design of all 3 sets of experiments: when using a consistent range of feature values to sample from, different *relations* can have a *wider* or *narrower* range of stimulus pairs that can instantiate them. This means

sented with a pair of lamps and a set of category labels (*museum* or *office*); they were given the correct answer after making their guess. After subjects completed a minimum of 20 trials, the experiment was ended early *if* they correctly categorized 19 of their last 20 trials, where it was assumed they had or would have reached the solution.

Test phase: The test phase was similar to the training phase, except: (1) there was only 1 block of training, and (2) there was no corrective feedback.

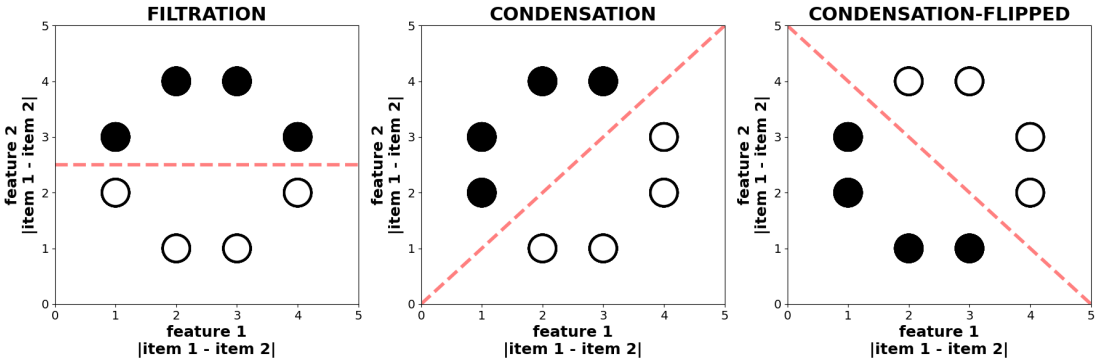

Figure 21: Visualization of the stimulus space in experiment 2. Each point represents a pair of stimuli. The x-axis represents the difference in the stimulus pair on feature 1, while the y-axis represents the difference on feature 2. Each structure consists of 2 categories – labeled by shading (black or white). Unlike the *condensation* and *condensation-flipped* condition, the *filtration* condition can be solved on the basis of a single featural difference.

Results & Discussion

The purpose of experiment 2 was to demonstrate whether evidence of selective attention pressures exist in a learning domain characterized by learning multiple, featural relations – a classic phenomenon in the category learning literature (Kruschke, 1993; Shepard et al., 1961). Of interest is whether the *filtration* structure –

featural information could be used as a predictive cue for categorizing *some* of the stimulus pairs in the category structure (though not all). This issue is addressed somewhat in the computational experiments discussed later in this manuscript, where models are given (a) features, (b) relations, or (c) *both* features and relations.

which can be solved on the basis of 1 featural relation – is learned faster and more accurately than the *condensation* condition. Test phase generalization accuracy (after training) showed that accuracy in the *filtration* condition (M = 64.73%, SD = 24.1%) was not significantly higher than in the *condensation* condition (M = 55.8%, SD = 22.5%), t = 1.406 , p = 0.165. Interestingly, accuracy in the *condensation-flipped* condition (M = 68.53%, SD = 20.4%) was significantly higher than in the *condensation* condition (M = 55.8%, SD = 22.5%), t = 2.198 , p = 0.032. Finally, there was no significant difference in the *filtration* condition (M = 64.73%, SD = 24.1%) compared to the *condensation-flipped* condition (M = 68.53%, SD = 20.4%), t = -0.632 , p = 0.53. The average learning curves are shown in Figure 22.

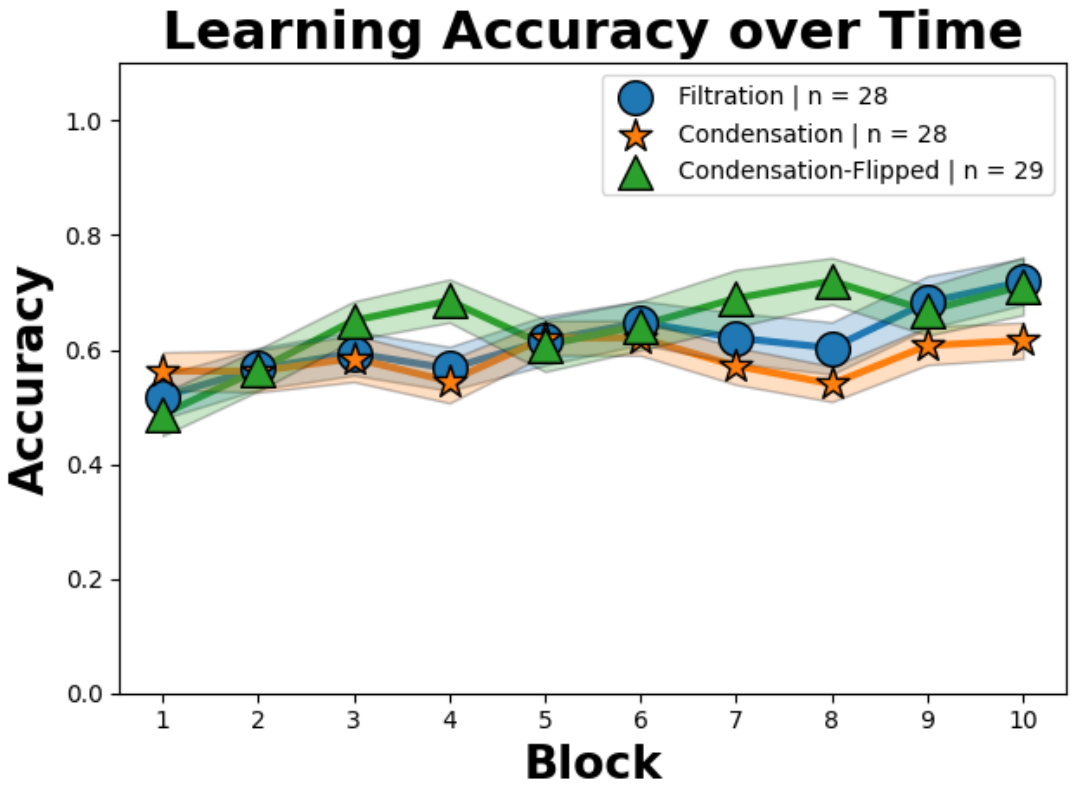

Figure 22: Plot of the learning accuracy for each block of training in each condition of experiment 2. Width of the error bands represent -/+ the *standard error of the mean.*

The test phase accuracy alone does not provide strong support for a potential effect of attention. However, when looking at the learning curves of each subject in each condition, there is little evidence that anyone figured out the problem solution in the *condensation* condition (see Figure 23). In contrast, a relatively meaningful number of subjects in the *filtration* condition seemed to have 'solved' the problem, evidenced by subject learning trajectories that reached sustained, ceiling performance. An even greater number of subjects appeared to have reached ceiling in the *condensation-flipped* condition. While the *condensation-flipped* condition couldn't be solved on the basis of one featural relation alone, it could have been learned by considering the *total similarity* of the presented stimulus pair (unlike in the standard *condensation* condition).

The notion of whether or not subjects consider *total similarity* (or *family resemblance*; Rosch & Mervis, 1975) has been a major debate in the classic category learning literature (Medin et al., 1987; Wills et al., 2013). Many theories invoke *family resemblance* as an important explanatory construct, arguing that many real-world category structures have a *family resemblance* structure. However, there is very little evidence that humans leverage *total similarity* in laboratory experiments without heavy-handed support where stimuli have few, independent, easily identifiable features (Medin et al., 1987; Wills et al., 2013). However, the present work differs from a standard category learning task in that the *relational features* that subjects were encouraged to leverage were all instantiations of the same latent variable: *similarity*. It could be that human learners cannot easily consider *total similarity* in situations where the latent variables that make up the learning cues are distinct (which is the case in a standard category learning

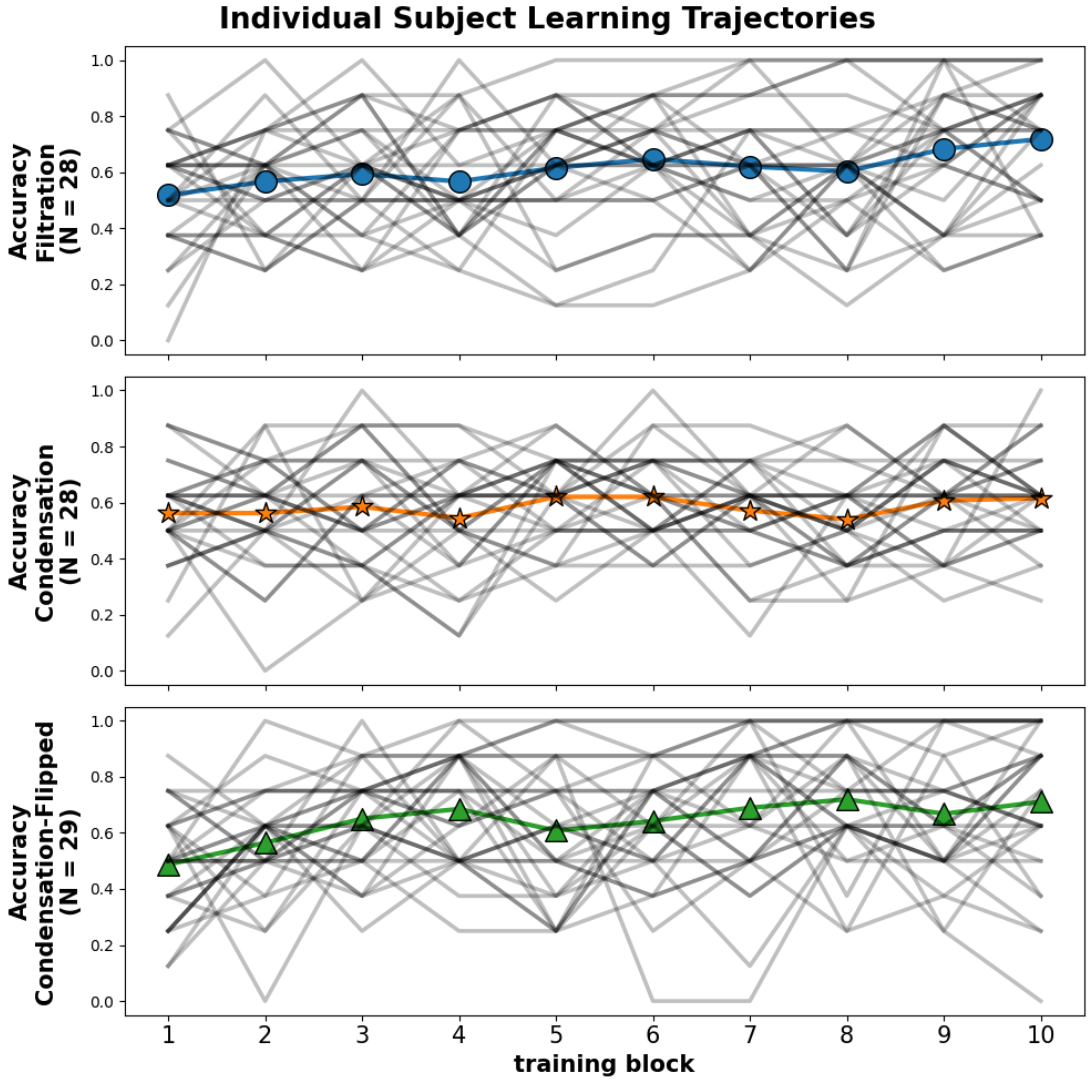

Figure 23: Plot of the learning accuracy for each subject in each block of training in each condition of experiment 2.

experiment)[33].

Experiment 3: Relational Extension of Shepard et al. (1961)

A critical finding from the category learning literature comes from Shepard et al. (1961) – replicated by Nosofsky, Gluck, et al. (1994), Rehder & Hoffman (2005) and Lewandowsky (2011), and partially by Kurtz, Levering, et al. (2013). In Shepard et al. (1961)'s demonstration, subjects were trained to classify stimuli with 3 binary features into 2 mutually exclusive categories – using 6 different category structures. Human learning on these 6 category learning problems (or, *structures*) has been a widely used benchmark for testing computational models of categorization (Crump et al., 2013; Kurtz, 2007; Kurtz, Levering, et al., 2013; Lewandowsky, 2011; Love et al., 2004; Nosofsky, Gluck, et al., 1994), namely because they produce a well-replicated order of learning difficulty with human subjects. The difficulty of learning the 6 problems can often be predicted using a number of interesting complexity metrics, such as: boolean complexity (Feldman, 2000) or information entropy (Pape et al., 2015).

In experiment 3, the 6 category structures from Shepard et al. (1961) (referred to as the 6 SHJ structures) were instantiated using relation *feature comparisons* instead of absolute features values. Subjects classified pairs of stimuli that vary on

[33]Importantly, we do not know for sure that subjects learning the *condensation-flipped* structure in experiment 2 are leveraging a recognition of *total similarity*. It may instead be the case that subjects are leveraging a single-feature hypothesis testing approach that only focuses on the predictive value of a single relational feature (like we expect in a classic category learning experiment; Nosofsky, Palmeri, et al., 1994). In order to tease apart what strategy subjects are employing, future work might benefit from a self-report questionnaire simply asking them.

3 binary dimensions; category membership was defined by feature *relations* instead of features themselves. One goal of experiment 3 was to provide an empirical benchmark for computational mechanisms of relational learning. Another goal was to investigate whether the ordering of learning difficulty shifts from what's expected in a feature-based learning preparation– adding to the general discussion of mechanistic overlap across both featural and relational domains.

Participants

Subjects consisted of 117 undergraduate students from Binghamton University; students were compensated with partial credit towards a course research-participation requirement.

Materials & Design

Subjects completed the experiment in an in-person laboratory, meaning the monitor size and color profile was exactly consistent across all subjects. The stimuli used for experiment 3 are objects varying on 3 binary features (lamp shades varying on their shape, shading, and size); this results in 8 unique possible stimuli. Each stimulus is represented based on the state of each binary feature it instantiates. There are 36 unique pairwise stimulus combinations (order invariant) that can be derived from the 8 stimuli. Each stimulus pair can be recoded according to which features match and mismatch; in this paper, we use 0 to indicate that a feature matches, and 1 to indicate that a feature mismatches. There are 8 possible recodings of stimulus pairings, each of which can be instantiated with 4 unique pairs (with the exception of the identity pairing, 000, which can be instantiated

with 8 unique pairs). In the traditional SHJ category structures, the category of a stimulus is determined by its featural representation. In the present work, we instead define the category of a stimulus pair according to its *comparison recoding*, using the same category assignment structure originally used by Shepard et al. (1961). The critical measure is how easy each category structure is to learn throughout and after training.

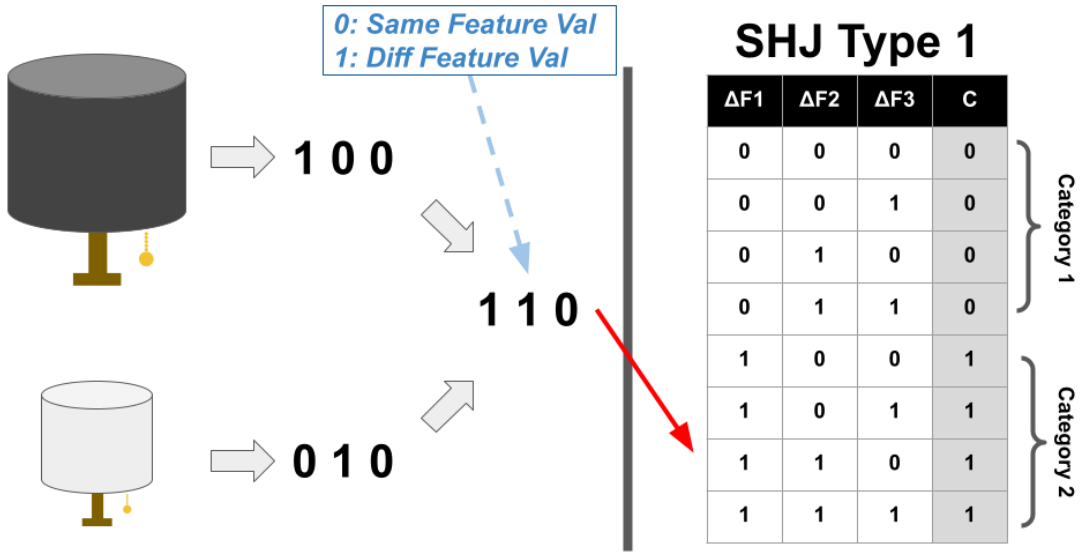

Figure 24: Example of how the stimulus pairs from experiment 3 were mapped onto the original category structures used by Shepard et al. (1961). Each stimulus pair is recoded based on whether the stimuli match on any of their 3 features. The recording is what was used to determine whether or not the stimulus pair belongs to one of 2 possible categories. Note: the 3 features (size, shading, shape) were counterbalanced randomly for each subject to mitigate the effect of any particular feature being more or less salient than another. In the table to the right, each feature column represents whether the stimulus pair matches or mismatches on that feature; the C column represents the category the stimulus pair belongs to.

Procedure

Subjects were told that they were playing the role of an interior designer, and that their task was to determine if each pair of lamps would make an "acceptable

arrangement" (see Figure 25)[34]. They were also told that after a series of training trials, they would participate in a test phase to see how well they learned.

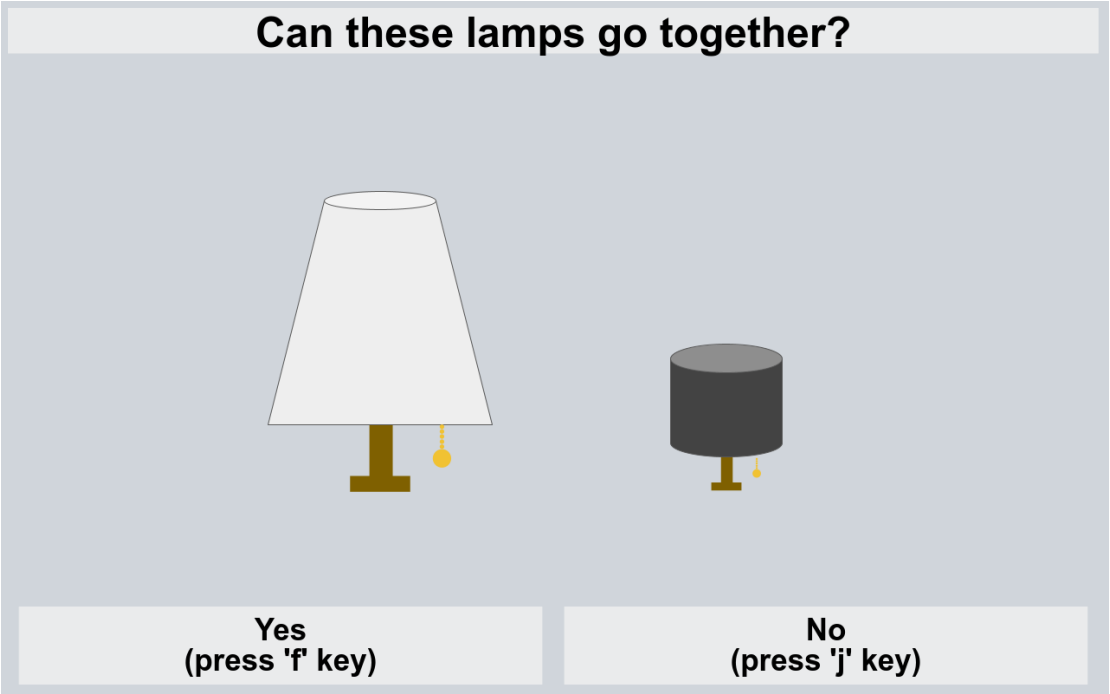

Figure 25: Screenshot of a training trial in experiment 3; taken on a screen with ~ 13 inch (~ 33cm) display.

Training Phase: In the training phase, subjects engaged in 10 blocks of supervised classification training (using the exact same preparation as experiment 2). Each block consisted of 8 trials; on each trial, a stimulus pairing was sampled from 1 of the 8 possible *comparison recodings*. All 8 *comparison recodings* were shown once per block. After subjects completed a minimum of 40 trials, the experiment ended early *if* they correctly categorized 37 of their last 40 trials, where it was assumed they had or would have reached the solution.

Test Phase: The test phase consisted of 1 block of classification; the procedure was the same as the training phase except no feedback was provided.

[34]This is sometimes referred to as an A / ~ A task, since there aren't necessarily 2 categories being learned (rather, subjects learn that lamp pairs are either a positive or negative instance of "acceptable arrangement"). Note that this design choice differs slightly from previous feature-based investigations where subjects learned to distinguish between 2, unique categories (Kurtz, Levering, et al., 2013; Nosofsky, Gluck, et al., 1994; Shepard et al., 1961).

Table 2: Stimulus features and corresponding category labels for each of the 6 category structures (last 6 columns). Each feature column (F) represents whether or not the stimulus pair shares the same value (0: same, 1: different).

F1 diff	F2 diff	F3 diff	**Type 1**	**Type 2**	**Type 3**	**Type 5**4	**Type 5**	**Type 6**
0	0	0	**0**	**0**	**0**	**0**	**1**	**0**
0	0	1	**0**	**0**	**0**	**0**	**0**	**1**
0	1	0	**0**	**1**	**1**	**0**	**0**	**1**
0	1	1	**0**	**1**	**0**	**1**	**0**	**0**
1	0	0	**1**	**1**	**0**	**0**	**0**	**1**
1	0	1	**1**	**1**	**1**	**1**	**1**	**0**
1	1	0	**1**	**0**	**1**	**1**	**1**	**0**
1	1	1	**1**	**0**	**1**	**1**	**1**	**1**

Results & Discussion

If the cognitive mechanisms theoretically invoked in the classic category learning literature are *similar to* the mechanisms involved in relational inductive learning, then there should be overlap in the *learning difficulty ordering* of the 6 category structures learned in a featural and relational stimulus domain. Upon inspection, this seems to *partially* be the case. The easiest structures learned appear to be *type 1* (which is almost always the easiest SHJ type to learn in the category learning

literature) and *type 4* (which is often much more difficult than type 1). The hardest structure to learn was *type 6*, which is similar to what's found in the category learning literature. Types 2, 3, and 5 seemed to be learned at roughly the same rate – slower than types 1 and 4 but faster than type 6. A one-way ANOVA testing against the hypothesis of no difference between the *test phase* accuracy for the 6 category structures was significant, $F(5, 112) = 9.501$, $p < 0.001$.

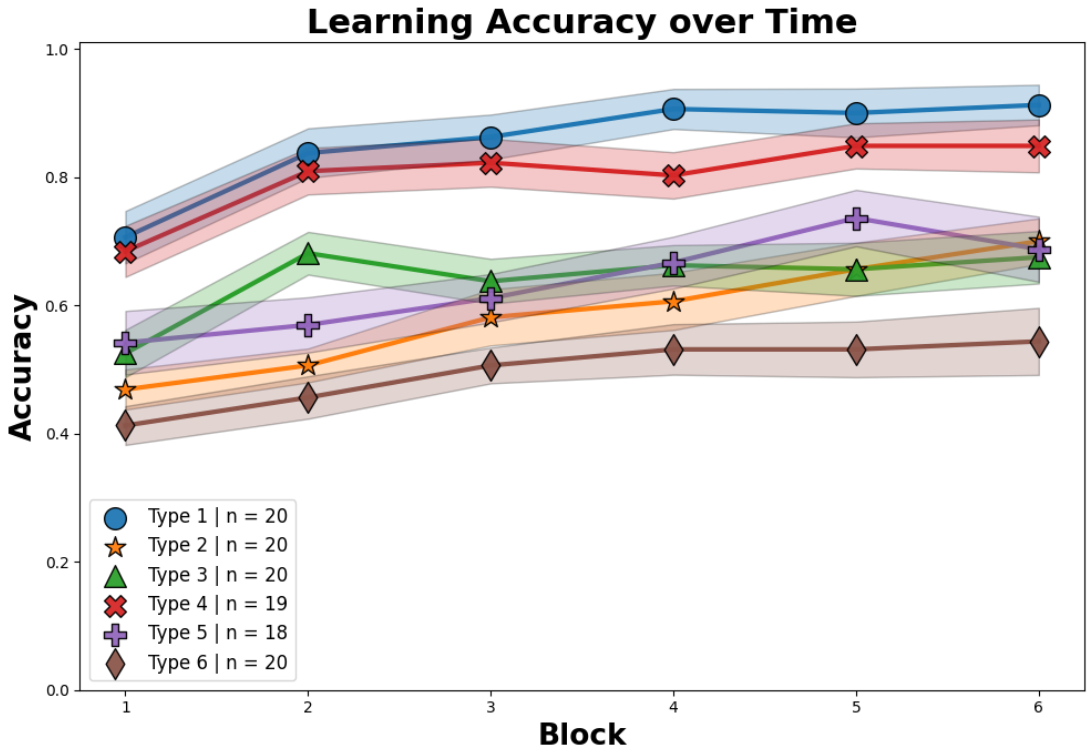

Figure 26: Plot of the learning accuracy for each block of training in each condition of experiment 3. Width of the error bands represent -/+ the *standard error of the mean*.

While partially mirroring what's found in the classic category learning literature, the present findings deviate from expectations in 2 important ways. First, we see that the type 2 structure – which is usually easier to learn than types 3, 4, 5, and 6 and harder to learn than type 1 – is not learned very quickly. Instead, the type 2 structure is learned relatively poorly (at roughly the same rate as types 3 and 5). In the classic category learning literature, the type 2 structure can be solved by

learning an *exclusive-or* (XOR) function mapping feature values to category labels – which many subjects seem to learn somewhat quickly. However, when subjects are not given *explicit instructions* to consider logical rules during learning, they tend to learn type 2 much slower (Kurtz, Levering, et al. (2013)). This might suggest that subjects in the present learning domain are not inclined to use rules as a basis for learning (though that contradicts the evidence from experiments 1a, 1b, and 1c). Another possibility is that the rule required to solve type 2 is relatively complex, and extrapolating a complex rule using *featural relations* (instead of raw feature values) is beyond the working memory capacity of subjects in our experiment.

Another way in which the data from the present work deviate from the expected ordering is that the type 4 structure was learned very quickly. One hypothesis is that subjects were quickly learning to leverage *total similarity* (or, *family resemblance*) to solve the type 4 structure. This instantiation of the type 4 structure[35] had the specific property that stimulus pairs that were *mostly similar* were in one category, and stimulus pairs that were *mostly different* were in the contrasting category. This explanation mirrors what may have happened in the *condensation-flipped* condition in experiment 2. Again, the use of *total similarity*– while theoretically popular – is very difficult to find in human learners during a standard laboratory category learning experiment. Though as discussed in experiment 2, the *relational features* in the present work embody the same latent construct (similarity), which might be a condition required for humans to leverage *total similarity* in an inductive learning task[36].

[35]which had the potential of being instantiated in other ways

[36]One way to test whether total similarity was leveraged when learning the type 4 structure in experiment 3 is to instantiate type 4 in a way where *total similarity* is no longer predictive of category membership. If learning accuracy worsens, this suggests that *total similarity* was the latent construct subjects were using to classify the stimulus pairs. Another useful approach

might be to just use subjects' own self-report of their learning strategy.

Computational Simulation

The behavioral investigation of the present work aimed to describe how human subjects learn categories in the presence of featural and relational cues. In some cases, the behavioral data suggest that subjects were leveraging *relational* feature comparisons as the basis for classification decisions. In other cases, subjects may have been attending to featural cues to solve the task, possibly in conjunction with the relational cues as well. In other cases further, subjects may have been attending to *aggregate similarity* across feature comparisons. This computational investigation tested whether existing computational accounts of feature learning could predict these empirical results.

First, 2 distinct connectionist models of category learning (Kruschke, 1992; Kurtz, 2007) were selected as a way to test a range of theoretical mechanisms that have been used to explain human feature -> category induction. Second, the present work tested the predictions of 3 potential computational mechanisms of stimulus representation: (1) a purely featural representation, (2) a purely relational representation that represents differences in feature values, and (3) a mix of both representations (concatenated). The empirical plausibility of each model and each representational scheme were assessed by the relative success at predicting the generalization decisions of human subjects from the present work.

Note: in the Computational Simulation chapter, the Models and each Procedure subsections were written before any simulations were conducted.

Models

Both of the models discussed are feed-forward Connectionist networks (Rumelhart et al., 1986). This particular framework was chosen for two reasons: (1) connectionist networks have had a long history of instantiating computational theories from the category learning literature (Gluck & Bower, 1988; Kruschke, 1992; Kurtz, 2007), and (2) these connectionist models in particular can be easily extended to use any of the 3 candidate stimulus representations in the present work. In all cases, the models are provided with some numeric representation of a pair of stimuli; this information is then propagated through some architecture of network weights to eventually produce some classification decision. The weights of the network are optimized with backpropagation to minimize prediction error (i.e., the numeric difference between the model's decision and the correct answer). The models' accuracy is then matched with human subjects to assess empirical fit.

Architectures

ALCOVE (Kruschke, 1992)

In exemplar models of concept learning, a theoretical subject first learns to associate a number of different points in feature space (i.e., stimuli) with a set of possible category labels. Then, when a new stimulus is encountered, it is classified based on its similarity to remembered *reference points* in memory, aggregated across all stimulus features. Often, the similarity computation is augmented to allow

greater emphasis on features that are highly diagnostic of category membership (i.e., selective attention). The relatively simple process that exemplar models propose has been widely successful at predicting various phenomena in the category learning literature (Medin & Schaffer, 1978; Nosofsky, 1986).

ALCOVE (Kruschke, 1992) is a connectionist implementation of exemplar models that has been tested in a large number of empirical investigations (Kruschke, 1993, 2008; Lee & Navarro, 2002). ALCOVE represents stimuli as a set of feature values in some psychological space. When making a classification decision, the feature values are propagated through a set of attention weights to a hidden layer of exemplar nodes with radial basis activation functions, given by the equation:

$$a_j^h = e^{-c(\sum_i \alpha_i |h_{ji} - x_i|^r)^{q/r}}$$

where c is a specificity parameter that defines the sharpness of the hidden nodes' radial basis functions, α_i is the attention weight for feature i, h_{ji} is the *ith* feature of the *jth* exemplar, q defines the shape of the radial basis function (1 resulting in an exponential shape, 2 resulting in a *Gaussian* shape), and r determines the distance metric used for the similarity computation (1 results in *city-block* distance, 2 results in *Euclidean* distance). The hidden activation values correspond to the similarity between each stored exemplar and the test item being presented. These values are then propagated through another set of association weights, producing an output that corresponds to classification responses:

$$a_k^o = \sum_j w_{kj} a_j^h$$

where w_{kj} is the weighted connection between hidden node j and category k. The probability of classifying a test item to category k is given by the Luce-choice equation using the output of each category node:

$$P(K) = \frac{e^{\phi a_K^o}}{\sum_k e^{\phi a_k^o}}$$

where ϕ is a *response-mapping* parameter that modulates the strength of the category node with the highest probability. Throughout training, ALCOVE's weights are updated via the backpropagation algorithm to minimize the error between ALCOVE's classification decision and a modified description of the correct response, given by the equation:

$$E = \frac{1}{2}\sum_k (t_k - a_k^o)^2$$

where the value t_k takes on the value of 1 or a_k^o (whichever is greater) if k is the correct category, and -1 or a_k^o (whichever is lower) if k is the incorrect category.

ALCOVE is an ideal candidate model for this investigation for a number of reasons. First, ALCOVE instantiates 2 key mechanisms that have been central to the category learning literature: similarity-based generalization (Shepard, 1987) and selective attention (Nosofsky, 1986) – both of which inspired the behavioral investigations of the present work. Second, ALCOVE can serve as a representative instance of exemplar theories of classification, which have had much historical success in the category learning literature thus far. Third, ALCOVE is instantiated as a connectionist network, which neatly aligns with the methodological aims of the present computational investigation.

DIVA (Kurtz, 2007)

Many models of categorization make classification decisions by directly learning which regions of a *feature space* correspond with a set of category labels (Ashby & Alfonso-Reese, 1995; Jäkel et al., 2009; Kruschke, 1992; Nosofsky, 1986; Rosseel, 2002) – mediated through a set of reference points in regions of *feature space* that are populated with stimuli from previous experience. These models essentially ask the question, given a presented stimulus, what category does it likely belong to? An alternative approach – sometimes called a *generative* approach to concept learning (Hsu & Griffiths, 2010; Kurtz, 2015; Ng & Jordan, 2001) – asks the reverse: given a particular category, how likely is it to have produced the presented stimulus? *DIVA* (Kurtz, 2007) is one model in particular that embodies this framework. DIVA is a feed-forward connectionist model that treats classification as a process of assessing how well a learned category representation *explains* (or, *reconstructs*) a presented stimulus – using a modified version of the classic *autoencoder* architecture (Rumelhart et al., 1985).

Like other models of category learning, DIVA represents stimuli as a set of feature values. When making a classification decision, DIVA first propagates a set of feature values x through a layer of *encoding weights* (w_c) and *bias weights* (b_c):

$$a^h = \sigma(w_e x + b_e)$$

where σ is a *sigmoidal activation function* (squashing all values between 0 and 1). The output of this process – the hidden layer's *activation* – is then propagated

through a set of category-specific *decoding weights*, given by the equation:

$$a_k^o = \sigma^o(w_k a^h + b_k)$$

where σ^o can be either a sigmoidal or linear activation function. There is one set of decoding weights w_k for each category in the learning domain, and the outputs are treated as a set of reconstructed stimulus features[37]. The outputs of each category *channel* are then used to determine which category best *reconstructed* the test stimulus. *Reconstructive success* is operationalized as the *sum squared error* between the stimulus features and the category *channel* output values; the total reconstructive success is then used to make a classification decision using an inverse of the Luce-choice rule:

$$P(K) = \frac{SSE(K)^{-1}}{\sum_k SSE(k)^{-1}}$$

where $P(K)$ is the probability that the test stimulus belongs to category K. DIVA is also augmented with a set of *focusing weights* that bias the impact of features *with the most diverse values* across category channel outputs. The *diversity* of a feature is calculated by taking the sum of the absolute pairwise differences between category channel outputs. Before training, DIVA's encoding and decoding weights are initialized as random values – meaning that early reconstructive success will be very poor across all category channels. Throughout learning, the encoding weights and decoding weights (for the correct category) are optimized via the backpropagation algorithm to minimize the reconstructive error (SSE), given by

[37] i.e., with the same dimensionality as the input

the equation:

$$E = \sum_i (x_i - a_{ki}^o)^2$$

.

DIVA serves as a useful candidate architecture in the present work for a number of reasons. First, DIVA provides an alternative to the *reference point* account of categorization, opting instead to learn the statistical relationships between features instead of specific regions of stimulus space directly. Second, DIVA has provided both *post-hoc* and *a priori* predictions about human categorization behavior in a number of investigations (Conaway, 2016; Conaway & Kurtz, 2017; Kurtz, 2007, 2015). One prediction in particular is that the degree of statistical regularities between features within a set of exemplars impacts the ease at which certain category structures are learned (Conaway, 2016), suggesting that subjects may incorporate some knowledge of feature covariation into their classification decisions. Whether this mechanistic principle has usefulness in a relational learning domain has implications for the present work. Finally, DIVA is instantiated as a connectionist network, which helps ensure that any contrast between predictions made by ALCOVE results from differences in theoretical architecture – rather than a particular learning framework.

Stimulus Representation Format

DIVA and ALCOVE were chosen for the present investigation to test the predictiveness of a wide breadth of computational principles from the category learning literature. However, the primary concern of the present work was not necessarily to demonstrate which model (of the two) provides a better description

of human behavior in a learning domain that they weren't originally designed to explain. Rather, the present work more specifically asked how humans leverage *relational* information during a category learning task. To address this question, the present work manipulated the *input representations* of each model to include *purely featural*, *purely relational*, and *both featural and relational* (mixed) information about a presented stimulus pair. This required no essential change to ALCOVE or DIVA's architecture or fundamental assumptions, and only required a change to the input representation.

In the *purely featural* representation, stimulus pairs were be represented as a concatenated vector of feature values for each stimulus in the pair. In the *purely relational* representation, stimulus pairs were represented as the *absolute difference* of stimulus A's feature values and stimulus B's feature values. For example, if stimulus A were represented as the vector $x_A = [0.9, 0.3, 0.5]$, and stimulus B were represented as $x_B = [0.7, 1.0, 0.5]$, then the *purely relational* representation would be $|x_A - x_B| = [0.2, 0.7, 0.0]$. In the *mixed* representation, stimulus pairs were represented as the concatenation of the *purely featural* and *purely relational* representations. For example, if stimulus A were represented as the vector $x_A = [0.9, 0.3, 0.5]$, and stimulus B were represented as $x_B = [.7, 1.0, .5]$, then the *mixed* representation would be $[0.9, 0.3, 0.5, 0.7, 1.0, 0.5, 0.2, 0.7, 0.0]$. The *mixed* representation was used to assess whether *featural* and *relational* information are jointly integrated into concept learning[38].[39]

[38]This could explain the findings from experiment 1b that demonstrated some shift in aggregate generalization behavior when a region of feature space was missing during training.

[39]Interestingly, some work from the *relational similarity perception* literature suggests that human subjects integrate either featural or relational information *exclusively* in their similarity judgements (Goldstone et al., 1991) – though it is to be determined whether that phenomenological finding extends to the present work.

Exp 2 Behavioral Fits

Procedure

The simulation procedure was designed to match the behavioral investigations as closely as possible. On each initialization, a model was presented with the same number of randomly sampled stimulus pairs and training blocks as human subjects in experiment 2. The model then made a probabilistic response corresponding to a classification during each trial. Backpropagation was used to update the model's weights (once per trial) to improve accuracy for the next trial. The model's predictive success was operationalized as the *sum squared difference* at each block of training between humans and models, averaged across individual subjects and model initializations (respectively).

Hyperparameter Search

DIVA and ALCOVE have 4 and 5 (respectively) experimenter-defined *hyperparameters* that influence their learning and generalization behavior. Traditionally, a computational search is conducted to find the best fitting hyperparameters that explain the behavioral data in question (Levering et al., 2020; Nosofsky, Gluck, et al., 1994). In the present work, we used a *random search* (Bergstra & Bengio, 2012), where a random set of 10,000 hyperparameter value combinations were selected from a predefined range. The range for each hyperparameter is listed in tables 3 and 4. For each random sample of hyperparameter values, 500 model initializations of the simulated experiment were conducted and eventually aggregated into a single prediction about the average learning at each block. The choice of 500

initializations was selected as an optimal tradeoff between consistency of model prediction and time required to feasibly conduct the computational experiment.

Table 3: Minimum and maximum value of the search range for each hyperparameter in ALCOVE (brackets indicate a discrete set of values).

Hyperparameter	Search Range
specificity	[1, 9]
attention learning rate	[0.001, 3]
association learning rate	[0.00001, 2]
response mapping parameter	[1, 10]
distance metric	{1, 2}

Table 4: Minimum and maximum value of the search range for each hyperparameter in DIVA.

Hyperparameter	Search Range
learning rate	[0.0001, 3]
weight range	[0.00001, 5]
number of hidden nodes	[1, 15]
focusing parameter	[0, 10]

Results & Discussion

In some sense, both models generally fail to capture the qualitative pattern of problem difficulty in the human subjects (Figures 27 & 28). When using a purely *featural* input encoding, ALCOVE predicts a slight advantage of the *filtration* structure over the *condensation* and *condensation-flipped* structures[40]. Beyond that, neither of the models regardless of input representation were able to capture the advantage of the *filtration* and *condensation-flipped* category structures. This is particularly interesting because when the models use a *relational* input encoding, it directly matches the preparation of Kruschke (1993). Consequently, we would expect that the *filtration* condition would be much easier to learn than the *condensation* condition – a testament to likely *attention* effects in human subjects. In fact, when assessing both models fit to *just* the *filtration* and *condensation* structure, they both replicate the prediction that *filtration* will be easier to learn (see Figure 30; Appendix B).

The key difficulty here does not seem to be that the models do not predict any category structure being easier or harder to learn than any other. The key issue appears to be the conflict in using *attentional* mechanisms that boost performance on *filtration* but inhibit performance on *condensation-flipped* (a structure that our human subjects learned relatively well)[41]. This is likely why the best fitting version of ALCOVE fails to predict the classic *filtration* advantage that it should have produced – particularly in the *relational* input representation. To do so would

[40]To a small extent, DIVA with a *mixed* input encoding makes a similar prediction.

[41]Another possibility is the models' difficulty in fitting the behavioral stem from the fact that our human subjects learn relatively slower than we would expect them to in a standard feature-based learning study. While this might be an important reflection of the working memory constraints present in a relational learning task, it may have the unintended consequence of reducing meaningful variability in the behavioral signature we're asking the models to predict.

require using a high *attention learning rate*; but, that would have the undesirable effect of making the *condensation-flipped* condition harder to learn. ALCOVE's limited ability isn't necessarily surprising given ALCOVE was designed for the *featural* learning domain where features are treated as numeric coordinates that each represent distinct latent variables[42]. However, in the present case where the featural *relations* instantiate the same latent construct (*degree of sameness*), humans seem to leverage *aggregate* similarity in the way ALCOVE does not predict.

DIVA also has difficulty showing a strong advantage of the *filtration* and *condensation-flipped*. A likely tension for DIVA might be in predicting a difference between the *condensation* and *condensation-flipped* condition. Because the *condensation-flipped* and *condensation* conditions are identical beyond their single flipped axis, DIVA wouldn't directly predict that the *condensation* condition would be the hardest structure for human subjects to learn. DIVA could ramp up its attentional mechanism to slow its acquisition of the *condensation* structure, but that would conflict with how fast the *condensation-flipped* condition was learned. In summary, DIVA is caught in the same catch-22 as ALCOVE: *attention* is unable to predict fast learning in the *filtration* condition without decreasing learning speed in the *condensation-flipped* condition. Further, when observing DIVA and ALCOVE's best fit to just the *condensation* and *condensation-flipped* conditions exclusively, neither model show an advantage of the *condensation-flipped* condition (see Appendix C).

[42]And in this learning domain specifically, ALCOVE makes very compelling behavioral fits to empirical phenomena.

Exp 3 Behavioral Fits

Procedure

This simulation used the same procedure used to test the empirical predictions in experiment 2.

Hyperparameter Search

The simulation of experiment 3 used the same hyperparameter search procedure as the simulation for experiment 2. The same hyperparameter search ranges were also used, listed in tables 3 and 4.

Results & Discussion

Unlike in experiment 2, both models seemed to capture the qualitative pattern of problem difficulty in experiment 3, generally regardless of input representation scheme. The biggest exception is that both models failed to predict an obvious type 4 advantage – with the exception of ALCOVE using a *relational* encoding. Instead, both models seemed to be generally predicting the revised ordering of learning difficulty observed during feature-based category learning (Kurtz, Levering, et al., 2013). The strongest quantitative fits occured when both models used either a *relational* and *mixed* input representation, while fits seemed relatively poor for the *featural* input representation. This makes sense given that feature-values were not intended to be predictive for solving the category structure, which would support the idea that human learners are leveraging *relations* exclusively or in addition to *features*. Overall, DIVA using a *relational* encoding scheme had the

best *quantitative* fit to the data in experiment 3 – despite failing to predict the strong learning acquisition rate of the *type 4* structure.

The failure of both models to predict the type 4 advantage was somewhat expected given the results from experiment 2: neither model has an explicit mechanism for integrating *aggregate similarity* into a classification decision – which is arguably what our human subjects may have been doing. The only exception could be ALCOVE's prediction using a *relational* input representation. In that case, ALCOVE does seem to predict that type 4 would be easier to learn than types 3, 4, 5, and 6, while also predicting the rapid acquisition of the *type 1* structure. The attention learning parameter for ALCOVE's best fit was relatively low (.00244), which makes sense given that heavy attentional weighting wouldn't work as a solution for learning the *type 4* structure (which has no immediate attentional solution)[43]. This very low attention learning rate may have been a factor in reducing learning speed overall for ALCOVE with a *relational* input representation, which would explain the relatively poor *quantitative* fit (despite somewhat capturing the *qualitative* order of learning difficulty across the 6 structures).

Another interesting observation was that – somewhat unintuitively – both models predict a type 1 advantage *even when* using the *featural* representation format. Given that category membership was based on *featural difference*, it is unclear why feature-based models of category learning would find type 1 easier to learn when stimulus features themselves don't provide any immediate predictive value. In the case of DIVA, it could be that gradient descent was able to find some particular weight scheme that outputs a comparison of feature values, which can

[43]It is somewhat interesting that the type 1 structure was learned the fastest for ALCOVE even when attention was very low.

then be leveraged during DIVA's reconstruction objective (such a mechanism is discussed further at the end of the General Discussion). In the case of ALCOVE, it may be something particular about the interaction of *attentional weighting* and representing exemplars in a relatively high dimensional space (6 dimensions, in the case of the *featural* representation format).

Table 5: Each column (besides the index) lists the *sum-squared error* (or, difference) between aggregate model accuracy and aggregate human accuracy at each block of training; each cell represents the model's sum-squared error for its best fitting hyperparameters.

Model	Experiment 2	Experiment 3
ALCOVE *featural*	.0952	.3895
ALCOVE *relational*	.1242	.3077
ALCOVE *mixed*	.1328	.2946
DIVA *featural*	.0927	.5035
DIVA *relational*	**.0902**	**.1341**
DIVA *mixed*	.092	.2811

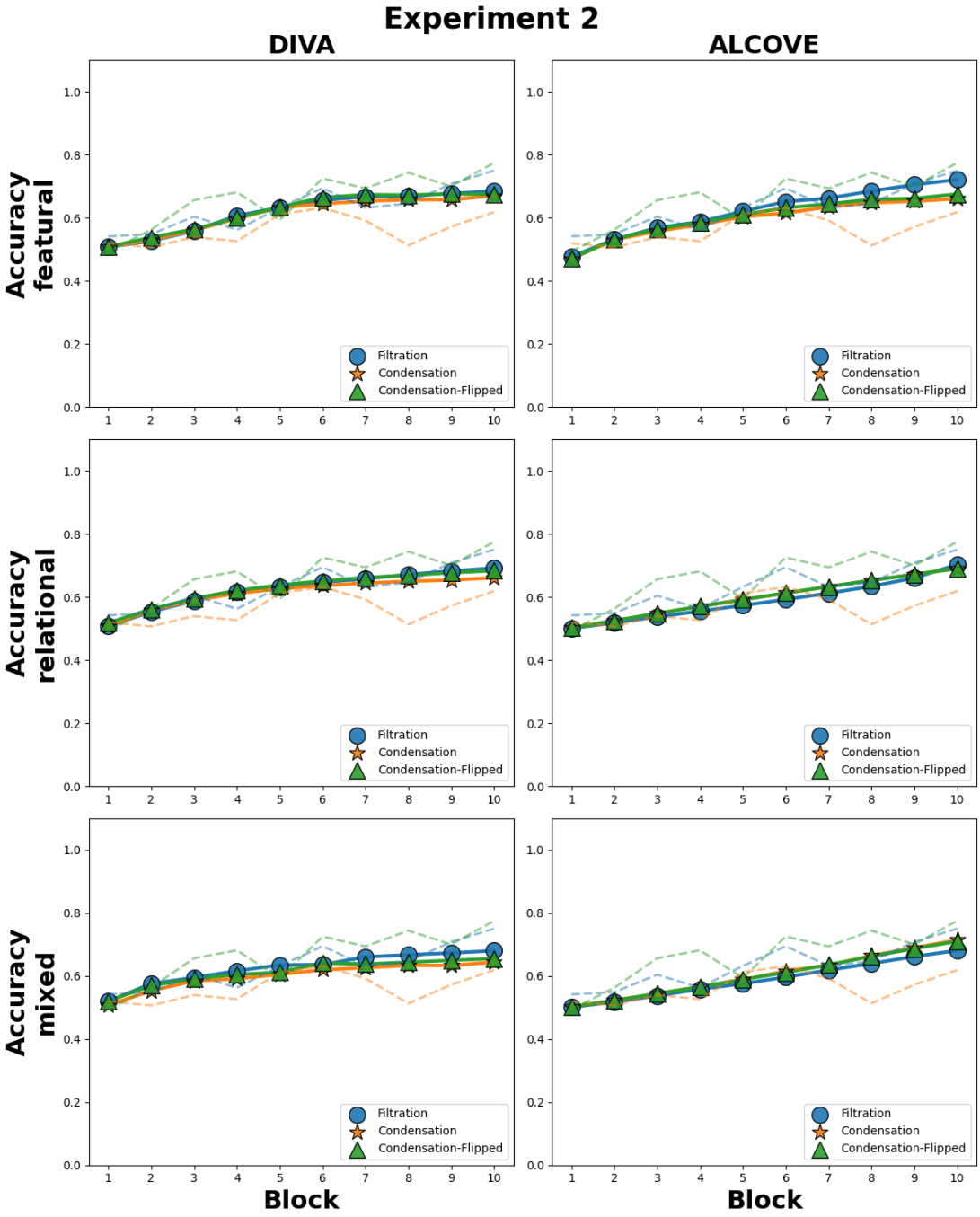

Figure 27: The grid of plots shows the learning curves for humans (dashed lines) and models (solid lines) for each category structure and each stimulus representation in experiment 2.

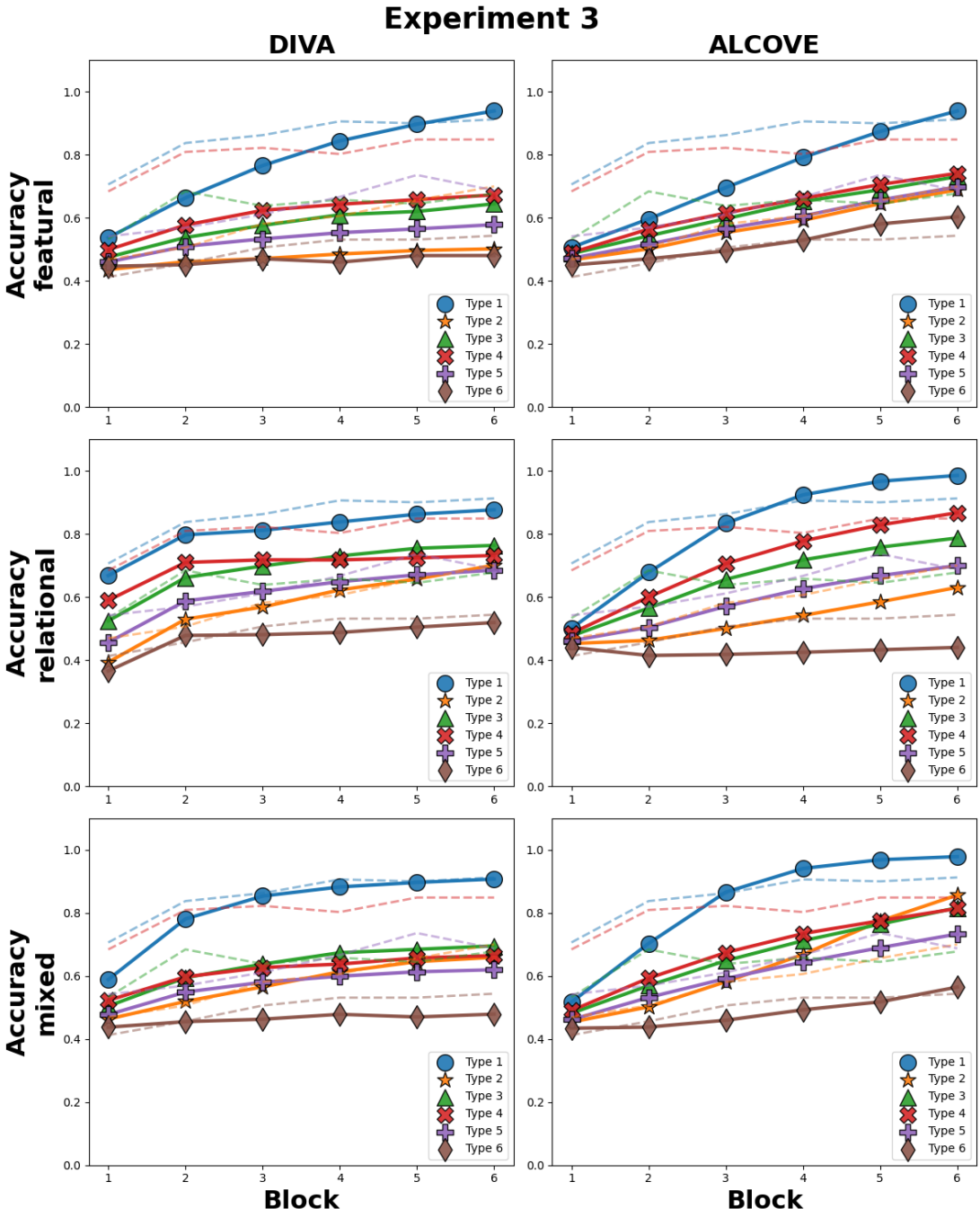

Figure 28: The grid of plots shows the learning curves for humans (dashed lines) and models (solid lines) for each category structure and each stimulus representation in experiment 3.

General Discussion

The purpose of the present investigation was to test whether the phenomenological and computational principles from the category learning literature have any explanatory value in a relational learning domain. Subjects learned to categorize stimulus pairs using the same general preparation of a standard feature-based category learning experiment. The category structures themselves were defined by the *difference in feature* values between each stimulus pair.

Experiments 1a, 1b, and 1c demonstrated that *aggregate* generalization profiles in a *relational* feature space partially conform to the predictions of previous category learning experiments. In addition, experiments 1a, 1b, and 1c provided evidence that learners employ *rule-like* learning strategies in a relational domain – particularly when they're trained to classify two regions in a relational feature space. In experiment 2, we found light evidence of selective attention pressures across featural relationships of multi-featured stimulus pairs. We found even stronger evidence that human learners were sensitive to *total similarity*. Finally, experiment 3 partially replicated the classic learning difficulty ordering of Shepard et al.'s (1961) 6 category structures, but with categories defined by stimulus relations.

The computational investigations revealed a number of interesting phenomena that highlight what mechanisms subjects might have leveraged when provided

with *featural relations* during an inductive learning task. The *relational* stimulus representation had the best quantitative fits for DIVA, while ALCOVE's best fits utilized the *featural* (experiment 2) and *mixed* (experiment 3) representations. In addition, both ALCOVE and DIVA demonstrated difficulty in predicting the more rapid learning of the *condensation-flipped* and *type 4* category structures from experiments 2 and 3 (respectively). This supports the idea that *total similarity* might be an additional conceptual tool our subjects leveraged here that they wouldn't have leveraged in a standard feature learning preparation.

Interestingly, the notion of *total similarity* does somewhat align in principle with Tversky (1977)'s *contrast model* of similarity – where some combination of the shared properties of two objects is leveraged to make a similarity judgement[44]. This is further supported by Goldstone et al. (1991), who demonstrated that human learners likely leverage combined relational similarity as a basis for making similarity judgements. Interestingly, Goldstone et al. (1991) also found evidence that when human learners are faced with both *features* (attributes) and *relations*, the weighting of featural and relational similarity might shift depending on whether a pair of multi-featured stimuli are *mostly featurally* similar or *mostly relationally* similar. That is, the weighting of features and relations are distinct but *nonindependent* (Goldstone et al., 1991).[45] They also found that a sizeable majority of subjects seemed to weight *relations* more heavily than *features* during similarity judgements. Though there are methodological differences between the *similarity* literature

[44]This similarity judgement could then form the basis for a classification decision.

[45]Though an alternative, less-supported hypothesis is that the distinction of *features* and *relations* is because *features* in Goldstone et al. (1991) are represented as distinct latent constructs, while the *relations* all represent the same construct (similarity). This could be tested by using stimulus features that all share the same latent construct instantiated as unique, independent perceptual objects – or as Goldstone & Medin (1994) describes: *matches out of place*.

(Tversky, 1977, p. @goldstone1991relational) and the present work, this does provide a possible explanation for how subjects might be integrating both *featural* and *relational* information when making forced-choice classification decisions.

Connection to *Abstract Learning*

Overall, this work makes potentially important incremental contributions to multiple cognitive science literatures. Researchers have found that nonhuman animals (in addition to humans) can learn to associate *featural differences* with some *outcome / reward* (Gonzalez et al., 1954; Köhler, 1938; Lazareva, 2012). In addition, under limited conditions, animals can even seem sensitive to stimuli that embody *sameness or difference* concepts (Blaisdell & Cook, 2005; Wasserman et al., 2004; Zentall & Hogan, 1974; Zentall et al., 2008). This has been demonstrated using various learning tasks, such as discrimination (Gonzalez et al., 1954; Köhler, 1938), match-to-sample (Fagot et al., 2001; Zentall & Hogan, 1974), and classification (wasserman1995pigeons; Wasserman et al., 2001; Young & Wasserman, 2001). Similarly, learners in the present work were trained to map featural relations to category *labels*. One crucial observation in the present work was the finding that generalization profiles in experiments 1a, 1b, and 1c were much sharper when subjects observed a category defined by *identicality* – where fish pairs had the same body length. This finding appears to be related to a similar phenomenon observed by Young & Wasserman (2001).

Young & Wasserman (2001) trained human subjects to categorize stimuli that consisted of arrays of up to 16 unique visual icons. The stimuli (icon arrays) varied in terms of the number of unique icons present in the display – ranging

from 1 set of the same icon (maximally similar arrays) or 1 set of 16 unique icons (maximally different arrays). Subjects learned via supervised classification training to discriminate between stimulus arrays embodying *sameness* or *difference*. A proportion of subjects in Young & Wasserman (2001)'s experiments showed a generalization profile where the probability of generalizing the *sameness* response gradually decreased for stimulus arrays with more variability – which Young & Wasserman (2001) refer to as a *continuous* response. In contrast, Young & Wasserman (2001) also observed a proportion of subjects whose generalization behavior was much sharper. After learning the *sameness* concept, these subjects were less likely to generalize the *sameness* response to any stimulus array with any amount of variability (or, entropy) – which Young & Wasserman (2001) refer to as a *categorical* response. Fascinatingly, pigeons and baboons trained on the same task do not show the *categorical* generalization behavior (Wasserman et al., 1995, 2001).

In Young & Wasserman (2001), *dissimilarity* was defined by the variability between the unique, non-matching icons in a single stimulus array. In the present work, *dissimilarity* in experiment 1 was defined by the difference in length between two fish in a pair. Despite the difference in the stimulus domains, the similarities between the generalization behavior of subjects in experiment 1 and subjects in Young & Wasserman (2001) are striking. When learning categories defined exclusively by *identicality*, subjects in the present work generalized very similarly to the *categorical* generalizers in Young & Wasserman (2001). In contrast, when categories of fish pairs were defined by some non-zero difference in length, subjects in experiment 1 generalized much more similarly to Young & Wasserman (2001)'s

continuous generalizers. This might suggest humans utilize similar generalization strategies when learning same/different concepts defined by either qualitative or quantitative stimulus features. However, Young & Wasserman (2001) emphasize that the *continuous* generalization profile in humans and non-human animals could be accurately predicted using Shannon-Wiener et al. (1949)'s measure of information *entropy* – which they convincingly argue subjects might be sensitive to. At present, it is unclear (to the author) how an *entropy* measure could be extended to explain the generalization behavior of *continuous* generalizers in the present work given that stimuli in experiment 1 varied based on a continuous feature (length). Nevertheless, this would be an important theoretical question to address.

Connection to *Feature-Based Category Learning*

The category learning literature has demonstrated a number of consistently replicated empirical phenomena useful towards testing mechanistic explanations of category identification, discrimination, and generation. The present work was designed to test whether some of the explanatory principles from the category learning literature have predictive value when categories are defined by *relational values* instead of *feature values*. The results partially support this hypothesis. Selective attention pressures *appear* to be present when human learners are tasked with mapping *multiple featural relations* to category labels. The category structure in experiment 2 that was separable by a single *relation* had more subjects reaching ceiling accuracy – except in the case where a global measurement of *family resemblance* could be leveraged as a predictive cue.

Selective attention was also apparent in experiment 3, where the structure that was separable by a single *relation* (*type 1*) was learned the fastest. In experiment 3, we also partially replicated the learning difficulty ordering of Shepard et al. (1961)'s 6 category structures. This might suggest that some of the same mechanistic (Nosofsky, Gluck, et al., 1994) and information-theoretic (Feldman, 2000; Pape et al., 2015; Vigo, 2013) constraints that describe *feature-based* learning might also describe learning of featural *relations*.

In experiments 1a, 1b, and 1c, we found evidence that learners employed *rule-like* decision strategies – something commonly seen in both category learning (Nosofsky, Palmeri, et al., 1994) and reinforcement learning (Lee et al., 2018) with human subjects. We further replicated the finding that *classification* learning results in more *discriminative* learning behavior (Chin-Parker & Ross, 2002, 2004; Levering & Kurtz, 2015). Further, we extended the finding that aggregate generalization behavior is well-fit by functions from the exponential family (Nosofsky, 1985; Shepard, 1987). However, one critical limitation is that generalization decisions in the test phase were discrete (taking the value 0, .5, or 1). A key theoretical issue is whether or not this accurately describes gradation in an individual learner's confidence in category membership. This could be addressed by leveraging a more continuous classification response[One option is to make subjects provide a large, repeated number of decisions for the same stimuli – this was avoided in the present work to reduce subject fatigue.].

Though discrete generalization responses reduced our ability to speculate about the specific, fine-grained response profiles of individual subjects, it still allowed us to speculate about subjects' particular generalization strategies. For example, the

diff -1 and *diff-0* structures were highly similar in that category membership was defined by either (1) a very small feature difference, or (2) identicality (respectively). However, the *diff-0* condition produced qualitatively different generalization profiles than the *diff -1* condition. When category membership was defined by *0* difference in length – or, *exact sameness in length* – subjects seemed to employ an *all-or-nothing*, rule-like response strategy. When a small amount of *difference* was added to stimuli in the target category of the *diff -1* condition, response profiles became much more graded. This could indicate a relational inductive bias that privileges the state of *exact sameness* or *equality*. Intuitively, this seems useful for ecological fitness given that *equality* might be a powerful construct for an animal to leverage[46].

Connection to *Relational Cognition*

Relational reasoning is frequently stated to be a critical process in many theories of 'higher-order' cognitive processing. For instance, various theories of *analogical reasoning* posit that representations of complex stimuli encode relationships between latent variables – which serve as the basis for analogical *transfer* between highly disparate concepts (Gentner, 1983; Gick & Holyoak, 1980). Relational cognition is also commonly invoked in the theory learning literature – which describes how humans learn and represent complex concepts like *kinship systems* or *magnetism* (Kemp et al., 2010). In some computational accounts of theory learning, relations are *explicitly* embedded in stimulus representations and serve as the primary basis

[46]Interestingly, the frequentist statistical paradigm commonly leveraged by psychology researchers heavily leverages the concept of *exact sameness* or *equality* – often described as a *null hypothesis*. Given that statistics is used to predict natural events, it might be that *equality* is an inherently useful conceptual construct – especially in regards to human environments and culture.

for computation (Kemp et al., 2010; Kemp, 2012).

While propositional networks and relational graphs are frequently invoked as representational constructs in theories of higher-order cognition, there does not appear to be a resolved mechanistic explanation for how a scene is mapped *from* a collection of (often unlabeled) objects *to* the structured representations invoked in the existing literature. For instance, if a representation of a complex stimulus contains a relational property like *larger-than*, what specific size difference between 2 features is required for their relation to no longer be *same-size*? Further, how is a *learned* relational property generalized during new, similar experiences?

The present work examined human relational learning at a very early stage: learning magnitude or binary comparisons between two latent variables in a psychological space. The findings suggest that the principles of generalization that describe feature-based category learning – namely, selective attention, similarity-based generalization, and rule-like learning strategies – have predictive potential in a simple, *unitary* (Corral & Jones, 2017) relational learning domain. While the mechanisms that describe how objects are mapped to low-level or unitary relations may not have predictive value at later stages of higher-order cognition, they could still determine the representational constructs a person acquires before they could do any higher-order cognition to begin with. An explanation of this process would be a crucial component of any complete theory of the cognitive pipeline.

Lastly, a key theme in the literature on higher-order cognition is the use of richly structured, relational psychological representations (Gentner, 1983; Kemp, 2012). It could be argued that the mechanisms required to learn the relations in the present work are distinct from the types of relational systems commonly

studied in the higher-order cognition literature, and might only be relevant for very simple cases where subjects map featural *magnitude differences* to arbitrary category labels. Future work could try addressing whether or not the present findings extend to more complicated relational learning preparations.

However, the mechanisms in the present work might be relevant for a theoretically distinct type of relational learning process. Corral & Jones (2017) propose the distinction between (a) *compositional* representations – which encode complex structure between elements of a representation, and (b) *unitary* representations – which consist of a *relational* attribute describing a set of elements or a single-value *difference* between elements[47]. The empirical investigations in present work showcase an instance of *unitary* relational learning, and have implications for theoretical accounts of how *unitary* relations are learned and generalized in a laboratory setting. Specifically, the present work finds evidence of *selective attention* pressures during learning of *unitary features*. Interestingly, work by Corral et al. (2018) also proposed that selective attention guides *unitary* relation learning – based on evidence that subjects can learn to selectively attend to a relevant *feature difference* in order to predict an arbitrary category label.

Connection to *Graph Learning*

Researchers have studied how people learn in domains that can be described as interconnected networks of objects or events that share a relation (i.e., a *graph*). *Graphs* have been frequently invoked as important data structures for cognitive representations across many different branches of cognitive psychology, e.g. *language*

[47]Interestingly, recent evidence suggests that encouraging subjects to perceive a stimulus *unitarily* leads to quicker learning (Corral & Jones, 2017)

(Chan & Vitevitch, 2010; Vitevitch, 2008; Vitevitch et al., 2012), *semantic cognition* (Collins et al., 1969; Collins & Loftus, 1975), *function learning* (Wu et al., 2020), and *higher-order cognition* (Gentner, 1983; Griffiths et al., 2010). The stimulus domain in the present work can also be described as a graph, where individual stimuli are represented as *nodes* and their classification label represents the *connection* between them. The *graph* learning literature has typically explored the impact of structural properties of a graph (e.g., *node-centrality* or *clustering coefficient*) on processing, learning, and generalization (Chan & Vitevitch, 2009; Lynn & Bassett, 2020; Solomon et al., 2019), often leveraging stimuli with perceptually unalignable features with no predictive value (Karuza et al., 2017; Kemp et al., 2007, 2010)[48].

However, in the present work, *featural differences* are explicitly linked to the task subjects are required to learn. There seems to be many real-world scenarios where humans are required to use *featural differences* to make some decision or classification. For example, it might be considered unfashionable to wear two primary articles of clothing that *match* on their color dimension; recognizing *acceptable* fashion would require computing some difference in color value between two stimuli (e.g., a shirt and pants). While this example may seem *low-stakes* at first glance, fashion may be one of the most crucial variables in the fitness landscape of modern humans. Given that humans can make these choices with relative ease, a complete mechanistic account of how humans learn in graph structured learning domains should be able to explain how people leverage *feature comparison* to identify connections between objects.

[48]Though, see Corral & Jones (2014).

A Connectionist Mechanism for Relation -> Category Induction

The category learning literature has been fruitful in producing a number of theoretical principles, frameworks, and models that predict human feature learning. In the present work, we obtained evidence of phenomenological overlap between featural and relational induction – which could be construed as evidence for some mechanistic overlap between the two learning domains. Specifically, the present work suggests that the theoretical mechanisms of category learning can operate on *featural differences* instead of raw features alone. This was embodied in the computational investigation that used *featural difference* encodings as the input to classic models of feature-based category learning. One limitation of this approach is that the *featural encoding* was hard-coded by the experimenter (author), and no account is given for how featural encodings emerge from learning and experience.

Interestingly, work by K. J. Kurtz (personal communication, 2020) demonstrated that a very simple connectionist network could learn – via supervised training – to accurately map a set of features to a relational category label reflecting the difference between the two features. The network initially consists of two random weights, each applied to a set of two features to be compared. When trained to correctly identify whether feature *1* is larger than feature *2*, the network eventually falls into a solution in weight space where one weight is positive, and the other is negative – but both are of relatively equal size[49]. Figure 29 shows the error landscape plotted as a function of network weight strengths. Importantly, there

[49]The model resembles the structure of an *analog comparator* that outputs the difference between two incoming voltages (Malmstadt et al., 1981).

are no obvious local minima, and the network should almost always converge towards the same solution using any particular weight updating scheme[50]. While this approach still requires setting a supervised learning objective, it serves as a useful proof-of-concept for how a simple, relational computation mechanism can emerge in a connectionist framework (at least under certain conditions).

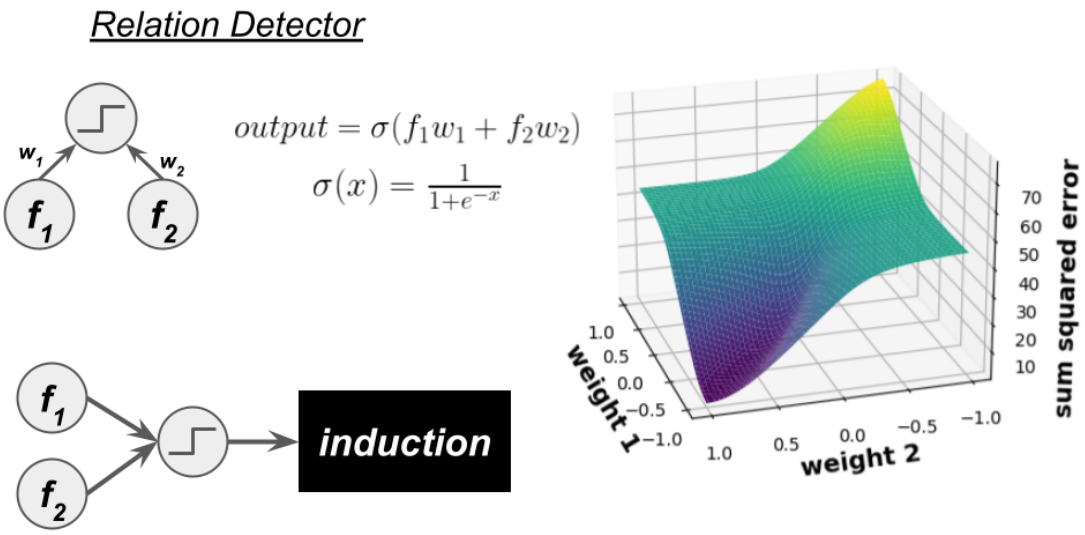

Figure 29: Visualization of the (left) *relation detector* network described by K. J. Kurtz (personal communication, 2020), and (right) a visualization of the weight space against the model's learning error (*sum squared error* between model prediction and relational category label).

Conclusion

Many researchers have stressed the importance of *relational reasoning* in human cognition – some have even gone as far to say that it is a fundamental delimiter of human and animal intelligence. The present work sought to explore whether the

[50]To the extent that a multiplicatively weighted network (e.g., a connectionist model) is an adequate representation of actual biological neurons, it may be that the computational mechanisms needed to compute unitary relations are both (a) built into hardware of animal cognition, and (b) relatively easy for any weight-updating optimization procedure to converge on (including whatever weight-updating algorithm is most biologically plausible).

phenomenological and theoretical principles from the human and animal feature-based learning literature have any predictive value in describing human relational learning. The findings provided partial support for the hypothesis that *unitary* relations may be learned and generalized via the same cognitive mechanisms that describe feature-based learning. Specifically, we found evidence that selective attention pressures and similarity-based generalization may apply to *relations* defined by feature differences. Additionally, the present work yielded evidence that humans may leverage *total similarity* when classifying a compound stimulus pair. These findings provide phenomenological markers of human behavior that can be used to test computational theories of *unitary* relational learning. Finally, the present work serves as another attempt to extend the traditional classification paradigm towards a *relational* learning domain (Corral et al., 2018; Kurtz & Boukrina, 2004), hopefully advancing the goal of uncovering universal principles of human cognition *across* a wider span of conceptual learning domains.

Appendix A

Table 6: Best fitting hyperparameters for ALCOVE for experiment 2.

Hyper Parameter	*featural*	*relational*	*mixed*
specificity	2.938082	4.666213	8.604288
attention learning rate	2.594525	0.027507	0.00169
association learning rate	1.814507	0.021289	0.085691
response mapping parameter	9.243057	1.542445	6.700929
distance metric	2	1	2

Table 7: Best fitting hyperparameters for ALCOVE for experiment 3.

Hyper Parameter	*featural*	*relational*	*mixed*
specificity	3.749629	1.30468	2.690876
attention learning rate	0.006408	0.002438	0.0017
association learning rate	0.892996	0.075602	0.802086
response mapping parameter	9.572896	4.274555	1.621296
distance metric	2	1	1

Table 8: Best fitting hyperparameters for DIVA for experiment 2.

Hyper Parameter	*featural*	*relational*	*mixed*
learning rate	0.240058	0.071007	0.451934
weight range	0.179094	3.354325	2.624257
number of hidden nodes	9	5	6
focusing parameter	9.384847	3.833945	9.80778

Table 9: Best fitting hyperparameters for DIVA for experiment 3.

Hyper Parameter	*featural*	*relational*	*mixed*
learning rate	0.089783	1.552884	0.185936
weight range	0.46569	0.090213	0.643656
number of hidden nodes	13	2	2
focusing parameter	7.665502	6.38872	8.494504

Appendix B

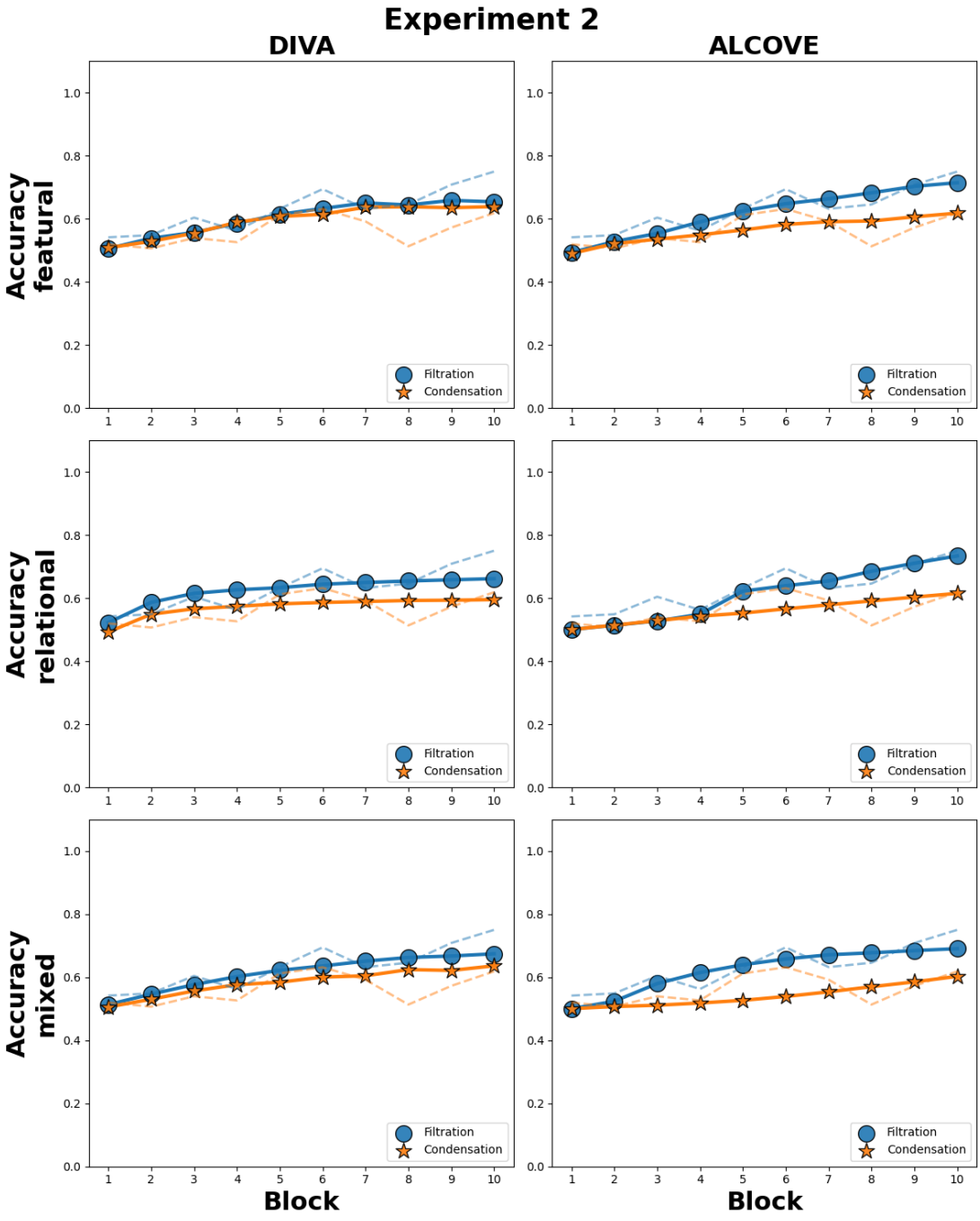

Figure 30: Learning curves using ALCOVE and DIVA's best fitting hyperparameters when predicting *just* the *filtration* and *condensation* results from experiment 2.

Appendix C

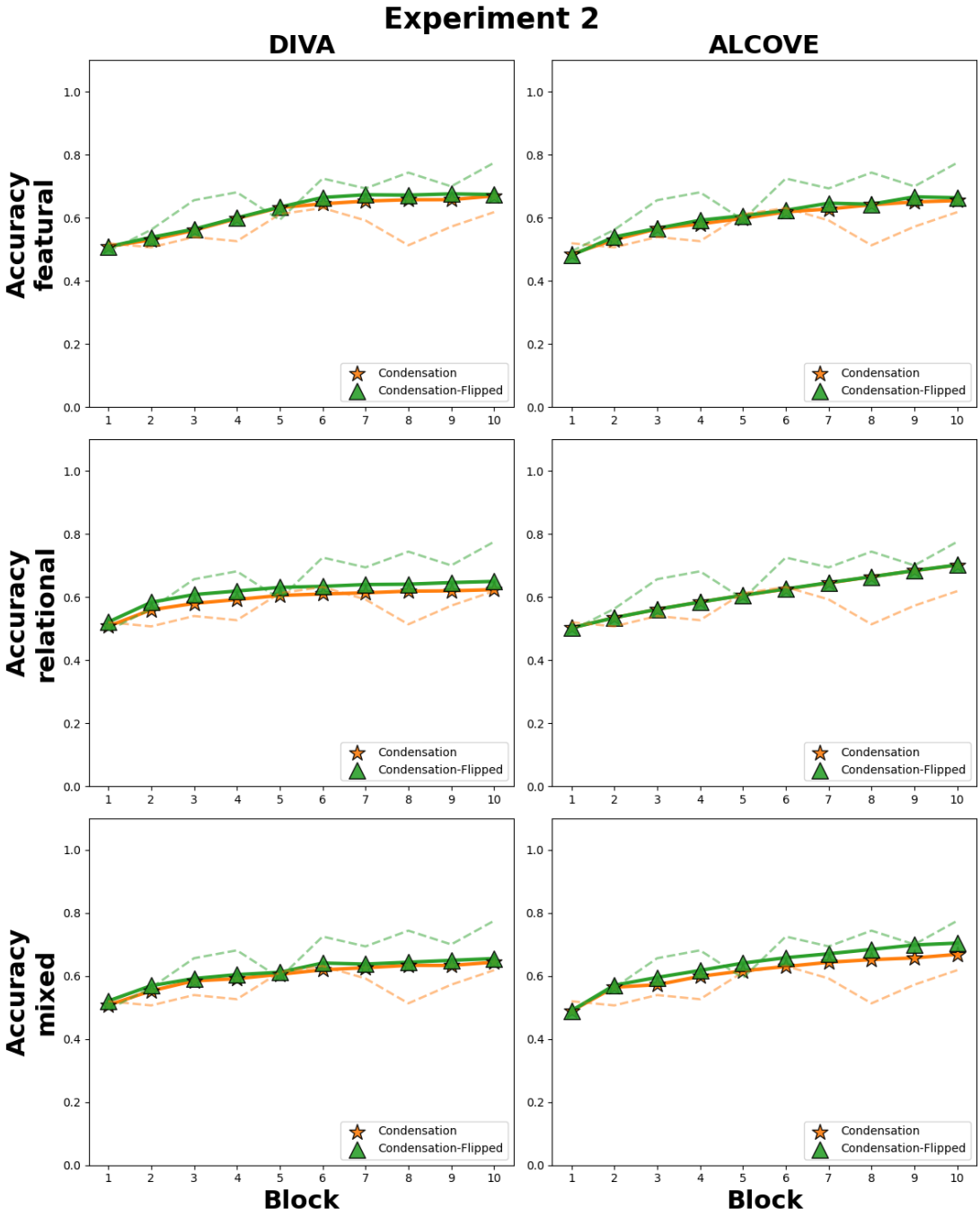

Figure 31: Learning curves using ALCOVE and DIVA's best fitting hyperparameters when predicting *just* the *condensation* and *condensation-flipped* results from experiment 2.

References

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81–121.

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

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., & others. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33.

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: Presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 414.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Blaisdell, A. P., & Cook, R. G. (2005). Two-itemsame-different concept learning in pigeons. *Animal Learning & Behavior*, *33*(1), 67–77.

Bodily, K. D., Katz, J. S., & Wright, A. A. (2008). Matching-to-sample abstract-concept learning by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*(1), 178.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. *New York: John Wiley & Sons, Inc*, *14*, 330.

Busemeyer, J. R., Myung, I. J., & McDaniel, M. A. (1993). Cue competition effects: Empirical tests of adaptive network learning models. *Psychological Science*, *4*(3), 190–195.

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

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(3), 185–211.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive Science*, *34*(4), 685–697.

Cheng, K. (2000). Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science*, *11*(5), 403–408.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, *30*(3), 353–362.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 216.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Collins, A. M., Quillian, M. R., & others. (1969). *Retrieval time from semantic memory*.

Conaway, N. (2016). *Re-evaluating the reference point view of human classification learning*. State University of New York at Binghamton.

Conaway, N., & Kurtz, K. J. (2017). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic Bulletin & Review*, *24*(4), 1312–1323.

Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition*, *132*(3), 280–300.

Corral, D., & Jones, M. (2017). Learning relational concepts through unitary versus compositional representations. *CogSci*.

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within-versus between-category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Davis, E. (2014). *Representations of commonsense knowledge*. Morgan Kaufmann.

Domjan, M. (1993). *The principles of learning and behavior*.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.

Fagot, J., Wasserman, E. A., & Young, M. E. (2001). Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons (papio papio) and humans (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *27*(4), 316.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). *The structure-mapping engine* (Vol. 1275). Department of Computer Science, University of Illinois at Urbana-Champaign.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 47–59.

Gentner, D., & Kurtz, K. J. (2005). *Relational categories.*

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 29.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*(2), 222–262.

Goldstone, R. L., & Son, J. Y. (2012). *Similarity*. Oxford University Press.

Gonzalez, R. C., Gentry, G. V., & Bitterman, M. E. (1954). Relational discrimination of intermediate size in the chimpanzee. *Journal of Comparative and Physiological Psychology*, *47*(5), 385.

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

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*(1), 79.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT press.

Holyoak, K. J., & Hummel, J. E. (2001). Toward an understanding of analogy within. *The Analogical Mind: Perspectives from Cognitive Science*, 161.

Honig, W. K., & Urcuioli, P. J. (1981). The legacy of guttman and kalish (1956): 25 years of research on stimulus generalization. *Journal of the Experimental Analysis of Behavior*, *36*(3), 405–445.

Hsu, A. S., & Griffiths, T. E. (2010). Effects of generative and discriminative learning on use of category variability. *32nd Annual Conference of the Cognitive Science Society*.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*(9), 381–388.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*(1), 1–9.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*(4), 685.

Kemp, C., Goodman, N., & Tenenbaum, J. (2007). Learning and using relational theories. *Advances in Neural Information Processing Systems*, *20*.

Kemp, C., & Jern, A. (2014). A taxonomy of inductive problems. *Psychonomic Bulletin & Review*, *21*(1), 23–46.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Koggdal, J. (n.d.). *OCanvas* (Version 2.10.0) [Computer software]. http://ocanvas.org

Köhler, W. (1938). *Simple structural functions in the chimpanzee and in the chicken.*

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797–2822.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175.

Kruschke, J. K. (2008). Models of categorization. *The Cambridge Handbook of Computational Psychology*, 267–301.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1083.

Kurtz, K. J. (2007). The divergent autoencoder (diva) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In *Psychology of learning and motivation* (Vol. 63, pp. 77–114). Elsevier.

Kurtz, K. J., & Boukrina, O. (2004). Learning relational categories by comparison of paired examples. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*.

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Kurtz, K. J., & Wetzel, M. T. (2021). On the generalization of simple alternating category structures. *Cognitive Science*, *45*(4), e12972.

Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, *27*(5), 781–794.

Lazareva, O. F. (2012). Relational learning in a context of transposition: A review. *Journal of the Experimental Analysis of Behavior*, *97*(2), 231–248.

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1955.

Lee, M. D., & Navarro, D. J. (2002). Extending the alcove model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*(1), 43–58.

Levering, K. R., Conaway, N., & Kurtz, K. J. (2020). Revisiting the linear separability constraint: New implications for theories of human category learning. *Memory & Cognition*, *48*(3), 335–347.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266–282.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309.

Lynn, C. W., & Bassett, D. S. (2020). How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, *117*(47), 29407–29415.

Malmstadt, H. V., Enke, C. G., Crouch, S. R., & Crouch, S. R. (1981). *Electronics and instrumentation for scientists*. Benjamin-Cummings Publishing Company.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*(4), 517–535.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592.

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279.

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *Neuroimage*, *46*(1), 319–326.

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. *Formal Approaches in Categorization*, 40–64.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289.

Ng, A., & Jordan, M. (2001). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, *14*.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, *38*(5), 415–432.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*(4), 279–290.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, *43*(1), 25–53.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & Cognition*, *22*(3), 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53.

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64*, 66–75.

Patterson, J. D., & Kurtz, K. J. (2014). Performance pressure and comparison in relational category learning. *Proceedings of the Annual Meeting of the Cognitive*

Science Society, *36*.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109–130.

Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, *6*(1), 125–136.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1.

Rattermann, M. J., & Gentner, D. (1998). The effect of language on similarity: The use of relational labels improves young children's performance in a mapping task. *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, *274282*.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41.

Rips, L. J. (1989). *Similarity, typicality, and categorization.*

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Rumelhart, D. E., Hinton, G. E., McClelland, J. L., & others. (1986). A general framework for parallel distributed processing. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *1*(45-76), 26.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation.* California Univ San Diego La Jolla Inst

for Cognitive Science.

Schlegelmilch, R., Wills, A. J., & Helversen, B. von. (2021). A cognitive category-learning model of rule abstraction, attention learning, and contextual modulation. *Psychological Review*.

Shannon-Wiener, C., Weaver, W., & Weater, W. (1949). The mathematical theory of communication. *The Mathematical Theory of Communication. EUA: University of Illinois Press, Urbana*.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

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.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior Research Methods*, *51*(4), 1717–1736.

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*(5), 430.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123.

Vigo, R. (2013). The gist of concepts. *Cognition*, *129*(1), 138–162.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, *67*(1), 30–44.

Wasserman, E. A., & Castro, L. (2021). Assessing attention in category learning by animals. *Current Directions in Psychological Science*, 09637214211045686.

Wasserman, E. A., Fagot, J., & Young, M. E. (2001). Same–different conceptualization by baboons (papio papio): The role of entropy. *Journal of Comparative Psychology*, *115*(1), 42.

Wasserman, E. A., Hugart, J. A., & Kirkpatrick-Steger, K. (1995). Pigeons show same-different conceptualization after training with complex visual stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(3), 248.

Wasserman, E. A., Young, M. E., & Cook, R. G. (2004). Variability discrimination in humans and animals: Implications for adaptive action. *American Psychologist*, *59*(9), 879.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *Quarterly Journal of Experimental Psychology*, *66*(2), 299–318.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1),

102.

Wu, C. M., Schulz, E., & Gershman, S. J. (2020). Inference and search on graph-structured spaces. *bioRxiv*.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776.

Young, M. E., & Wasserman, E. A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 278.

Zentall, T., & Hogan, D. (1974). Abstract concept learning in the pigeon. *Journal of Experimental Psychology*, *102*(3), 393.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K., & Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition & Behavior Reviews*.