THE ROLE OF ENTROPY DETECTION IN JUDGMENTS OF SIMILARITY AND DIFFERENCE OF VISUAL STIMULI

A DISSERTATION TO BE SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

PSYCHOLOGY

AUGUST 2014

By

Samantha M. O’Hanlon

Dissertation Committee:
Scott Sinnett (Chairperson)
Patricia Couvillon
Leonidas Doumas
Brandy Frazier
Sean Callahan

Keywords: categorization; cognitive development; computational models; difference; entropy detection; similarity.
I would like to thank the members of my committee for assisting me in the completion of my degree and my labmate, Margeaux Ciraolo, for her careful reading and thoughtful comments on the many iterations of drafts of this manuscript. Special thanks are owed to Patrick Videau, not only for taking such good care of me during the homestretch of graduate school, but for inspiring me to pursue my Ph.D. in the first place.
ABSTRACT

The ability to make judgments about degrees of similarity is fundamental to cognition, human or otherwise. Reasoning about the ways in which two things are similar allows people to make logical inferences, create and understand analogy, and organize the world into meaningful categories. Non-human animals are also sensitive to degree of similarity as evidenced by over 100 years of research into associative learning mechanisms. Entropy detection is an appealing explanation for non-human performance on same-different discrimination tasks because it is a zero-parameter model that does not rely on abstract concepts of “same” and “different.” The research presented herein indicates that the choices humans make when judging the similarity or difference of visual stimuli, of varying complexity and spatial alignment, are consistent with perceived degree of similarity and of difference being functions of entropy. Additionally, this work demonstrates that one need not invoke the labels of “same” or “different” to prompt humans to sort items based on entropy or to use levels of entropy to construct higher-order relations of sameness and difference. These results offer a highly parsimonious account of human same-different discrimination because they are consistent with the animal literature and, therefore, suggest a conserved mechanism that is potentially involved in many aspects of both human and non-human animal cognition.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................... i

ABSTRACT ............................................................................................................................ ii

LIST OF TABLES ....................................................................................................................... v

LIST OF FIGURES ................................................................................................................... vi

CHAPTER 1: INTRODUCTION ................................................................................................. 1

CHAPTER 2. BACKGROUND .................................................................................................. 5

  Geometric Models of Similarity and Multidimensional Scaling (MDS) ................. 5

  Featural Models of Similarity ........................................................................................... 11

  Alignment-Based Models of Similarity .......................................................................... 17

  An Entropy-Based Account of Perceptual Similarity .................................................... 28

CHAPTER 3. SUMMARY OF MASTER’S WORK AND DISSEPTION AIMS ......................... 30

  Methods ............................................................................................................................... 32

  Results & Discussion ........................................................................................................ 37

  Aims of the Present Research ........................................................................................ 39

CHAPTER 4. EXPERIMENT 1 – DIFFERENCE WITH COMPLEX STIMULI AND REANALYSIS OF MASTER’S WORK ................................................................. 41

  Methods ............................................................................................................................... 41

  Results ................................................................................................................................. 42

  Discussion .......................................................................................................................... 47
LIST OF TABLES

Table 1 – Feature codes for generation of complex stimuli.............................................. 32
Table 2 – Fill codes for generation of stimuli..................................................................... 35
Table 3 – Summary of MA results.......................................................................................... 37
Table 4 – Combined effect of predictors in Exp. 1................................................................. 44
Table 5 – Effects of predictors in isolation in Exp. 1 ............................................................. 44
Table 6 – Combined effect of predictors in Exp. 2................................................................. 53
Table 7 – Effects of predictors in isolation in Exp. 2 ............................................................. 53
Table 8 – Average percent error on Exps. 1 and 2................................................................. 55
Table 9 – Icon distributions from Young & Wasserman (2001) ............................................ 61
Table 10 – Statistics calculated in Exps. 3A and 3B .............................................................. 63
Table 11 – Correlations between statistics calculated in Exp. 3A....................................... 65
Table 12 – Correlations between statistics calculated in Exp. 3B....................................... 67
Table 13 – Correlations between statistics calculated in Exps. 3A and 3B ......................... 67
LIST OF FIGURES

Figure 1 – Block letters used by Tversky (1977) ................................................................. 6
Figure 2 – The Triangle Inequality ............................................................ 8
Figure 3 – Plant stimuli used by Tversky and Gati (1982) .............................................. 10
Figure 4 – Two pairs of block letters used by Gati and Tversky (1984) ......................... 16
Figure 5 – Schematic depicting the comparison of two two-object scenes in SIAM ...... 21
Figure 6 – Comparison of firing patterns in LISA and DORA ........................................ 25
Figure 7 – Developing structured representations from unstructured input ................ 27
Figure 8 – Complex stimuli used in Exp. 1 ........................................................................ 31
Figure 9 – Simple stimuli used in Exps. 2 and 3 of my Master’s work .......................... 34
Figure 10 – Performance of predictors in Exp. 1 ............................................................... 46
Figure 11 – Non-aligned stimuli used in Exp. 2 ............................................................... 51
Figure 12 – Performance of predictors in similarity conditions of Exps. 1 and 2 ....... 52
Figure 13 – Performance of predictors in difference conditions of Exps. 1 and 2 ....... 54
Figure 14 – Average error on congruent trials of Exps. 1 and 2 ................................. 55
Figure 15 – Grids used in Exps. 3A and 3B ................................................................. 59
Figure 16 – Comparison of participants who did and did not pass training in Exp. 3B... 66
CHAPTER 1: INTRODUCTION

In general terms, cognition is the process by which animals take in information through their senses, transform that information into neuronal impulses, and then process it in some manner. The results of this processing can ultimately affect the animal’s behavior, thus the ability to acquire information about the world and make use of it is of tremendous adaptive significance. Therefore, researchers have worked for centuries to describe the method(s) by which animals interact with and learn about the world. Over time, many theories on the nature of cognition have been proposed but none of them has adequately defined all of the principle components of cognition. Similarity plays a central role in many theories of human cognition, specifically, because it is a key component of a wide range of cognitive operations including language acquisition (e.g., Kemler Nelson, 1989), categorization (e.g., Rosch, 1973; Smith & Heise, 1992), and relational reasoning (e.g., Taylor & Hummel, 2009). Consequently, many theories of cognition are dependent upon similarity. Despite its centrality, there is no grand consensus on what similarity is, how it is detected and judged, or even whether it is a useful construct. Goodman (1972) and others argue that identifying two things as “similar” requires specifying the ways in which they are similar, and that this referencing of shared properties is the actual information-bearing part of any statement about similarity. Medin, Goldstone, and Gentner (1993) challenge this argument by claiming that the respects in which two things can be similar are constrained by the comparison process itself. Although what it means for two things to be “similar” may be disputed, research has shown that, over the course
of cognitive development, humans’ abstract concept of *sameness* is refined from more concrete precursors.

The ways in which similarity judgments are made changes over the course of development; children initially process objects holistically (e.g., Kemler Nelson, 1989) and gain flexibility in their understanding of similarity though experience and learning (e.g., Gentner & Rattermann, 1991). In congruence with evidence in favor of holistic processing, children tend to assume that novel labels belong to whole objects and infer that these labels can be applied to other similar objects (E.M. Markman, 1990). When deciding which objects one can generalize a new label to, children are sensitive to linguistic context. In a study with 36-month-olds, Landau, Smith, and Jones (1988) demonstrated that toddlers preferentially attend to the shape of a novel object when it is paired with a new word presented as a count noun, and to color when the new word is presented as an adjective. Likewise, when the novel object has eyes, children as young as two to three years of age have been shown to focus on shape (Landau et al., 1988). Together, the constraints that context places on meaning and inference allow children to use assessments of perceptual similarity to acquire language at an astonishing rate and to create meaningful categories of conceptual similarity with which to organize the world around them (Medin et al., 1993; Smith & Heise, 1992).

There are features in the world that tend to co-occur such that sets of correlated features may result in the emergence of natural categories (Rosch, 1973). The presence of a given feature or combination of features is conceptually relevant to the extent that it is predictive of other features. It is these correlations between features that underlie the development of categories and ground concepts in reality (Smith & Heise, 1992).
Experience results in both implicit knowledge about relations between perceptual features and explicit knowledge about features relevant to category membership. Due to the capacity-limited nature of attention, one can only attend to a subset of features in the environment. Thus, knowledge of features and of their relationships to one another contribute to which features one attends to in different contexts; the fact that attention can be shifted in a context-dependent manner allows perceptual similarity to be dynamic. Additionally, the ability to perceive similarities is informative because it allows one to make inferences rather than having to learn everything about each object, event, relationship, etc. in isolation (Goldstone & Son, 2012). The ability to gauge the extent to which two things are similar is useful because the probability of making a correct inference increases as a function of similarity (i.e., the more similar two things are, the more likely it is that one would be correct in inferring that if one of them has some property, so does the other; Tenenbaum, 1999). Without the ability to detect similarity in a degree-sensitive manner, there could be no generalization because it would be impossible to compare two situations, which are never exactly identical, or even two instances of interacting with the same object, which can never occur in exactly the same context. Generalization is an essential component of learning; therefore, because it depends upon the ability to detect similarities, similarity is a topic of considerable importance within many areas of psychology.

Because similarity is crucial for early learning and we rely on the ability to perceive similarity and the knowledge we gain from this process throughout our lives, understanding human similarity processing is a fundamental component of understanding human cognition. As a consequence, many models of human cognitive processes (e.g.,
similarity-based models of categorization; Goldstone, 1994; Nosofsky, 1984) rely on some type of similarity judgment. To that end, there is a long history of attempts to generate a formal description of how similarity is computed. Prior to the widespread availability and use of computers, cognitive theorists explored mental processes via simple(r) mathematical modeling. Thus, we will begin our discussion of the ways in which various computational models of human cognition assess similarity by first addressing the early conceptualizations of how the mind represents items and judges their similarity.
CHAPTER 2. BACKGROUND

In order to be tested, hypotheses regarding the underpinnings of human cognition are translated into formal models, which make predictions that are then compared to experimental data. Early research on judgments of similarity led to the development of geometric models of similarity, which were later supplanted by set theoretical models based on shared and unique features. The advent of computers facilitated advanced modeling techniques that have integrated increasingly complex processes to approximate human cognitive processes more accurately than ever before. As each model has built upon the tenets of the previous, they will be presented chronologically.

Geometric Models of Similarity and Multidimensional Scaling (MDS)

Geometric models were among the first to provide a formal description of judgments of similarity. In this family of models, items are represented as points in \( n \)-dimensional space, wherein each spatial dimension corresponds to some featural dimension. Multidimensional scaling (MDS) is a group of techniques which use algorithms to fit a geometric model to a set of proximity data (Steyvers, 2002) collected from subjective measures of similarity, thereby creating a representation of relative similarities of the stimuli (Goldstone & Son, 2012). In these representations, the geometric distance between points is inversely proportional to the similarity of the two stimuli they represent (i.e., the farther away two items are, the less similar they are; A.B.
Markman, 1999). The discrepancy between a model’s predictions and the actual observed data is indicative of how well that model accounts for the data, or its goodness of fit.

The goodness of fit of an MDS-derived model is limited by its number of usable dimensions. As the number of dimensions increases, the fit improves; however, the model also becomes more difficult to interpret because what each dimension may represent becomes less and less obvious (Goldstone & Son, 2012). Dimensions that cannot be explained are neither meaningful nor informative; thus most MDS-derived models used in psychology have three useful dimensions, and very rarely more than five (A. B. Markman, 1999). Though geometric models remain highly influential in the study of similarity, they have been heavily criticized.

**Tversky’s criticisms.** Geometric models rely on two fundamental assumptions: (1) that items can be represented by a point in n-dimensional space, and (2) that dissimilarity can be calculated as a metric distance in this space. Tversky (1977) challenged both of these assumptions by claiming that dimensional representations are not appropriate for some stimuli and by experimentally demonstrating that there are cases in which three axioms of metric distance functions are violated.

![Figure 1](image)

*Figure 1 – Block letters used by Tversky (1977). Participants are more likely to confuse two stimuli when one includes the other. For example, block letter B2 contains A2 but not C2. Therefore A2 and B2 are more likely to be incorrectly labeled as “same” than are B2 and C2.*
The minimality axiom. The minimality axiom states that the distance between a point and itself is zero, \( d(x, x) = 0 \); the distance between a point and another point is greater than zero, \( d(x, y) > 0 \); and that this holds true for all points, \( d(x, y) > d(x, x) = d(y, y) = 0 \). In the case of similarity, this means that the similarity between Object A and itself is the same as the similarity between Object B and itself, which is the same as the similarity between any other object and itself. In an experiment in which participants were shown two block letters (see Figure 1) in sequence, Tversky (1977) demonstrated that certain pairs of stimuli are more likely than others to be incorrectly labeled as identical. These results violate the minimality axiom because, if the similarity between an object and itself is consistent across all objects, then the probability of incorrectly labeling two items as identical should likewise be consistent.

The symmetry axiom. The distance between point \( x \) and point \( y \) can be defined as the length of the line \( \overline{XY} \). According to the symmetry axiom, the distance from point \( x \) to point \( y \) is equivalent to the distance from point \( y \) to point \( x \); \( d(x, y) = d(y, x) \). Applied to a geometric model of similarity, this translates as the relation of sameness being non-directional. When making statements about similarity (e.g., \( x \) is like \( y \)), however, the subject (\( x \)) and the referent (\( y \)) are not interchangeable; switching them can change both the implied meaning of the statement and how similar the two items are rated to be. For example, flipping a simile affects which features appear relevant to the comparison and hence alter its meaning. The statement, “that surgeon is like a butcher,” for instance, means something very different from the statement, “that butcher is like a surgeon” (Medin et al., 1993). In the first case, it is implied that the surgeon is careless, rough, and
perhaps inclined to do more cutting than is necessary; whereas the second statement is a
compliment to the butcher, suggesting that he is very skilled and precise.

To demonstrate the asymmetrical nature of judgments of similarity, Tversky
(1977) presented participants with pairs of statements following the format of, “Country A is like Country B,” and “Country B is like Country A,” and asked them to choose the statement that they preferred to use. Each pair of countries was constructed such that they differed in how well known they were to the participants. Tversky found that for each of the 21 pairs of countries he presented, the majority of participants selected the statement in which the more well-known country served as the referent and the lesser-known country was the subject than vice versa. For example, when given the statements, “North Korea is like Red China, and “Red China is like North Korea,” 95.7% of participants chose the first statement as the one they preferred.¹

---

¹ At the time that this experiment was conducted (ca. 1977), China was more prominent than North Korea due to the recent occurrence of the Chinese Cultural Revolution (1966-1976) and death of Chairman Mao Zedong (1976).
The triangle inequality. The third and final condition of a metric distance function that is violated by the similarity data is the triangle inequality (see Figure 2). This axiom states that, for any three points in space, the distance from point \( x \) to point \( y \) cannot be more than the sum of the distance from point \( x \) to point \( z \) and the distance from point \( y \) to point \( z \); \( d(x, y) \leq d(x, z) + d(y, z) \). The only instance in which the distance from point \( x \) to point \( y \) is equal to the distance from point \( x \) to point \( z \) plus the distance from point \( y \) to point \( z \) is when point \( z \) lies on the line \( XY \). When point \( z \) is not on \( XY \), the length of \( XZ \) plus the length of \( ZY \) has to be greater than the length of \( XY \).

In the case of similarity, this rule of metric distance has been interpreted to mean that if \( x \) is similar to \( y \), and \( y \) is similar to \( z \), then \( x \) cannot be very dissimilar to \( z \). Experiments have shown that this axiom is violated because, in some cases, \( x \) and \( y \) are not necessarily similar in the same way that \( y \) and \( z \) are similar, thus \( x \) and \( z \) may, in fact, be very different. One classic example used to illustrate this point involves comparing three countries two-at-a-time (Tversky, 1977). First, consider Jamaica and Cuba – one might think of these two countries as rather similar because they are both tropical islands and are in close geographical proximity to one another. Next, consider Cuba and Russia – at the time when this example was coined, they were both Communist nations and were therefore considered similar on the basis of their political affinity. Finally, compare Jamaica and Russia – not very similar. So, although Jamaica is similar to Cuba (\( d(x, y) \) is small), and Cuba is (was) similar to Russia (\( d(y, z) \) is also small), Jamaica and Russia are not similar (\( d(x, z) \) is large) because the ways in which the countries are similar differs and thus these comparisons violate the triangle inequality. Judgments of the similarity of visual stimuli have also been shown to violate the triangle inequality.
In one such experiment with visual stimuli, Gati and Tversky (1984; Tversky & Gati, 1982) presented participants with line drawings of plants that varied along two separable dimensions: the shape of their pot and the shape of their leaves (see Figure 3). Participants were shown two plants at a time and were asked to rate their dissimilarity on a 20-point scale. The authors found that pairs of stimuli that differed in both their pot shape and leaf length (e.g., plants ap and cr) were rated as more dissimilar than would be expected given the dissimilarity ratings given to pairs of stimuli that only differed along one dimension (e.g., plants ar and cr and plants ap and ar). In response to the many failings of geometric models, Tversky (1977) proposed the Contrast Model, which belongs to the larger class of featural models.

Figure 3 – Plant stimuli used by Tversky and Gati (1982, Fig. 6). Participants rate plants ap and cr (solid red lines) as more different than would be expected given their ratings of each in comparison to ar (dashed red line).
Featural Models of Similarity

In featural models, items are represented as sets of features and degree of similarity is determined by a matching function. Unlike geometric models of similarity, featural models account for variability in feature salience and predict asymmetries in ratings of similarity (Tversky & Gati, 1978; Goldstone, 1993). In these models, changing the weights associated with various features (i.e., making certain properties more or less salient) in a context-dependent manner makes similarity dynamic. Experimental evidence indicates that shared features are weighted more heavily than unshared features when stimuli are verbal rather than visual (Gati & Tversky, 1984), or cohesive rather than non-cohesive (Ritov, Gati, & Tversky, 1990); when making a judgment of similarity rather than a judgment of difference (Tversky, 1977); and when there are a large number of unshared features rather than shared features (Gati & Tversky, 1984). Tversky’s (1977) Contrast Model was designed to be able to account for these experimental phenomena.

The Contrast Model. Tversky’s (1977) Contrast Model considers items as sets of features and computes the similarity of two items as a weighted function of shared \((A \cap B)\) and unique \((A - B)\) and \((B - A)\) features; the weights \((\theta, a, \text{ and } \beta)\) vary with feature salience.

\[
S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)
\]

---

2 A matching function is any function that satisfies the assumptions of matching and monotonicity. Matching assumes that the similarity of items \(a\) and \(b\) \((S(a,b))\) is a function of the features they share \((A \cap B)\), the features that \(a\) has that \(b\) does not \((A - B)\), and the features that \(b\) has that \(a\) does not \((B - A)\). Monotonicity assumes that similarity increases as the number of shared features increases and as the number of unshared features decreases (Tversky, 1977).
Types of properties. There are two types of properties that contribute to feature salience. Intensive properties are those that pertain to the intensity of various aspects of the stimulus such as brightness and loudness. Intensity is based on perception and is therefore relatively stable. Diagnostic properties are those that carry information relevant to the task at hand and are therefore context-dependent. For example, when trying to figure out whether a dog is a fish or a mammal, noting the presence or absence of fur is quite useful. However, when trying to determine whether this dog is a pug or a spitz, the feature has fur is no longer informative. In the same pivotal publication in which he introduced the Contrast Model, Tversky (1977) also brought attention to another featural model called the Ratio Model.

The Ratio Model. The Ratio Model, which is a formalization of the Property Theory proposed by Sjöberg (1972; Tversky, 1977), is very similar to the Contrast Model in that both models represent items as sets of features and calculate similarity as a weighted function of shared and unique features in which the weights vary as a function of feature salience. The defining difference is that, in the Ratio Model, shared features are divided by total features (i.e., shared plus unique features).

\[
S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}
\]

This operation results in the output of the function being bound between 0 (completely different) and 1 (identical). Restricting the range of possible values of a measure in this way is useful because it allows one to directly compare across instances. Frequency Theory was proposed to expand the utility of Property Theory by providing an account of
how the context in which a judgment of similarity is being made affects the judgment itself.

*Frequency Theory.* An alternative to Property Theory is Frequency Theory (Sjöberg, 1972), which states that similarity is not only a function of shared and unshared properties, but also of the context in which the similarity judgment is being made (i.e., the *universe*). By this view, the two stimuli being compared belong to a *common class* defined as the set of all items that possess the same shared properties that are present in the two stimuli. By taking context into account, Frequency Theory predicts that (1) the larger the common class, the *less* similar any two members of that class will be judged to be; and (2) the larger the universe, the *more* similar any two members of that class will be judged to be. The second prediction of the frequency theory has been confirmed in many experiments and is now called the *extension effect* (e.g., Tversky, 1977). When the set under consideration (i.e., the universe) is expanded to include members that do not possess some feature that is common to all members of the original set, said feature gains diagnostic value. This additional diagnostic value causes the feature to become more salient, thereby resulting in the members of the original set being perceived as more similar to each other than they were originally (Tversky & Gati, 1978; Medin et al., 1993).

For instance, let us consider the common class *crayons* in the context of the universe of *writing implements*. Starting with a set comprised of just two crayons, both belonging to the common class *crayons*, they initially seem somewhat similar – they are the same general size and shape, both are made of wax and have some paper around them; but, at the same time, they may be different colors, have different wording on their
wrappers, and be worn down to different degrees. If the set under consideration was expanded to include dozens of additional members of the common class crayons, this same pair of crayons would seem less similar to each other in that context than in the previous one in which they were viewed in isolation. On the other hand, if the context was instead broadened to include all members of the universe of writing implements (e.g., pens, markers, pencils, styli, chisels, etc.), not just crayons, these two crayons would seem even more similar to each other than they did initially. This example illustrates that the perceived similarity of two items can change based on context, which not only includes the size of the set under consideration, but the frequency of that feature within the set.

Both of the effects exemplified above occur because the salience of a feature is the product of its diagnostic value, and the diagnostic value of a feature varies as a function of its frequency of occurrence in the set under consideration. For example, let us assume that when comparing the two crayons in isolation, the shared feature made of wax contributes substantially to their perceived similarity. However, when the common class is expanded to fifty crayons, all of the members of this set have the feature made of wax, and thus this feature loses diagnostic value. Conversely, when the set under consideration shifts from being entirely made up of crayons to also containing non-crayon writing implements, the fact that the non-crayon members of this universe lack the made of wax feature results in this feature gaining diagnostic value.

Tversky (1977) provided a demonstration of context affecting subjective ratings of similarity through yet another experiment in which participants were asked to rate the similarity of countries. In this case, participants were presented with eight pairs of
countries; the countries within each pair shared a border. The different pairs of countries
within a participant’s set could be from one continent (e.g., all eight pairs of neighboring
countries were European) or two (e.g., some pairs were European and others were South
American). If similarity was independent of context, one would expect the similarity
ratings of Countries A and B (e.g., Italy and Switzerland) to be consistent across groups.
This was not the case – participants who received pairs from two continents rated
Countries A and B as more similar than did the participants who received countries that
all lie on the same continent. These results are an instance of the extension effect in that,
when all of the countries under consideration are on the same continent, the feature
continent (e.g., Europe) has no diagnostic value; expanding the universe to include
countries on different continents makes continent a relevant feature. The results presented
by Tversky support the notion that judgments of similarity are influenced by the size of
the set in question and the relations between the members of the set.

**Criticisms.** While featural models of similarity have many advantages over
traditional geometric models, they are not without shortcomings. For one, there are two
conditions that must be met in order for something to be considered a matching function:
monotonicity and matching. Matching assumes that the similarity between two items is a
function of their shared and unshared features\(^3\) (Tversky, 1977). However, this
calculation fails to take into account that mutually unshared features could also be
interpreted as a commonality (Goldstone, Medin, & Gentner, 1991). For example, when

\( S(a,b) = f(A \cap B, A-B, B-A) \)

\(^3\)
considering a crayon and a pencil in the context of other writing implements (e.g., pens and markers), not capped could be seen as a shared feature.

Moreover, Tversky’s (1977) Contrast Model assumes independence, meaning that in judgments of similarity, the combined effect of two features is independent of some third invariant feature. This assumption does not account for features combining to create global/ emergent features (Gati & Tversky, 1984; Goldstone et al., 1991). For example, consider the block letters in Figure 4: Pair A is identical to Pair B, save for the additional horizontal line shared by the items in Pair B. If features are independent of each other, then Pair B should be judged to be more similar because those items share one more common feature than Pair A do. However, participants judge Pair A to be more similar because the addition of the horizontal line closes one of the figures, thereby giving rise to the global features of open versus closed and confounding the distinction between simple shared and unique features (Gati & Tversky, 1984). One method of accounting for the effects of global or emergent properties is to provide a way to encode and evaluate information pertaining to the relationships between features. Alignment-based models of similarity accomplish this by operating on richly structured representations of stimuli.

![Figure 4](image)

**Pair A**

**Pair B**

*Figure 4* – Two pairs of block letters used by Gati and Tversky (1984). Participants judge Pair A to be more similar than Pair B, even though Pair B has more common features than Pair A.
Alignment-Based Models of Similarity

Perhaps the most significant drawback to models that treat all items under consideration as simple collections of features, as featural models do, or values along various dimensions, as geometric models do, is that they do not take the structure of the items (i.e., the relationships between those features) into account. Computational models of analogy (e.g., Gentner, 1983) were among the first to make use of structured representations (Goldstone, 1994, 2001). Alignment-based models of similarity have taken inspiration from these models of analogy and built upon models of similarity that use feature matching. As a result, unlike geometric and featural models, alignment-based models make similarity based on relations possible by employing matching functions that are sensitive to both features and to the roles that these features are playing (Falkenhainer, Forbus, & Gentner, 1989).

Structural Alignment. Structured representations can be viewed as being composed of feature Possessing objects playing roles (Goldstone, 1994). In alignment-based models, comparing two such representations involves mapping elements of one representation to elements of the same type belonging to the other representation (e.g., object-to-object) via structural alignment. Structural alignment is a process that takes two structured representations and attempts to find the most structurally consistent way to match them (Gentner, 1983, 1989). This means favoring one-to-one correspondences – mapping each element of one representation to no more than one element in the other representation – and parallel connectivity – mapping both a role and its associated objects to a corresponding role and objects (Gentner & A. B. Markman, 1994). One notable
achievement of models that use structural alignment is their success on cross-mapped analogies.

A cross-mapped analogy is an analogy in which the two scenes (or representations) under consideration contain both an object match and a relational match and these two matches are in conflict, i.e., two elements in one representation match the same one element in the other representation. For example, when comparing a scene in which a dog is chasing a cat to one in which a cat is chasing a mouse, a purely featural model would choose the cat in the first scene as being most similar to the cat in the second. However, alignment-based models are capable of overcoming object matches because they prefer role-based similarity to featural similarity (Goldstone & Son, 2012). In this case, that would result in the dog in the first scene being mapped to the cat in the second based on their role as chaser and the cat in the first scene being mapped to the mouse in the second based on their role as chased. The ability to perform cross-mappings is significant because finding relation-based correspondences allows one to make more powerful inferences (Holyoak, 2012). For example, if one knows that the mouse is running from the cat because the cat will harm the mouse if the cat catches it, mapping the mouse in this instance of the chases relation to the cat in the other instance would result in inferring that the cat is running from the dog because the dog will likewise harm the cat if the cat is caught. This process of structural alignment results in matches, alignable differences, and non-alignable differences.

4 Cross-mapping a dog chasing a cat and a cat chasing a dog would be even more difficult because, in this case, both objects are identical; their roles are the only thing that has changed.
Matches can be aligned elements or aligned systems of interconnected elements forming a common structure (e.g., a stool and a chair are both designed to be sat upon). Alignable differences are those that are related to these aligned systems, i.e., they are aligned elements that do not match (e.g., a cat is covered in fur and a fish is covered in scales). Non-alignable differences are those that are separate from aligned systems, i.e., they are elements that are not placed in correspondence with other elements (e.g., a helicopter has doors and a frog has ???. A. B. Markman & Gentner, 1993a, 1993b). Differentiating between alignable and non-alignable differences offers an attractive explanation for the non-symmetrical nature of judgments of similarity and difference.

Tversky (1977) noted that when presented with two pairs of stimuli, participants would often choose the same pair as being “most similar” or “most different,” depending upon the condition. For example, when participants were given West Germany and East Germany versus Ceylon and Nepal, 67% of participants in the similarity condition and 70% of participants in the difference condition selected West Germany and East Germany. Tversky explained this effect by claiming that making similarity judgments highlights commonalities, whereas making difference judgments highlights differences. Goldstone et al. (1991) expanded upon this idea, suggesting that relations are more important in similarity and attributes are more important in differences. However, if similarity and difference are both calculated via structural alignment, relations should be equally important in computing differences (A. B. Markman, 1996). Differentiating between the two types of differences offers an explanation for this asymmetry in that alignable differences are hypothesized to be more salient than non-alignable differences (Gentner & A. B. Markman, 1994; A. B. Markman & Gentner, 1996). This claim has
gained further support from the finding that participants have an easier time listing differences for more similar pairs (presumably because those differences are alignable; Gentner & A. B. Markman, 1994) and that alignable differences are more memorable than non-alignable differences (A. B. Markman & Gentner, 1997). Some models that make use of alignment-based similarity apply an analogous distinction to aligned verses non-aligned matches.

**Similarity, Interactive Activation, and Mapping (SIAM).** Goldstone’s (1994) *Similarity, Interactive Activation, and Mapping* model (SIAM) offers an explanation for why mismatches have a greater impact on similarity when in better-aligned pairs, i.e., why a pair of matching features influences similarity to the extent that it is consistent with the overall pattern of alignment. SIAM is a network of nodes representing potential correspondences between scenes; it compares two scenes by finding correspondences between their objects, the features of those objects, and the roles that those objects are playing. There are three types of correspondences: feature-feature, object-object, and role-role. Nodes representing correspondences that are consistent with each other (i.e., yield one-to-one mappings) are mutually excitatory whereas those that are inconsistent (i.e., yield many-to-one mappings) inhibit each other.

The schematic presented in Figure 5 depicts the object-object nodes (grey cells) and feature-feature nodes (white cells) of two scenes with two objects each (role-role nodes not pictured). The first cell in the first row represents Object A from Scene 1 corresponding to Object C from Scene 2. This node has inhibitory connections to the node representing Object A corresponding to Object D and the node representing Object B corresponding to Object C because the mappings A-D and B-C are inconsistent with the
mapping A-C. The node representing Object A corresponding with Object C has an excitatory connection to the node representing Object B corresponding to Object D because the mapping B-D is consistent with the mapping A-C in that if A corresponds to C, then B must correspond to D in order to avoid a non-one-to-one mapping. The influence that an individual featural match exerts on overall similarity is dependent upon whether it is consistent with other correspondences.

**Scene 2**

![Diagram](#)

**Figure 5** – Schematic depicting the comparison of two two-object scenes in SIAM (adapted from Goldstone, 1994, Fig. 8). Each cell is a node that represents a potential correspondence. Each node shares excitatory connections with all of the other nodes that represent correspondences consistent with it, and shares inhibitory connections with those that represent correspondences that it is inconsistent with. For simplicity, the only connections shown here are object-object node to object-object node.

When considering featural matches, Goldstone (1994) differentiates between *matches out of place* (MOP’s) and *matches in place* (MIP’s). MOP’s are features that match but belong to non-aligned objects; MIP’s are featural matches that belong to aligned objects, i.e., are consistent with existing object correspondences. SIAM makes a
number of experimentally supported predictions regarding MIP’s and MOP’s that purely featural models do not. For example, because nodes that represent inconsistent mappings inhibit each other, and those that are consistent excite each other, SIAM predicts that the impact of MIP’s on similarity is greater than that of MOP’s and that the effect of MOP’s is negated entirely when they compete against MIP’s. (For additional predictions and empirical support, see Goldstone, 1994). Although SIAM and models like it provided some initial insight into the process of comparing stimuli and judging their similarity, they were highly specialized for making certain kinds of comparisons and were therefore not broadly applicable.

**LISA & DORA.** The early alignment-based models are sensitive to roles in some sense (e.g., relative spatial position of objects in scenes), and are therefore a great improvement over models that do not take any relational information into consideration. However, there are many aspects of more complex types of comparisons that these models do not address. For example, the capacity for relational thought (i.e., thinking about things in terms of the roles that they play) is a cornerstone of human cognition (Penn, Holyoak, & Povinelli, 2008). Thinking about a relation requires that said relation be represented independently of the objects playing the roles in the relation and the features of those objects. In order to fulfill this requirement in a model, the relation (or roles in the relation) must correspond to an element and this element has to be the same across instances of this relation (e.g., the relation *above* must correspond to the same element regardless of what is above what; Hummel & Biederman, 1992; Hummel & Holyoak, 1997). The element corresponding to the relation also has to be able to dynamically bind to the elements playing the roles in the relation (Doumas, Hummel, &
Sandhofer, 2008). Furthermore, a complete account of relational thinking must explain where these relational elements originate. Relations never occur in isolation (e.g., you cannot experience *aboveness* without objects), therefore the thinker must be able to detect relational invariants and isolate them from features in order to represent relations as independent entities.

*How LISA represents relations.* The LISA (Learning and Inference with Schemas and Analogies) model created by Hummel and Holyoak (1997) solved many of the aforementioned problems by recasting multi-place predicates as linked single-place predicates (i.e., role-filler bindings) and representing them in a symbolic connectionist architecture (see Hummel & Holyoak, 1997). For example (see Figure 6a), the two-place predicate *cooks*(Sam, tamale pie) can be decomposed into *cooker*(Sam) and *cooked*(tamale pie). In LISA, elements corresponding to predicates and objects are connected to units corresponding to their features (semantic units); unlike many other models, predicates and objects share the same pool of semantic units, therefore the same features are used for both (e.g., red as a feature or as a predicate is represented by the same unit). Predicates (*cooker*) and objects (Sam) are bound to each other through sub-proposition units\(^5\) (Sam+*cooker*). Sub-proposition units are linked into higher *arity* structures (i.e., structures that can take more arguments) by proposition units. As with previous models, comparison is done via the process of structural alignment.

Representing complex relational structures in this manner reduces the problem of learning them into learning single-place predicates (i.e., roles or properties) and linking

---

\(^5\) In DORA, sub-proposition units are called role-binding units.
them. LISA encodes binding information via synchrony of firing (see Figure 6a). After a proposition unit fires, the first sub-proposition and everything below it (the predicate and object units to which it is bound and semantic units to which they are bound) fire in synchrony, then the second sub-proposition unit and everything below it fire in synchrony.

What DORA does differently. One weakness of the LISA model is that binding by synchrony precludes learning predicates from objects because, when predicate and object units and the semantic units to which they are bound fire together, the model cannot differentiate between semantic units that are bound to the predicate, the object, or both. In the case of our “cooker(Sam) and cooked(tamale pie)” example, LISA would not know that the feature human belongs to both Sam and the role cooker. This lack of role-filler independence is the basis for LISA’s inability to learn that human is a feature of cookers from examples of humans cooking. Doumas et al.’s (2008) DORA (Discovery of Relations by Analogy) model overcomes this limitation by using systematic asynchrony of firing (see Figure 6b). After a proposition unit fires in DORA, the first role-binding unit\textsuperscript{6} and its role (i.e., predicate unit), and the semantic units to which that role is bound, fire together. Then that same role-binding unit fires with its object and its associated semantic units. After that, the same sequence of firing occurs in the second role-binding unit and everything below it. This singular difference between LISA and DORA allows DORA to do everything that LISA can do, plus learn structured representations from unstructured input.

\textsuperscript{6} In LISA, role-binding units are called sub-proposition units.
Figure 6 – Comparison of firing patterns in LISA and DORA (adapted from Doumas et al., 2008, Fig. 3). (a) In LISA, role-binding units and all of the units they are associated with fire in synchrony. (b) When a role-binding unit fires in DORA, its predicate and object units fire in succession.
When comparing two propositions, binding by systematic asynchrony of firing allows DORA to isolate and predicate shared properties. Some of these shared properties may be irrelevant to the concept; the process of iterative comparison refines predicates because the relevant features are invariant and the irrelevant ones are different across instances of the concept. For example, let us say that the first time Margeaux encounters a tamale pie, it is a cold Pyrex baking dish of chili with a layer of cornbread baked on top. All of these features are part of her initial representation of tamale pie. The next time Margeaux has tamale pie, it is fresh out of the oven; in an aluminum baking pan; and contains chili, cheese and cornbread. The features of tamale pie that are common to both of these instances – namely, a baking vessel with chili and cornbread – will be more strongly associated with Margeaux’s concept of tamale pie and the unique features (e.g., contains cheese) will be less so. Over the course of encountering successive instances of tamale pie, Margeaux will continue to refine her representation of tamale pie based on the presence or absence of features and by comparing instances of tamale pie stored in her memory. This refinement process will result in the development of a concept of what it means for something to be a tamale pie and Margeaux will be able to reason about tamale pie in increasingly complex ways. For example, Margeaux might think of tamale pie as playing an analogous role to tuna casserole when planning a meal: when tuna casserole is served as a main dish, there is no starchy side dish because the casserole contains noodles, which are a starch; tamale pie contains cornbread, which is also a starch, and should therefore not require a starchy side dish. Margeaux thinking of tamale pie in terms of its role in a meal is an example of relational reasoning, a capacity that is acquired during development.
One well-established trend in cognitive development is the gradual shift from reasoning about things in a holistic fashion to being capable of relational thinking (see Smith, 1989; Gentner & Ratterman, 1991). Very young children have difficulty reasoning about properties independent of the objects to which they belong, suggesting that children’s initial mental representations are holistic. By the age of about three, children have explicit representations of object properties, as evidenced by their ability to notice partial similarities (e.g., the ball and the apple are both red). These representations develop into functional single-place predicates (e.g., a concept of redness that can be applied to anything) and are eventually linked to form structured representations of whole relations (e.g., redder than).

Like children, DORA begins with unstructured representations of objects, which are compared in order to develop abstract (i.e., predicated) concepts (see Figure 7). It is important that any model of human cognition

Figure 7 – The process of developing structured representations from unstructured input, as predicted by DORA.
not only succeed at the same things humans do, but fail in the same ways as well. At intermediate points in the process of comparison-based predication, DORA’s performance on various tasks matches that of children in the same stage of development. Because DORA’s development follows the same trajectory as children’s, DORA is the first computational model to provide an account of how structured representations can arise from unstructured input (Doumas et al., 2008). Although successive generations of models have gained explanatory power in regard to how sameness is judged (e.g., via structural alignment) and elucidated the pivotal role of judgments of similarity in human cognition and cognitive development (e.g., DORA), explicitly defining the perceptual properties that these processes act on has gone relatively unaddressed.

An Entropy-Based Account of Perceptual Similarity

Another answer to the question of how similarity is processed has arisen in the animal literature, and makes predictions that are often compatible with those made by representational models that calculate similarity as a function of proportion of shared features (e.g., the Ratio Model). Several studies have examined whether non-human animals are capable of acquiring and using conceptual knowledge, especially the categories of “same” and “different.” Wasserman, Young, and colleagues have conducted many such cross-species studies of performance on visual same-different discrimination tasks and have concluded that pigeons and baboons – and perhaps humans as well – make use of some form of entropy detection when making decisions about sameness and difference (Wasserman, Young, & Cook, 2004; Young & Wasserman, 2001). The relative uniformity of these results across the disparate species tested suggests that the
ability to detect and form categories based on variation in levels of entropy is likely to be present in other species (Wasserman, Fagot, & Young, 2001).

Entropy is a measure of disorder within a system; in the context of the study of similarity, it is a measure of diversity within a set. Calculating entropy for a set of stimuli incorporates both the number of categories of elements within the set and the relative frequencies of said categories (Young & Wasserman, 2001). Studies of entropy detection generally quantify entropy using the following equation (Shannon & Weaver, 1949): 

$$E_A = - \sum_{a \in A} p_a \log_2 p_a$$

where $a$ is a category of variable $A$, and $p_a$ is the number of elements of category $a$ divided by the number of elements within the set; there are no free parameters. Entropy is maximized when all of the elements are different (i.e., belonging to different categories) and is zero when all elements are identical (see Appendix A for example calculations).

Entropy detection offers a powerful and parsimonious explanation of same-different categorization in non-humans because its use does not require an explicit concept of same or different, and works as a zero-parameter model (unlike the Contrast and Ratio Models, which have three and two free parameters, respectively). It is possible that an entropy-based mechanism may underlie not only the capacity to detect basic same and different (i.e., degrees of uniformity), but also serves as the foundation for the development of explicit representations of same and different in humans (see e.g., Smith, 1989).
CHAPTER 3. SUMMARY OF MASTER’S WORK AND DISSERTATION AIMS

Although some models of human cognition (e.g., DORA) offer solutions to how a system might isolate subsets of features in order to develop structured representations of concepts, what these features are and where they come from are problems that remain to be addressed. A considerable body of research spanning multiple decades suggests that same and different are very likely to be among the most elemental of features; entropy detection, a zero-parameter model, provides an elegant and parsimonious account of where these features might originate. Additionally, entropy detection comes at no extra cost to systems that represent items as distributed patterns (brains for example) and perform comparisons (as many successful models of learning do), making it a desirable alternative to featural models which are not so transparently instantiable.

Traditional featural models (e.g., the Contrast Model and the Ratio Model) and alignment-based models all represent stimuli as sets of features and calculate similarity via some matching function, yet the predictions they make are not identical. The Contrast Model predicts that the greater the number of features shared by two items, the more similar they will be judged to be; whereas the Ratio Model predicts that the greater the proportion of features shared, the more similar two items will be judged to be. When comparing two pairs of stimuli, both of these heuristics often predict the same result because the pair that has the greater number of features in common is likely to also have the greater proportion of shared features. Likewise, predictions made in accordance with entropy-based accounts tend to be congruent with those that are ratio- and number-based.
My Master’s thesis focused on how and under what circumstances ratio-, number-, and entropy-based metrics are used in judgments of visual similarity and difference. Experiments 1 and 2 explored judgments of the similarity of pairs of complex and simple visual stimuli, respectively. Experiment 3 used the same simple stimuli as Experiment 2 in judgments of difference. The initial purpose of this research was to determine which of two ways of quantifying similarity provided a better account of participants’ choices: number, as predicted by the Contrast Model, or ratio, as predicted by the Ratio Model and some alignment-based models. The idea of exploring entropy as a possible predictor arose sometime in the middle of this project; therefore entropy-based analyses were post-hoc.
Methods

Participants. Undergraduate participants were recruited from psychology courses at the University of Hawai‘i at Mānoa and were given the option of volunteering to participate in one thirty-minute session in our lab. There were no specifications on age, gender, or race and no personal information beyond their responses was recorded.

Participants were given a consent form (see Appendix C) and asked to take a moment to read it and ask any questions that they might have, then either verbally agree or not agree to participate in the study. Each participant was notified that s/he may refuse to participate before the experiment begins and that s/he may choose to withdraw from the study at any time thereafter without prejudice. Seventy-three (N = 73) undergraduates participated in Experiment 1 (similarity with complex stimuli), one hundred forty-eight (N = 148) participated in Experiment 2 (similarity with simple stimuli), and one hundred twenty-four (N = 124) participated in Experiment 3 (difference with simple stimuli).

Table 1 – Feature codes for generation of complex stimuli.

<table>
<thead>
<tr>
<th>Cell</th>
<th>Feature</th>
<th>Absent (0)</th>
<th>Option 1 (1)</th>
<th>Option 2 (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>head</td>
<td>-</td>
<td>round</td>
<td>square</td>
</tr>
<tr>
<td>1</td>
<td>horns</td>
<td>-</td>
<td>curve</td>
<td>bolt</td>
</tr>
<tr>
<td>2</td>
<td>antenna stalk</td>
<td>-</td>
<td>straight</td>
<td>curved</td>
</tr>
<tr>
<td>3</td>
<td>antenna tip</td>
<td>-</td>
<td>ball</td>
<td>star</td>
</tr>
<tr>
<td>4</td>
<td>arms</td>
<td>-</td>
<td>1 joint</td>
<td>2 joints</td>
</tr>
<tr>
<td>5</td>
<td>hands</td>
<td>-</td>
<td>pincers</td>
<td>fingers</td>
</tr>
<tr>
<td>6</td>
<td>legs</td>
<td>-</td>
<td>1 joint</td>
<td>2 joints</td>
</tr>
<tr>
<td>7</td>
<td>feet</td>
<td>-</td>
<td>solid</td>
<td>fluffy</td>
</tr>
<tr>
<td>8</td>
<td>eyes</td>
<td>-</td>
<td>lens</td>
<td>compound</td>
</tr>
<tr>
<td>9</td>
<td>nose</td>
<td>-</td>
<td>arc</td>
<td>trapezoid</td>
</tr>
<tr>
<td>10</td>
<td>mouth</td>
<td>-</td>
<td>square</td>
<td>round</td>
</tr>
<tr>
<td>11</td>
<td>mouth feature</td>
<td>-</td>
<td>tongue</td>
<td>teeth</td>
</tr>
<tr>
<td>12</td>
<td>wings</td>
<td>-</td>
<td>butterfly</td>
<td>fruit fly</td>
</tr>
<tr>
<td>13</td>
<td>body width</td>
<td>N/A</td>
<td>thin</td>
<td>fat</td>
</tr>
<tr>
<td>14</td>
<td>body segmentation</td>
<td>N/A</td>
<td>1 segment</td>
<td>2 segments</td>
</tr>
<tr>
<td>15</td>
<td>color</td>
<td>N/A</td>
<td>blue</td>
<td>green</td>
</tr>
</tbody>
</table>
Creation of stimuli. The stimuli used in these experiments were 100%-alignable in order to control for the known effects of relations and alignability. In Experiment 1, the stimuli being compared were images of pairs of cartoon bugs (see Figure 8) displayed on a computer. To create these stimuli, pairs of sixteen-item arrays (i.e., feature vectors) that met the following conditions were randomly generated: indexes 0-12 could take the values 0, 1, or 2; and indexes 13-15 could only take the values 1 or 2. Each cell coded for a feature (see Table 1); 1’s and 2’s code for different variants of the same features (e.g., round versus square head shape) and 0’s code for the absence of features. Restricting the possible values of indexes 13-15 resulted in every bug having, at minimum, a body of some shape, size, and color. To translate these arrays into images, an image was made of each feature option, and these component images were layered on top of each other according to the feature vectors generated.

The feature vectors were generated such that the bugs within any given pair are missing the same features (i.e., a feature cannot be present in one bug and absent in the other) and each feature that is present is either common to the pair or unique. These conditions result in the paired bugs not having any non-alignable differences. To facilitate subsequent analysis, the stimuli are labeled \( n_x \) where \( n \) is the number of shared features, and \( x \) is the total number of features present in each member of the pair.
The features of the stimuli used in Experiment 1 were of varying salience and, although the similarities and differences were alignable, they were not easily separable. To obtain more reliable results, the stimuli used in Experiments 2 and 3, which were spatially aligned pairs of column arrays (see Figure 9), were composed of features of more consistent salience; these features were both discrete and spatially aligned in order to facilitate comparison. A similar but much simpler process than that used to create the stimuli for Experiment 1 was used to generate these stimuli. In this case, arrays ranging in size from two to eight items were generated and each index could take a value between 1 and 16. Each one of these values coded for a specific fill pattern (see Table 2) and each index corresponded to a specific position in the stimulus pair. Again, all matches were spatially aligned and one-to-one. Once these arrays of values were made, an empty pair of array stimuli of the appropriate length was copied and its cells were filled with the indicated patterns.

**Figure 9** – Simple stimuli used in Experiments 2 and 3 of my Master’s work. (a) A congruent trial: Pair A shares 2 of 6 features and Pair L shares 1 of 5 features. (b) An incongruent trial: Pair A shares 3 of 8 features and Pair L shares 1 of 2.
Data collection and analysis. Participants sat at a desk on which there was a computer screen and a keyboard. Two pairs of visual stimuli were presented on the screen simultaneously and participants were asked to choose one of the pairs according to some criterion by pressing the key corresponding to his/her selection. In Experiments 1 and 2, participants were asked to choose the pair of stimuli that was the “most same”; in Experiment 3, they were asked to choose the pair that was the “most different.” To select the pair on the left (hereafter designated as Pair A), they pressed the “A” key; to select the pair on the right (hereafter designated as Pair L), they pressed the “L” key. (See Appendix D for experimenter script.) Each participant went through 100 trials. Responses and reaction times were recorded automatically; no information tying specific responses to specific participants was recorded.
The data from all of the participants within a given experiment were aggregated and, for each trial, the pairs of stimuli’s difference in number \((N)\), ratio \((R)\), absolute global entropy \((E_A)\), and relative global entropy \((E_R)\) were calculated.\(^7\) To address our original research question of whether \(N\) or \(R\) is the better predictor of judgment of similarity and difference, incongruent non-endpoint trials (i.e., trials in which neither of the pairs of stimuli matched completely or mismatched completely and \(N\) and \(R\) predicted different outcomes) were isolated from the data set. The proportion of trials that was chosen in accordance with \(R\) \((p(ratio))\) was calculated for each combination of \(R\) and \(N\)\(^8\) and linear regressions were run using the R statistical package (R Development Core Team, 2010) to determine which metric best accounted for the observed data.

The error rate of each experiment was calculated for congruent non-endpoint trials (i.e., those in which number and ratio make the same predictions) in order to facilitate the comparison of the relative difficulty of the tasks and to serve as a reference point for looking at the choice-rates in the incongruent trials.

Investigating the roles of absolute and relative entropy was not part of our original plan; therefore the only analysis that was conducted in regard to entropy was determining which of the two metrics best accounted for each complete data set. Prior to fitting logistic regression models to each data set, \(E_R\) was divided into bins of 0.06 for both types of stimuli and \(E_A\) into bins of 0.5 for the complex stimuli and 0.4 for the simple

\(^7\) See Appendix A for how each measure is calculated, example calculations, and a detailed explanation of how the measures relate to each other.

\(^8\) Combinations of \(R\) and \(N\) were grouped by their absolute value. For example, the trials in which \(R = 0.5\) and \(N = -2\) were grouped with trials in which \(R = -0.5\) and \(N = 2\).
stimuli, then fitting a logistic regression model to each. These binning ranges were chosen to make the number of bins in each analysis roughly equal, thereby minimizing the problems that can arise from the models having very different numbers of degrees of freedom.

**Results & Discussion**

Participants’ choices were consistent with judging similarity and difference as a function of the ratio of shared features in the majority of trials in all three experiments (see Table 3). In all three experiments, participants’ choices on incongruent non-endpoint trials had a significant positive correlation with the difference in the pairs’ ratio of shared features ($R; p < .001$) and no significant correlation with the pairs’ number of shared features ($N$). The differences in the error rates of Experiments 1 and 2 indicated that judging the similarity of pairs of the simple stimuli (arrays) is easier than judging the similarity of pairs of the complex stimuli (bugs). Additionally, Experiment 3, the difference task with simple stimuli, was more difficult than either of the similarity tasks.

**Table 3** – Summary of MA results. In all three experiments, $R$ outperformed $N$ in predicting participants’ behavior on incongruent trials. The error rates in congruent trials suggest that Experiment 3 was the most difficult, followed by Experiment 1, then Experiment 2. When applied to the entire data set for each experiment, $E_R$ consistently outperformed $E_A$.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Incongruent Trials</th>
<th>Congruent Trials</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$ (%)</td>
<td>$N$ (%)</td>
<td>Error (%)</td>
</tr>
<tr>
<td>1</td>
<td>73.40</td>
<td>26.60</td>
<td>24.55</td>
</tr>
<tr>
<td>2</td>
<td>62.67</td>
<td>37.33</td>
<td>15.08</td>
</tr>
<tr>
<td>3</td>
<td>81.01</td>
<td>18.99</td>
<td>27.84</td>
</tr>
</tbody>
</table>
Both absolute entropy ($E_A$) and relative entropy ($E_R$) were shown to be excellent predictors of participants’ choices ($p < .001$ for both), with $E_R$ being able to account for a greater proportion of the data. When these data were analyzed for my Master’s thesis, the two entropy-based metrics were not compared to ratio. However, this comparison was eventually conducted for the similarity data only and it was discovered that, on incongruent non-endpoint trials, ratio ($R$) was a much better predictor for the simple stimuli (75.11% versus 24.89%) whereas relative entropy was a moderately better predictor for the complex stimuli (42.47% versus 57.53%). Although this finding initially appears to be in opposition to an entropy-based account of judgments of similarity, it may be due to the stimuli being fully alignable and spatially aligned and can be accounted for by entropy. Specifically, when comparing a pair of stimuli whose features are all in one-to-one correspondence, it is possible to consider each pair of features in isolation, yielding two possible outcomes for said pair: either the two features match ($E_A = E_R = 0$) or they do not ($E_A = E_R = 1$). The entropy of each corresponding pair of features can then be averaged to yield the pair’s pairwise entropy ($E_{PW}$). This possible interpretation of participants’ preference for $R$ provided the impetus for further experimentation to determine whether this same metric is capable of accounting for judgments of difference and to ascertain whether manipulating the spatial alignment of stimuli would compel participants to use global measures of entropy ($E_A$ or $E_R$).
Aims of the Present Research

Because similarity plays a central role in cognition, I wanted to continue to investigate metrics that humans may use to make judgments of sameness and difference. The possibility of recasting ratio ($R$) as an entropic function ($E_{PW}$) in order to explain participants’ preferences provided the impetus for further experimentation attempting to better elucidate the role of entropy detection in human judgments of similarity and difference. Experiments similar to those in my Master’s thesis were carried out to ascertain whether manipulating the spatial alignment of stimuli would compel participants to use global measures of entropy ($E_A$ or $E_R$). Experiment 1 of the work presented herein involves conducting a difference task with complex stimuli (bugs). The results from Experiment 1 were analyzed with the previously completed work in hopes of gaining a more complete understanding of how the type of stimuli (simple versus complex) and the nature of the task (similarity versus difference) interact to influence the ways participants make judgments of sameness and difference.

One potential shortcoming of the experiments described above is that the stimuli used are fully alignable and spatially aligned. As such, participants’ apparent preference for a ratio-based metric can be accounted for by pairwise entropy ($E_{PW}$). One issue with this interpretation is that pairwise entropy and ratio are complements of each other (e.g., if a pair shared five of their seven features, their proportion of features shared would be $5/7$ and $E_{PW} = 2/7$), therefore the predictions that these models make are entirely

---

9 All of the experiments presented herein have been approved for exempt status by the UH Human Studies Program (see Appendix B for formal documentation).
congruent. Experiment 2 addressed this weakness by extending these findings to stimuli that are not spatially aligned.

Experiment 3 of the present research focused on participants’ ability to first categorize stimuli based on their entropy without invoking the labels of “same” and “different” and then use entropy to form higher-order relationships. Together, these three experiments provide evidence not only for humans being able to detect entropy, as other animals have been shown to do, but also to use entropy as the basis for developing the uniquely human abstract concepts of *sameness* and *difference*.
CHAPTER 4. EXPERIMENT 1 – DIFFERENCE WITH COMPLEX STIMULI AND REANALYSIS OF MASTER’S WORK

Experiment 1 built upon the work presented in my Master’s thesis by adding a difference experiment with complex stimuli (bugs). I reexamined the data collected for my Master’s thesis and compared it to the data from Experiment 1 in order to form a more complete story about how people make judgments of similarity and difference of these types of stimuli.

Methods

Participants. A total of 89 undergraduates participated in this experiment. Participants were recruited and asked to consent to participate in the study in the same manner described in Chapter 3.

Creation of stimuli. As in Experiment 1 of my Master’s work (see Chapter 3), the stimuli used in this experiment were images of pairs of cartoon bugs (see Figure 8) displayed on a computer.

Data collection and analysis. Participants were asked to choose the pair of stimuli that was the “most different” and data were collected in the same manner described in Chapter 3. Each participant’s error rate on congruent trials was calculated and compared to chance via a chi-square test. Participants whose scores were not
different from chance \((p > .05)\) and those who performed significantly below chance were excluded from analysis.

All data were aggregated and, for each trial, the pairs of stimuli’s difference in number \((N)\), ratio \((R)\); and, consequently, pairwise entropy, \(E_{PW}\), absolute global entropy \((E_A)\), and relative global entropy \((E_R)\) were calculated.\(^{10}\) \(R\) was rounded to one decimal place; \(E_A\) and \(E_R\) were binned by ranges of 0.4 and 0.06, respectively. Endpoint trials (those in which at least one stimulus pair is a complete match or complete mismatch) and symmetrical trials (those in which both pairs shared the same \(n\) of \(x\) features) were excluded from analysis and like trials were grouped.

A series of logistic regressions were run using the R statistical package (R Development Core Team, 2010) to determine which metric(s) best accounts for the observed data. Data originally presented in my Master’s thesis were reanalyzed in the same manner as described above.

**Results**

The results of the analyses of the new and the previously collected data were compared to determine whether the same metric(s) accounts for participants’ choices across all four conditions or whether they differed in a systematic way (e.g., one metric

---

\(^{10}\) Please see Appendix A for how each measure is calculated, example calculations, and a detailed explanation of how the measures relate to each other. The logic presented therein is critical for understanding the theoretical foundations on which all of the experiments presented in this manuscript were based.
provides the best account for similarity judgments whereas another provides the best account for difference judgments). The large number of degrees of freedom within each analysis results in inflated levels of significance, thereby rendering $p$-values rather uninformative. Therefore, I have elected to compare models by presenting the percent of variance they account for as a measure of goodness of fit (see Table 4 for associated $p$-values). In these models, the four predictor variables were considered both in isolation, to determine the relative ability of each to account for the data collected, and in combination to ascertain the predictors’ overall ability to account for the data and to investigate possible instances of multicollinearity. Accounting for variance in this manner facilitates direct comparisons across conditions to elucidate the method of detection used by participants to detect similarity in these trials.

Similarity with simple stimuli (arrays). Of the 148 participants who took part in this experiment, data from two were discarded because they performed at or below chance ($n = 146$). Using all four predictors to model this data set accounted for 86.65% of the variance in the array similarity data. The only significant predictor in this model was ratio ($R$), with the other three not reaching significance in the context of the other predictors (see Table 4). Considering all four predictors in isolation, $R$ was the best, accounting for 84.39% of the variance. This is just 2.26% less than what was accounted for by using all four predictors; thus, adding additional predictors does not significantly

---

11 Usually this statistic, the coefficient of determination, would be denoted as $R^2$. However, I have chosen to use the longer, descriptive label throughout this manuscript due to the potential for confounding $R^2$ with $R$ (ratio, one of our four predictor variables).
improve the fit of the model. Interestingly, although $E_A$ was the second-best predictor when considering all four predictors together, it accounted for the least amount of variance in isolation, indicating that the variance accounted for by $N$ and $E_R$ has a high degree of overlap with $R$ (see Table 5).

Table 4 – Combined effect of predictors in Experiment 1. The $p$-values associated with each predictor and the total percentage of variance that is accounted for when using all four predictors are presented here.

<table>
<thead>
<tr>
<th></th>
<th>Array Similarity</th>
<th>Array Difference</th>
<th>Bug Similarity</th>
<th>Bug Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>$&lt; 2 \times 10^{-16}$</td>
<td>$&lt; 2 \times 10^{-16}$</td>
<td>$2.21 \times 10^{12}$</td>
<td>.171</td>
</tr>
<tr>
<td>$N$</td>
<td>.085</td>
<td>$2.25 \times 10^{-4}$</td>
<td>.819</td>
<td>.004</td>
</tr>
<tr>
<td>$E_A$</td>
<td>.057</td>
<td>$4.91 \times 10^{-13}$</td>
<td>$3.37 \times 10^{-6}$</td>
<td>$6.94 \times 10^{-15}$</td>
</tr>
<tr>
<td>$E_R$</td>
<td>.878</td>
<td>.168</td>
<td>.554</td>
<td>$1.67 \times 10^{-4}$</td>
</tr>
<tr>
<td>Total %</td>
<td>86.65</td>
<td>87.83</td>
<td>67.00</td>
<td>63.00</td>
</tr>
</tbody>
</table>

Table 5 – Effects of predictors in isolation in Experiment 1. The percent of the total variance that is accounted for by each predictor in isolation is presented here; $p < 2 \times 10^{-16}$ unless otherwise noted.

<table>
<thead>
<tr>
<th></th>
<th>Array Similarity</th>
<th>Array Difference</th>
<th>Bug Similarity</th>
<th>Bug Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>84.39</td>
<td>69.14</td>
<td>63.51</td>
<td>54.22</td>
</tr>
<tr>
<td>$N$</td>
<td>42.82</td>
<td>0.01, $p = .472$</td>
<td>32.84</td>
<td>23.59</td>
</tr>
<tr>
<td>$E_A$</td>
<td>9.95</td>
<td>64.65</td>
<td>20.49</td>
<td>29.09</td>
</tr>
<tr>
<td>$E_R$</td>
<td>51.46</td>
<td>81.95</td>
<td>62.84</td>
<td>60.24</td>
</tr>
</tbody>
</table>

**Difference with simple stimuli (arrays).** Of the 149 participants who took part in this experiment, data from 15 were discarded because they performed at or below chance ($n = 134$). Using all four predictors accounted for 87.83% of the total variance in the data set. The most significant predictor in this model was ratio ($R$), followed closely by absolute entropy ($E_A$). Relative entropy ($E_R$) was not significant in the context of the
other predictors (see Table 4). Surprisingly, $E_R$ was the best predictor when each was applied in isolation, accounting for 81.95% of the total variance. $E_A$ and $R$ in isolation accounted for 64.65% and 69.14% of the variance, respectively, whereas $N$ captured less than 1% (see Table 5). Using both $E_R$ and $E_A$ as predictors accounted for 82.53% of the variance, which is less than a 1% improvement over $E_R$ in isolation. The same lack of improvement was seen when adding $R$ as a predictor; $E_R$ and $R$ together account for 83.37% of the variance, less than 1.5% more of the variance than $E_R$ does in isolation. Interestingly, $R$ and $E_A$ – the second- and third-best when looking at predictors in isolation – together account for 87.47% of the variance, which is just 0.36% less than all four predictors together. Adding $E_R$ to this model does not increase the percent of the variance accounted for at all, indicating that $R$ and $E_A$ together capture all of the variance accounted for by $E_R$.

**Similarity with complex stimuli (bugs).** All 73 of the participants who took part in this experiment performed above chance ($n = 73$). When applied together, all four predictors accounted for 67.00% of the variance. The most significant predictor in this model was ratio ($R$), followed by absolute entropy ($E_A$). Neither relative entropy ($E_R$) nor number ($N$) was significant in the context of the other predictors (see Table 4). Examining each predictor in isolation, ratio ($R$) and relative entropy ($E_R$) performed almost identically, accounting for 63.51% and 62.84% of the total variance, respectively (see Table 5). Using both $R$ and $E_R$ as predictors accounted for 65.61% of the variance, which is about 2% more than $R$ in isolation. This lack of increase in variance accounted for indicates that $R$ and $E_R$ are capturing the same variance. $R$ is slightly better at
capturing this variance than is $E_R$, thus explaining the non-significance of $E_R$ in the four-predictor model.

![Figure 10](image)

**Figure 10** – Relative performance of predictors in Experiment 1. Dividing the percent of the variance accounted for by each predictor in isolation by the total variance that is accounted for by all four predictors together makes the performance of each predictor directly comparable across conditions. For example, of the variance that was accounted for by all four predictors, ratio ($R$) alone was able to account for 97.39% and 78.72% in the Array Similarity and Array Difference conditions, respectively.

**Difference with complex stimuli (bugs).** Of the 89 participants who took part in this experiment, data from 17 were discarded because they performed at or below chance ($n = 72$). Collectively, all four predictors accounted for 63.00% of the variance. The most significant predictor in this model was absolute entropy ($E_A$), followed by relative entropy ($E_R$) and number ($N$). Contrary to what was expected, given the results of the previous three experiments, ratio ($R$) was not significant in the context of the other predictors (see Table 4). Considering each predictor in isolation, relative entropy ($E_R$) was the best, accounting for 60.24% of the variance, followed closely by ratio ($R$), which accounted for 54.22% of the variance. Absolute entropy ($E_A$) and number ($N$) performed comparably, accounting for 23.59% and 29.09% of the variance, respectively (see Table
5). Using all four predictors only accounted for 2.76% more of the variance than did $E_R$ in isolation; thus; the addition of other predictors does not significantly improve the fit of the model.

**Discussion**

Previous analysis of the similarity data collected for my Master’s thesis indicated that participants’ choices are best predicted by difference in ratio of shared features ($R$). I therefore predicted that $R$ would be the best predictor of participants’ choices across all four conditions. The use of $R$ is congruent with the use of pairwise entropy ($E_{PW}$) because they are complements of each other. Due to the alignability and spatial alignment of the stimuli, I did not expect participants to use either global entropy metric ($E_A$ or $E_R$). Surprisingly, upon reanalysis, this was only partially the case (see Figure 10). $R$ remained the best predictor for both of the similarity tasks, both in isolation and in combination with the other predictors. Also consistent across both similarity conditions was the finding that, although $E_A$ was the second-best predictor in the context of all of the other predictors, it was the worst in isolation; $E_R$ was the second-best predictor in isolation in both similarity tasks. Together, these findings indicate that there is a high degree of multicollinearity between $R$ and $E_R$ when accounting for judgments of similarity of spatially aligned stimuli.

In the difference with simple stimuli condition, in which participants chose the pair of arrays that they judged to be the most different, $R$ and $E_A$ were likewise the first- and second-best predictors when viewing the effects of all four predictor variables together. Contrastingly, in the difference with complex stimuli task, in which participants
chose the pair of bugs that they judged to be the most different, $E_A$ was vastly more significant than the other three predictors when they were applied together; $R$ did not reach significance. When investigating the difference task predictors in isolation, the relationships between predictor variables was similar, albeit somewhat reversed, compared to what was observed in the similarity conditions. $E_R$ was the best predictor in isolation for both difference tasks, followed by $R$. The consistent preference for $R$ and $E_R$ across all four conditions suggests that these two measures may be interrelated, e.g., that $R$ is being calculated as a function of entropy (i.e., $E_{PW}$).

These results exhibit an interesting pattern in which the metrics that best explain the participants’ behavior progressively shift as a function of task difficulty. Judgments of similarity are easier to make than are judgments of difference; in both similarity tasks, $R$ was the best predictor in combination and in isolation. Of the two difference tasks, the one with the array stimuli was easier due to the decreased variability in feature salience and increased separability of features. In this condition, $R$ was the best predictor in combination while $E_R$ was the best in isolation. In the most difficult of the four conditions, the difference task with bug stimuli, $E_A$ was the best predictor in combination and $E_R$ was the best predictor in isolation. One possible explanation for this observed shift from using ratio ($R$)/pairwise entropy ($E_{PW}$) to using global measures of entropy ($E_A$ and $E_R$) is that, in terms of ratios, difference is the antithesis of similarity; thus, difference is a more difficult relation to compute because it involves two steps: judging which is the most similar, then choosing the opposite. In contrast, global measures of entropy can be thought of as features that exist along a smooth continuum, with “same” and “different” representing opposite ends of the spectrum. Thus, deciding which of two pairs of stimuli
is “most similar” or “most different” is reduced to judging their relative position on this continuum, which is just one step. Additionally, this effect may be amplified in the bug condition because the features of the array stimuli are discrete and easily separable, whereas the bugs’ features are not.

The initial impetus for the creation of the array stimuli was the high degree of variability in the salience of the features of the bug stimuli (e.g., wings are a much more salient feature than are antennae tips and therefore have a larger impact on overall similarity of stimuli). The features of the array stimuli, on the other hand, are uniform in their size and shape and therefore should make more consistent contributions to same/different judgments. As in my initial analysis, it was expected that the more variable the feature salience is, the more variable the data would be. The finding that a greater proportion of the total variance was able to be accounted for in conditions using the array stimuli compared to those using the bug stimuli (86.65% and 87.83% versus 67.00% and 63.00%) provides support for the idea that not all features contribute equally to judgments of similarity and difference. Because the use of standardized features successfully reduced the variability of the data collected and facilitated the comparison process (as indicated by the lower error rates compared to when the bug stimuli were used, see Chapter 3), I elected to conduct a follow-up experiment with stimuli constructed from these same features.
CHAPTER 5. EXPERIMENT 2: SIMILARITY AND DIFFERENCE WITH NON-ALIGNED STIMULI

Experiment 2 attempted to differentiate between the use of entropy-based metrics and difference in ratio of shared features \((R)\) by inhibiting the use of pairwise entropy \((E_{PW})\). To accomplish this, stimuli comprised of features of uniform salience that were not spatially aligned were used in both a similarity (Experiment 2A) and a difference (Experiment 2B) task.

Methods

Participants. A total of 135 undergraduates (73 in Experiment 2A and 62 in Experiment 2B) participated in this experiment. The data of one of the participants in Experiment 2B were excluded from analysis due to failure to follow instructions. Participants were recruited and asked to consent to participate in the study in the same manner described in Chapter 3.

Creation of stimuli. The stimuli used in this experiment were images of pairs of groups of tiles (see Figure 11) displayed on a computer; groups ranged in size from two to eight tiles. The program I have written to generate these stimuli does so such that the two groups of tiles are separated spatially and that none of the tiles within a group occlude each other. Tiles are not spatially aligned, but all matches between groups are one-to-one (i.e., a tile in one collection cannot match more than one tile in the other.
The stimuli are labeled \( n_x \) where \( n \) is the number of shared tiles, and \( x \) is the total number of tiles in each collection; five stimuli were generated for each possible \( n_x \), yielding 252 stimuli total.

**Data collection.** Data were collected in the same manner described in Chapter 3. In Experiment 2A, participants were asked to choose the pair of stimuli that is the “most similar”; in Experiment 2B, they were asked to choose the pair of stimuli that is the “most different” (see Appendix D for experimenter script). New data were analyzed and compared to the results of Experiment 1, which included the reanalysis of the data originally presented in my Master’s thesis, in the same manner as in Experiment 1 (see Chapter 4).

**Figure 11** – Non-aligned stimuli used in Experiment 2. (a) A congruent trial: Pair A shares 2 of 6 features and Pair L shares 1 of 5 features. (b) An incongruent trial: Pair A shares 3 of 8 features and Pair L shares 1 of 2.
Results

Tiles similarity. Of the 73 participants who took part in this experiment, data from 15 were discarded because they performed at or below chance \( (n = 58) \). Fitting a model to the data from Experiment 2A, in which participants were asked to select the pair of collections of tiles that was the “most similar,” using all four predictors accounted for 55.92% of the variance, with only ratio \( (R) \) reaching significance (see Table 6). Looking at all four predictors in isolation, \( R \) was the best, accounting for 55.46% of the variance (see Table 7), which is less than 1% less than the percentage of the variance that was accounted for by all four predictors combined. Thus, adding other predictors does not significantly improve the fit of the model. Consistent with the patterns observed in Experiment 1, relative entropy \( (E_R) \) was also a good predictor in isolation, accounting for 40.39% of the total variance (see Figure 12).

![Figure 12](image-url) – Relative performance of predictors in similarity conditions of Experiments 1 and 2.
Table 6 – Combined effect of predictors in Experiment 2. The $p$-values associated with each predictor and the total percentage of variance that is accounted for when using all four predictors are presented here.

<table>
<thead>
<tr>
<th></th>
<th>Tiles Similarity</th>
<th>Tiles Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>$2.14 \times 10^{-10}$</td>
<td>$2.13 \times 10^{-4}$</td>
</tr>
<tr>
<td>$N$</td>
<td>.128</td>
<td>.008</td>
</tr>
<tr>
<td>$E_A$</td>
<td>.615</td>
<td>.347</td>
</tr>
<tr>
<td>$E_R$</td>
<td>.665</td>
<td>.570</td>
</tr>
<tr>
<td><strong>Total %</strong></td>
<td><strong>55.92</strong></td>
<td><strong>57.20</strong></td>
</tr>
</tbody>
</table>

Table 7 – Effects of predictors in isolation in Experiment 2. The percent of the total variance that is accounted for by each predictor in isolation is presented here; $p < 2 \times 10^{-16}$ unless otherwise noted.

<table>
<thead>
<tr>
<th></th>
<th>Tiles Similarity</th>
<th>Tiles Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>55.46</td>
<td>38.93</td>
</tr>
<tr>
<td>$N$</td>
<td>7.03, $p = 1.07 \times 10^{-12}$</td>
<td>0.81, $p = .029$</td>
</tr>
<tr>
<td>$E_A$</td>
<td>18.00</td>
<td>46.07</td>
</tr>
<tr>
<td>$E_R$</td>
<td>40.39</td>
<td>53.88</td>
</tr>
</tbody>
</table>

Tiles difference. Data from 33 of the 60 participants who took part in this experiment were discarded because they performed at or below chance ($n = 27$). Of the total variance present in the data from Experiment 2B, in which participants were asked to select the pair of collections of tiles that was the “most different,” $57.20\%$ was accounted for when using all four predictor variables; ratio ($R$) and number ($N$) were the only predictors in this model to reach significance (see Table 6). In isolation, relative entropy ($E_R$) performed comparably to the four-predictor model, accounting for $53.88\%$ of the total variance. In contrast to the previous five experiments, absolute entropy ($E_A$) was the second-best predictor in isolation; ratio ($R$) was third-best (see Table 7 and Figure 13).
Figure 13 – Relative performance of predictors in difference conditions of Experiments 1 and 2.

Discussion

Because spatial alignment facilitates the mapping of corresponding elements (see Chapter 2), the lack of spatial alignment of the component features of the stimuli used in Experiment 2 (tiles) was expected to make pairwise entropy ($E_{pw}$, the complement of $R$; see Chapter 3) more difficult to calculate and thereby suppress its use. Therefore, if participants’ apparent use of ratio ($R$) when judging the similarity of aligned stimuli (as seen in Experiment 1) was actually due to the use of $E_{pw}$, as hypothesized given the results of the initial reanalysis of the data collected for my Master’s thesis (see Chapter 4), they should have abandoned this tactic in favor of a global entropy metric ($E_A$ or $E_R$). Consistent with the similarity conditions of Experiment 1 and contrary to expectations, ratio ($R$) was the best predictor in Experiment 2A, both in the context of the other three predictors and in isolation. One possible explanation for why participants continued to use $R/E_{pw}$ in Experiment 2A is that being explicitly asked to make judgments of similarity prompts people to seek out shared features (Gati & Tversky, 1984). Thus, the
fact that all of the features shared between pairs of collections of tiles were in one-to-one correspondence (i.e., all matches were paired) may have led participants to make pairwise comparisons even though the stimuli were not spatially aligned.

Table 8 – Average percent error on Experiments 1 and 2. The error rates of the difference tasks were consistently higher than those of the similarity tasks, indicating that, for all three stimulus types, the difference condition was the more difficult one. When comparing pairs of spatially aligned stimuli, using discrete features of relatively uniform salience makes the task easier. This effect is outweighed, however, by removing spatial alignment, as indicated by the higher error rates observed in the conditions that employed stimuli that were not spatially aligned.

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>Similarity</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Aligned Stimuli (Arrays)</td>
<td>4.82</td>
<td>6.16</td>
</tr>
<tr>
<td>Complex Aligned Stimuli (Bugs)</td>
<td>11.54</td>
<td>15.98</td>
</tr>
<tr>
<td>Simple Non-Aligned Stimuli (Tiles)</td>
<td>14.17</td>
<td>18.74</td>
</tr>
</tbody>
</table>

* These values differ from those presented in Table 3 because, here, “congruence” is defined as the condition in which all four predictors are in agreement, whereas, in Chapter 3, “congruence” only took $R$ and $N$ into consideration.

Figure 14 – Average error on congruent trials of Experiments 1 and 2. (Error bars = SEM)
As in the difference condition with simple stimuli (arrays) of Experiment 1, the variance in the data collected in Experiment 2B was best captured by ratio (R) when all four predictors were used. Consistent with both of the previous difference tasks, relative entropy ($E_R$) was the single best predictor in isolation. Interestingly, although ratio ($R$) was the second-best predictor in isolation in both of the previous difference tasks, absolute entropy ($E_A$) was the second-best predictor in isolation in Experiment 2B. Given that the tiles difference task was the most difficult out of all six same/different conditions (as indicated by the error rate on congruent trials, see Table 8 and Figure 14), this finding may indicate another shift in the preferred metric as the difficulty of the judgment increases (e.g., due to using stimuli that are not spatially aligned rather than stimuli that are spatially aligned). Using features of less uniform salience increases task difficulty when comparing spatially aligned stimuli (e.g., those of the bug stimuli versus those of the arrays; see Table 8); therefore, this hypothesis could be tested by conducting another pair of same/different experiments using stimuli composed of highly variable features that are not spatially aligned.

Together, the results of Experiments 1 and 2 extend the findings of Young, Wasserman, and colleagues in multiple ways. First, whereas their work was restricted to a categorization task involving one type of stimulus (e.g., Young & Wasserman, 1997, 2001; Wasserman, Young & Nolan, 2000; Wasserman, Fagot, & Young, 2001; Castro, Young, & Wasserman, 2006) the work presented here demonstrates that measures of entropy can account, at least in part, for comparison tasks with pairs of several different types of stimuli. Another new (and perhaps more compelling) finding is that, under certain conditions, participants will continue to respond to differences in entropy in a
degree-sensitive manner (as indicated by the consistent ability of entropy-based metrics to account for the data collected across the entire range of values tested), even when explicitly instructed to make judgments of similarity and difference. This is in contrast to Young and Wasserman’s (2001) explanation for why participants who were trained using endpoints (i.e., displays containing icons that were either all identical or all unique) failed to treat entropy as a continuous dimension when presented with intermediate cases and instead would respond categorically – treating all stimuli that were not completely homogenous as one category. Specifically, the authors had concluded that this dichotomous response pattern was attributable to the properties of the training stimuli inadvertently invoking the labels of “same” and “different.” Thus, the methodology and findings presented here may represent a novel way of approaching and explaining humans’ use of entropy detection in future studies of judgments of similarity and difference.
 CHAPTER 6. EXPERIMENT 3: ENTROPY-BASED CATEGORIZATION

Young & Wasserman (2001) found that, in a same-different categorization task, the majority of the college students they tested performed in accordance with using absolute entropy, whereas a smaller number appeared to utilize relative entropy. In Experiment 3A, I attempted to verify this preference as well as determine the extent to which participants are sensitive to degrees of difference in entropy by testing their ability to sort high- and low-entropy stimuli without invoking the relational category labels of *same* and *different*.

The goal of Experiment 3B, was to determine whether participants who successfully completed 3A could use entropy to form higher-order relationships (i.e., relationships between relationships) of *sameness* and *difference*. This ability is of interest because the development of the concept of *same* in humans is characterized by a stepwise progression from concrete types of *sameness* to increasingly abstract types of *sameness* (see, e.g., Smith, 1989; Gentner 1988). It has been argued that the ability to make use of higher-order relational similarity is one of the fundamental aspects of human cognition that sets it apart from the cognitive abilities of other animal species (Penn et al., 2008).
Methods

Participants. Thirty-six undergraduates participated in Experiment 3A; all but two of these participants also participated in Experiment 3B. One participant’s data were excluded from analysis for failure to follow directions. Participants were recruited and asked to consent to participate in the study in the same manner described in Chapter 3.

Creation of stimuli. The same fills used in Experiment 2 (see Table 2) were used to form five-by-five grids (25 slots total) with tiles in 16 slots (see Figure 15). The program that I wrote for conducting this experiment generates these stimuli during run time by choosing one of the compositions (i.e., “icon distributions”) used by Young and Wasserman (2001; see Table 9 below), selecting a fill for each “icon type,” and placing it in an available slot; fill selection and tile placement are randomized. These grids were used individually in Experiment 3A and in pairs in Experiment 3B.

Figure 15 – Grids used in Experiments 3A and 3B. (a) A high entropy grid for which $E_A = 3.0$ and $E_R = 0.75$. (b) A low entropy grid for which $E_A = 1.0$ and $E_R = 0.25$. 
**Data collection and analysis.** Experiments 3A and 3B were conducted together. Participants sat at a desk on which there was a computer screen and a keyboard. At the beginning of each Experiment, instructions were given on the computer screen (see Appendix D). Once the participant read and understood the instructions, visual stimuli appeared on the screen and s/he categorized them by pressing the key corresponding to his/her selection. Responses and reaction times were recorded automatically; no information tying specific responses to specific participants was collected. Both experiments used software of my design. Only those participants who successfully completed Experiment 3A were asked to complete Experiment 3B.

Prior to the start of Experiment 3A, the program chose two grid types for which $E_A$ equals 1.0 ($E_R = 0.25$) and two grid types for which $E_A$ equals 3.0 ($E_R = 0.75$; see Table 9, Grid Types 02-04 and 08-10, respectively); these four grid types served as the training set. The program also randomly assigned one of the two response keys to be the “high entropy” category (i.e., correspond to the $E_A = 3.0/E_R = 0.75$ grids during training) and the other to be the “low entropy” category (i.e., correspond to the $E_A = 1.0/E_R = 0.25$ grids during training). In the training phase, participants were presented with single grids from the training set and asked to categorize them. They were not told what the categories are; they instead received visual feedback (see Appendix D) and learned the categories via trial and error. The training phase ended when the participant was correct on ten consecutive trials. If the participant had not learned the rule after 10 minutes, the experiment was terminated.
Table 9 – Icon distributions from Young & Wasserman (2001; Table 1). Each numbered column represents a fill pattern chosen at random; the numbers in those columns indicate how many times those fill patterns occur in a given Grid Type. Grid Types 00-12 all contain 16 tiles each. Grid Types 13-16 are each made up of tiles that are all different, i.e., non-repeating. Grid Types 17-20 are each made up of tiles that are all the same.

<table>
<thead>
<tr>
<th>Grid Type</th>
<th>Icon Distribution</th>
<th>$E_A^*$</th>
<th>$E_R^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>16</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>01</td>
<td>14 2</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>02</td>
<td>12 3 1</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>03</td>
<td>13 1 1 1</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>04</td>
<td>8 8</td>
<td>1.0</td>
<td>0.2</td>
</tr>
<tr>
<td>05</td>
<td>11 1 1 1 1 1</td>
<td>1.5</td>
<td>0.4</td>
</tr>
<tr>
<td>06</td>
<td>4 4 4 4</td>
<td>2.0</td>
<td>0.5</td>
</tr>
<tr>
<td>07</td>
<td>8 1 1 1 1 1 1 1</td>
<td>2.5</td>
<td>0.6</td>
</tr>
<tr>
<td>08</td>
<td>6 1 1 1 1 1 1 1</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>09</td>
<td>2 2 2 2 2 2 2 2</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>10</td>
<td>5 3 1 1 1 1 1 1</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>11</td>
<td>4 1 1 1 1 1 1 1</td>
<td>3.5</td>
<td>0.9</td>
</tr>
<tr>
<td>12</td>
<td>1 1 1 1 1 1 1 1</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>17</td>
<td>1 1</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>18</td>
<td>1 1 1 1</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>19</td>
<td>1 1 1 1 1 1 1 1</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td>20</td>
<td>1 1 1 1 1 1 1 1</td>
<td>3.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

* Although the values presented here are rounded to the tenths place for ease of presentation, those calculated by my program were truncated at 12 digits.

In the testing phase, participants were asked to categorize grids of all 21 types (see Table 9); they only received feedback on trials in which the grid is of the same type as those used in the training phase. For example, if a participant’s training set was composed of grids of types 03, 04, 08, and 10, s/he would receive feedback on a trial with a grid of type 10, but not on a trial with a grid of type 09, even though the $E_A$ and $E_R$ of these stimuli are equivalent. This periodic feedback served to remind the participants of
the sorting rule without providing reinforcement on any of the trials of interest. The midpoint between the entropy levels of “high entropy” and “low entropy” stimuli used in the training set is $E_A = 2.0$ ($E_R = 0.5$); therefore, testing stimuli for which $E_A$ is greater than 2.0 and $E_R$ is greater than 0.5 were considered congruent “high entropy” and stimuli for which $E_A$ is less than 2.0 and $E_R$ is less than 0.5 were considered congruent “low entropy.” For example, if an $E_A = 2.5/E_R = 0.6$ grid is presented, pressing the key that corresponds to “high” (i.e., the one that was correct for $E_A = 3.0/E_R = 0.75$ in the training phase) would be counted as a correct response whereas pressing the key that corresponds to “low” (i.e., the one that was correct for $E_A = 1.0/E_R = 0.25$ in the training phase) would be counted as an incorrect response.

In Experiment 3B, data were collected in the same manner as in Experiment 3A, except that stimuli were pairs of grids rather than individual grids and the categories were “same” (i.e., either both grids are high entropy or both grids are low entropy) and “different” (i.e., one grid is high entropy and the other is low entropy). Data from these experiments were not aggregated. Rather, each participant’s scores were compared within and between experiments in order to examine differences between individuals and whether an individual’s score on one task predicted his/her score on the other task. Four statistics were calculated for each participant in Experiment 3A (see Table 10); these statistics were examined in pairs to determine which, if any, are correlated. Similarly, three statistics were calculated for each participant in 3B; these, too, were examined in pairs. Statistics 1-3 were also compared across experiments to determine whether participants’ behavior was consistent across tasks.
Table 10 – Statistics that were calculated in Experiments 3A and 3B. All four statistics were calculated for each participant in Experiment 3A and Statistics 1-3 were calculated for each participant in Experiment 3B.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of trials required to train to criterion</td>
</tr>
<tr>
<td>2</td>
<td>Percent correct on congruent trials</td>
</tr>
<tr>
<td>3</td>
<td>Percentage of incongruent trials in which the response was consistent with $E_A$</td>
</tr>
<tr>
<td>4</td>
<td>Percentage correct on $E_R$ only trials</td>
</tr>
</tbody>
</table>

Results and Discussion

In both experiments, Statistic 1 was the number of trials required to train to criterion. Requiring participants to complete a minimum of twenty trials before moving on to the test phase caused a floor effect. Therefore, the trial on which each participant answered correctly for the tenth consecutive time was recorded as his/her number of trials to criterion. Statistic 1 was not correlated with any of the other measures in either experiment. It was, however, the only one of the four statistics calculated that was correlated across experiments ($r = .46$, $p < .05$; see Table 10). The finding that Statistic 1 was not predictive of any of the other statistics calculated is surprising because participants who are more sensitive to entropy should have learned the task more readily, therefore performance on congruent trials (Statistic 2) was expected to be negatively correlated with the number of trials required to reach criterion (Statistic 1). If this prediction had been borne out, it would have provided evidence for individual differences.

---

12 For example, if a participant’s tenth consecutive correct answer was on trial 17, and she “passed” training at trial 20, her score on Statistic 1 would be 17.
in entropy detection not only affecting the rate of category learning, but also performance once the categories had been learned.

Statistics 2 and 3 were the participant’s percentage correct on congruent trials (i.e., trials in which $E_A$ and $E_R$ were not in conflict) and the percentage of incongruent trials (i.e., trials in which $E_A$ and $E_R$ were in conflict) in which s/he categorized the stimulus in accordance with using $E_A$, respectively. Testing stimuli for which $E_A$ equals 2.0 (see Table 9, Grid Type 18) are not categorizable based on $E_A$ because they fall exactly half way between the $E_A$ of the training categories (1.0 and 3.0). As such, if participants are categorizing these stimuli by $E_A$ alone, they should be equally likely to assign them to one category as the other. Conversely, these stimuli are categorizable based on $E_R$ because they have the maximum level of relative entropy ($E_R = 1.0$). Therefore, the fourth and final statistic calculated in Experiment 3A was the percentage of these trials that the participant’s choice was consistent with having learned the categories based on $E_R$.

Inconsistent with the findings of Young and Wasserman (2001), significantly more participants favored $E_R$ over $E_A$ on the incongruent trials of Experiment 3A ($\chi^2(1, n = 32) = 12.5, p < .001$) and of Experiment 3B ($\chi^2(1, n = 27) = 4.5, p < .05$). Additionally, the results of Experiment 3A indicate that there is a significant negative correlation between the percentage of incongruent trials in which participants’ decisions were consistent with using absolute entropy, $E_A$ (Statistic 3), and their performance on congruent trials (Statistic 2; $r(31) = -.39, p < .05$; see Table 11). Likewise, there was also a significant negative correlation between participants’ preference for $E_A$ on incongruent trials (Statistic 3) and their performance on the trials that were only categorizable based
on relative entropy, $E_R$ (Statistic 4; $r(31) = -.43, p < .05$). Consistent with these findings, participants’ performance on the $E_R$ only trials (Statistic 4) was positively correlated with their performance on congruent trials (Statistic 2; $r(31) = .43, p < .05$). To test whether participants’ preferences for one entropy metric over the other in incongruent trials (Statistic 3) was consistent with their behavior on congruent trials, I first grouped congruent trials by grid type and calculated the proportion of trials in which the participant responded “high,” $p(\text{high})$, for each group. I then calculated the proportion of variance in $p(\text{high})$ that was accounted for ($r^2$) by $E_A$ and by $E_R$. I then took the difference of these two values ($r^2_{EA} - r^2_{ER}$) and found that it has a significant positive correlation with Statistic 3 ($r(31) = .43, p < .05$), indicating that participants who used $E_A$ more frequently on incongruent likewise used $E_A$ on congruent trials.

Table 11 – Correlations between statistics calculated in Experiment 3A. All $p$-values shown here are for two-tailed $t$ tests with 31 degrees of freedom.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>$r = -.09, p = .62$</td>
<td>$r = .09, p = .61$</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>$r = -.39, p = .03$</td>
<td></td>
<td>$r = .43, p = .01$</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>$r = -.43, p = .01$</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Seven of the 34 participants who successfully completed Experiment 3A did not pass the training phase of Experiment 3B; one of these seven was excluded from either analysis for failure to follow instructions. Compared to those who passed, participants who did not pass the training phase of Experiment 3B had performed significantly worse on the congruent trials (Statistic 2; $t(31) = 6.12, p < .001$), and the $E_R$ only trials (Statistic
4; $t(31) = 3.30, p < .01$, of Experiment 3A. As expected, most participants were able to perform the task they were given in Experiment 3B. However, the average percent correct on congruent trials (Statistic 2) was lower than in Experiment 3A (90% versus 73%).

Contrary to my predictions, none of the statistics calculated for Experiment 3B were significantly correlated with each other and only Statistic 1 correlated across experiments (see Tables 11 and 12). One explanation for these findings is that, whereas Experiment 3A only required participants to detect and categorize based on a perceptual feature (i.e., entropy), Experiment 3B required them to detect two perceptual features (i.e., the entropy of each grid), realize that these features form a relation (i.e., same or different), and use this relation to place stimuli pairs into categories. Therefore, not only did

**Figure 16** – Comparison of participants who did and did not pass the training phase of Experiment 3B. Participants who did not pass the training phase of Experiment 3B performed significantly worse on the congruent and relative entropy ($E_R$) only trials of Experiment 3A. (Error bars = SEM)
Experiment 3B involve a greater number of individual components, but also involved reasoning about relations, which is known to be more difficult than reasoning about perceptual features alone.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1 ( r = -.04, p = .84 )</th>
<th>2 ( r = -.29, p = .14 )</th>
<th>3 ( r = -.73, p = .47 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 12** – Correlations between statistics calculated in Experiment 3B. All \( p \)-values shown here are for two-tailed \( t \) tests with 25 degrees of freedom.

**Table 13** – Correlations between analogous statistics calculated in Experiments 3A and 3B. All \( p \)-values shown here are for two-tailed \( t \) tests with 25 degrees of freedom.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>1 ( r = .46, p = .02 )</th>
<th>2 ( r = .29, p = .14 )</th>
<th>3 ( r = .30, p = .13 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although many of the predicted relationships were not observed in these two experiments, the fact that most of those who participated in Experiment 3A were able to perform the single-grid categorization task indicates that humans are sensitive to degrees of entropy. Given that human judgments of similarity and difference are also known to be degree-sensitive, this finding is consistent with my hypothesis that entropy detection is the basis of said judgments. Furthermore, participants’ overall success at detecting relationships of similarity and difference in the entropy levels of pairs of grids and using this information to perform the two-grid same/different categorization task in Experiment 3B lends additional support to this hypothesis in that it provides an example of relational similarity emerging from featural similarity.
In Experiment 1, the ability of a number of different zero-parameter models to account for human similarity and difference judgments was compared. It was predicted that, in all four tasks, participants would use the difference in ratio \((R)\) and, consequently, pairwise entropy \((E_{PW})\) more frequently than they would use the difference in number of shared features \((N)\), difference in absolute entropy \((E_A)\), or difference in relative entropy \((E_R)\) to determine which of two pairs of stimuli was the most similar. This proved to be the case in the two similarity conditions, but did not hold for the difference conditions. In the difference with simple stimuli (arrays) condition, \(R\) managed to capture the greatest amount of variance in the context of the other three predictor variables. \(E_R\) was the single best predictor when each was tested in isolation, signaling that there is a great deal of overlap in the variance that these variables are accounting for, i.e., there is a high degree of multicollinearity. \(E_R\) was likewise the best predictor in isolation for the difference with complex stimuli (bugs) condition, with \(E_A\) capturing the most variance when all four predictors were used. Taking these findings into consideration, I then hypothesized that this shift toward employing measures of global entropy in judgments of difference may be the result of the increasing difficulty of the tasks (see Chapter 4); this hypothesis was further supported by the findings of Experiment 2B.

One limitation of Experiments 1 and 2 is that the use of fully alignable stimuli precludes differentiating between pairwise entropy \((E_{PW})\) and ratio \((R)\) because they are complements of each other and, therefore, they always make the same prediction.
Experiment 2 attempted to overcome this concern via comparison tasks using stimuli that were not aligned spatially (tiles). These tasks were designed to interfere with pairwise comparison, thereby encouraging the use of global entropy measures ($E_A$ and $E_R$).

Consistent with the results of Experiment 1 and contrary to my original expectations, $R$ continued to be the best predictor for the similarity condition (Experiment 2A) and global measures of entropy ($E_R$ and $E_A$) were responsible for the majority of the captured variance in the difference condition (Experiment 3B). This continued shift from using $R$ to using $E_R$ and $E_A$ of difference tasks supports my hypothesis that participants’ preferences vary as a function of the difficulty of the task. Although the lack of spatial alignment of the stimuli used in Experiment 2 (tiles) failed to prevent the use of $R/E_{PW}$ in the similarity condition, similarity may still be a function of entropy in that ratio can be cast as an entropic function (see above). The observed pattern of transitioning to the use of global measures of entropy as task difficulty increases lends support to this entropy-based interpretation of the similarity data.

Similarity is a fundamental component of many aspects of cognition, human or otherwise; therefore, gaining a better understanding of the mechanisms underlying judgments of similarity has the potential to provide great insight into the inner workings of a plethora of cognitive processes. There is a substantial body of research conducted by Young, Wasserman, and colleagues indicating that the ability to detect differences in levels of entropy is common to such disparate species as humans, pigeons, and baboons. Experiments 3A and 3B replicated some of the findings of Young and Wasserman (2001) and expanded upon their work by demonstrating that entropy-based metrics (i.e., $E_A$ and $E_R$) are able to account for human judgments of similarity and difference using multiple
types of visual stimuli, and that participants are not only capable of sorting visual
displays based on their entropy, but that they can use entropy to form the higher-order
relationships of sameness and difference without being explicitly prompted with the
labels of same and different. Participants’ success in these tasks provides further evidence
that entropy may be the foundation on which abstract concepts of sameness are built.
Together, these results have the potential to provide evidence for a highly plausible,
simple mechanism for judging similarity and difference in humans.
APPENDIX A: EQUATIONS AND EXAMPLE CALCULATIONS

Each of the equations for a specific entropy metric is based on the Shannon-Weaver equation for absolute entropy \( E_A \) of a set (Shannon & Weaver, 1949), in which \( a \) is a category of variable \( A \), and \( p_a \) is the number elements of category \( a \) divided by the number of elements within the set.

\[
E_A = - \sum_{a \in A} p_a \log_2 p_a
\]

To calculate the relative entropy of a set, divide the set’s absolute entropy by the maximum entropy of a set that size. For example, consider a set of 16 elements, 2 of which are unique and 14 of which are identical:

\[
E_A = - \left[ \left( \frac{1}{16} \cdot \log_2 \frac{1}{16} \right) \cdot 2 + \left( \frac{14}{16} \cdot \log_2 \frac{14}{16} \right) \right] \approx 0.7
\]

\[
\max E_A = - \left[ \left( \frac{1}{16} \cdot \log_2 \frac{1}{16} \right) \cdot 16 \right] = 4.0
\]

\[
E_R = \frac{E_A}{\max E_A} = \frac{0.7}{4.0} \approx 0.175
\]

13 All calculations shown in this appendix are for single sets (pairs). In Chapters 3-5, these variables are meant to denote the differences in the two sets’ (pairs’) values on these metrics. For example, when discussing the adequacy of \( R \) as a predictor variable, we are referring to the stimuli’s difference in values of \( R \), as calculated here (i.e., \( R = R_A - R_L = n_A/x_A - n_L/x_L \) ).
This is how the entropy-based metrics are calculated for the Grids stimuli used in Experiments 3A and 3B. Note that this expression would increase in length for each additional category that contains any matches (e.g., for a Grid with 3 of one category, 8 of a second category, and 5 unique elements). When considering a pair of stimuli that share $x$ of $n$ features (elements), these equations can be further simplified; these alternate forms are used for the Bugs and Tiles stimuli used in Experiments 1 and 2, respectively.

$2n = \text{total number of elements}$

$x = \text{number of categories that contain 2 elements}$

$2n - 2x = \text{number of categories that contain 1 element}$

$$E_A(x, n) = -\left[\left(\frac{2}{2n} \cdot \log_2 \frac{2}{2n}\right) \cdot x + \left(\frac{1}{2n} \cdot \log_2 \frac{1}{2n}\right) \cdot (2n - 2x)\right]$$

$$= -\left[\left(\frac{x}{n} \cdot \log_2 \frac{1}{n}\right) + \left(\frac{n - x}{n} \cdot \log_2 \frac{1}{2n}\right)\right]$$

$$= \log_2 2n - \frac{x}{n}$$

$$\max E_A(x, n) = -\left[\left(\frac{1}{2n} \cdot \log_2 \frac{1}{2n}\right) \cdot (2n)\right]$$

$$= -\log_2 \frac{1}{2n} = \log_2 2n$$
For example, consider a pair of stimuli that share 2 of their 4 features:

\[
E_R(x, n) = \frac{E_A(x, n)}{\max E_A(x, n)} = \frac{\log_2 2n - \frac{x}{n}}{\log_2 2n} = 1 - \left(\frac{x}{n} \cdot \log_2 2n\right)
\]

For example, consider a pair of stimuli that share 2 of their 4 features:

\[
E_A(2,4) = - \left[\left(\frac{2}{4} \cdot \log_2 \frac{1}{4}\right) + \left(\frac{4 - 2}{4} \cdot \log_2 \frac{1}{8}\right)\right] = 2.5
\]

\[
\max E_A(x, n) = - \log_2 \frac{1}{2 \cdot 4} = 3.0
\]

\[
E_R(2,4) = \frac{2.5}{3.0} = 0.833
\]

There is an additional entropy-based metric that can be calculated for pairs of fully alignable stimuli which we have designated as pairwise entropy. Pairwise entropy is calculated as the average of the absolute entropy for each pair of features.

\[
E_{PW}(n, x) = \frac{\sum E_A}{x} = \frac{E_A(1,1) \cdot x + E_A(0,1) \cdot (x - n)}{x}
\]
For a pair of features (a set containing two elements), they either match (belong to the same category) or do not match (belong to different categories).

\[ E_A(1,1) = - \left[ \left( \frac{1}{1} \cdot \log_2 \frac{1}{1} \right) + \left( \frac{1 - 1}{1} \cdot \log_2 \frac{1}{2} \cdot 1 \right) \right] = \log_2 1 = 0 \]

\[ E_A(0,1) = - \left[ \left( \frac{0}{1} \cdot \log_2 \frac{1}{1} \right) + \left( \frac{1 - 0}{1} \cdot \log_2 \frac{1}{2} \cdot 1 \right) \right] = \log_2 \frac{1}{2} = 1 \]

Thus, the equation for pairwise entropy can be further simplified:

\[ E_{PW}(n, x) = \frac{0 \cdot x + 1 \cdot (x - n)}{x} \]

\[ = \frac{x - n}{x} \]

The output of this expression is equivalent to the proportion of non-matching pairs of features within the stimulus pair. Thus, pairwise entropy is the complement of the proportion of matching pairs of features within the stimulus pair, which itself is used as a predictor designated as \textit{ratio} (R).

\[ R(n, x) = \frac{n}{x} \]
For example, consider again a pair of stimuli that share 2 of their 4 features:

\[
E_{PW}(2,4) = \frac{4 - 2}{4} = \frac{2}{4} = 0.5
\]

\[
R(2,4) = \frac{2}{4} = 0.5
\]

The final predictor used was simply the number \(N\) of shared features:

\[
N(n, x) = n
\]
APPENDIX B: IRB APPROVALS

Approval of Exempt Status for Experiments 1 and 2

May 12, 2012

TO: Samantha O'Hanlon
Principal Investigator
Psychology

FROM: Ching Yuan Hu, Ph.D.
Interim Director
Human Studies Program
Office of Research Compliance
University of Hawai'i, Mānoa

Re: CHS #20352- “Judgments of Similarity”

This letter is your record of the Human Studies Program approval of this study as exempt.

On June 12, 2012, the University of Hawai'i (UH) Human Studies Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The authority for the exemption applicable to your study is documented in the Code of Federal Regulations at 45 CRF 46 (2).

Exempt studies are subject to the ethical principles articulated in The Belmont Report, found at: http://www.hawaii.edu/irb/html/manual/appendices/A/belmont.html

Exempt studies do not require regular continuing review by the Human Studies Program. However, if you propose to modify your study, you must receive approval from the Human Studies Program prior to implementing any changes. You can submit your proposed changes via email at uhirb@hawaii.edu. (The subject line should read: Exempt Study Modification.) The Human Studies Program may review the exempt status at that time and request an application for approval as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so. Signed consent forms, as applicable to your study, should be maintained for at least the duration of your project.

This approval does not expire. However, please notify the Human Studies Program when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the Human Studies Program at 956-5007 or uhirb@hawaii.edu. We wish you success in carrying out your research project.

Office of Research Compliance
Human Studies Program
Approval of Exempt Status for Experiment 3

July 19, 2013

TO: Samantha O'Hanlon
Principal Investigator
Psychology

FROM: Denise A. Lin-DeSheleer, MPH, MA
Director

SUBJECT: CHS #21437. "Visual Categorization"

This letter is your record of the Human Studies Program approval of this study as exempt.

On July 19, 2013, the University of Hawai'i (UH) Human Studies Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The authority for the exemption applicable to your study is documented in the Code of Federal Regulations at 45CFR 46.101(b)(Exempt Category 2).

Exempt studies are subject to the ethical principles articulated in The Belmont Report, found at http://www.hawaii.edu/irb/html/manual/appendices/A/belmont.html.

Exempt studies do not require regular continuing review by the Human Studies Program. However, if you propose to modify your study, you must receive approval from the Human Studies Program prior to implementing any changes. You can submit your proposed changes via email at whsbr@hawaii.edu. (The subject line should read: Exempt Study Modification.) The Human Studies Program may review the exempt status at that time and request an application for approval as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so. Signed consent forms, as applicable to your study, should be maintained for at least the duration of your project.

This approval does not expire. However, please notify the Human Studies Program when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the Human Studies Program at 956-5007 or whsbr@hawaii.edu. We wish you success in carrying out your research project.
APPENDIX C: CONSENT FORMS

Consent Form for Experiments 1 and 2

University of Hawai‘i

Consent to Participate in Research Project:
Judgments of Similarity

My name is Samantha O’Hanlon and I am a graduate student at the University of Hawai‘i at Mānoa (UH), in the Department of Psychology. As part of my course of study, I must conduct research. The purpose of my current research is to better understand how people judge things to be similar or different.

Project Description - Activities and Time Commitment: If you participate in this research project, I or one of my assistants will seat you at a computer and you will be shown pairs of figures. You will either be asked to choose a pair based on some criteria given to you by the experimenter, or you will be asked to describe some aspect of the figures. Your responses will be recorded, but there will be no information identifying them as belonging to you. Your participation will last no more than 30 minutes.

Benefits and Risks: This study involves minimal risk to participants, i.e. you will be exposed to nothing that you would not encounter in your normal daily activities. While participating in this study provides little benefit to you beyond the learning experience afforded by participating in a psychology study, this study has the potential to better our understanding of how humans go about judging how similar or how different things are from each other, the threshold for detecting these similarities and differences, and how these processes inform decision-making.

Confidentiality and Privacy: All data and personal information will be stored in a secure location and will be kept confidential to the extent allowed by law. Several public agencies with responsibility for research oversight, including the UH Human Studies Program, have the authority to review research records. No information linking you to your responses will be recorded. If you would like a summary of the results of this study once completed, please email me at sohanlon@hawaii.edu.

Voluntary Participation: Participation in this research project is entirely voluntary; whether or not you will participate is your decision alone. In addition, you may withdraw your consent or discontinue participation at any point without penalty.

Questions: If you have any questions about this project, please ask now or contact me at email at sohanlon@hawaii.edu. If you have any questions about your rights as a research participant, contact the University of Hawai‘i Human Studies Program by phone at (808) 956-5007 or by email at uhirb@hawaii.edu.

78
Consent Form for Experiment 3

University of Hawai‘i
Consent to Participate in Research Project: Visual Categorization

My name is Samantha (Sam) O’Hanlon and I am a doctoral student at the University of Hawai‘i at Mānoa (UH), in the Department of Psychology. As part of my course of study, I must conduct research. The purpose of my current research is to better understand how people form categories.

Project Description

Activities and Time Commitment: If you participate in this research project, you will be seated at a computer and will be shown figures and asked to categorize them. Your responses will be recorded, but there will be no information identifying them as belonging to you. Your participation will last no more than 45 minutes.

Benefits and Risks: This study involves minimal risk to participants, i.e., you will be exposed to nothing that you would not encounter in your normal daily activities. While participating in this study provides little benefit to you beyond the learning experience afforded by taking part in a psychology study, this study has the potential to better our understanding of how humans go about forming categories and factors that influence this process.

Confidentiality and Privacy: All data and personal information will be stored in a secure location and will be kept confidential to the extent allowed by law. Several public agencies with responsibility for research oversight, including the UH Human Subjects Program, have the authority to review research records. No information linking you to your responses will be recorded. If you would like a summary of the results of this study once completed, please email me at sohanlon@hawaii.edu.

Voluntary Participation: Participation in this research project is entirely voluntary; whether or not you will participate is your decision alone. In addition, you may withdraw your consent or discontinue participation at any point without penalty. If, at any point, the researcher suspects that you are failing to follow instructions (e.g., pressing response keys without doing the task), the experiment will be terminated and you will be dismissed.

Compensation: Your participation in this study will be reported to the instructor of your choosing and research participation credit will be applied in accordance with his/her course policies. Your instructor will only receive information about your participation; he/she will not have access to any information collected during the course of the experiment (e.g., your responses, experimental results).

If you have any questions about this project or would like a copy of this consent form, please ask now or contact me at via email at sohanlon@hawaii.edu. If you have any questions about your rights as a research participant, contact the University of Hawai‘i Human Subjects Program by phone at (808) 956-5007 or by email at uhhrb@hawaii.edu.
APPENDIX D: INSTRUCTIONS

Experimenter Script for Experiments 1 and 2

You are going to be shown pairs of _______________________________ and your job is to decide which of the two pair is most ____________________. If it is the pair on the left [gesture to the left side of the screen], you are going to hit “A” [point at the “A” key]; if it is the pair on the right [gesture to the right side of the screen], you are going to hit “L” [point at the “L” key]. Do you have any questions?

Instructions for Experiment 3A

This experiment is broken into three parts.

Part 1
- You will be presented with images and your job is to decide whether they belong in Category A (press the A key) or Category L (press the L key).
- We will begin with a Training Phase during which you will be provided with feedback. ✓ means you were correct; ✗ means you were incorrect.
- Once you have learned how to categorize the images, we will move on to the Testing Phase during which you will receive minimal feedback.
- Once you have completed Part 1, you will receive further instructions.

If you have any questions, please ask the experimenter now.
Once you are ready, press the SPACEBAR to begin.
Instructions for Experiment 3B

The instructions for Part 3 are the same as they were for Part 1. Note: The images and categories have changed.

Part 3
- You will be presented with images and your job is to decide whether they belong in Category A (press the A key) or Category L (press the L key).
- We will begin with a Training Phase during which you will be provided with feedback. ✓ means you were correct; ✗ means you were incorrect.
- Once you have learned how to categorize the images, we will move on to the Testing Phase during which you will receive minimal feedback.
- Once you have completed Part 3, the program will exit automatically and you will be given a brief questionnaire to fill out.

If you have any questions, please ask the experimenter now.
Once you are ready, press the SPACEBAR to begin.
REFERENCES


