

APPROACHES TO CATEGORIZATION

The topic of categorization has prompted much work among psychologists, philosophers, linguists, and anthropologists over the years. Historically, categories were thought to be distinct units in which membership was “all-or-none.” Possession of a certain set of features would ensure each item a complete and equivalent degree of membership (Rosch & Mervis, 1975). This classical theory of concepts implies that most concepts are made up of complex representations that can be broken down into smaller representations. These representations can be thought of as definitions or requirements. Novel exemplars are put into certain categories when they meet a certain concept’s specified criteria (Laurence & Margolis, 1999).

Although this theory may seem somewhat obvious on the surface, when considering categorization further, it seems that things are not so straightforward. There have been several criticisms of this theory. To begin with, the term concept has not been clearly defined. There is yet to be a definition that has consensus from all who study this area. There also appears to be some debate about the difference between classes, categories and concepts (e.g., Zentall, Galizio & Critchfield, 2002).

A behavioral approach to concepts was developed by Keller and Schoenfeld (1950) who took the approach that a class is formed when a group of stimuli control the same response. The terms class, category, and concept can be used synonymously under this definition, and for the purpose of this paper, these terms will be used interchangeably. However, according to some, there are different definitions for these terms. In cognitive psychology, a category is thought to be a group of stimuli that

“belong together” and have been grouped accordingly, while a concept is considered to be the knowledge or mental representation that facilitates this categorization process (Zentall, Galizio, & Critchfield, 2002).

In addition to the difficulties in defining the subject of categories and concepts, many concepts themselves do not have a clear definition. This is known as “Plato’s Problem,” and can be seen when considering many basic categories like chairs, games, or birds. For example, what defines a chair? A chair could be defined as something that has four legs and a back. This seems like good definition, but is not always appropriate. A bean bag chair does not have four legs. Like chairs, many basic concepts are too ambiguous or complex to be clearly defined.

Another problem arises when we are mistaken about a concept; it is possible to know the definitions, but still be inaccurate about the concept. For example, consider the properties of a disease like smallpox. People used to believe diseases like these were caused by evil spirits. Today we know that these people were mistaken about the causes of smallpox. Their concept of smallpox had fundamental beliefs that were inaccurate. The concept has changed over time and the classical theory cannot account for this change. Finally, the classical theory cannot account for items that do not clearly belong to one particular concept or why some items are better representations of a concept than others (Laurence & Margolis, 1999).

Most modern approaches to categorization agree that “all-or-none” membership is not likely in most categories. Most categories do not appear to have clear, definite properties. For example, some theorists may define the category of furniture structurally as common items found in a house. In addition, furniture can also be defined

functionally as something we sit on, lie on, (couch, bed) or something we sit objects on (table, desk). It is easy to see that some examples of furniture fit the above definitions 'better' than others. Most would agree that a bed or a table is furniture, but what about a lamp? Is a lamp considered furniture? It does not meet the functional definition but does meet the structural definition. Thus, a couch seems to be, in some respects, a "better" example of furniture than a lamp. In most categories, some members possess more of the typical and relevant features and are therefore considered to be more representative of the category.

These are the sorts of observations that led to the family-resemblance theory of categorization. Rosch (1975) emphasizes the mental representation of the structure of features of categories. Some categories, like families, have features that are common, but do not occur in every example. Some members of a category are more representative than others because they possess more of the common features. There is no set definition of which features category members should have, but there is an overall understanding of the common features (Rosch, 1975).

Rosch and Mervis (1975) tested the family-resemblance theory in semantic or natural language categories and in artificial or laboratory-created categories. They hypothesized that the degree to which a particular item possessed a family resemblance to other members of the same category would be positively correlated with ratings of what is most typical of a category. These ratings are known as prototypicality ratings. In addition, they hypothesized that members with more of the common features would be responded to differently (e.g., faster responding and learning).

To look at semantic categories, 400 participants were asked to list attributes for nouns in six distinct categories (furniture, vehicle, fruit, weapon, vegetable, and clothing). 20 nouns were used for each of the six categories. Each subject was asked to make an attribute list for six items, one from each category. Each category had predetermined features or attributes of typicality. For each of the nouns, the number of attributes was totaled, and a typicality rating obtained. In support of their hypothesis, findings showed that subject's ratings of representativeness of a category member depended on how many attributes it shared with other members. For example, in the category of vehicle, car and truck were seen as better examples than skates or an elevator. In the category of fruit, orange and apple were seen as better examples than a tomato or an olive.

When using artificial categories, Rosch and Mervis found similar results. Arbitrary letter and number sequences were used to increase experimental control. The letter strings with different degrees of feature overlap were compared with control conditions with no feature overlap. There were six letter strings in each category, with five or six letters in each string. Each letter string had a family resemblance score that was derived from how many letters the string shared with the other letter strings. For each category of six letter strings, the strings were paired, by twos, in three groups. The groups, high, medium, and low were determined by the rating. Two experimental stimulus designs possessed varying degrees of family resemblance. The symmetric categories contained two central strings of letters that shared the most letters in common with other letter strings in the category. The asymmetrical categories had one letter that was contained in five strings, another letter that was contained in four strings, another in three strings, another in two, and another in one. The degree of typicality was measured by the number of strings the letter was in. Subjects learned to identify two categories of letter strings in

one of the three conditions. After the letter strings were learned, the subjects were given six cards for each category and asked to rank according to typicality. Acquisition and responding was more rapid and typicality ratings were higher for the stimuli that were more representative of the category. These studies led Rosch to a theory of categorization in which stimuli are categorized by comparisons with a prototype.

Another theory proposed to explain categorization is the psychological essentialist theory. Psychological essentialism suggests that people behave as if they have pre-wired concepts, and are biased to form categories based on what is essential. A mental picture of the “essence” of a category exists in the mind (Medin & Ortony, 1989). Humans are biased to form categories on the basis of particular properties. A study that illustrates this theoretical perspective showed that when a child is shown a picture of a raccoon and then asked to imagine that it has a painted white stripe on it, implying that it is now a skunk; the child still insists that it is a raccoon. The child recognizes that the paint is not an essential feature and that it is not included in the essence of a skunk (Keil, 1989).

The exemplar theory of concepts rejects the idea that a single representation in the mind defines a concept. Instead, exemplar theory suggests that a person’s concept of something includes different examples of that thing that they remember. Since there is no “summary representation” that stands for a concept, there must be many examples included. For example, a person’s concept of dogs would include a set of dogs that the

person remembers (Murphy, 2002).

Cognitive theories, including the ones mentioned above, have been the dominant approach to categorization research. While much can be learned from these cognitive theories, which focus on the structural mental representations of categories, it is important to understand that other theories exist. For example, behavior analytic approaches to categorization emphasize the functional aspects of categories, but it seems they are not well represented in the literature (Laurence & Margolis, 1999). The present study is an exploration of the value of such a functional approach derived from the work of Murray Sidman. Sidman (1994) has greatly contributed to this area with his work in stimulus equivalence. Before discussing some of his findings, a review of stimulus equivalence is needed.

UNDERSTANDING STIMULUS EQUIVALENCE

Behavior-analytic approaches attempt to understand complex behavior by breaking it down into simpler units. For example, Sidman's (2000) theory explains symbolic behavior in terms of reinforcement contingencies. To begin with, a two-term contingency, including a response and reinforcer, would be the most basic unit. A three-term contingency occurs when a discriminative stimulus is added to the response and reinforcer. Now this two-term contingency comes under the control of a discriminative stimulus. That is, a discriminative stimulus sets the occasion for a response to occur, which then results in a reinforcer. For example, a child learns that when an adult is around (discriminative stimulus) throwing a tantrum (response) will lead to candy (positive reinforcer), so through differential reinforcement, the child throws tantrums

when adults are around to get his way. The three-term contingency can be thought of as a simple unit of operant behavior.

A four-term contingency can be formed by adding another term. This new term is a condition that determines whether or not a three-term contingency is to be activated. In this situation, the response may produce reinforcement in the presence of either of two discriminative stimuli. The stimulus choice that is reinforced will depend on the conditional stimulus that is present. For example, a child learns that in the presence of his mother, screaming and crying will produce candy, but in the presence of his father, kicking the floor will produce candy. The child has learned that in order to be reinforced, the type of tantrum he throws should depend on which adult he is around. This type of four-term contingency is also known as a conditional discrimination.

A procedure known as match-to-sample (MTS) employs these four-term contingencies. In MTS training, the subject is given a sample stimulus, and then comparison stimuli, the subject responds by matching the sample to one of the comparisons, and then receives a consequence. There are several types of matching that can occur. Identity matching involves matching the sample to a physically identical comparison. Oddity matching occurs when the sample is matched to a comparison that is different. As well as many other researchers, Sidman uses what is known as arbitrary or symbolic matching. This type of matching occurs when the sample is matched to a comparison that does not share a physical relation. This arbitrary MTS procedure is often used in establishing equivalence classes. It is this procedure that has opened up new directions of study in behavior analysis.

Defining Stimulus Equivalence

The defining properties of stimulus equivalence are best explained through mathematical terms. Reflexivity, symmetry, and transitivity must emerge after conditional discriminations have been trained. Reflexivity is matching a stimulus to itself. This can be thought of as a form of identity matching. Symmetric relations demonstrate the ability to reverse the learned conditional discrimination. Transitivity involves matching stimuli that were paired with a common stimulus, but were never paired together in training. For example, if two conditional discriminations were trained (match A to B and match B to C) through differential reinforcement, then reflexivity (A to A and B to B), symmetry (B to A and C to A), and transitivity (B to C and C to B) should emerge without being directly trained and without being reinforced. When a subject demonstrates all of these relations, the stimuli are said to be equivalent and can be used interchangeably. This interchangeability defines an equivalence class.

These basic properties of stimulus equivalence can be seen in Sidman and Tailby's (1982) seminal study of stimulus equivalence. They demonstrated equivalence-class formation from conditional discrimination training. The eight children in their study were trained with three conditional discriminations through symbolic MTS. The children were trained, using the Greek letters lambda, xi, and gamma, to match dictated letters to three different sets of printed letters. An alpha-numeric notation is often used in equivalence research to help understand and simplify the description of training and class organization. Typically the letters represent different stimuli that appear together as comparisons or as samples that appear separately. The numbers represent the particular class. Each of the three classes consisted of the dictated letter (A), the uppercase letter

(B), the lowercase letter (C) and another set of printed Greek letters (D). To train the A to B relation, an auditory sample (A1, A2, or A3) was played and the subjects had a choice between each of the uppercase letters (B1, B2, B3). To train the A to C relations, auditory samples were matched to lowercase letters (C1, C2, C3). The printed letters in set D were trained to the printed letters in set C. The subject's responses were only reinforced when correct. For example, choosing B1 or C1 was reinforced for A1, choosing B2 or C2 was reinforced for A2, and choosing B3 or C3 was reinforced for A3. In addition, choosing C1 was reinforced for D1. After acquisition of the three conditional discriminations, reinforcement was gradually reduced to help prepare the subjects for the upcoming probe trials in which responses were not reinforced.

Probe trials were used to test for the formation of equivalence classes, and consisted of trials that tested for reflexivity, symmetry, and transitivity. The probe trials were mixed in with baseline trials to maintain a certain degree of familiarity and reinforcement. Baseline scores were maintained throughout testing by all of the subjects. Reflexivity trials presented A, B, or C stimuli as the sample and the identical counterpart were included in the comparisons. In symmetry trials, C stimuli were samples and D stimuli were the comparisons. Transitivity trials presented either one of the B stimuli as the sample and C stimuli as the comparisons, or one of the C stimuli as the sample and B stimuli as the comparisons. Probe tests showed that most subjects had formed equivalence classes based on the conditional discrimination training.

Why Does Equivalence Occur?

There are several ways to approach equivalence research. There is some debate about whether equivalence classes require special processes or are derived from more

basic ones. Sidman (2000) argues that when a direct reinforcement contingency is put in place, equivalence classes that include all the elements involved in the contingency will automatically follow. That is, equivalence is viewed as a fundamental process. According to Sidman, two-term contingencies, three-term contingencies and four-term contingencies should all generate equivalence classes. Also, the reinforcer and the response should both become members of the equivalence class. Two other leading theories take the stance that equivalence is not fundamental or automatic, but rather, requires learning.

Relational Frame Theory considers equivalence relations to be learned. Relational frame theorists believe that equivalence relations are not an automatic reaction to reinforcement contingencies, but that people have the capacity to learn many relations, and equivalence is one of these relations. In a related theory, Horne and Lowe (1996) take the position that equivalence classes are learned through naming. This naming can be overt or covert. They claim that by the use of naming, a child's behavior that has been established with a single stimulus comes to generalize to other objects that are physically unrelated. Eventually, the child responds to these different objects similarly. This emergent behavior, though not directly reinforced or trained, is a consequence of different stimuli being part of the same name relation. Furthermore, Horne and Lowe believe that the successes humans have with match-to-sample procedures are directly contingent on naming the stimuli.

A study by Lowe and Beasty (1987) shows how naming can facilitate subjects' ability to pass equivalence tests. The verbalizations of the children were recorded while performing visual-visual MTS task. Children were first taught to match a vertical line to

green (A1 B1) and a horizontal line to red (A2 B2). In the second phase children learned to match a vertical line to a triangle (A1 C1) and a horizontal line to a cross (A2 C2). Of the 29 subjects, 17 passed the tests for equivalence. Recordings of the subjects' verbalizations showed that all stimuli were named. All of the children who passed equivalence tests had named the correct pairs during training. For example, for the relation of A1B1, some of the children responded with "up green" and "up triangle" for the A1C1 relation. Additionally, during training some of the children named all three members of the class even though only two of the members were presented at a time. For example, when presented an A1B1 trial the subject might say "up green, up triangle." Eventually, when the subject is exposed to any of the members of this class, they will respond with "up green triangle." Through this naming, subjects associate up, green, and triangle, and have formed an equivalence class.

Review of Equivalence Research

Practical Applications

An example that illustrates the practical applications of stimulus equivalence is Sidman and Cresson's (1973) study. They used conditional discriminations to teach reading comprehension to two severely retarded Down's syndrome boys. Using twenty common words (e.g., bee, box, car, hat), the subjects were trained with the dictated words (A), the pictures (B), and the printed words (C). The subjects were first trained with identity matching and were taught to match printed words to each other (C to C). Additionally they were trained to match dictated words to pictures (A to B). They were next taught to name the pictures of the samples aloud (B to A) and then to select the printed word that matched the dictated word (A to C). After teaching auditory receptive reading by matching the picture to the printed word (B to C) the subjects demonstrated simple reading comprehension by matching the printed word to the picture

(C to B) and the printed word to the dictated word (C to A). Although time consuming, the results of this study provided evidence that persons with severe retardation can be taught some basic reading comprehension if the right method is employed.

Training Structure

Another area that has attracted interest is the training structure. Training structure refers to the sequence in which conditional discriminations are presented and the arrangement of common stimuli (Saunders & Green, 1999). There are several ways in which the structure of MTS training can differ. One common training structure is known as one-to-many or sample-as-node (SaN) training. In this structure, the sample stimulus serves as the link, or node, between all other stimuli. Training A to B and then A to C would be considered one-to-many training because the A (the sample in both conditional discriminations) relates to both B and C. The reverse of this training structure is the many-to-one or comparison-as-node (CaN) training structure. An example of this would be training B to A and then C to A. In this case, A is still the node; however, this time it is the comparison and not the sample. Last is a procedure known as a linear structure. In this structure, each stimulus, with the exception of the first and last, is a node. An example of this structure is training A to B and B to C. The node serves as the comparison in one conditional discrimination and as the sample in the next. The equivalence literature treats each of these training structures as “equally likely to produce emergent performances when the number of nodes equals one” (Saunders, Saunders, Williams, & Spradlin, 1993). However some studies have shown that with different populations and under different conditions some structures prove to be more effective than others.

Saunders, Drake, and Spradlin (1999) looked at training structure differences with normally developing preschool children. They compared many-to-one (CaN) and one-to-many

(SaN) procedures in five-member equivalence classes. The number of sessions required to learn the conditional discriminations in the one-to-many (SaN) group ranged from 11 to 74 sessions, with an average of 24. The sessions required for the children trained with the many-to-one (CaN) ranged from 22 to 48 with an average of 37. Only 2 of the 6 children trained with one-to-many (SaN) structure showed equivalence classes, while all of the children trained with the many-to-one (CaN) structure showed equivalence.

A study by Fields, Hobbie-Reeve, Adams, and Reeve (1999) compared one-to-many (SaN) and many-to-one (CaN) training structure with normal adults across 5 and 7 member classes. Results showed that those trained with many-to-one (CaN) structure required more trials before showing acquisition. All subjects eventually showed formation of equivalence classes. However, with the seven-member classes, the relations were much slower to emerge following one-to-many (SaN) training than with many-to-one (CaN) training. This effect has also been found in 8 to 14 year olds (Saunders et al., 1993).

The many-to-one (CaN) structure has also proven more effective with subjects with mild mental retardation. Spradlin and Saunders (1986) found that mentally retarded subjects trained with many-to-one (CaN) performed at or near 100% correct on equivalence probes, while those trained with one-to-many (SaN) performed at best 70%. Saunders et al. (1993) looked at the effects of instructions on training structure with mildly retarded adolescents and adults. They found that instructions facilitate performance with those trained with many-to-one (CaN) structure but not in one-to-many (SaN).

EQUIVALENCE CLASSES AND CATEGORIZATION

An important question is whether or not equivalence classes represent a good model for studying natural language classes. Some critics of this approach have argued that equivalence

research is not relevant to language categories because the classes formed are based on arbitrary relations, unlike language. The relation between the word and the “referent” is arbitrary. For example, the word “tree” is not the only word that represents this object. The words “arbol” and “baum” also represent this same object even though there is no physical or perceptual relation between these words. This single, symbolic relation is very much like an equivalence relation. However, the objects included in the “tree class,” like maples, oaks, and palms, are not arbitrary. They share some perceptual features, which allow classification of novel stimuli. Training equivalence classes in which stimuli share perceptual features would result in equivalence classes that are better models of natural language classes.

An aspect of language that makes categories so powerful is that novel things are responded to in adaptive ways. For example, seeing a picture of a shark is enough to help you identify a real live shark in a novel situation. In addition, learning about the potential danger of a shark will likely lead to your avoidance of a shark in novel situations. Even though you have never seen a real shark, the picture and information you previously learned will transfer to other situations. For example, if someone screams shark at the beach, this word alone is enough to warrant action. Language helps us to adapt and act appropriately in new situations. Most arbitrary relations that are learned in equivalence classes are unlike language because nothing beyond the single relation is learned. That is, the arbitrary relations do not generalize to novel examples. Equivalence classes are thought by some to lack this strong and powerful property and therefore have not always been considered a good model of language classes (Harnad, 1996).

In consideration of some of these arguments, Fields, Reeve, Adams and Verhave (1991) looked at effects of generalization of line lengths on equivalence classes. They trained nonsense syllables to a long line and a short line. After equivalence classes were established,

generalization was tested by using six variations of the short line and six variations of the long line. Generalization to the novel line lengths was shown by all subjects. For example, once subjects were trained to match a 3cm short line to a nonsense syllable, they also matched other lengths of short lines (2cm and 4cm) to the same nonsense syllable. This generalization demonstrated the possible open-endedness of equivalence classes. However, this study still lacks the evidence needed to show equivalence classes as good models of language. Although this study was successful in showing some degree of generalization, Harnad may argue that a category of simple lines cannot be compared to a complex lexical categories. There is quite a difference between being able to respond identically to different line lengths and being able to respond identically to more complex stimuli. For example, the lines only differed on one dimension, the length, while stimuli in many natural categories can differ on many more than one dimension (e.g. length of hair, color of hair, height, weight). Consider a comparison between a German Shepard and a Chihuahua. While both may be responded to as members of the dog category, it is not clear that simple stimulus generalization can account for this.

Galizio, Stewart, and Pilgrim (in review) sought to train subjects with a stimulus set that would result in the formation of equivalence classes that shared more properties with natural language classes. Particularly, they wanted to address the problem of generalized equivalence classes and how abstraction can be used to categorize. They also wanted to determine whether typicality effects could be observed in equivalence-class members. They used stimuli that varied on more than one dimension to make them more complex than simple lines.

Adult subjects learned a total of eight conditional discriminations during baseline training. A nonsense syllable (A-jom, wug, or niz) served as the sample and was trained with abstract comparison stimuli with different combinations of features (B-I). The stimuli in Figure

1 show that the comparison stimuli had a number of features, but only four of the features were relevant. The B set of stimuli had all four relevant features. These included the base, appendages, insert, and fill. Each class (wug, jom, or niz) included a particular variant of these four class-consistent features. For example, the Jom class had a diamond base, squiggly appendages, straight lines insert, and dot fill. The C and D stimuli had three of the relevant features, which included the base, insert and appendages or the fill, appendages, and the insert. The E stimuli had two of the relevant features, which included the appendages and the fill. The F, G, H, and I stimuli had only one relevant feature, either the base, appendages, insert, or fill. Each stimulus also included a number of irrelevant features. The irrelevant features were counterbalanced

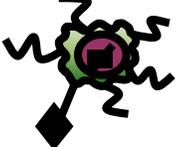
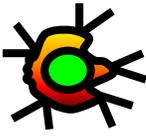
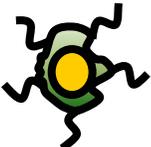
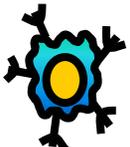
	Class 1	Class 2	Class 3
A	WUG	JOM	NIZ
B (4F)			
C (3F-BIA)			
D (3F-IFA)			
E (2F-BF)			
F (1F-B)			
G (1F-I)			
H (1F-F)			
I (1F-A)			

Figure 1. Stimulus set used during baseline training, symmetry, and transitivity probes.

across classes so that relevant features were shared with class members and irrelevant features were found on members from all three classes. Subject's responses were reinforced only when they picked the comparison with the class-consistent features. For example, when A1 (wug) was the sample, a response choosing B1, C1, D1, E1, F1, G1, H1 or I1 was reinforced. Through baseline training, subjects learned that these contingencies defined each of the classes. Subjects continued baseline training until meeting criteria (22/24) on two consecutive blocks of trials. Reinforcement was gradually reduced, from 100% to eventually 50%, to prepare subjects for upcoming unreinforced probe trials.

After baseline training, subjects were exposed to probe trials, including tests for symmetry and transitivity. Symmetry trials would present the abstract stimuli (B-I) as samples and the trigrams as comparisons (A). Transitivity trials would present the abstract stimuli (B-I) as samples as well as comparisons. Probe trials also included novel probes which presented new combinations of features not before seen by subjects. Trials presented the tri-gram as the sample and a stimulus with a new combination of features as the comparison. The only way to identify this novel stimulus was to pay attention to the relevant features learned during baseline training. Lastly, subjects were given cards with pictures of all the stimuli and were asked to sort them into three groups and then to rate them from the one they thought most represented the category to the one they thought was least representative.

Analysis of probe trials provided evidence that all subjects had formed generalized equivalence classes. Symmetry probes were responded to with 94 to 100% accuracy. Transitivity probes ranged from 90-100% accurate and novel probes had 98-

100% class-consistent responding. While the symmetry and transitivity probes were important in showing equivalence classes had formed, the novel probes were of particular interest. The subjects' class-consistent responding to novel probes suggested that the behavior had come under the control of the relevant features, permitting class-appropriate responding to stimuli never presented in training.

Typicality effects were found by analyzing errors and latencies. Galizio et al. found that subjects made more errors with one-and two-feature stimuli than with three-and four-feature stimuli in baseline training. Also, latencies were longer for one-and two-feature stimuli. That is, three-and four-feature stimuli were responded to more quickly. In sorting tests administered after testing, subjects rated four-feature stimuli as most representative and one-feature stimuli as least representative members of their classes.

These findings are thought to be strong evidence of the validity of equivalence classes as models of natural language classes for several reasons. Galizio et al. directly trained 8 conditional discriminations. Through symmetry and transitivity 54 more relations emerged and three eight-member equivalence classes were demonstrated. Novel probes indicated that another 51 stimuli were class members, implying that these classes were open-ended in nature. As long as a stimulus possessed one or more of the relevant features, it was included in the class. These findings seem to closely model the powerful generalization that occurs in natural language classes.

However, the results from the previous study were obtained using a one-to-many, or sample-as-comparison training structure. The trigrams were always used as the sample, with each of the combinations of features of the abstract stimuli as comparisons.

As mentioned before, there is some evidence that training structure affects acquisition and equivalence-class formation. Another important point to consider is the possible effect of a word-like trigram in the training structure. Horne and Lowe might argue that because equivalence comes from naming, a word, or a name, as a node might have been essential in a subject's acquisition of the complex classes. On the other hand, theories such as Sidman (2000) would argue that the same results would be found with or without the use of trigrams as nodal stimuli.

An additional issue in the Galizio et al. study is the role of the instructions. Behavior in many situations seems to be greatly influenced by instructions. Early studies of human operant behavior found unusual differences between studies with minimal and extensive instructions. Instructions have proved to be a very important part in developing and controlling human behavior (Baron & Galizio, 1983). The instructions used in the Galizio et al. study were fairly explicit. For example, the subject was instructed to "select the object that goes with the syllable in the center of the screen and click on it," and to respond as rapidly as possible. In many equivalence studies, minimal instructions are used to ensure performance is due to contingencies and not instructions (Pilgrim & Galizio, 1990; Pilgrim & Galizio, 1995). For example, the subject is instructed to simply pick one of the comparisons and is not informed to respond in a certain time. Thus, the Galizio et al. instructions provided more detail as to how to respond and the consequences of responding. There is a possibility that these detailed instructions were necessary to produce the effects found.

PURPOSE OF PRESENT STUDY

The present series of experiment were systematic replications of the Galizio et al. study with different training structures and instructions. The experiments examined the effects of training structure and instructions on typicality effects and the formation of generalized equivalence classes. Additionally, the role of tri-gram naming was examined by taking it out of the sample position. Experiment 1 was conducted to see if a many-to-one training structure, the reverse of the Galizio et al. study, would produce changes in acquisition, the formation of equivalence classes, and typicality effects. Thus, the samples in this experiment could be any of the one, two, three, or four feature abstract shapes and the comparisons were always one of the three trigrams.

EXPERIMENT 1

Participants

There were 8 participants in Experiment 1. The participants were undergraduate students from the University of North Carolina at Wilmington. Participants received credit to fulfill course requirements in psychology classes.

Apparatus

Participants were individually trained and tested on color Macintosh computers in a quiet room. Match-to-sample software, developed by Dube (1991), was used. Stimuli consisted of black nonsense trigrams approximately 0.75 X 2.0 cm in size and color abstract objects. (See Figures 1-3). The abstract stimuli were approximately 4.0 X 4.0 cm in size.

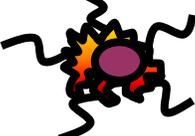
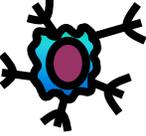
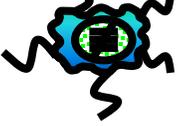
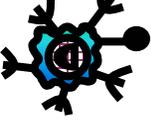
	Class 1	Class 2	Class 3
A	WUG	JOM	NIZ
J (1F-B)			
K (1F-F)			
L (1F-I)			
M (1F-A)			
N (2F-FB)			
O (2F-FA)			
P (2F-IB)			

Figure 2. Stimulus set used in unequal-feature probe trials.

	Class 1	Class 2	Class 3
A	WUG	JOM	NIZ
Q (2F-IA)			
R (3F-IAB)			
S (3F-FIA)			
T (3F-FIB)			
U (3F-FAB)			
V (4F-FIAB)			
W (1A)			
X (1I)			
Y (1F)			
Z (1B)			
AA (2FB)			

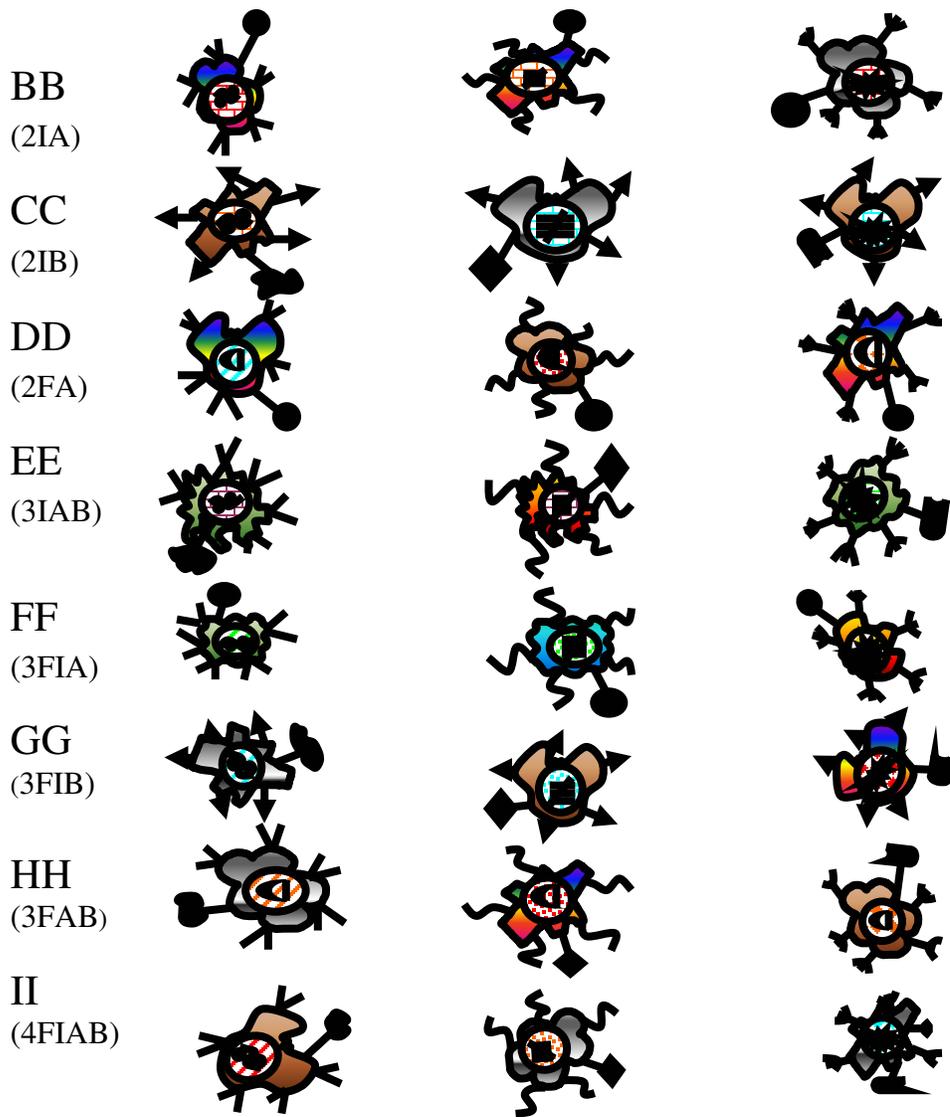


Figure 3. Stimulus set used in equal-feature probe trials.

Stimulus Set

Subjects were trained with nonsense syllables (Row A in Figure 1) and abstract stimuli (Rows B-I in Figure 1). Figure 1 shows that the abstract stimuli (B-I) differed along eight dimensions. Four of these dimensions, or features, were relevant and class-consistent. These features were the fill, insert, appendages and base. Training stimuli could have one, two, three, or four of these features. For example, the prototype stimulus (Row B in Figure 1) had all four relevant features. On the other hand, some of the training stimuli had only one of the relevant features (Rows F-I in Figure 1). For example, Row F in Figure 1 shows the one-feature stimuli with the relevant base. Note that the base is found on stimuli in Rows B, C, and E, but, like all the relevant features, the bases are class consistent. That is, class one has a different base from class two, which has a different base from class three. The irrelevant features were the shape of the figure, position of the base, the color of the inner circle, and the color of the outer shading around the inner circle. The irrelevant features were distributed quasi-randomly among classes to ensure they are not the basis for correct responding. Like the relevant features, the irrelevant features had three variants, and each of these appeared two or three times in Class 1 (Wug), Class 2 (Jom) and Class 3 (Niz). For example, as seen in Figure 1, the yellow/orange outer shading and the purple inner circle color can be seen in all three classes.

Procedure

The procedure included four phases. First, participants completed a pre-sorting task. Next, participants received baseline training and probe trials. Finally, participants completed the sorting task again and also rated the stimuli.

Pre-Sort Task

The pre-sort task was conducted to see how subjects would categorize stimuli without conditional discrimination training. This task determined what properties of the stimuli controlled categorizing before training. Subjects were presented with three sheets of paper with one of the tri-grams (Jom, Wug, Niz) printed at the top middle of each page. They were also given cards with pictures of each of the abstract stimuli to be used in training (as in Figure 1). Participants were asked to sort the cards into three groups, in any way that they deemed appropriate. Subjects' responses were recorded by the experimenter.

Baseline Training

Subjects were read the following instructions:

Please read these instructions along with me as I read them out loud.

This is an experiment in learning – it is not a psychological test. We are investigating certain aspects of the learning process that are common to all people.

More specifically, we are interested in finding out how many points you will be able to make on a learning task. In other words, your job in this experiment will be to make as many points as you can. The way in which you can make points works like this:

When this experiment begins, a stimulus figure will appear in the center of the screen. The stimulus in the center of the screen will always be the example stimulus. After you have looked at the example stimulus in the center of the screen, use the mouse to position the cursor on it and click. Other stimuli will

appear on the screen. These are your choices. You should use the mouse to select the object that goes with the stimulus in the center of the screen and click on it. Sometimes after your choice, colored stars will appear on the screen accompanied by music. Each time this occurs, 1 point is added to your total score. However, sometimes you will hear a buzzer, and this will subtract 1 point from your score. Also, please note that on some trials there will be no feedback (no stars or buzzer) to tell you if you've made the correct choice.

We are also interested in how rapidly you can make your choice. In the beginning, you will not know which choice is correct. However, once you learn which objects go together, it is essential that you make your choice as quickly as you can.

If you have any questions, please ask them at this time. I will not be able to answer any questions or make any comments once we've begun.

Again, remember that your job is to earn stars as often as you can.

Baseline training began following the subject's completion of the instructions.

Baseline training consisted of many-to-one (CaN) training. One of the abstract stimuli, either a four (Row B, Figure 1), three (Rows C and D in Figure 1), two (Row E in Figure 1), or one-feature shape (Rows F-I in Figure 1) was presented as the sample. In this experiment, the comparisons were always the three trigrams (Row A in Figure 1). An example of a trial would present a shape with a diamond base as the sample (e.g., Stimulus F2) and the trigrams Jom (A2), Wug (A1), and Niz (A3), as the comparisons.

When the trial began, subjects first had to click on the sample stimulus. So, to use the example from above, stimulus F2 would appear and subjects were required to click on

it. This served as the observing response and prompted the trigrams to appear in three of the four corners of the screen. Participants then clicked on one of the comparisons.

Comparison locations were randomized. Correct responses, which were reinforced, were followed by stars flashing on the screen and a jingle. So, for this example, selecting the trigram Jom (A2) would produce reinforcement. Selections of Wug (A1) or Niz (A3) would be incorrect responses, and the subject would hear a buzzer. After each trial, the screen went blank for a 1.5 sec inter-trial interval.

As Figure 1 shows, only eight of the possible fifteen arrangements of the four relevant features were used for training. Using these arrangements ensured that each relevant feature was presented an equal number of times. The three-feature stimuli were presented in two combinations. As Figure 1 shows, the C set included base, appendages, and insert, and the D set included appendages, insert, and fill. The two-feature stimuli, as seen in Figure 1, appeared in just one combination of the base and fill. Figure 1 also shows that each of the features served as a one-feature stimulus. The F set had the relevant base, the G set had the relevant insert, the H set had the relevant fill, and the I set had the relevant appendages.

Training was arranged in blocks of 24 trials. Figure 1 shows all the trial types used in training where each of the stimuli in Rows B-I served as the samples with the trigrams as comparisons (Row A in Figure 1). Responding to the trigram with the class number corresponding to the sample was reinforced (e.g., given B1 as a sample, choosing A1 was reinforced). Each block contained one of each of the possible trial types in random order. The comparisons were presented in a randomized order with no position containing the correct comparison for more than two consecutive trials. The sample

stimulus was never the same on two consecutive trials, and one of the four possible positions was never reinforced more than twice in a row. Subjects cycled through training blocks until the mastery criterion (22/24) was met on two consecutive blocks. Two additional baseline trial blocks were presented with reduced reinforcement to prepare subjects for upcoming unreinforced test trials. Subjects were required to meet criteria (22/24) on a block with 75% reinforcement, where 6 of the trials were randomly chosen to be non-reinforced, and a block with 50% reinforcement, where 12 trials were randomly chosen to be non-reinforced. Mastery of these two reduced reinforcement blocks allowed subjects to move to the next phase in the experiment.

Probe Trials

Following acquisition of baseline conditional discriminations, probe trials were introduced. Probe trials tested the properties of symmetry and transitivity, as well as generalized control by class-consistent features. Each trial block consisted of only one type of probe and was intermixed with baseline trials.

Symmetry trial blocks contained a total of 48 trials composed of 24 baseline trials and 24 probe trials. The 24 symmetry trial types were each presented once for a total of 24 probe trials. Responses on symmetry trials were never reinforced, thus, an overall block reinforcement rate of 50% was maintained. Symmetry trials presented a structure opposite that of the baseline training in that they presented one the tri-grams (A) as the sample, and the comparisons were either four-feature prototype stimuli (e.g., B1, B2 and B3), three-feature stimuli (from Rows C and D of Figure 1), two-feature stimuli (Row E) or one-feature stimuli (Rows F-I).

Transitivity trials involved one of the abstract shapes (from rows B-I of Figure 1) presented as the sample and three others as comparisons. During baseline training these shapes had never served as comparisons, but were potentially the nodal trigram stimuli (A1, A2 and A3). These nodal stimuli, the trigrams, did not appear on transitivity trials. An example of a transitivity trial type would be a one-feature shape with the base (Stimulus F1 in Figure 1) as the sample and the comparisons would be the three-feature shapes with insert, pattern, and appendages (Stimuli D1, D2, and D3 in Figure 1). The transitivity trials only included trials that did not allow for identity matching with critical features. So, if a sample was a two-feature shape with a relevant base and fill (Stimulus E1 in Figure 1), a comparison might be the one-feature shape with the relevant insert (Stimulus G1 in Figure 1) or the one-feature shape with relevant appendages (Stimulus I1 in Figure 1). In this case, comparisons would not include the one-feature shape with a relevant base (Stimulus F1 in Figure 1) or the one-feature shape with a relevant fill (Stimulus H1 in Figure 1) because this would allow subjects to identity match features. Transitivity trial blocks included 24 probe trials and 24 baseline trials. Again, a 50% reinforcement density was maintained due to the non-reinforced transitivity trials.

To test for generalization of class-consistent features, novel probes were used. Figure 2 shows that some of these probes included novel stimuli with new combinations of relevant and irrelevant features. Like the training stimuli, these novel examples varied in number of features and will be referred to as unequal-feature feature probes. As seen in Figure 2, these novel stimuli included new combinations of irrelevant features. For example, Stimulus O1 in Figure 2 shows a stimulus with a new color combination. This stimulus has a rainbow color in the outer area surrounding the inner circle that subjects

had never seen before. In order for subjects to match these trials in a class-consistent manner, they must ignore these new examples of the irrelevant features, and respond to the relevant features they learned in baseline training. Note that Stimulus O1 in Figure 2 has a relevant fill and relevant appendages that would allow the subject to match class consistently. Trials were similar to baseline trials except that reinforcement was not provided. In these novel trials the abstract stimuli that served as samples were replaced with novel examples. An example of this trial type would present an unequal-feature feature stimulus as the sample (Stimulus P1 in Figure 2) and the trigrams (A1, A2, and A3 in Figure 1) would be the comparisons. Trial blocks included 20 unequal-feature probe trials, 18 reinforced baseline trials and 6 non-reinforced baseline trials to maintain a reinforcement density of 50%.

To control for stimulus complexity of novel probes, a second set of novel probes was created. These stimuli, known as equal-feature probes and shown in Figure 3, assured an equal number of features for each stimulus by adding irrelevant variants of the relevant features to four, three, two, and one-feature probe stimuli. For example, Figure 3 shows a one-feature stimulus may have a class-consistent insert, but would also have a irrelevant base, appendages, and fill. An example of this shape is Stimulus X1 in Figure 3. This new set of novel stimuli varied in the number of class-consistent features, but were equal in terms of complexity, i.e., they all had eight features. As seen in Figure 3, Stimulus X1, although it has only one relevant feature, does not differ from Stimulus II1, which has all four of the relevant features, in terms of complexity. Since these features were irrelevant, they can be seen in all three classes as Figure 3 shows. An example of this trial type would include an equal-feature stimulus (Stimulus T1 in Figure 3) as a

sample and the trigrams as comparisons (A1, A2, and A3 in Figure 1). Trial blocks included 20 equal-feature probe trials, 18 reinforced baseline trials and six non-reinforced baseline trials to maintain a reinforcement density of 50%.

Subjects cycled through the following sequence of probe-trial blocks: one trial block of Novel Probes (Unequal-Feature), one trial block of Novel Probes (Equal-Feature), one trial block of Novel Probes (Unequal-Feature), one trial block of Novel Probes (Equal-Feature), 2 trial blocks of symmetry, 2 trial blocks of transitivity, one trial block of Novel Probes (Unequal-Feature), one trial block of Novel Probes (Equal-Feature), one trial block of Novel Probes (Unequal-Feature), one trial block of Novel Probes (Equal-Feature). The subject continued this sequence until it was completed or until their 50-min session was up. At the beginning of every session the subject started at the beginning of the program with the baseline trials blocks.

Post-Sort Task

Upon completion of the computer task in the final session, subjects were given a sorting task identical to the one administered prior to training. Subjects were given the pictures of the stimuli and the three sheets of paper with the trigrams at the top and were asked to sort the pictures of the stimuli in three groups. However, after sorting, they were asked to rate each of the stimuli within each group. Subjects were asked to arrange the stimuli in order starting with the most representative of the category, which was given a rating of 8, and ending with the least representative of the category, which was given a rating of 1. Subjects' responses were recorded by the experimenter.

Results

Acquisition

Number of trials required to meet criterion ranged from 192 to 912. Total number of trials and errors during acquisition are shown in Table 1. As Table 1 shows, it took subjects an average of 520 trials to meet acquisition criteria. Additionally, subjects made an average of 233 errors. The possibility of a typicality effect in acquisition was assessed by analyzing errors made on different trial types as a function of number of relevant features. The number of trials correct was summed for each of the trials containing one, two, three, or four-feature stimuli. Because subjects had no basis for matching on the first block of trials, errors made on the initial trial block were not included in this sum. The sums were then divided by the total number of trials for each feature number producing a percent correct score (correct trials/correct trials + error trials). For example, if a subject had 100 trials with one-feature stimuli and got 84 of the trials correct then they would have a score $84/100$ or 84% correct. A typicality effect would be indicated by a direct relation between percent correct and number of relevant features. Figure 4 shows something of a typicality effect with higher percent correct scores for trials containing three and four-feature shapes than for trials with one and two-feature shapes. A single-factor within-subjects ANOVA confirmed statistical significance [$F(3,21)=5.8$, $p<.01$] of this typicality effect. A Tukey's post hoc test confirmed that the One-feature and Two-Feature condition were significantly different from the Four-Feature condition but were not significantly different from each other ($p<.05$). Also, the Two-Feature condition was significantly different from the Three-Feature condition ($p<.05$). Table 1

Table 1.

Number of trials required to meet criterion, total errors and percent correct on baseline errors by number of features for subjects in Experiment 1.

<u>Subject</u>	<u>Trials</u>	<u>Errors</u>	<u>%correct</u> <u>1F</u>	<u>%correct</u> <u>2F</u>	<u>%correct</u> <u>3F</u>	<u>%correct</u> <u>4F</u>
I-1	479	177	80	82	93	95
I-2	784	366	57	28	96	100
I-3	912	345	67	67	91	87
I-4	192	73	91	95	96	97
I-5	620	366	67	64	84	85
I-6	432	225	79	85	87	87
I-7	552	216	78	80	93	95
I-8	192	97	90	92	92	88
Mean	520.34	233.13	76.13	74.13	91.5	91.75

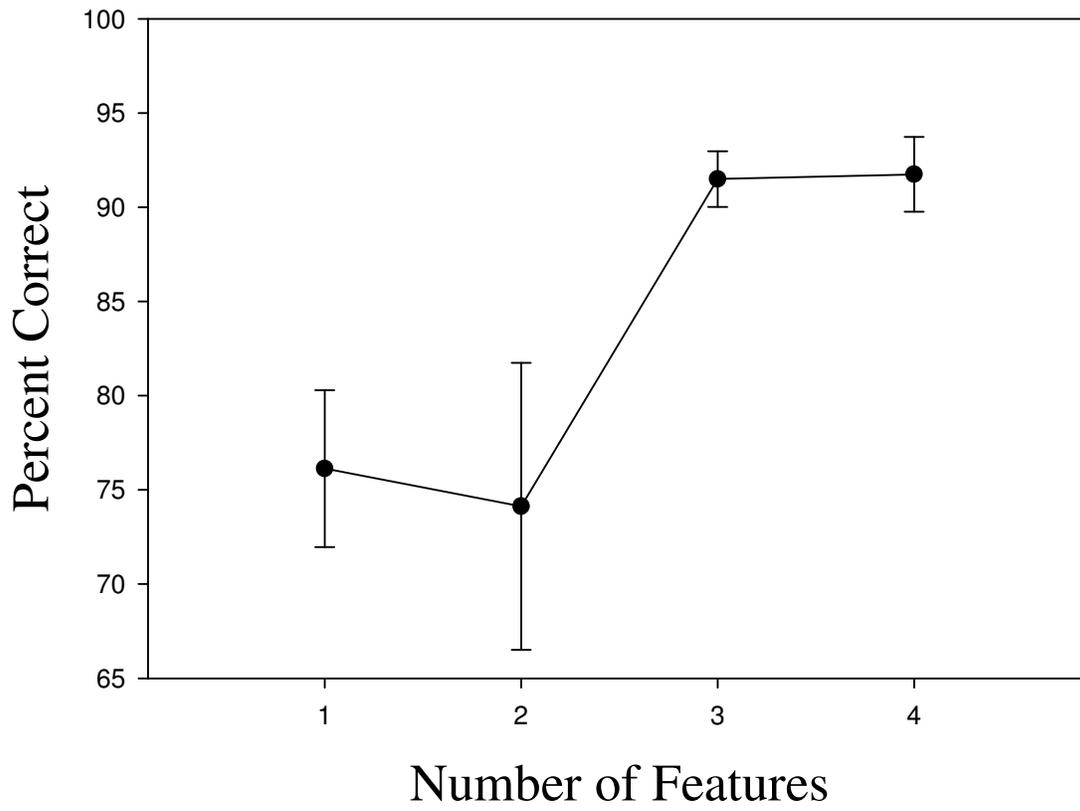


Figure 4. Percent correct on baseline errors for each number of features for subjects in Experiment 1.

shows that these effects were evident at the level of individual subjects, with seven out of eight subjects showing a higher percentage correct on four-feature trials than on one-feature trials. Only subject # I-8 failed to show more errors as the number of features increased.

Baseline Reaction Times

Reaction times, which measured the amount of time from the presentation of the comparison stimuli to the subject's response, were summed independently for trials with comparisons of each feature number. These sums were divided by the total number of trials in order to assess typicality effects on reaction time. Reaction times for initial trial blocks as well as incorrect trials were not included. To reduce the skew of the reaction time distribution, a reciprocal transformation was used to convert reaction time to speed scores.

Figure 5 presents baseline speed scores for sessions 1 and 2. Session 3 speed scores were excluded based on past findings from Galizio et al. that showed that by session 3 subjects' response speeds had reached asymptote, and that typicality effects thus dissipated. As Figure 5 shows, responding on one and four-feature stimuli was slightly slower than responding on two and three-feature stimuli, but there was considerable variability. This finding is not indicative of a typicality effect. Closer examination of individual speed scores, as shown in Table 2, reveal little variation within subjects. Subjects # I-5 and # I-7 appear to be the only subjects that may have shown typicality effects, with relatively slow responding to one-feature stimuli. However, most subjects did not respond significantly faster on trials with three and four-feature stimuli than on one and two-feature stimuli trials. A single-factor within-subjects ANOVA

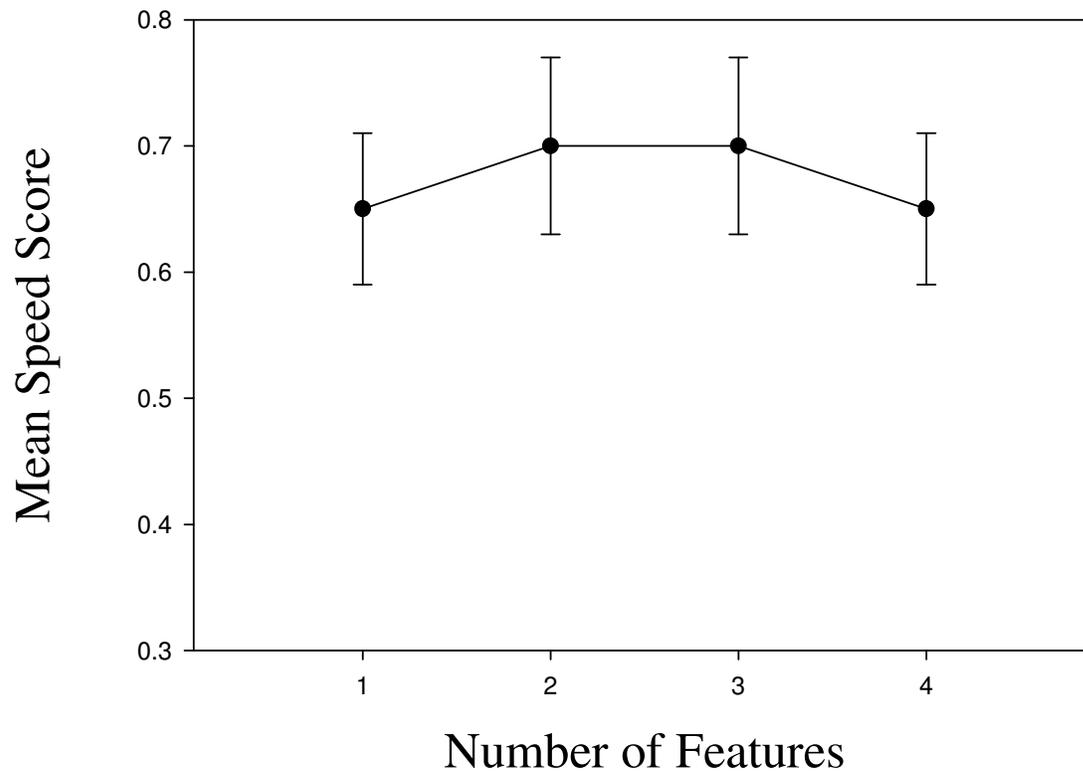


Figure 5. Mean speed scores on baseline trials for each number of features for subjects in Experiment 1.

Table 2.

Speed scores for baseline trials and unequal and equal probe trials for subjects in Experiment 1.

<u>Subject</u>	Baseline Trials				Equal Probe Trials				Unequal Probe Trials			
	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>
I-1	0.45	0.4	0.5	0.43	0.6	0.5	0.5	0.56	0.7	0.91	0.67	0.63
I-2	0.83	0.8	0.9	0.59	0.9	0.8	0.8	0.77	0.6	0.83	0.91	0.83
I-3	0.56	0.6	0.6	0.59	0.5	0.5	0.6	0.56	0.5	0.45	0.63	0.56
I-4	0.67	0.7	0.6	0.63	0.6	0.7	0.6	0.71	0.7	0.71	0.67	0.67
I-5	0.59	0.6	0.8	0.77	0.8	0.8	0.7	0.91	0.7	0.67	0.83	0.77
I-6	0.91	0.9	0.8	0.83	0.7	0.9	0.8	0.91	0.8	0.91	0.91	0.83
I-7	0.77	0.9	0.9	0.91	0.7	0.7	0.9	0.83	0.8	0.91	0.91	0.83
I-8	0.42	0.5	0.5	0.45	0.4	0.5	0.5	0.56	0.5	0.48	0.48	0.45
Mean	0.65	0.7	0.7	0.65	0.6	0.7	0.7	0.73	0.7	0.73	0.75	0.7

confirmed that there was no significant effect of Features on baseline speeds [$F(3,21)=1.92, p>.05$].

Probe Trials

Performances on symmetry probes and transitivity probes demonstrated the formation of equivalence classes for all of the subjects. As shown in Table 3, symmetry performances ranged from 94% to 100% correct with an average of 98% correct. Transitivity performances averaged 98% correct and ranged from 96% to 100% correct. All but two of the subjects who formed equivalence classes also matched class-consistently on unequal-feature and equal-feature probe trials. Table 3 shows unequal-feature probe performances ranged from 75% to 100% correct and equal-feature probe performances ranged from 76% to 100% correct, both with an average of 90% correct. Thus, six of the eight subjects included novel stimuli with one or more relevant features in the class even though they were not directly trained during baseline trials. Subjects # I-2 and # I-5 did not show this trend. Although they had high scores on symmetry and transitivity trials, their scores on the unequal-feature probe trials did not reflect the same level of mastery. Their scores show that they were responding above chance. This pattern of scoring likely suggests that these subjects were missing one type of trials.

Probe trials were also assessed for typicality effects. Individual speed scores, shown in Table 2, seem to follow a trend similar to the baseline speed scores in that most subjects do not show patterns consistent with typicality effects. Figure 6 shows that subjects responded at more or less comparable speeds regardless of feature number on both unequal-feature and equal-feature probes. A two-factor within-subjects ANOVA

Table 3.

Percent correct on probe trials for subjects in Experiment 1.

<u>Subject</u>	<u>Unequal</u>	<u>Equal</u>	<u>Symmetry</u>	<u>Transitivity</u>
I-1	92	90	100	98
I-2	75	79	96	100
I-3	97	97	100	96
I-4	100	100	100	100
I-5	77	76	94	96
I-6	93	91	99	98
I-7	91	90	96	96
I-8	97	97	100	100
Mean	90.25	90	98.13	98

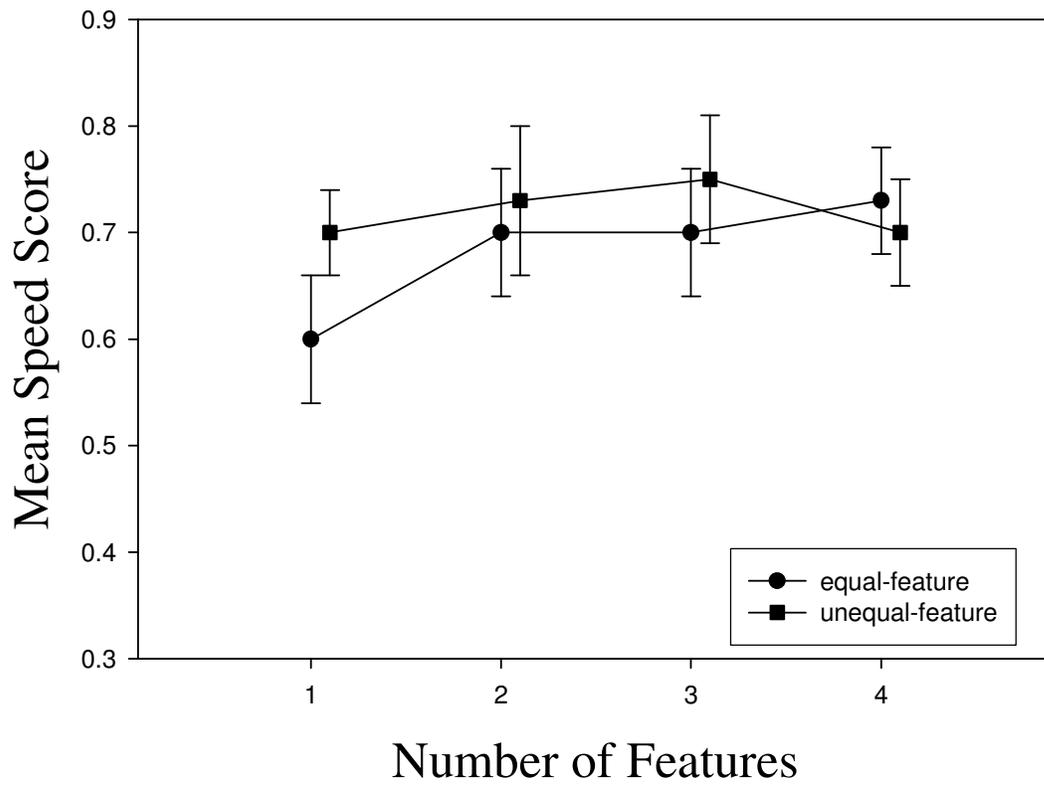


Figure 6. Mean speed scores on novel probe trials for each number of features for subjects in Experiment 1.

verified there was no typicality effects on unequal-feature probe trials [$F(3,21)=2.72$, $p>.05$].

Sort and Rating Task

As seen in Table 4, all the subjects sorted in a class-consistent way on the post-sort task, with an average of 99% class consistent. This was a dramatic improvement over 38% class-consistent during the pre-sort task.

For the rating task, rankings of the trained stimuli were weighted (i.e., 8=most typical; 1=least typical) for each subject and summed across classes at each level of feature number. The sum for each feature number was then divided by the total number of stimuli included for that level (i.e., 3 four-feature, 6 three-feature, 3 two-feature, and 12 one-feature stimuli), producing a mean typicality rating. Mean ratings are shown in Figure 7 and indicate a strong typicality effect. Subjects' ratings of the stimuli increased as the number of relevant features increased. A single-factor within-subjects ANOVA confirmed a main effect of Features [$F(3,21)=50.94$, $p<.01$]. A Tukey's post hoc test found the One-Feature and Two-Feature Groups to be significantly different from the Three-Feature and Four-Feature Groups, as well as a significant difference between the Three-Feature and Four-Feature Group ($p>.05$).

Discussion

To summarize the results of Experiment 1, the formation of equivalence classes was demonstrated through high levels of performance on symmetry and transitivity tests in all eight subjects. Performances on novel probes showed that not all subjects had formed generalized equivalence classes; however, six of the eight subjects did show categorization of novel stimuli according to the features that were relevant in baseline

Table 4.

Sort scores and ratings for subjects in Experiment 1.

<u>Subject</u>	<u>Pre-sort</u>	<u>Post-sort</u>	<u>1-Feature</u>	<u>2-Feature</u>	<u>3-Feature</u>	<u>4-Feature</u>
I-1	38	100	3	4.7	6.2	7
I-2	38	96	3.8	5.7	4.7	6
I-3	38	100	2.8	4	6.5	8
I-4	38	100	2.5	5	7	7
I-5	38	96	2.3	4	7.2	6.7
I-6	37	100	2	3.7	5.2	7.3
I-7	38	100	2.2	3	7	7
I-8	38	100	2.5	5	6.5	8
Mean	37.89	99	2.64	4.39	6.29	7.13

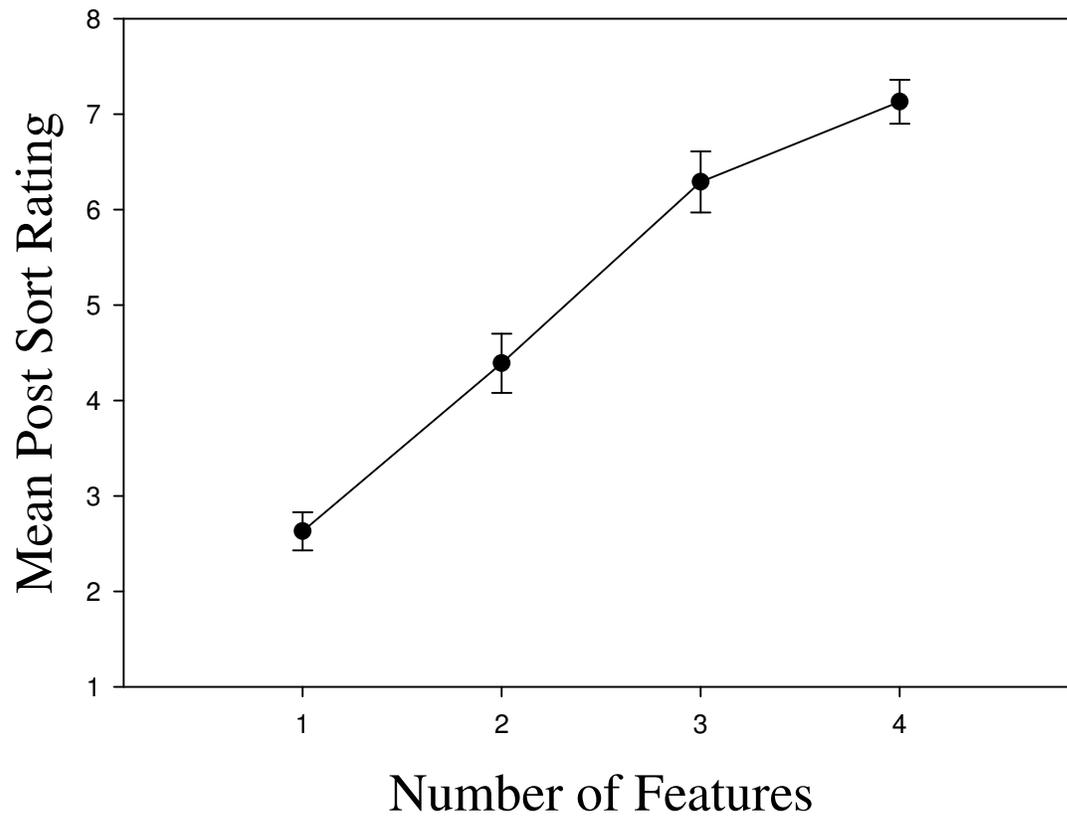


Figure 7. Mean post sort ratings for each number of features for subjects in Experiment 1.

training. Thus, a many-to-one training structure can produce generalized equivalence classes using the Galizio et al. “family resemblance” procedures. Typicality effects were not as robust as in the Galizio et al. study, which used one-to-many training. In the present study, typicality effects were found only in the analysis of baseline errors and the post-sort rating task. That is, subjects made significantly more errors on baseline trials that contained one and two-feature stimuli than on baseline trials with three and four-feature stimuli. In the rating task, subjects rated the one-feature stimuli as least typical of the category, the two-feature stimuli as the next to least typical, the three-feature stimuli as the next to most typical, and the four-feature stimuli as the most typical of the category. While these results parallel more closely the findings of Galizio et al., other findings differed in that typicality effects were not found in the latencies of baseline or probe trials.

The results of Experiment 1 showed that equivalence classes were acquired with a many-to-one training structure. Thus, it appears that the trigram does not have to be the sample in order for subjects to show equivalence. However, no typicality effects as a function of latencies were found. This finding requires further exploration of whether this was due to the many-to-one training structure or due to the absence of the trigram in the sample position. Results from the Galizio et al. study and the present study show that when the trigram is the sample or the node, generalized equivalence is acquired (although it should be noted that this was less consistent in Experiment 1). However, in both these studies the trigram served as the nodal stimulus, whether as sample (Galizio et al.) or as comparison (Experiment 1). The next step would be to remove the trigram from the nodal position completely. In Experiment 2, a one-to-many training structure was used,

but the sample was replaced with the prototype (four-feature stimulus) of each category. The comparisons included the trigrams, but also included all the other abstract shapes (the three, two, and one-feature shapes—see Figure 1).

Experiment 2 also addressed the question of the role of instructions. In the Galizio et al. study detailed instructions were used to facilitate the acquisition of the large number of baseline conditional discriminations in the short period of time the subjects were available to the experimenters. In Experiment 1, these same detailed instructions were used. In both studies, the subjects learned the conditional discriminations fairly readily. Experiment 2 used the detailed instructions, but also studied a group that received more minimal instructions to see if any group differences would be found.

EXPERIMENT 2

Participants

There were 16 participants in Experiment 2. The participants were undergraduate students from the University of North Carolina at Wilmington. Participants received credit to fulfill course requirements in psychology classes.

Procedure

Experiment 2 employed the same general procedure as Experiment 1. This experiment differed in terms of the instructions given to the subjects. A minimal set of instructions was introduced in this experiment. Procedures for baseline training and probe tests differed from Experiment 1 only in the training structure that was presented.

Instructions

Half (eight) of the subjects in Experiment 2 were given the same detailed instructions used in the Experiment 1. The remaining eight subjects were read the following minimal instructions:

Please read these instructions along with me as I read them out loud.

This is an experiment in learning – it is not a psychological test. We are investigating certain aspects of the learning process that are common to all people.

When this experiment begins, you are going to see a picture in the middle of the screen. Use the mouse to move the cursor to the picture and click on it. When you click, you're going to see three other pictures in the corners of the screen.

These are your choices. Please pick one and click on it.

Sometimes after your choice, colored stars will appear on the screen. Sometimes you will hear a buzzer. Your job is to earn stars as often as you can, but please note that on some trials there will be no feedback (no stars or buzzer) to tell you if you've made the correct choice.

If you have any questions, please ask them at this time. I will not be able to answer any questions or make any comments once we've begun.

Again, remember that your job is to earn stars as often as you can.

Pre and post-sort task

Procedures for the pre-sort task and post-sort task were the same as in Experiment

1.

Baseline training

One of the three prototypes (four-feature stimuli, see Stimuli B in Figure 1) was always presented as the sample. Otherwise, training trial types and other procedures were the same as Experiment 1 except that one of the comparison sets included the three trigrams.

Probe trials

Subjects followed the same probe sequence (novel, symmetry, and transitivity) as in Experiment 1 after mastering baseline conditional discriminations. The sample on symmetry trials was always one of the stimuli used as comparisons in training, including the trigrams. The comparisons were always the prototypes.

On transitivity trials, subjects matched the trigrams and the three, two, and one-feature shapes. There were no prototype stimuli presented in the transitivity trials. The sample and comparisons could be any of the trigrams, three, two, or one-feature stimuli. Again, trials that permitted feature-identity matching were not included.

Unequal-feature and equal-feature novel probe trials followed the same structure as baseline trials in that a prototype was always the sample. However, the comparisons were novel examples of the four, three, two, or one-feature stimuli (see Figures 2 & 3). Trigrams were not included in these trial types because there are no novel examples of them.

Results

Acquisition

Total number of trials and errors during acquisition are shown in Table 5. As table 5 shows, number of trials required to meet criteria on baselines ranged from 48 to 733. Subjects in the Minimal Instruction Group required an average of 69 trials to meet

Table 5.

Number of trials to meet criterion, total errors and percent correct on baseline errors by number of features for subjects in Experiment 2.

<u>Subject</u>	<u>Trials</u>	<u>Errors</u>	<u>%correct</u> <u>0F</u>	<u>%correct</u> <u>1F</u>	<u>%correct</u> <u>2F</u>	<u>%correct</u> <u>3F</u>
II-1-M	72	23	90	98	99	100
II-2-M	48	36	69	99	100	98
II-3-M	48	4	98	100	100	100
II-4-M	48	11	91	100	100	100
II-5-M	145	52	74	97	96	99
II-6-M	48	15	89	99	100	100
II-7-M	72	17	89	97	100	100
II-8-M	72	10	95	99	100	100
Mean	69.13	21	86.88	98.63	99.38	99.63
II-1-D	72	104	90	99	99	99
II-2-D	72	20	85	99	100	100
II-3-D	48	5	99	100	100	99
II-4-D	733	94	96	82	96	100
II-5-D	360	58	93	92	95	100
II-6-D	72	53	93	98	100	100
II-7-D	72	32	81	99	99	100
II-8-D	96	98	35	96	100	100
Mean	190.63	58	84	95.63	98.63	99.75

baseline criteria with an average of 21 errors. The subjects in the Detailed Instructions Group acquired baseline trials in an average of 191 trials making 58 errors. This apparent mean difference was largely due to slow acquisition by just two of the subjects (II-4-D & II-5-D) and a one-factor between-subjects ANOVA confirmed that there were no significant differences between groups for trials [$F(1, 14)=1.99, p>.05$] and for errors [$F(1,14)=6.54, p>.05$].

The possibility of a typicality effect in acquisition was assessed for baseline errors for both groups. Figure 8 and Table 5 show that both the groups were similar in that the most errors were made on trials with zero-feature (trigrams) stimuli. A two-factor within-subjects ANOVA confirmed there was no main effect of Instructions [$F(1,7)=0.53, p>.05$] and no interaction [$F(3,21)=0.12, p>.05$]; however, there was a significant main effect for Features [$F(3,21)=11.53, p<.01$]. A Tukey's post hoc test found the Zero- Feature condition was significantly different from the One, Two, and Three-Feature condition ($p>.05$).

Baseline Reaction Times

As seen in Table 6, most subjects in the Minimal Instructions Group as well as most subjects in the Detailed Instructions Group responded slowest on trials with zero-feature stimuli. Mean baseline speed scores for sessions 1 and 2 are shown in Figure 9. As Figure 9 shows, responding on zero-feature and one-feature stimuli was the slowest for both groups. Speed scores for trials with two and three-feature stimuli were slightly higher than the speed scores of zero-feature and one-feature stimuli for both groups. A two-factor within-subjects ANOVA confirmed there was no significant difference

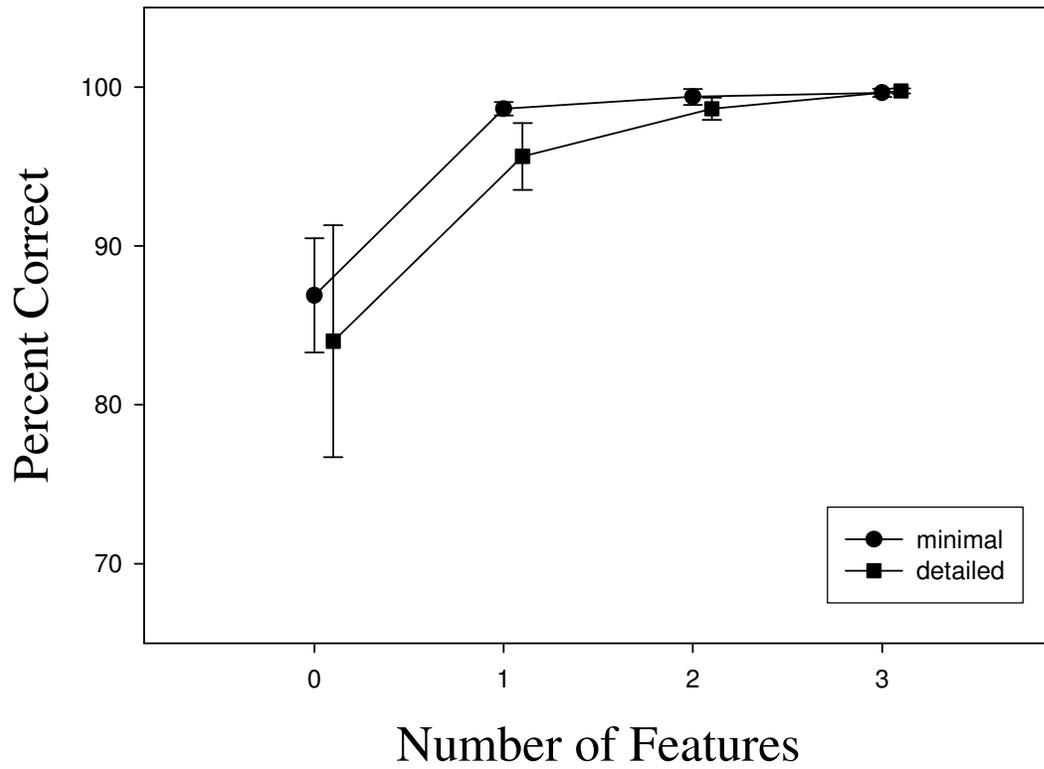


Figure 8. Percent correct on baseline errors for each number of features for subjects in Experiment 2.

Table 6.

Speed scores for baseline trials, equal probe trials and unequal probe trials for subjects in Experiment 2.

<u>Subject</u>	<u>Baseline Trials</u>				<u>Equal Probe Trials</u>				<u>Unequal Probe Trials</u>			
	<u>0F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>
II-1-M	.5	.5	.5	0.45	.3	.5	.4	0.4	.4	.38	.38	.48
II-2-M	.2	.3	.3	0.27	.2	.2	.2	0.45	.3	.26	.21	.13
II-3-M	0.4	0.4	0.5	0.45	0.3	0.3	0.4	0.37	0.3	0.4	0.45	0.43
II-4-M	0.6	0.5	0.7	0.56	0.3	0.4	0.5	0.56	0.5	0.59	0.5	0.67
II-5-M	0.2	0.5	0.4	0.59	0.3	0.4	0.5	0.53	0.4	0.45	0.5	0.53
II-6-M	0.3	0.5	0.5	0.5	0.3	0.4	0.5	0.59	0.5	0.45	0.48	0.59
II-7-M	0.4	0.5	0.4	0.42	0.2	0.3	0.4	0.34	0.4	0.38	0.4	0.42
II-8-M	0.3	0.4	0.5	0.48	0.2	0.3	0.4	0.48	0.4	0.42	0.42	0.34
Mean	0.4	0.4	0.5	0.47	0.3	0.3	0.4	0.47	0.4	0.42	0.42	0.45
II-1-D	0.6	0.8	0.9	0.77	0.4	0.6	0.6	0.91	0.7	0.77	0.67	0.83
II-2-D	0.4	0.6	0.8	0.63	0.3	0.4	0.6	0.83	0.5	0.56	0.63	0.91
II-3-D	0.6	0.6	0.6	0.45	0.4	0.4	0.5	0.56	0.5	0.48	0.43	0.53
II-4-D	0.5	0.5	0.5	0.59	0.2	0.4	0.6	0.77	0.4	0.56	0.48	0.59
II-5-D	0.5	0.5	0.5	0.67	0.3	0.5	0.5	0.71	0.6	0.56	0.5	0.63
II-6-D	.3	.4	.5	0.42	.2	.3	.4	0.53	.4	.43	.43	.53
II-7-D	.3	.3	.3	0.37	.2	.3	.3	0.36	.3	.37	.34	.45
II-8-D	.4	.4	.3	0.4	.3	.3	.4	0.5	.3	.37	.34	.45
Mean	.5	.5	.6	0.54	.3	.4	.5	0.65	.5	.51	.48	.62

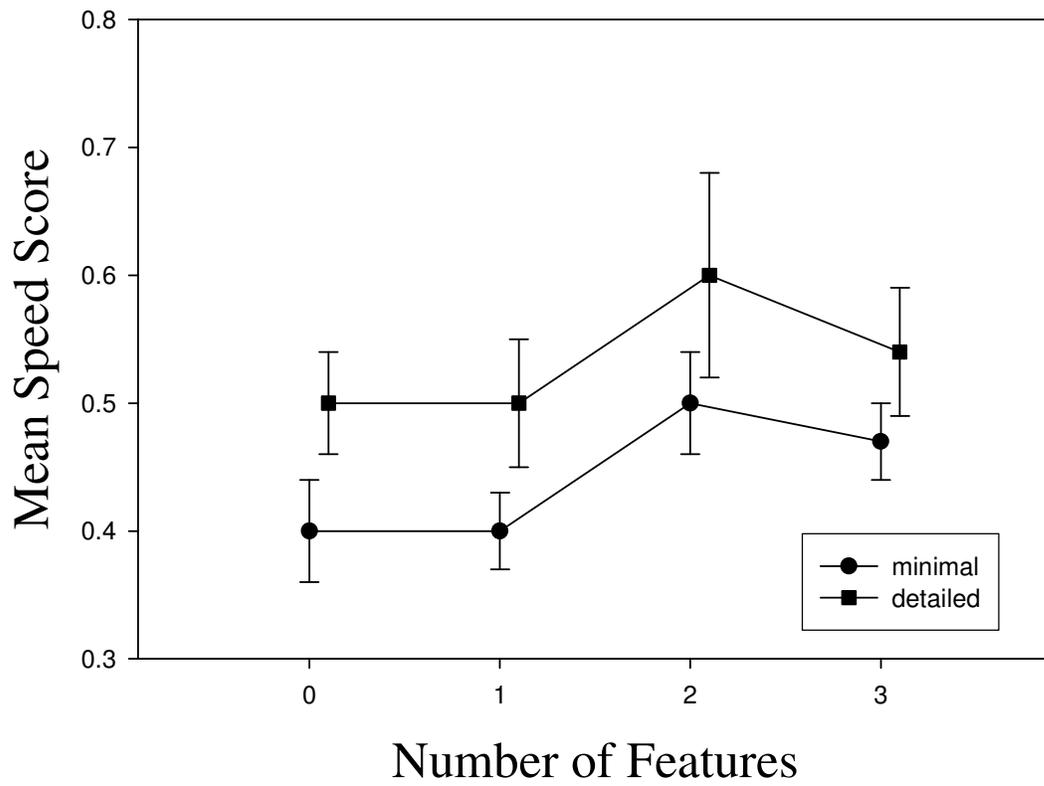


Figure 9. Mean speed scores on baseline trials for each number of features for subjects in Experiment 2.

between groups [$F(1,7)=1.75, p>.05$] and no interaction [$F(1,7)=1.75, p>.05$]. However there was a main effect of Features [$F(3,21)=5.62, p<.01$]. A Tukey's post hoc test found the Zero-Feature condition was significantly different from the Two and Three-Feature condition ($p>.05$).

Probe Trials

While performance on unequal-feature, equal-feature, and symmetry probes indicated class-consistent matching for many of the subjects, as seen in Figure 10, performance on transitivity probes did not indicate the same degree of class-consistent matching for subjects in the Minimal Instructions Group and in the Detailed Instructions Group. As shown in Table 7, only subject # II-3-M had a score high enough to suggest the formation of equivalence classes. All other subjects' scores suggest that equivalence classes had not formed.

Individual speed scores on probe trials, used to assess typicality effects, can be seen in Table 6. Individual speed scores, shown in Table 6, show both groups following similar patterns of responding for equal-feature probes. Figure 11 shows subjects' speed scores increasing as the number of features increase on equal-feature probe trials. The Detailed Instructions Group responded slightly faster than the Minimal Instructions Group. However, a two-factor within-subjects ANOVA verified there was no effect of Instructions [$F(1, 7)=3.47, p>.05$]. There was a main effect of Features [$F(3,21)=50.88, p>.01$] suggesting a typicality effect, and an interaction [$F(3,21)=6.37, p>.01$]. That is, subjects in the Detailed Instructions Group responded faster on trials with four-feature stimuli than did subjects in the Minimal Instructions Group. A Tukey's post hoc test

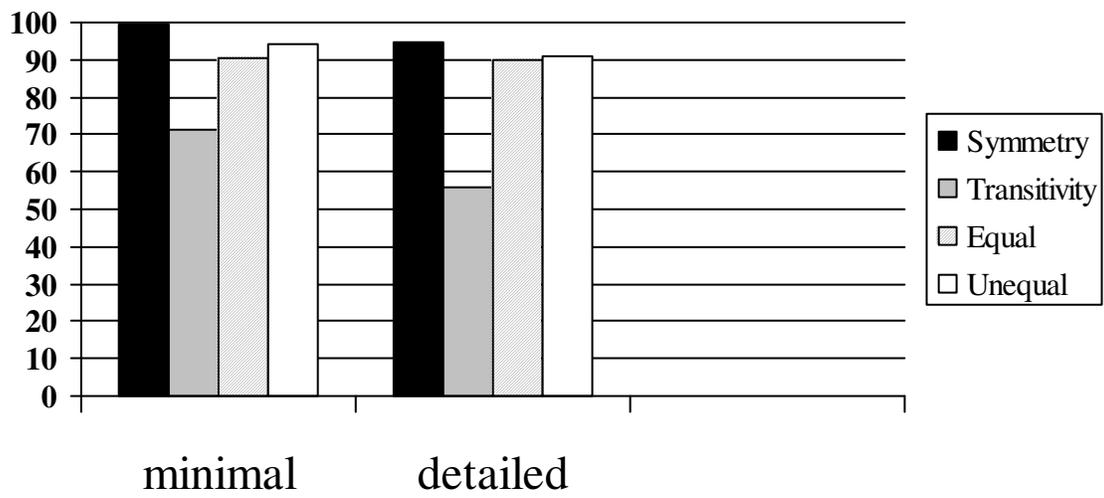


Figure 10. Percent correct on symmetry, transitivity, and novel probe trials for subjects in Experiment 2.

Table 7.
Percent correct on probe trials for subjects in Experiment 2.

<u>Subjects</u>	<u>Unequal</u>	<u>Equal</u>	<u>Symmetry</u>	<u>Transitivity</u>
II-1-M	92	72	100	43
II-2-M	93	94	96	83
II-3-M	94	91	100	98
II-4-M	97	93	100	58
II-5-M	94	92	100	65
II-6-M	95	94	100	85
II-7-M	93	91	98	61
II-8-M	94	95	100	78
Mean	94	90.25	99.25	71.38
II-1-D	91	90	86	54
II-2-D	90	88	96	54
II-3-D	92	89	100	52
II-4-D	89	86	95	67
II-5-D	88	88	92	67
II-6-D	92	92	92	46
II-7-D	95	94	100	48
II-8-D	92	91	96	58
Mean	91.13	89.75	94.63	55.75

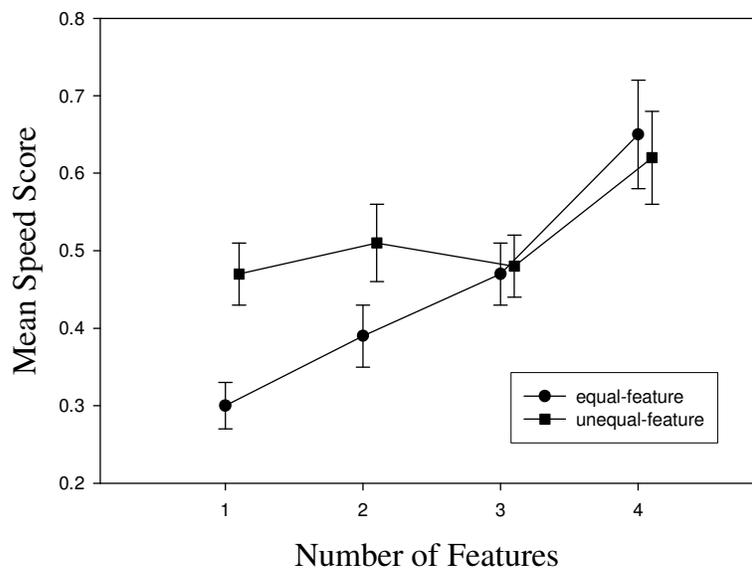
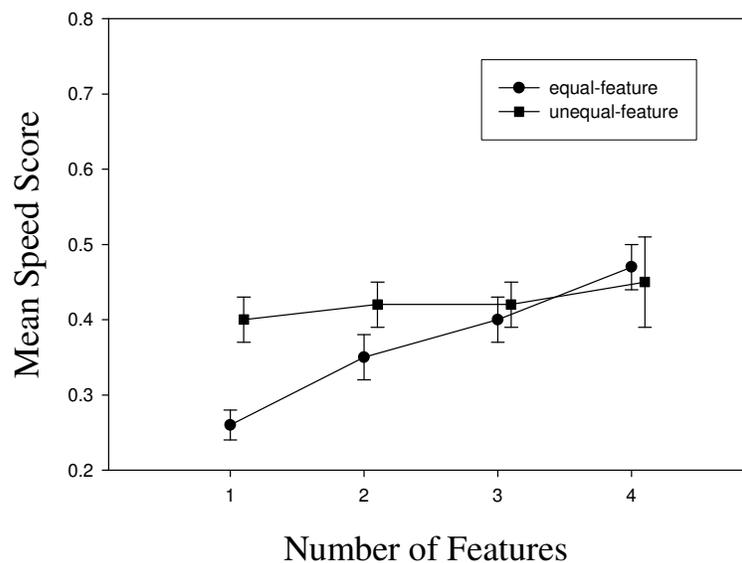


Figure 11. Mean speed scores on novel probe trials for each number of features for subjects in minimal instructions condition (top) and detailed instructions condition (bottom) in Experiment 2.

found there was a significant difference between all groups ($p < .05$) on number of features. That is, the Four-Feature condition was significantly different from the Three-Feature condition which was significantly different from the Two-Feature condition which was significantly different from the One-Feature condition.

Table 6 shows more variation between groups for unequal-feature probes. Figure 11 shows slower responding for subjects in the Minimal Instructions Group with a slight increase in responding on four-feature trials. The Detailed Instructions Group shows much faster responding on four-feature stimuli than on one, two, and three-feature stimuli trials. A two-factor within-subjects ANOVA confirmed a main effect of Features [$F(3,21) = 17.65, p < .01$], but there was no effect of Instructions [$F(1,7) = 2.28, p > .05$] and no interaction [$F(3,21) = 1.8, p > .05$]. A Tukey's post hoc test found the One, Two, and Three-Feature Groups were significantly different from the Four-Feature Group ($p > .05$).

Sort and Rating Task

As seen in Table 8, most subjects improved significantly in the post-sort task in comparison to the pre-sort task. A strong typicality effect is indicated in mean ratings plotted in Figure 12. Figure 12 shows that subjects' ratings of the stimuli increased as the number of relevant features increased. A two-factor within-subjects ANOVA confirmed a main effect of Features [$F(3,21) = 15.23, p < .01$]. There was no effect of Instructions [$F(1,7) = 0.9, p > .05$] and no interaction [$F(3,21) = 0.4, p > .05$]. A Tukey's post hoc test found the One and Two-Feature conditions were significantly different from the Three and Four-Feature conditions ($p > .05$).

Discussion

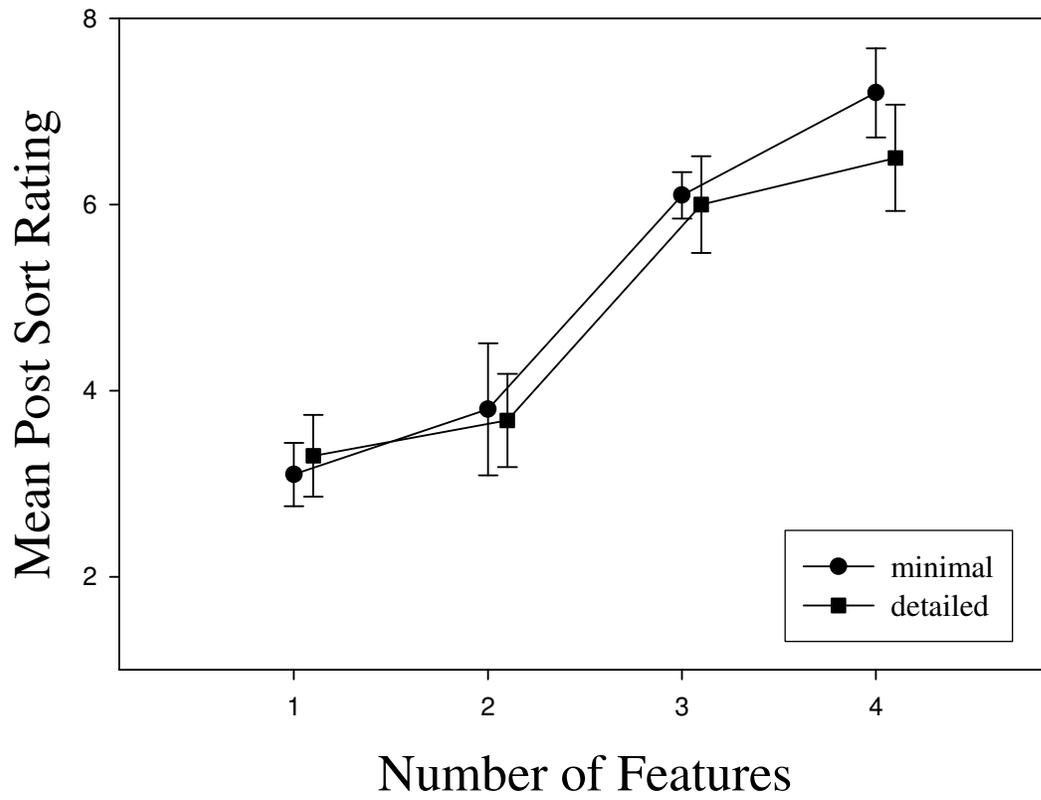


Figure 12. Mean post sort ratings for each number of features for subjects in Experiment 2.

Table 8.
Sort scores and ratings for subjects in Experiment 2.

<u>Subject</u>	<u>Pre-sort</u>	<u>Post-sort</u>	<u>1-Feature</u>	<u>2-Feature</u>	<u>3-Feature</u>	<u>4-Feature</u>
II-1-M	50	100	3	3	6.5	8
II-2-M	59	100	2.6	4.7	6.5	8
II-3-M	38	100	2.8	4	6.5	8
II-4-M	59	100	4	6.3	4.6	4.3
II-5-M	67	80	1.6	5.6	6.6	6
II-6-M	38	100	2.5	5	6.5	8
II-7-M	33	100	4.5	1	5.8	7.3
II-8-M	33	100	4	1	5.5	8
Mean	47.13	97.5	3.1	3.8	6.06	7.2
II-1-D	38	100	3	3.3	6.7	7.3
II-2-D	61	100	3.3	6.7	5	6.3
II-3-D	38	100	2.9	4.7	6.5	7
II-4-D	38	85	6.3	3	2.7	2.7
II-5-D	38	90	2.8	2.7	7	6.7
II-6-D	38	100	3.3	2.3	6.7	7.3
II-7-D	42	100	2.3	3	7	6.7
II-8-D	38	100	2.8	3.7	6.5	8
Mean	41.38	96.88	3.3	3.7	6.01	6.5

In Experiment 2, trigrams were removed from the sample and node position to see if equivalence and typicality would be affected. There was some evidence of result similar to those in the Galizio et al. study and in Experiment 1, such as relatively high levels of performance on symmetry, unequal-feature, and equal-feature probes. However, it appears that in Experiment 2, a critical feature is missing: the formation of equivalence classes. The subjects in this experiment appear to have formed some kind of class as evidenced by improvements in the post-sort task, but without the accurate performance on transitivity tests, these classes cannot be considered equivalence classes. This finding is puzzling and requires further consideration.

One possible explanation for these results is that the sample in this experiment facilitated identity matching among features. To review, in the Galizio et al. study and in Experiment 1, baseline training always involved matching a trigram to an abstract shape. In Experiment 2 the sample was always one of the three prototype abstract shapes. In this experiment on every trial, excluding the trials where the trigrams were comparisons, subjects could simply look at the sample, which always contained all four relevant features, and pick the comparison that had at least one of the same features. This may explain why subjects rapidly acquired the baselines and did well on symmetry, unequal-feature, and equal-feature probes, which also permitted identity matching, but failed to match in a class-consistent manner when it came to transitivity trials, which included only stimuli that shared no common features. This claim can be supported through inspection of Figure 8, which shows that subjects made a large number of errors on baseline trials that involved trigrams, but made only a small number of errors on the other trial types.

For this reason, it seems that the results of Experiment 2 may not be comparable to the results found in the Galizio et al. study and Experiment 1. In addition, it cannot be concluded that it was the training structure, per se, that resulted in the failure to form equivalence classes but rather, the possibility of feature matching. These results led to Experiment 3, which used the same procedures and training structure, but replaced the prototype samples of Experiment 2 with one set of one-feature abstract shapes, to reduce identity matching with features. With the sample containing only one of the relevant features, there were far fewer trials where identity matching could occur. For example, if the sample was a one-feature appendage (Stimulus E1 in Figure 1.), then comparisons including the other one-feature stimuli (Stimulus F1, G1, or H1 in Figure 1), the two-feature stimuli (Stimulus E1 in Figure 1), and the trigrams (Stimulus A1 in Figure 1) would not present an opportunity to identity match features. If the subjects were only matching identical features, they would never get enough trials correct within a block to move on to probe trials. Experiment 3 was conducted to address the same questions as Experiment 2 with the hope that the results would be clarified by reducing the likelihood of identity matching.

EXPERIMENT 3

Participants

There were 16 participants in Experiment 3. The participants were undergraduate students from the University of North Carolina at Wilmington. Participants received credit to fulfill course requirements in psychology classes.

Procedure

All aspects of the procedures were the same as Experiment 2 except that one-feature stimuli served as the samples for training and novel probe testing. There were four different one-feature training stimuli (see Figure 1). For this reason, four subjects received training in which the samples were one-feature appendage stimuli, four subjects received training with one-feature base stimuli, four subjects received training with one-feature fill stimuli, and four subjects received training with one-feature insert stimuli.

Results and Discussion

Acquisition

Number of trials to criterion for baseline conditional discriminations ranged from 72 to 821. Total number of trials and errors during acquisition are shown in Table 9. As Table 9 shows, it took subjects in the minimal instruction group an average of 364 trials to acquire baselines, making an average of 139 errors. The subjects in the Detailed Instructions Group acquired baseline trials in an average of 323 trials making 113 errors. A one-factor between-subjects ANOVA confirmed the absence of significant differences either for trials [$F(1,14)=0.04, p>.05$] or for errors [$F(1,14)=0.38, p>.05$].

Typicality effects were assessed for baseline errors for both groups. Figure 13 shows that the instructional groups were similar in that the most errors were made on trials with zero and one-feature stimuli. A two-factor within-subjects ANOVA confirmed there was no main effect of Instructions [$F(1,7)=0.71, p>.05$] and no interaction [$F(4,28)=2.09, p>.05$]. However, there was a significant typicality effect with a main effect on Features [$F(3,21)=30.90, p<.01$]. A Tukey's post hoc test found that the Zero-Feature and One-Feature Groups were significantly different from the Two-Feature, Three-Feature, and Four-Feature Groups ($p<.05$), which did not differ from one another.

Table 9.

Number of trials to meet criterion, total errors, and percent correct on baseline errors by number of features for subjects in Experiment 3.

<u>Subject</u>	<u>Trials</u>	<u>Errors</u>	<u>%correct</u> <u>0F</u>	<u>%correct</u> <u>1F</u>	<u>%correct</u> <u>2F</u>	<u>%correct</u> <u>3F</u>	<u>%correct</u> <u>4F</u>
III-1-M	264	84	92	82	100	93	98
III-2-M	264	87	88	85	85	98	98
III-3-M	72	23	96	95	98	99	98
III-4-M	821	339	70	72	80	85	90
III-5-M	192	79	80	84	100	98	100
III-6-M	312	178	92	75	100	92	98
III-7-M	432	153	77	79	100	86	99
III-8-M	557	168	92	85	99	90	100
Mean	364.25	138.88	85.88	82.13	95.25	92.63	97.63
III-1-D	228	186	88	65	87	99	99
III-2-D	456	98	95	82	100	93	100
III-3-D	679	224	56	70	70	99	99
III-4-D	408	157	75	76	99	91	99
III-5-D	192	41	90	89	97	100	100
III-6-D	168	41	83	70	90	95	98
III-7-D	192	77	77	80	85	98	99
III-8-D	264	82	75	76	95	94	99
Mean	323.38	113.25	79.88	76	90.38	96.13	99.13

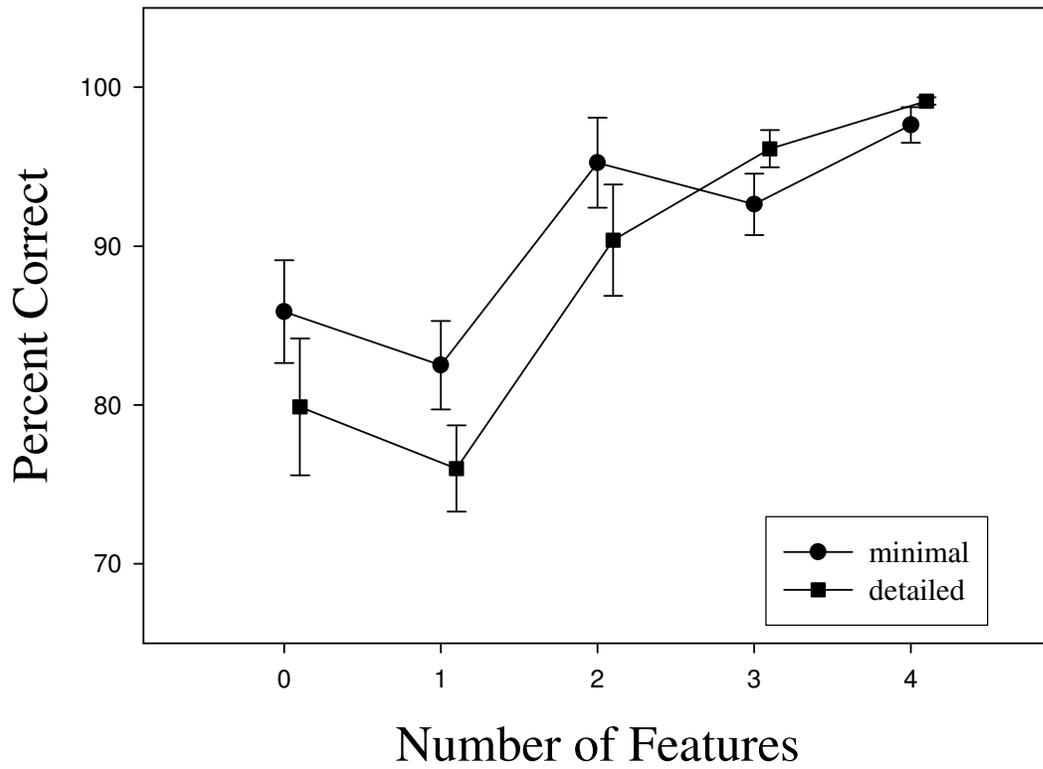


Figure 13. Percent correct on baseline errors for each number of features for subjects in Experiment 3.

Baseline Reaction Times

As seen in Table 10, most subjects in the Minimal Instructions Group as well as most subjects in the Detailed Instructions Group responded fastest to trials with four-feature stimuli. Mean baseline speed scores for sessions 1 and 2 are shown in Figure 14. As Figure 14 shows, responding became faster as number of features increased. A two-factor within-subjects ANOVA confirmed there was no significant difference between groups [$F(1,7)=1.39, p>.05$]. However there was a main effect of Features [$F(4,28)=13.12, p<.01$] demonstrating a typicality effect and an interaction [$F(4,28)=5.03, p<.01$] indicating that with detailed instructions, subjects responded faster on three and four-feature trials. A simple main effect of feature number was significant for both detailed and minimal instructions group. A Tukey's post hoc test found that the Minimal Instructions Group and the Detailed Instructions Group were significantly different on trials containing three and four-feature stimuli ($p<.05$).

Probe Trials

As shown in Table 11 and Figure 15, subjects in the Minimal Instructions Group had high accuracy on symmetry probes. However, on transitivity and novel probe trials, some subjects did not perform as well. In the Detailed Instructions Group, seven out of eight subjects formed equivalence classes. In contrast, only five out of eight subjects in the Minimal Group showed evidence of equivalence. Subjects # III-4-M and # III-8-M performed only slightly better than chance with scores of 46 and 43 on transitivity trials. Although, Subject # III-6-M did better than chance with a score of 79 on transitivity trials, it does not seem this score is high enough to meet most criteria for the formation of equivalence classes. In addition, half of the subjects (III-4-M, III-5-M, III-6-M, and III-

Table 10.

Speed scores for baseline trials and equal and unequal probe trials for subjects in Experiment 3.

<u>Subject</u>	<u>Baseline Trials</u>					<u>Equal Probe Trials</u>				<u>Unequal Probe Trials</u>			
	<u>0F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>	<u>1F</u>	<u>2F</u>	<u>3F</u>	<u>4F</u>
III-1-M	0.32	0.3	0.4	0.3	0.32	0.2	0.2	0.3	0.26	0.3	0.32	0.31	0.37
III-2-M	0.42	0.4	0.5	0.4	0.48	0.2	0.3	0.3	0.32	0.3	0.31	0.32	0.32
III-3-M	0.48	0.4	0.5	0.5	0.56	0.2	0.3	0.3	0.4	0.3	0.42	0.37	0.38
III-4-M	0.38	0.5	0.5	0.5	0.5	0.8	0.7	0.7	0.71	0.5	0.56	0.67	0.67
III-5-M	0.33	0.4	0.4	0.5	0.48	0.4	0.5	0.4	0.63	0.8	0.59	0.53	0.83
III-6-M	0.33	0.4	0.4	0.5	0.48	0.2	0.2	0.3	0.5	0.3	0.19	0.4	0.4
III-7-M	0.33	0.3	0.5	0.4	0.48	0.2	0.3	0.2	0.5	0.4	0.31	0.37	0.63
III-8-M	0.32	0.4	0.5	0.5	0.48	0.2	0.4	0.4	0.5	0.2	0.28	0.43	0.5
Mean	0.36	0.4	0.5	0.5	0.47	0.3	0.4	0.4	0.48	0.4	0.37	0.43	0.51
III-1-D	0.38	0.3	0.3	0.5	0.42	0.2	0.3	0.4	0.67	0.3	0.33	0.34	0.5
III-2-D	0.59	0.5	0.7	0.6	0.77	0.4	0.5	0.6	0.77	0.5	0.63	0.63	0.71
III-3-D	0.33	0.3	0.4	0.6	0.56	0.3	0.4	0.5	0.89	0.4	0.5	0.53	0.67
III-4-D	0.45	0.5	0.5	0.5	0.5	0.3	0.4	0.5	0.67	0.5	0.48	0.83	0.63
III-5-D	0.29	0.2	0.2	0.4	0.5	0.1	0.2	0.3	0.38	0.2	0.24	0.27	0.43
III-6-D	0.56	0.5	0.5	0.6	0.77	0.4	0.6	0.6	0.71	0.7	0.63	0.63	0.71
III-7-D	0.42	0.3	0.5	0.5	0.45	0.2	0.3	0.3	0.36	0.3	0.3	0.33	0.37
III-8-D	0.4	0.4	0.4	0.5	0.5	0.3	0.4	0.5	0.59	0.4	0.45	0.43	0.59
Mean	0.43	0.4	0.4	0.5	0.56	0.3	0.4	0.5	0.63	0.4	0.45	0.5	0.58

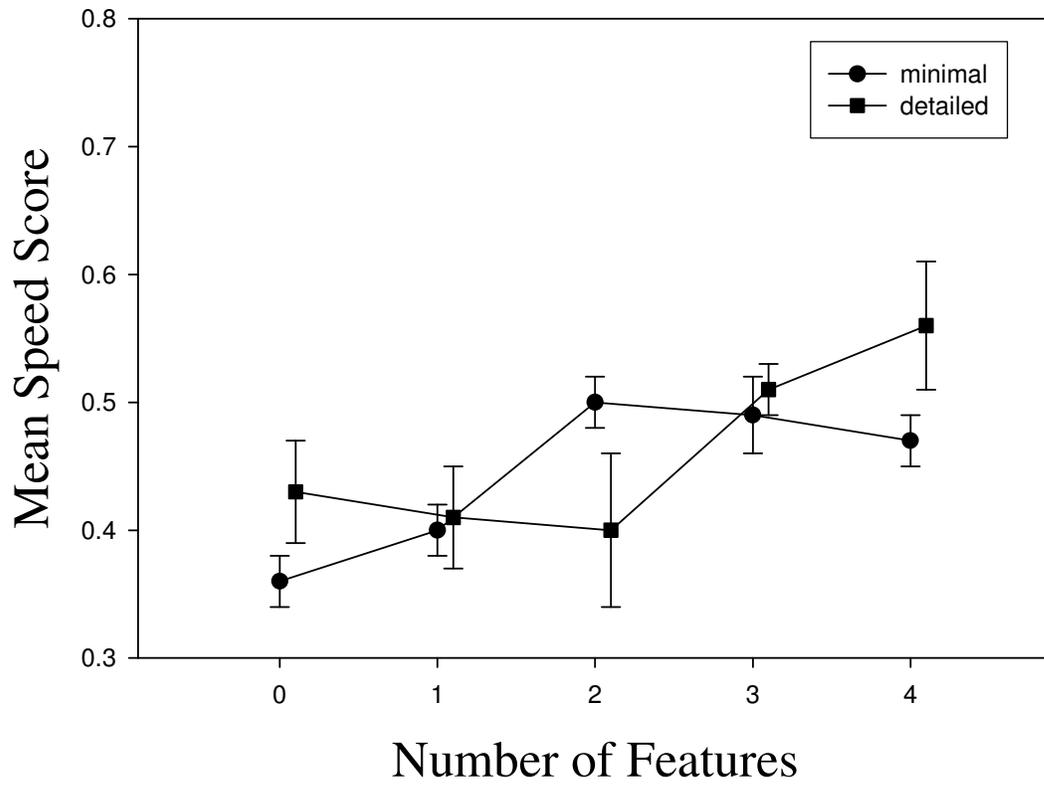


Figure 14. Mean speed scores on baseline trials for each number of features for subjects in Experiment 3.

Table 11.

Percent correct on probe trials for subjects in Experiment 3.

<u>Subject</u>	<u>Unequal</u>	<u>Equal</u>	<u>Symmetry</u>	<u>Transitivity</u>
III-1-M	93	93	100	98
III-2-M	94	94	100	100
III-3-M	90	89	100	97
III-4-M	67	68	82	46
III-5-M	86	88	96	92
III-6-M	84	85	92	79
III-7-M	93	91	98	100
III-8-M	78	75	77	43
Mean	85.63	85.38	93.13	81.88
III-1-D	84	78	94	71
III-2-D	94	95	100	100
III-3-D	94	94	100	96
III-4-D	95	92	98	98
III-5-D	94	92	100	100
III-6-D	97	97	100	100
III-7-D	93	92	100	100
III-8-D	95	94	98	98
Mean	93.25	91.75	98.75	95.38

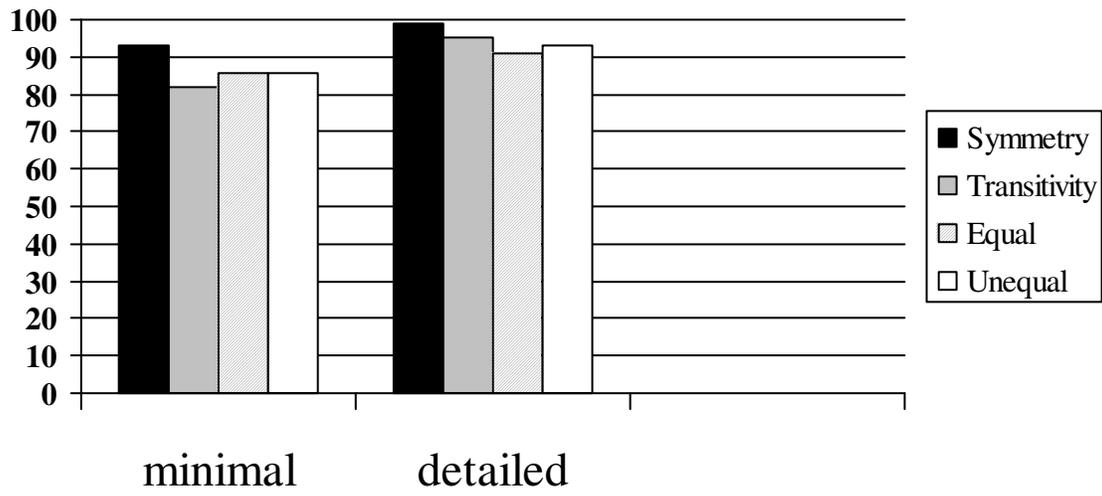


Figure 15. Percent correct on symmetry, transitivity, and novel probe trials for subjects in Experiment 3.

8-M) in the Minimal Instructions Group did not show generalization of class-consistent features. That is, subjects' novel probe scores did not reflect the same degree of mastery as did scores on symmetry trials. In the Detailed Instructions Group, all but one subject (III-1-D) showed mastery of unequal-feature and equal-feature probes, suggesting the formation of generalized equivalence classes.

Individual speed scores on probe trials can be seen in Table 10. Table 10 shows both groups following similar patterns of responding for probes where no zero-feature probe trials were responded to slower than four-feature trials. Figure 16 shows subjects speed scores increasing as the number of features increase, with the Detailed Instructions Group responding slightly faster than the Minimal Instructions Group on novel trials. A two-factor within-subjects ANOVA on equal-feature probe latencies verified there was no effect of Instructions [$F(1, 7)=0.41, p>.05$] and no interaction [$F(3,21)=2.37, p>.05$]. There was a main effect of Features [$F(3,21)=76.93, p>.01$] suggesting a typicality effect. A Tukey's post hoc test found the One-Feature condition was significantly different from the Two-Feature condition, the One-Feature and Two-Feature condition were significantly different from the Three-Feature and Four-Feature condition, and the Three-Feature condition was significantly different from the Four-Feature condition, ($p<.05$). A two-factor within-subjects ANOVA on unequal-feature probe latencies verified there was no effect of Instructions [$F(1, 7)=0.50, p>.05$] and no interaction [$F(3,21)=0.30, p>.05$]. There was a main effect of Features [$F(3,21)=9.19, p>.01$] suggesting a typicality effect. A Tukey's post hoc test found the One-Feature and Two-Feature condition to be significantly different from the Four-Feature condition, ($p<.05$).

Sort and Rating Task

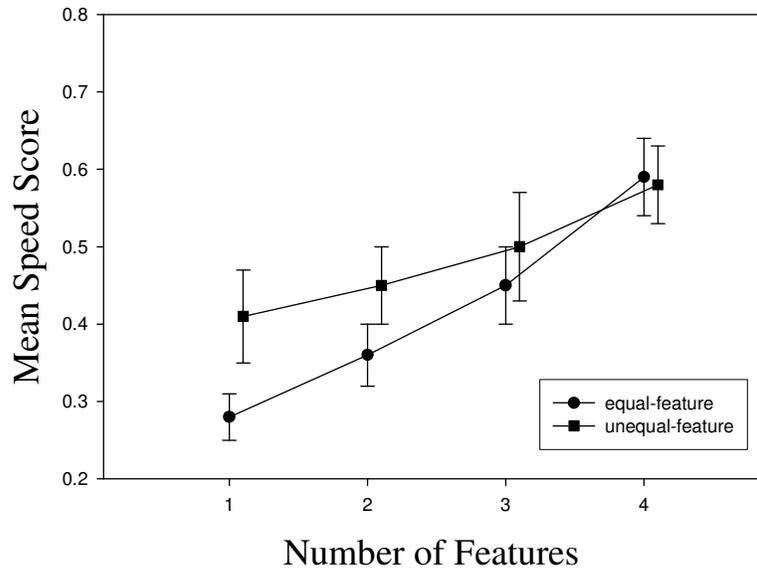
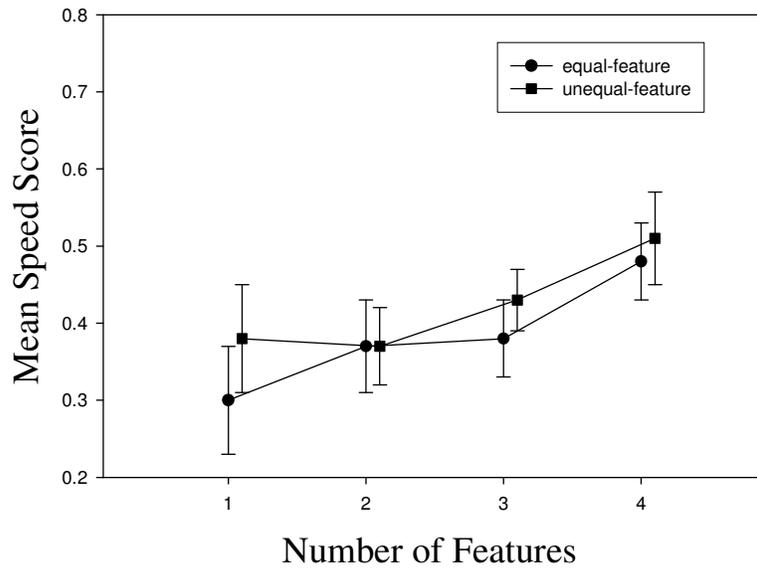


Figure 16. Mean speed scores by feature number on novel probe trials for subjects in the minimal instructions condition (top) and the detailed instructions condition (bottom) in Experiment 3.

As seen in Table 12, most subjects improved significantly on the post-sort task in comparison to the pre-sort task. Subjects were mostly random on the pre-sort task, but were nearly perfect on the post-sort task. A strong typicality effect is indicated by mean ratings plotted in Figure 17. Figure 17 shows subjects' ratings of the stimuli increased as the number of relevant features increased. A two-factor within-subjects ANOVA confirmed a main effect on Features [$F(3,21)=35.05, p<.01$]. There was no effect of Instructions [$F(1,7)=3.49, p>.05$] and no interaction [$F(3,21)=0.93, p>.05$]

In the Detailed Instructions Group, seven out of eight subjects showed generalized equivalence classes, and typicality effects were found in baseline errors, baseline and probe speed scores, as well as post sort ratings. While some of the subjects in the Minimal Instructions Group also showed these effects, it was much less consistent.

GENERAL DISCUSSION

This set of experiments was conducted to expand on the findings of the Galizio et al. study. There was interest in the effects of training structure and instructions on the subjects' success in forming equivalence classes and showing typicality effects. Experiment 1 used a many-to-one training structure to see what outcomes this opposite training would have. Results from Experiment 1 were that subjects successfully formed equivalence classes, but typicality effects were not as strong as in the Galizio et al. study. Results from Experiment 1 showed typicality effects in baseline errors and post-sort ratings, but unlike the Galizio et al. study, no typicality effects in baseline and probe-trial speed scores were obtained. The purpose of Experiment 2 was to replicate the Galizio et al. findings under several new conditions. First, a new training structure that involved removing the trigram from the node and sample position was used. This training

Table 12.

Sort scores and ratings for subjects in Experiment 3.

<u>Subject</u>	<u>Pre-sort</u>	<u>Post-sort</u>	<u>1-Feature</u>	<u>2-Feature</u>	<u>3-Feature</u>	<u>4-Feature</u>
III-1-M	38	100	2.7	4.3	6.5	8
III-2-M	37	100	2.7	4.3	6.5	8
III-3-M	61	100	2.8	4	6.5	8
III-4-M	38	38	4.8	4	4.2	4
III-5-M	27	100	2.8	4	7.5	6
III-6-M	38	92	2.9	3.3	6.8	7.3
III-7-M	33	100	3.4	2.3	6	8
III-8-M	38	38	5.2	5.3	3.5	2
Mean	38.75	83.5	3.4	3.9	5.9	6.4
III-1-D	38	59	3.8	5	5.2	5.3
III-2-D	34	100	2.6	4.3	6.5	8
III-3-D	42	96	2.7	4	6.5	8
III-4-D	61	100	2.9	3.3	7	7.3
III-5-D	42	100	2.8	4	8.3	6.7
III-6-D	51	100	2.8	5	6.5	8
III-7-D	38	100	2.5	5	6.5	8
III-8-D	38	100	3	3	6.3	8
Mean	43	94.38	2.9	4.2	6.6	7.4

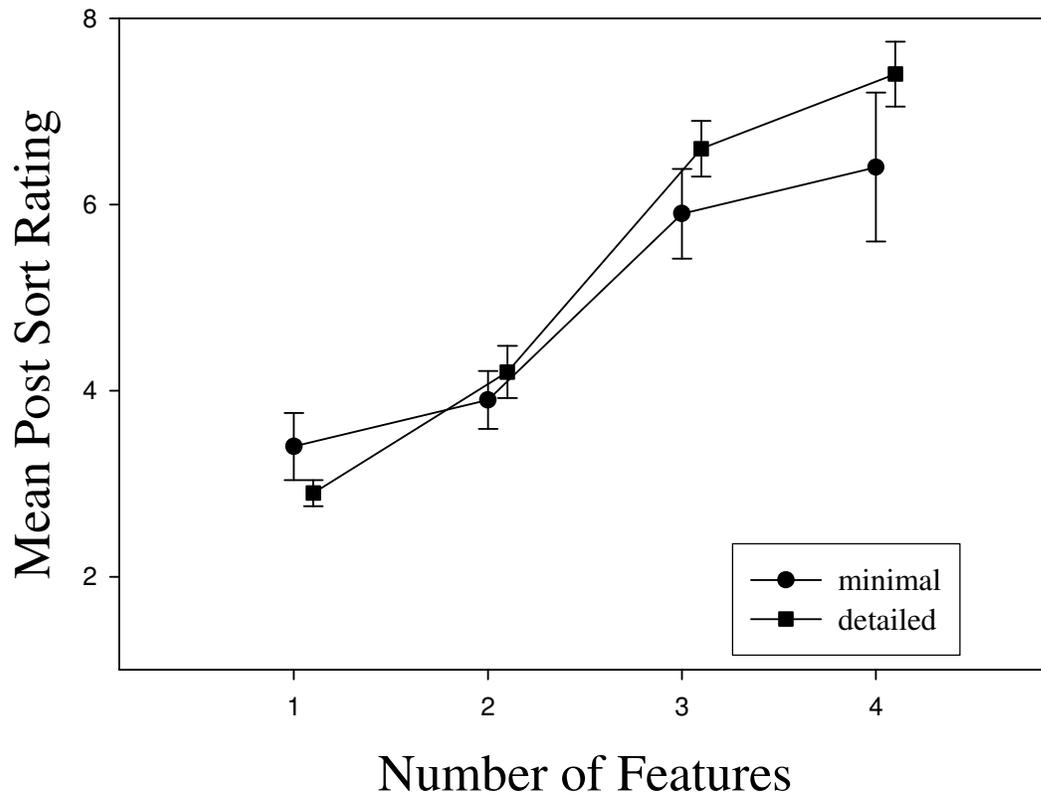


Figure 17. Mean post sort ratings for each number of features for subjects in Experiment 3.

structure was a one-to-many training structure, which placed the prototype in the sample/node position. Second, a minimal set of instructions was introduced. Findings from Experiment 2 revealed that subjects failed to form equivalence classes, perhaps due to feature identity matching developed during the training. Experiment 3 also used a one-to-many training structure, with one-feature abstract shapes as the sample to reduce control by identity matching. Subjects' performances in the Detailed Instructions Group in Experiment 3 closely paralleled Galizio et al.'s findings in several ways. Most subjects formed generalized equivalence classes, and showed typicality effects in baseline errors and latencies, probe latencies, and post-sort ratings. The success of these subjects confirmed that the trigram does not have to be the sample or the node in order for family resemblance classes to form and typicality effects to occur. Yet, it seems that the instructions may have influenced the subjects' performances. Unlike results obtained from subjects in the Detailed Instructions Group, results from subjects in the Minimal Instructions Group were not comparable to the Galizio et al. study in terms of the formation of equivalence classes.

As discussed earlier, the interpretation of the Experiment 2 results is somewhat problematic. It seems that Experiment 2 did not serve the purpose it was intended to and the questions that were proposed were not answered in a meaningful way. Due to the nature of the training structure, subjects seemed to have learned the baselines by simple identity matching of features. To review, subjects were given the prototype as the sample and were required to match it to the other abstract shapes and trigrams. On a large majority of trials (21/24), subjects could simply choose the comparison that had a feature that matched one or more of the sample's features. Their scores reflected mastery of the

conditional discriminations and subjects were moved on to probe trials. Subjects could continue to use this strategy of feature matching on symmetry, unequal-feature, and equal-feature probe trials, and show high performance levels. However, when subjects were given transitivity trials, this approach was no longer possible. Common features between sample and class-consistent comparisons were absent on transitivity trials, making it impossible on most trials to feature match. That subjects performed at very low levels on these trials made it clear that, in fact, these subjects had not formed equivalence classes.

Unlike Experiment 2, subjects in Experiment 1 and in the Detailed Instructions Group in Experiment 3 did form equivalence classes. Subjects in Experiment 1 were trained with a many-to-one structure in which the comparison, the trigram, was also the node. Subjects in Experiment 3 were trained with a one-to-many structure in which the one-feature stimuli served as the sample and the node. Subjects in both experiments showed evidence of generalized stimulus equivalence. The formation of equivalence classes in both experiments shows that several different training structures can be successful in training generalized equivalence. However, it seems that typicality effects, especially those assessed through latencies may depend on certain training structures. Additionally, instructions did produce differences among groups in forming equivalence classes with subjects in the Detailed Instructions Group forming equivalence at a much higher rate than the subjects in the Minimal Instructions Group.

As mentioned before, a surprising difference was found with regards to instructions in Experiment 3. With the exception of subject # III-1-D, the subjects in the Detailed Instructions Group performed at a much higher level on unequal-feature, equal-

feature, and transitivity trials. The performances by the subjects in the Detailed Instructions Group suggest that these subjects had formed generalized equivalence classes. Not only were subjects accurate on transitivity trials, but when subjects were presented with novel probe trials, they were successful in abstracting the relevant features and matching accordingly. Yet, only four of the eight subjects in the Minimal Instructions Group showed evidence of these generalized equivalence classes. Perhaps subjects that were given the detailed instructions were more likely to form classes because the instructions explicitly stated that the “objects go together.” It is possible that the generalized equivalence classes would have slowly emerged in subjects in the Minimal Instructions Group if more time was available. Their scores show that they were performing well above chance, so it could be that they were not incapable of learning generalized equivalence classes, however, they were slower to learn due to the instructions.

Additionally, these experiments speak to the question of the role of the trigram in the training structure. Although the trigrams were involved in the training structure of Experiment 3, they served neither as samples nor as nodal stimuli. Despite this, subjects were still able to form generalized stimulus equivalence classes, at least if they received detailed instructions. The issue of naming in the formation of stimulus equivalence classes can also be addressed by the success of Experiment 3. While naming cannot be ruled out as a possibility, it would seem that with this structure, naming would be much less likely. The trigram, which can be seen as a name for the class, was presented much less frequently. Furthermore, the trigram was not the nodal stimulus for the classes. It would seem, from the results of this training structure, that naming was not necessary for

subjects to form generalized stimulus equivalence classes as Lowe and Beasty (1987) suggest. On the other hand, the fact that instructions appeared to modulate this effect indicates that verbal factors may indeed play a role.

Future studies can further address this question. The success of subjects in this study suggests that subjects might also be able to form generalized stimulus equivalence classes with a training structure that eliminates the trigram. Future experiments could use the same general procedure and same training structure as Experiments 1 and 3 with one exception. In these experiments, the trigram could be completely taken out of the training structure by replacing it with an abstract shape. Using an abstract shape instead of a trigram would not provide subjects with a stimulus that could be used to name the class. Subjects could still come up with a name for the classes, but this training structure would not facilitate naming. If subjects could form generalized stimulus equivalence classes with this structure, it would provide further evidence that naming may not be necessary for generalized stimulus equivalence classes to form.

Just as the similarities of Experiments 1 and 3 provide us with a better understanding of the conditions necessary to form equivalence classes, the differences of these two experiments also further understanding. One difference involves acquisition of baselines. Subjects in Experiment 1 took much longer to learn baselines than subjects in Experiment 3. Subjects in Experiment 1 acquired baselines in an average of 520 trials making an average of 233 errors. Subjects in Experiment 3 acquired baseline in less time, an average of 343 trials, and made fewer errors, on average 126. These results support earlier findings (Saunders, Drake, and Spradlin, 1999; Fields, Hobbie-Reeve, Adams, and Reeve, 1999; and Spradlin and Saunders, 1986) that suggest subjects trained

with the one-to-many training structure require more trials than subjects trained with many-to-one training structure.

In addition, previous studies (Saunders, Drake, and Spradlin, 1999; Fields, Hobbie-Reeve, Adams, and Reeve, 1999; and Spradlin and Saunders, 1986) have found that subjects trained with a one-to-many training structure do not perform as well as subjects trained with many-to-one training structure on equivalence probes. The current study found this to be true. All subjects that were trained with many-to-one structure showed equivalence, while only 11 out of 16 subjects trained with one-to-many training structure showed equivalence. However, as mentioned earlier, this difference could be due to the different instructions that were used. In Experiment 1, which only used the detailed instructions, all of the subjects formed equivalence classes. Likewise, in Experiment 3, many of the subjects (seven out of eight) in the Detailed Instructions Group formed equivalence classes, but only four out of eight subjects in the Minimal Instructions Group formed equivalence classes.

Another important difference between Experiments 1 and 3 involved typicality effects. Like Galizio et al., Experiment 3 found typicality effects in baseline errors, baseline and probe response speeds, and post-sort ratings. In contrast, subjects in Experiment 1 showed typicality effects in baseline errors and post-sort ratings, but not on either of the response speed measures. It is possible that the training structure influenced this difference. Another possibility is that the structure of the probes had an influence on latencies.

Experiment 3, which used a one-to-many training structure in baselines, also used this structure in probe trials. This structure requires simultaneous discriminations

between the abstract stimuli. As described earlier, the abstract stimuli were very complex, making discriminations difficult and time consuming. That is, three of the abstract stimuli appeared as comparisons, which required subjects to discriminate between them and ultimately compare them to the sample. In Experiment 1, the trigrams served as comparisons, making it much less complicated for subjects to discriminate the comparison stimuli. It seems likely that discriminating between three relatively simple trigrams would take less time than discriminating between three complex abstract shapes. The complexity of the abstract stimuli as comparisons may have required several seconds to compare to each other as well as the sample to determine which comparison was the correct answer. It could be that the nature of the comparison display that affected latencies. Typicality effects, as shown in latencies, could have been masked due a ceiling effect: the very rapid response times in the many-to-one structure of Experiment 1. To check this prediction, future studies that maintained this structure for baseline trials, but changed to a one-to-many training for probe trials could be conducted. If this prediction is correct, results would show a lack of typicality in baseline trials, but typicality could be seen in probe-trial latencies.

To summarize, the similarities of the Galizio et al. study and Experiments 1 and 3 suggest that different structures, one-to-many with the trigram as the sample, many-to-one with the trigram as the comparison, and one-to-many with a one-feature shape as the sample, can be used to train generalized stimulus equivalence classes. The results show that subjects in Experiment 3, where the trigram was not the sample or the node, performed just as well as subjects in Experiment 1, where the trigram was the node. This suggests that neither the trigram nor a particular training structure is an essential part of

forming generalized stimulus equivalence classes. Furthermore, this finding speaks to the argument that naming may not be necessary to form equivalence classes.

Despite the differences across experiments, it appears that different training structures can be used to train generalized stimulus equivalence classes and produce typicality effects. The effects observed in the Galizio et al. study are not limited to one particular training structure. Furthermore, this evidence of generalization would suggest that family-resemblance classes derived from contingencies are a good model for natural language categories. These experiments indicate that several different training structures can be used to form language-like equivalence classes with typicality effects.