AGING AND CATEGORIZATION: USING GENERALIZED EQUIVALENCE CLASSES AND THEIR CHARACTERISTICS TO COMPARE OLDER AND YOUNGER ADULTS

Christine M. Engle

A Thesis Submitted to the University of North Carolina Wilmington in Partial Fulfillment of the Requirements for the Degree of Master of Arts

Department of Psychology

University of North Carolina Wilmington

2005

Approved by

Chair

Accepted by

Dean, Graduate School

This thesis has been prepared in the
Style and format consistent with the
American Psychological Association

TABLE OF CONTENTS

ABSTRACT

Previous literature has done little to bring together accounts of stimulus equivalence, the transfer of function among stimulus classes, and age-related changes associated with the creation of stimulus classes. This experiment explores these ideas using two participant groups, one consisting of younger adults and one consisting of adult volunteers over the age of 65. Participants were given training using nonsense syllables and eight sets of abstract stimuli. The stimuli differed on a number of features, four of which were class-consistent. Each stimulus contained a combination of one, two, three, or four of the class-consistent features, and the number of class-consistent features was used to identify the typicality of the stimulus within each class. Upon completion of the equivalence training and testing procedure, each participant was told that one of the stimuli from training carried a disease that infects 50 % of the animals or plants with which it comes into contact. Participants were then shown a series of stimuli from the testing phase of the equivalence procedure and asked to rate how likely each of these stimuli were to also infect plants or animals. Ratings from this phase determined the transfer of function within the stimulus classes created during the equivalence training procedure. Results showed that older adults need more training trials to master baseline criterion levels than younger adults did, but both groups demonstrated the formation of equivalence classes and typicality effects within those classes. Further, both groups also demonstrated transfer of function within the equivalence classes that was related to the typicality rating of each stimulus within a class.

ACKNOWLEDGEMENTS

I would like to thank the members of my defense committee, Dr. Carol Pilgrim and Dr. Nicole Alea, for all of their feedback and efforts on my behalf involving my research. I would like to give special thanks to my advisor, Dr. Mark Galizio, for all of his support and guidance in both my research and my classroom efforts. Thanks also go Dr. Joe Kishton and to my grandmother, Mrs. Agnes Ruppel, who helped to provide me with adult volunteers willing to participate in this experiment. More thanks also go to the stimulus equivalence lab group, who helped me keep my sanity by listening to my complaints and offering positive support. Finally, special thanks go to my parents, my brother and his family, and my fiancé, Todd. To my parents, thank you for all of your support, love, and advice you've both given during the duration of the experiment. To my brother and his family, thanks for all your help over Easter during the computer crisis. To Todd, thanks for always listening to my research chatter even when you may not have totally understood what I was talking about, and for giving your total support during my time in graduate school.

DEDICATION

I would like to dedicate this thesis to the stimulus equivalence lab group, as proof of a finishing point when your research seems like it may go on endlessly.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Most studies of equivalence classes have used young adults and children as subjects. However, the processes formed in childhood and practiced in young adulthood are the same processes that are used in older adults (see McDowd & Shaw, 2001 for a review). In light of some of the recent research in cognitive functions that decline with age, it is of interest to study class formation in older adults as well. With more and more research demonstrating the gaps between younger and older adults in areas such as language, memory, and attention (McDowd & Shaw, 2001), it seems important to determine if class formation also changes with age. Some research has already begun in this field using the stimulus equivalence method mentioned above (Wilson & Milan, 1995; Perez-Gonzalez & Moreno-Sierra, 2001) but with mixed results and further questions raised. Before reviewing this research and proposing some ideas to be included in new research, it is first necessary to review the principles and terminology of stimulus equivalence.

Stimulus equivalence is an approach to categorization that is based on the basic principles of operant conditioning. An example to consider throughout this explanation is one of a young child and his bottle. The child reaches for the bottle and gets milk. In this scenario, the child's behavior of reaching for the bottle is called the *response,* and the consequence of getting the bottle (and the milk) is called the *reinforcer*. The most basic unit of behavior is this two-term contingency, the relationship between a response and a reinforcer. When a response can produce a reinforcer and that reinforcer is unlikely to be produced in the absence of that response, then a contingency is said to exist between the two. Additionally, it is assumed that other responses do not reliably produce the same reinforcer. Reinforcement increases the likelihood that the behavior will occur again and punishment decreases the likelihood that the behavior will occur again. Thus, in the example mentioned earlier, the child will reach for the bottle more and more as each time he receives the reinforcement of the milk. Now imagine that the child is in a crib, surrounded by many objects including full and empty bottles. The child learns to discriminate between all other objects, reaching for the full bottle to get the milk. Reaching for the empty bottle or a blanket does not provide the same reinforcement of milk; thus, the full bottle has become a *discriminative stimulus*. The likelihood of the response producing the reinforcer changes depending on the presence or absence of the discriminative stimulus. Thus, the three-term contingency consists of a discriminative stimulus, a response, and a reinforcer. A fourth term, known as the *conditional stimulus*, can also be added to this process to produce conditional discrimination. Continuing with the previous example, suppose that the child's mother plays a game in which she says the word "bottle" or the name of some other object when the child is in the crib with many objects including the full bottle. If the mother permits the child to obtain the bottle only when the child reaches for it after hearing the word, the mother has created a four term contingency. Thus, the spoken word "bottle" becomes the conditional stimulus, determining which object to select from the crib to produce reinforcement. Conditional discriminations contain all four parts; the conditional stimulus, which cues the discriminative stimulus that, when selected, produces reinforcement. Conditional discriminations are the contingencies used in studying stimulus equivalence.

Understanding Stimulus Equivalence

Sidman's method of studying stimulus equivalence can be explained best using a pioneering study involving eight children between five and eight years old (Sidman & Tailby, 1982). Sidman and Tailby used a match-to-sample procedure in which participants are asked to match one item (the *sample*) to one of a group of items presented later (the *comparisons*). Sidman and Tailby initially presented dictated names of the Greek letters "lambda", "xi", and "gamma" as the samples, and the uppercase printed symbols of these letters as the comparisons. When the spoken word "xi" was the sample, the choice of the uppercase printed symbol "Ξ" was reinforced; when the spoken word "lambda" was the sample, the printed symbol "Λ" was reinforced, and when the spoken word "gamma" was the sample, the printed symbol "Γ" was reinforced. Participants were then given additional training using lowercase printed Greek letters as the comparison choices. At the end of this training, participants could correctly match the dictated Greek letter names to both their uppercase printed symbol and their lowercase printed symbol. To facilitate in explaining this procedure, Sidman and Tailby chose to label each group of stimuli (dictated name, uppercase letter, or lowercase letter) with single letters, followed by numbers to distinguish between stimuli within each group. For example, all dictated Greek letter names became known as "A", and the specific dictated name "lambda" became known as "A1", "xi" as "A2", and "gamma" as "A3". The uppercase printed symbols were grouped as "B" stimuli ("B1" being the uppercase symbol for lambda, "B2" the symbol for xi, and "B3" the symbol for gamma), and the lowercase printed symbols were grouped as "C" stimuli. Thus, the training procedures mentioned above trained A samples to B comparisons (dictated names to uppercase printed symbols) and A samples to C comparisons (dictated names to lowercase printed symbols). One can also see that this labeling system also determines which stimuli can be matched together; e.g., all of the "1" stimuli are matches for one another, regardless of which appears as the sample and which are the comparisons. Sidman and Tailby then introduced a fourth group of stimuli, printed symbols that were other Greek letters (labeled "D" in the notation) for the final training phase of the experiment. They presented the D stimuli as samples and the C stimuli as comparisons (Greek symbols to lowercase symbols) until participants performed at a 90% accuracy rate, and then intermixed these DC trials with the previously trained AB and AC trials. During the next phase of the experiment, Sidman and Tailby presented trials involving the same stimuli, but testing for untrained relationships. One of these types of testing trials used the C stimuli as samples with D comparisons (lowercase symbols to other Greek symbols) to determine if participants could correctly match these stimuli in reverse of the training procedure. Simply stated, if D1 goes with C1, then would C1 go with D1? This type of reversibility is known as *symmetry* in the equivalence literature, and is one of the relations necessary to produce equivalence relations. The other property necessary for equivalence relations are the transitivity relations, in which participants have not received any direct training linking two groups of stimuli but instead choose class-consistent comparisons based on the previous relations with the other stimuli. More simply stated, if A1 goes with B1 and B1 goes with C1, then would A1 go with C1? Sidman and Tailby tested transitivity relations with unreinforced probe trials using D samples with B comparisons, B samples with D comparisons, A samples with D comparisons, B samples with C comparisons, and C samples with B comparisons. Such tests have since come be called *equivalence* probes, because they involve the demonstration of both symmetry and transitivity.

Results from this experiment showed that six of the eight children correctly demonstrated the untrained relationships during the testing phases. Given only three types of training relationships (A to B, A to C, and D to C), these children showed five more emergent relationships (D to B, B to D, A to D, B to C, and C to B). These five relationships demonstrate the creation of a class via four-term contingencies. The children demonstrated formation of three distinct classes of stimuli, where all of the stimuli within a class were used interchangeably, though they were each physically different. This type of class formation is a model for the study of language, where different symbols (words or names) are used to represent items that are not present during conversation. Class formation also allows for the interchangeable use of different words with similar meanings. The authors noted the importance of these findings for the study of language development, saying "matching auditory to visual stimuli can represent simple auditory comprehension… matching visual stimuli to each other can constitute simple reading comprehension…and naming textual stimuli aloud can be simple oral reading" (Sidman & Tailby, 1982). Thus one can see the usefulness of using stimulus equivalence formation to learn more about the development of language and reading comprehension.

Theories

Sidman and Tailby's 1982 paper generated a field of research that has grown extensively in recent years. The various explanations for equivalence class formation account for different characteristics of equivalence classes in different ways. It is necessary to look at the three major theories of equivalence class formation to understand how characteristics of equivalence classes such as typicality effects and transfer of function are best explained. These and other characteristics will be discussed later in the paper.

In 2000, Sidman published an outline of his theory of equivalence class formation. He began by explaining that a reinforcement contingency can produce both analytic units and equivalence relations. These analytic units include two and three-term contingencies known as operant reinforcement and simple discrimination respectively, and four-term contingencies (described earlier as conditional discriminations). Equivalence relations are a different product of reinforcement contingencies, including the emergent relationships that are not directly reinforced. Thus, equivalence relationships are a fundamental outcome of reinforcement contingencies, not something that comes about from a learning history with those contingencies. The equivalence classes that are formed from these contingencies come to include not only the simple and conditioned stimuli, but also the response and the reinforcer. Thus Sidman's theory proposes that all elements of the contingencies are included in the equivalence classes. This idea of equivalence as a fundamental characteristic rather than a learned one set Sidman's theory apart from other proposed theories of class formation. Further, Sidman suggests that stimuli included in an equivalence class do not need a common name to become equivalent, as the reinforcement contingencies and emergent relations are enough.

While Sidman holds that naming is not a required element of equivalence classes, others have expressed differences in this viewpoint. Lowe, Horne, Harris, and Randle (2002) have theorized that naming is absolutely necessary for the formation of equivalence classes to occur. An earlier paper by Horne and Lowe, published in 1996, outlines the naming theory of equivalence. These authors explain that naming facilitates the formation of equivalence classes, and that a name is necessary to properly describe the relations between stimuli. Even if a word is not included in the stimuli used in the training and testing procedures, Horne and Lowe argue that participants will use verbal relations to form the class anyway. According to these authors, this naming ability is what makes equivalence relationships so uniquely human; only human participants have ever shown undeniable equivalence classes, although participants may or may not be able to produce the names given to each stimulus or class. Horne and Lowe argue that naming is essential to the complete formation of equivalence classes.

Lowe et al (2002) designed a group of three studies to determine if training a common name to several stimuli is enough to pull all of the stimuli into an equivalence class, without any other associations between the stimuli. Participants for Experiment 1 were 11 children between the ages of 2 and 4 years. The children were given two groups of stimuli, one including items familiar to the children and one of arbitrary, unfamiliar items. Stimuli used in the familiar group consisted of three different types of cups and three different types of hats. Stimuli in the arbitrary group consisted of six green wooden shapes. Children were first trained to vocally identify one of each kind of familiar item (denoted Cup1 or Hat1) when they were placed on a table in front of them. After reaching a criterion value (three correct answers of four trials), the children were given all six of the familiar stimuli at once and asked to name each one. These training trials insured that the children could correctly identify each item individually. The third phase of the training also included all six of the familiar items, but the children were asked to name one of the items (hat or cup) and to then give the experimenter the other stimuli with the same name; this was done until the children could complete three correct consecutive trials with each type of item. The same procedures were then repeated with the arbitrary wooden shapes, grouped into two classes of three shapes each and named either "zag" or "vek." Upon correct grouping and identification of the arbitrary stimuli, the children were then given reviews of all of the stimuli (familiar and arbitrary). The final phase of the experiment included testing sessions, in which the children were given one object (either arbitrary or familiar) and asked to name it and then to give the experimenter "the others". The testing was done first with all of the arbitrary stimuli, and then with all of the familiar stimuli. Results showed that four of the nine children that completed all of the phases of the experiment could correctly group both the familiar items and the arbitrary items by either using the names or by using only a sample. The other five children could correctly match in pairs and when asked to name, but not when given only a sample object and asked to choose the other matching objects. Lowe et al. noted that these five children could complete the matching trials with all of the items when prompted to name the objects, but not when naming was not required. Thus, "it may be necessary to ensure that the subject first emits the appropriate speaker behavior when he or she sees the sample before he or she can successfully select the others" (Lowe et al, 2002).

Experiment 2 was designed to determine if the previously created three-member classes of arbitrary stimuli could be expanded to include three more members in each class. Participants were two of the children used in the previous experiment, and training to identify the new objects as "zag" or "vek" was very similar to that described for

Experiment 1. Instead of being given only five objects as comparisons for the single sample, each child was given eleven items. Results from this experiment showed that class expansion could occur when the items were related by a common name, giving more evidence to the idea that the type of unbounded class structure that occurs in natural language categories comes about by assigning a name to the class. The third experiment was designed to determine how important spatial proximity of the stimuli was for correct categorization, and also introduced listener probes into the testing phase to determine if the children were equally able to identify the items when given the name as when given a sample item and asked to choose all of the matching items. Trials were conducted in the same manner as the previous two experiments. The items placed on the table during the training and testing phases were predetermined and counterbalanced to insure that spatial proximity of the items did not play any role in the children's classification of them. Results from this final experiment showed that the children were just as able to classify the items regardless of the location of the items on the table. Additionally, all of the subjects could correctly match the items together when given the name of the class (for example, "Where is the "vek"?") as when given a sample item. The authors conclude that Experiment 3 provided evidence that the arbitrary names given to the items entered into the created classes just as the objects had. Overall conclusions of the three experiments were that "simply training a common [vocal] response to each of a number of arbitrary stimuli establishes those stimuli as a class or category" (Lowe et al, 2002). The results showed that vocalizing a name for a group of related stimuli facilitated the learning of those arbitrary relationships, lending weight to the argument that a vocalization is an important part of a language-like category. However, this experiment does not provide evidence that naming is essential for this type of category formation, a point that other researchers have picked up on and followed up in other studies.

In an attempt to learn more about the necessity of naming in equivalence class formation, O'Donnell and Saunders (2003) reviewed studies of equivalence classes and language with 55 non-verbal individuals diagnosed with mental retardation. In this study, the results from fifty-five individuals used in previous equivalence class formation studies were reviewed. The individuals studied in this analysis had varying degrees of verbal skills (ranging from ability to complete short sentences to complete muteness). The types of equivalence classes formed and tested were grouped across studies according to the type of stimuli used (novel stimuli or arbitrary stimuli) and whether or not the reinforcement was given using a differential-outcomes procedure, to make all of the results comparable. With each of the individuals, the diagnosis, mental age, previous experience, naming ability, and equivalence class formation performance was entered into a table for comparison. The results showed that thirty-four of the 55 individuals showed equivalence class formation with accuracy above 90%, nine of the individuals showed no type of equivalence class formation, and sixteen showed progress on at least one of the relations necessary to form equivalence classes (symmetry, transitivity, or equivalence). These results held across language ability and previous experience with class formation. The authors explain that possible reasons for the failures were non-stringent accuracy criteria, which in turn gave participants experience with inaccurate trials but not much experience with accurate trials and reinforcement. The results show that even populations with significantly lower verbal skills are capable of forming equivalence classes, and in turn using those classes to learn basic reading comprehension skills. This conclusion is somewhat contradictory to the conclusions drawn from the Lowe et al (2002) study involving the necessity of naming in equivalence class formation. Thus, this paper demonstrates how equivalence classes can be used to study verbal ability and then used to help increase language and communication skills in individuals that were previously thought to be unreachable. One can begin to see how important the ability to form equivalence classes may be in both initial development of language ability and continued expansion of language with age.

Yet another area of research involved with class formation is Relational Frame theory. It is understood from Sidman's equivalence theory that all stimuli become interchangeable after entering into a class together; the stimuli are all equivalently related. Relational Frame theory proposes that the equivalence relation is only one of many possible relations that can form between stimuli within a class. In their 2001 paper outlining the details of Relational Frame theory, Hayes, Fox, Gifford, Wilson, Barnes-Holmes, and Healy describe other possible relations that might form, including "greater than", "opposite of", or " smaller than." Hayes et al. give the example of a child beginning to speak, during which a parent will often prompt the child to name different items or name items that the child requests. Soon the child learns the symmetric relationship between a name and an object, based on the reinforcement received from the parent; this relationship is bi-directional. Further exposure to symmetry trials until they are mastered allows the child to develop a relational response in which new exemplars control behavior in the same way as the previously trained relations. In a new setting, a familiar stimulus still has the same identity frame; when faced with a new question, such as "Which is bigger?" the child responds on the basis of past history with a reinforced response from a particular frame. Soon, the identity frame comes to include discrimination of other characteristics of the item, such as size. Drawing this theory further out leads to the idea of stimulus hierarchy within a class, in which relations between class members allow for ordering or ranking. For example, ordering all the stimuli in a class from “largest” to “smallest” is possible with Relational Frame theory. This notion has turned out to be quite important to class formation theories, and is discussed in more depth later in this paper. It is safe to say that equivalence theory, which relates all class members equally and does not require naming for class formation, has come up against some debate.

Categorization and equivalence

While the debate amongst the theories previously mentioned continues, there has been research to expand Sidman’s equivalence approach to language development. Some characteristics of language have been supported quite well using equivalence theory, including priming effects (Hayes and Bissett, 1998), hierarchical categorization (Griffee and Dougher, 2002), and transfer of function (Dymond and Rehfeldt, 2000). These areas of research have provided results that all theories of language must take into account, and so deserve some attention here.

One characteristic of lexical classes is the priming effect, where certain words will cue other words from the same category to be more easily recognizable. Hayes and Bissett (1998) set out to study the priming effect using equivalence classes, which have the advantage of using arbitrary stimuli rather than familiar stimuli. Hayes and Bissett explain that if equivalence classes are really like natural language classes, then they will show priming effects in the same way that those natural classes do. To test this, 14 participants were exposed to three-letter nonsense words in a match-to-sample procedure. The nonsense words were previously assigned to one of three arbitrary classes by the experimenters, labeled A, B, and C. Participants were told that the nonsense words were in fact foreign words, and they should attempt to learn which foreign words go together using the feedback given. Training trials included AB and AC relations. After reaching a 90% accuracy criterion with the training trials, participants were then tested for the equivalence relations, CB. Hayes and Bissett specifically did not test for the symmetric relations to allow for exposure control for the final task. After testing for equivalence relations, participants were then given trials in which two of the nonsense words appeared on the screen at the same time. They were asked to determine whether or not the two words were both "foreign" words exposed during training. Trials here included combinations of nonsense words that had been directly related during training, nonsense words that were related during the equivalence testing trials, nonsense words that were symmetrically related from the training but had not yet been exposed together, and nonsense words that were from different equivalence classes. Results from the experiment showed that participants had faster reaction times and fewer errors with those pairs that had been directly related during training and those pairs that had been related during the equivalence probe tests. Participants were slowest to respond and committed more errors with those words that were from different arbitrary classes. These results show that overall, arbitrary equivalence classes created in a laboratory can show the same priming effects that natural language categories show. Hayes and Bissett acknowledge how important of a step this is for the study of category formation using arbitrary classes, noting that cognitive literature in the same area has demonstrated extensive priming effects with natural categories. Priming effects are just one of the characteristics of natural categories that must be evident in any model of categories; another characteristic to be considered is hierarchical structure.

Griffee and Dougher (2002) looked at the hierarchical structure of categorization using arbitrary category formation in younger adults. These authors noted that within natural category formation, there are stimuli that are physically related and stimuli that are functionally related. The different relations among stimuli within a class develop at different times, such that young children most often associate stimuli that are physically similar, but over time learn that functional relationships between stimuli can also be important. Thus, there is a hierarchy that exists among related stimuli that allows the individual to determine the level of distinction that is most useful given a certain context. Griffee and Dougher identified three levels of categorization; the *superordinate* level is the most general, followed by the more specific *basic* level, and finally, the most specific *subordinate* level. For example, if one thinks of the category "living things", the superordinate level might include a distinction between plants and animals. The basic level makes a further distinction in the "animal" category, say between mammals, fish, and reptiles. The subordinate level would make an even further distinction; say between dogs, cats, and humans. The level of the category that is necessary is determined by the contextual control of the situation; if a child asks his mother what kind of pet he can get, her response of "animal" or "mammal" may not make as much sense as the more specific "dog." This particular kind of structure does not initially lend itself to the structure of stimulus equivalence classes because of the lack of symmetry in the hierarchy. It is fair to say that all dogs are mammals and that all mammals are animals, but not all animals are mammals and not all mammals are dogs. Further, the stimuli from the same level of hierarchy are not necessarily equivalent; dogs are not exactly the same as humans or cats. Griffee and Dougher propose a solution to this by noting that the contextual control by the situational cues (such as surrounding animals or a particular pet) is what cues the appropriate level of categorization to use. By including contextual control as a term in the contingencies created by reinforcement, these authors relate hierarchical categorization to previous stimulus equivalence literature. Their 2002 study attempted to determine if arbitrary classes created in a laboratory setting would exhibit the same hierarchical structure that natural language categories do.

Griffee and Dougher used fourteen triangles with varying top-most angle measures (from 12° to 168°) and corresponding side lengths, three background colors, and seven nonsense syllables. Younger adult participants were instructed to choose one of seven buttons presented at the bottom of a computer screen upon presentation of a sample. Initial training sessions matched four of the triangles with background colors, and participants learned that the correct button choice was dependent on both the shape and the color presentation. Thus, one button was always correct when the background was green, but when the background color was red, the correct button choice depended upon the sample given such that two shapes produced one correct button choice while two other shapes produced a second correct button choice. With the third background color (yellow), the correct button choice was also dependent on the shape presented, such that each of four shapes was paired with a separate correct button choice. This scenario thus created three levels of discrimination; with green, one button was always correct (the superordinate level); with red, two buttons could be correct and the triangle must be observed (the basic level, with one button assigned to the narrower triangles and a different button assigned to the wider triangles), and with yellow, four buttons could be correct and the triangle must be even more closely observed (the subordinate level, where each button choice was reinforced for a different sample triangle). Testing for this phase of training was similar to the training but used all of the colors and all fourteen of the triangles created, including the four triangles used in training. Participants were given the same instructions, to press whichever button they felt was most appropriate given the sample and the previous training. The second phase of the experiment was conducted to see if participants could learn to generalize the initial training to nonsense syllables. For the second phase of the training, each of the buttons across the bottom was labeled with one of the nonsense syllables and participants were given the same instructions. Correct choices were again determined by the background color and the shape of the triangle presented in the same manner as they were in Phase 1. Testing was conducted in the same manner as in Phase 1 as well. Participants were then given a final test of transitivity, to determine if they could correctly match the nonsense syllables with the buttons used in the first phase of training. Thus, nonsense words and background colors were presented as samples in varying combinations, and participants were instructed to indicate how well the color and the word went together using the buttons across the bottom of the screen. Results showed that participants were successful in learning the discriminations of Phase 1 and applying them to the novel triangles not presented in training. Thus, participants could correctly deem the appropriate level of discrimination (superordinate, basic, or subordinate) based on the background color of the sample. Further, participants also correctly learned the discriminations trained and tested in Phase 2, and all but one of the participants performed without error on the transitivity tests. Thus, participants were most accurate during the trials for the superordinate phase, selecting the correct button or word every time. Trials for the basic phase were less accurate, as the discriminations were no longer based only on background color, but also on triangle shape. Participants were most likely to miss the trials involving the triangles in the center of the spectrum, where the triangles were divided into the two classes. Finally, participants were least accurate during the subordinate phase trials, as the correct button or word choices were dependent on a more specific distinction between the triangle shapes. The authors conclude that this procedure was successful in creating hierarchical categorization based on contextual control, and that this type of categorization may be quite useful when applied to the natural language categories that are created outside of a laboratory.

The hierarchical model of categorization blends quite well with equivalence theory and the natural language classes that can be observed in people. It also provides an explanation of how certain stimuli may be related and thus occasion one another during free recall, a phenomenon known as clustering. *Clustering* occurs when related stimuli are grouped together, and thus recalling one stimulus from the group cues the recall of the other stimuli within the group. For example, the word “bread” may cue words such as “bake”, “toast”, or “sandwich”, but not words such as “wrench”, “sister”, or “midnight”. Clustering has been most often observed in free recall associations that rely on categories that already exist for the participant. Galizio, Stewart, and Pilgrim (2001) demonstrated clustering in artificial categories. Galizio and colleagues hypothesized that clustering may be similarly observed in arbitrary classes created using the stimulus equivalence match-to-sample procedure. Twelve nonsense syllables were generated and grouped into three arbitrary classes (A, B, and C) of four syllables each (1, 2, 3, and 4). Participants were exposed to AB, AC, and CD training. Testing procedures probed for the existence of reflexive, symmetric, and transitive relations between all of the stimuli within each class. The testing procedures first looked for the symmetric relations of BA, CA, and DC; transitive and equivalence tests probed for BC, AD, and BD relations. Participants were then given a sheet of paper and asked to recall the nonsense syllables; recalled words were scored based on their proximity to other words from the same class. Results showed that participants formed the equivalence classes A, B, C, and D, and most often recalled nonsense syllables in clusters based on these classes. Participants recalled the words from each of the equivalence classes in groups, listing one class before moving onto the next. Thus, clustering was indeed observed using arbitrary classes and therefore is a fundamental feature of categorization of any kind, not just of natural categories.

Both hierarchical categorization and clustering have thus been shown to exist in the arbitrary classes created via stimulus equivalence procedures, just as they exist in natural language categories. These results provide strong evidence that categories created using stimulus equivalence procedures share properties with categories created naturally, lending credit to the stimulus equivalence explanations of natural language categories. However, there are other features of natural language categories that must be explained, including the transfer of functions from one stimulus to another class-consistent stimulus. This feature of natural language categories has been studied using arbitrary classes, and those studies are reviewed here now.

One feature of natural categories is the transfer of function, that is, certain functions trained to one category member may generalize to other members as well. For example, suppose that an individual is bitten by a spider, and so becomes afraid of all things called “spider”, because it is likely that all things called “spider” will bite, just as the first spider did. Now suppose that the individual encounters a new insect that a friend labels “spider”; the fear will transfer to the new insect as it becomes classified into the category of spiders. Dougher, Augustson, Markham, Greenway, and Wulfert (1994) studied this transfer of functions between stimuli within equivalence classes using fear (measured through galvanic skin response) as the trained function. Using the match-to-sample procedure described previously, participants learned three arbitrary classes of outlined shapes. The shapes within each class were tested for equivalence relations, which were evident. Dougher et al then paired one stimulus from one class with a level of shock. Galvanic skin responses showed that participants began to fear not only the stimulus paired with the shock, but also the other stimuli included in the equivalence class. Stimuli in a different equivalence class taught at the same time as the first class did not elicit similar fear responses. Thus, participants learned to fear all stimuli within a class based on training with only one stimulus within the class, and did not fear other stimuli not included in the class. In a second experiment, Dougher et al showed that by extinguishing fear in one class member, extinction developed to all class members. Further, by retraining fear to one class member, fear reemerged in all other class members as well. These results support the transfer of function as a trained characteristic of equivalence classes, just as it exists for natural language classes.

Dymond and Rehfeldt (2000) reviewed transfer of function in light of the class formation theories that have been proposed thus far, namely equivalence theory and relational frame theory. These authors began first with transfer of function via symmetric relations, and note that this type of transfer "should not be considered genuine derived control" because of the direct training of symmetric contingencies. They point out that with symmetric relations the stimuli have been paired together during training, and thus no feature of this relationship could be considered derived because of this training exposure. The stimuli involved in transitive relations are not directly paired together during any training procedure, and as such may be a better example of derived control. However, the review of equivalence relations suggested that the transfer of function observed in equivalence classes is indeed a derived relationship, and they cite several studies that have obtained positive results in this area. Many different responses have been shown to transfer within classes, including clapping, waving, and pressing a button. Further, many different subject populations have demonstrated transfer of function within classes, from very young children to younger adults. However, there are several instances of failure of transfer of function, potentially because of training structure or verbal instructions given (see Dymond and Rehfeldt, 2000, for a more complete review).

Thus, transfer of function is another feature of natural language classes that has been shown to also be evident in arbitrary class formation. Along with clustering and hierarchical class formation, the existence of these features in arbitrary classes provides support for the study of natural categories using arbitrary classes formed in the laboratory. Any theory that proposes a model of class formation must take these features into account. It must also account for the classification of novel stimuli, those previously not encountered and thus not yet classified. Consideration should be given to just how these novel stimuli may be incorporated into existing classes.

Novelty

A critical aspect of categorization that may appear problematic for equivalence theory is the generalization of response to novel stimuli. Murphy (2002) addresses the novelty issue from a more cognitive perspective, but his points may be considered in a behavioral light as well. He identifies a feature of categories called induction, in which novel stimuli may be brought into a category based on similarity of the novel stimulus and the other class members. Thus, novel stimuli that are more similar to existing categories are easier to categorize than novel stimuli that are not as similar. Murphy notes that similarity of stimuli does not necessarily have to be physical similarity, but may also include functional similarity (telephone and email) and thematic similarity (golf balls and golf clubs). He also cites several studies that show that this inductive power of categories does not seem to have capacity limitations or repetitive limitations; novel stimuli may be put into any category regardless of the size of that category, and may also be categorized in several classes at once. Murphy's overview of the induction feature of categories identifies yet another characteristic of natural categories that must be considered when using arbitrary classes in a laboratory.

An equivalence account for this inductive aspect of categorization comes from a study by Fields, Reeve, Matneja, Varelas, Belanich, Fitzer, and Shamoun (2002). In an earlier study done by Fields, Reeve, Adams, Brown, and Verhave (1996), equivalence classes were used to study the concept of stimulus generalization, in which novel items are classified based on their physical similarity to familiar items that have already been classified. Participants were given two lines, identified as long and short. They were then given several more lines of intermediate lengths and asked to classify them with either the long line or the short line. Results from this study showed that classification can be gradated, such that stimuli most similar to the long line were grouped together, stimuli most similar to the short line were grouped together, and stimuli furthest away from either class were grouped as "neither class." In their later study, these authors looked at how physically similar stimuli had to be in order to still be classified together into one group. Fields et al also proposed that certain variables such as the training design and the number of different types of categories trained would be necessary for generalized categorization. Fields et al created six gradated stimulus classes, each with two distinctive images that mark the "ends" of the class. For example, one class consisted of a photo of a female face at one "end" (known as the *anchor*), and a photo of a male face at the other "end" (known as the *base*); another example used a truck and a car. Within each class, the other class members were combined pictures of the two end stimuli, creating a spectrum of images moving from one endpoint to the other. In the male/female example, the male and female faces were combined in different ways to create some faces that were closer to the female face but with increasingly masculine features, until the face began to more closely resemble the male face with a few female features. The stimulus in the center of the spectrum was named the *midpoint*. Fields et al taught these classes to 36 younger adults using four different match-to-sample training designs that included (a) one or three class members as samples, (b) one or three class members as comparisons, and (c) stimuli from either one or two classes. The testing procedure was designed to see if, after learning to categorize the training stimuli via one of the training designs, individuals would create novel categories using stimuli not seen during training. Results showed that participants given the training with only one class member as a sample and one class member as a comparison in one class were the only participants that did not definitively show generalized categorization in the testing phase. Participants in the other three training groups all showed some degree of categorization with the novel stimuli, with the strongest performance coming from the group given three class members as samples with two different classes. Overall, the results from this study and the previous one support the idea that novel stimuli will be categorized by physical similarity, and that classification will be gradated based on the physical similarity of the class members.

Fields et al used stimulus generalization to explain the inclusion of novel stimuli into existing classes. Their study demonstrated that novel stimuli that are physically similar to familiar stimuli are likely to be grouped into the same class. One can see this kind of pattern in several example classes. For example, given the class "birds", one structural component could be "has wings." All novel stimuli encountered that have wings could be structurally classified with other birds. However, it is easy to see how simply classifying all things by structure is not necessarily accurate; airplanes have wings as well. Further, the wings of a penguin do not physically resemble the wings of a robin, making it hard to argue that physical similarity puts these two items into a single class. This type of classification poses a problem for the stimulus generalization aspect of novel classification, which relies on the physical similarity between class members. Stimuli may also be classified by functional ability. In this case, airplane may be grouped into the "transportation" class, as it is a form of transportation but in no way resembles a bus.

It is also possible to conceive of some things that may be classified as "birds" or "transportation" that do not necessarily meet the first definition. For example, a penguin has wings but cannot fly, so it may not be as representative of the class "birds", even though it meets the structural definition. However, penguins are still classified into the category "birds" with relatively little trouble. Rosch and Mervis (1975) propose that this is possible due to the family resemblance within the category of "birds". Although penguins may not physically resemble a more prototypical bird (a robin) in many ways, they do share one feature; they both have wings. Rosch and Mervis argue that sharing one feature is enough to relate two stimuli into the same class. With family resemblance, it is possible for two class members to share no common features, as long as each has a feature that is associated with to the class. However, the stimuli that have the most features in common with other stimuli within the class might become better examples of the class, a characteristic known as "typicality". Typicality effects are demonstrated when classes can be organized from most representative to least representative; those class members with more features common to all of the stimuli within the class are rated as more representative than others with fewer common features.

Rosch and Mervis investigated typicality and family resemblance in their 1975 study. Participants were given an item from one of six categories and asked to list all of the features that came to mind when thinking of that item. Those features listed for items from the same category were considered the "more typical" features of items of that particular category, features that the prototypical item of the category might have. By correlating the number of common features with a previously determined measure of resemblance, Rosch and Mervis verified that the items rated most typical contained more category-related features. In a second experiment, participants were given an item and asked to list all of the possible categories under which the item might fall. Results from this experiment showed that the more typical features of a category an item had, the less likely it was to be named as a member of another, different category. Thus, those items with greater numbers of typical features are easier to be classified and less likely to be put into other categories. Thus, regardless of familiarity with an item, participants are still more likely to group items together by relevant features. Those items with more category-relevant features are considered closer to the "prototype" of the category.

Galizio, Stewart, and Pilgrim (2004) attempted to provide an equivalence analysis of family resemblance classes. In their first experiment, eighteen participants were exposed to several arbitrary stimuli via a match-to-sample procedure. Stimuli used in this experiment were grouped into three classes by experimenters, with eight stimuli in each class (identified as stimuli B-I). Each class had four "relevant" features that were class-consistent and several "irrelevant" features that were not class consistent. Stimuli within a class could have between one and four relevant features. For example, one of the relevant features was a printed design in the center of the stimulus; each class had a different center design, but not all of the stimuli within a class had to have the center design to be deemed a class member. Stimuli without the center design could be classified using one of the other relevant features (center design, appendage design, and base shape). Irrelevant features consisted of shape of the stimulus, figure outline thickness, and position of the base figure. These features were not class-consistent, and so varied across all the classes equally. The training stimuli consisted of three stimulus that included all four relevant features, six stimuli that included three of the four relevant features in combination, three stimuli that included two of the four relevant features in combination, and twelve stimuli that included one of each relevant feature. This was done to measure typicality effects based on number of features; it was predicted that those stimuli with the greatest number of class-consistent features would be considered the prototype of the class. Each participant was given match-to-sample training, in which the sample was always a nonsense trigram (A) and the comparisons were the stimuli mentioned above (B-I). Baseline training continued until the participant had reached an accuracy criterion of 90%. Participants were then exposed to non-reinforced trials to test for symmetry, transitivity, and equivalence relationships, as well as the classification of stimuli not yet seen during training. For example, a symmetric trial tested for BA relations, and a transitive trial tested for CB relations. Trials involving novel stimuli (those not seen during training) were also given. Novel stimuli (J-Z) consisted of relevant features that had all been seen during training trials, but several irrelevant features that were not included in training. Novel probe trials consisted of nonsense trigrams (A) as samples and novel stimuli (J-Z) as comparisons. These non-reinforced trials were intermixed with training trials to ensure that participants had maintained previous accuracy levels during the testing phases. At the end of the equivalence testing procedure, participants were given note cards with pictures of the abstract training stimuli printed on them and asked to sort the cards into categories in whatever way they felt was best, and to then sort each category "from most representative picture to least representative picture". Results from this experiment showed that all participants demonstrated the formation of symmetry and equivalence relationships between the stimuli. Further, participants were able to correctly classify novel stimuli based on the class-consistent features seen during training. Participants also responded fastest to trials that included the stimuli with all four of the class-consistent features and most often rated these stimuli as "most representative"; they responded slowest to trials that included the stimuli with only one of the class-consistent features and most often rated these stimuli as "least representative" of the category. These results show that novel stimuli can be classified into arbitrary classes as they are with natural categories, and typicality effects can be demonstrated using arbitrary classes and equivalence class procedures. However, the design of the stimuli left the results open to discussion. Although each "relevant" feature had three possible designs (for example, there were three different insert designs, one for each arbitrary category), the "irrelevant" features had more than three designs (for example, eight stimulus shapes were used in this experiment). Thus, when instructed to sort the stimuli into three categories, participants may have immediately disregarded the "irrelevant" features because there were simply too many options for only three categories. The "relevant" features may have been more salient for sorting the stimuli into three groups. To control for this scenario, Galizio et al designed a second experiment with a new set of stimuli.

Their second experiment used the same basic stimulus design as in Experiment 1, with each stimulus having several "relevant" and "irrelevant" features. However, instead of having several different variations for the "irrelevant" stimuli, each "irrelevant" feature had only three variations just as the "relevant" features did. Thus, all of the features presented in these stimuli could exist in one of three forms; only the reinforcement given during training trials allowed for discrimination of the class-consistent features. Participants were also given a pre-sort task before beginning the match-to-sample procedure that was very similar to the post-sort task described for Experiment 1, in which participants were given cards to sort into three categories. Giving a sort task before beginning training provided a measure of comparison against the post-sort measure and allowed the experimenters to determine if any of the "relevant" stimulus features were more salient than any of the "irrelevant" features. After completing the pre-sort task, participants were exposed to the same training, testing, and post-sort procedure as described previously. Results from this experiment showed that all twelve of the participants demonstrated equivalence class formation and the classification of novel stimuli based on class-consistent features. Further results replicated the results of the first experiment in that participants responded faster during trials that included the stimuli with four class-consistent features and were more likely to sort those stimuli as "most representative" of the category. Participants also responded slowest to those trials that included the stimuli with only one class-consistent feature and were more likely to sort those stimuli as "least representative" of the category. These results showed that equivalence classes can be used to study the classification of novel stimuli, one of the first behavioral examples of family resemblance classes. Results also supported predictions made about typicality effects within created equivalence classes, and showed that typicality effects do not exist only in natural classes, but also in arbitrarily created classes. The number of common features a stimulus has does impact how quickly it is classified and how well it exemplifies the class members. Galizio et al also brought together many of the ideas in classification literature that had yet to be unified in one study. Trained equivalence classes with family resemblance characteristics did show properties generally seen in natural language categories.

The experiments described above demonstrate that arbitrary classes do take on some of the characteristics of natural language classes, such as a generalized hierarchy and the transfer of functions and typicality effects within a class. These experiments have also demonstrated that stimulus equivalence procedures have been useful in creating arbitrary classes in a laboratory setting. It has also been shown that arbitrary classes are useful in studying natural language classes and how they are formed. However, the beginning of this paper introduced the idea of age-related changes in category formation and some behavioral approaches that may be useful for learning more about these changes, and more must be said about this area of research. Now that arbitrary classes and equivalence class procedures have been established as a productive way of learning about natural language class formation, it is possible to turn to age-related research in this field.

Aging

Many studies have been done in the area of human aging, not the least of which have included focus on declines that may occur. As the body and mind grow older, both cognitive abilities and physical abilities show a trend of slowing. Cognitive age differences have been shown with memory, attentional span, and language ability (see McDowd & Shaw, 2001 for a review). It is conceivable that declines may also occur in category formation as well, or in ability to classify novel stimuli into previously existing categories. One can only begin to imagine the impact this may have on daily life, especially in light of rapidly changing medical situations that may result in new diagnoses or altered living environments. Physical age differences have been demonstrated in walking speed, activity endurance, and reaction times. It is reaction time differences that may most impact the studies of equivalence classes and typicality effects, as measures of reaction time are often used in the study of equivalence class formation and typicality measurements.

Baron, Menich, and Perone (1983) compared reaction times in older and younger adult men using match-to-sample procedures. Participants were asked to respond to two stimuli in a chain procedure such that responses to the second stimulus were coded as either "correct" or "incorrect." Although initially participants were given as much time as necessary to respond to the second stimulus, they were encouraged to respond as quickly as possible and intervals between the stimuli were progressively shortened throughout the experiment. The testing took place in two phases to compare the impact of practice on reaction time; each phase consisted of an untimed phase followed by a training phase in which times were progressively shortened. Results showed that both age groups committed more errors as the time interval was shortened, and both groups benefited from more practice. However, the gap in response times between the two age groups that was initially found before training was still evident after training and practice; thus, although both groups improved with practice, it was not enough improvement to equate the response times of the two groups. The authors note a distinct overlap in the response time distributions between the two groups, mentioning that although the average response times were quite different between the two groups, the overlap between the distributions could indicate less of a real difference between the older and younger response times. They suggest that perhaps the physical response impacted the reaction time differences, rather than a slowing of cognitive abilities with age. Regardless of the difference between the groups, these results advocate for rehearsal times to ensure the least amount of difference between older and younger adult response times in a match-to-sample procedure.

Baron and Journey (1989) designed a study to determine if response topography played a role in age-related response time differences. Older and younger adults were given multiple comparisons at once and told to respond to the correct comparison as quickly as possible. Participants were given one, two, or four comparisons at one time and asked to locate and respond to the correct comparison. Responses were given either vocally or by a joystick in an effort to determine if response topography would impact reaction time. Results showed that all participants responded faster with the joystick than vocally, and all participants reacted more slowly when given more comparisons. Overall, older adults were slower to respond than younger adults regardless of the response required, but the difference did not increase as the number of comparisons increased or as the topography of the response changed. This study suggests that older adults are no slower in processing information than younger adults, but are simply slower to physically move to react to that information. The results of these two studies indicate that although reaction times may differ between older and younger adults in match-to-sample procedures and comparison procedures, these differences may be attributed to physical declines and not cognitive declines.

Very little research has been done with older adults and class formation, and conclusions from the few studies done have been conflicting. It is important to understand how class formation abilities may change with age, especially in light of recent discoveries about the onset of age-related declines. Hess and Slaughter (1986) looked at older and younger adults abilities to categorize stimuli around a prototypical stimulus. The hypothesis predicted concept learning to be an automatic process rather than a higher-level processing skill, and thus not a process that would be impacted severely by aging. Participants were shown a series of stimuli that varied across three features (color, shape, and number printed in the center), of which only two (shape and number) were class-consistent. Participants were given trials to learn which features were class-consistent until they performed 20 errorless trials in a row. Following the trials, participants were asked to write down the rule they used to classify the objects shown. Participants were scored on the number of errors made, the latest trial that resulted in an error, and ability to correctly identify the classification rule. Results showed that older adult participants made more total errors and had their last error trial later than did younger adults. Only 52% of the older adults were able to correctly identify the classification rule compared with 96% of younger adults. The authors conclude that categorization abilities are more impacted by age than previously thought. This could mean that categorization abilities are not as automatic as initially predicted by Hess and Slaughter, and thus potentially more impacted by age-related declines in cognitive abilities rather than just physical abilities.

Johnson, Hermann, and Bonilla (1995) also looked at typicality effects in older adults, comparing older adults diagnosed with Alzheimer's disease to normally functioning older adults. The experimenters created a list of concepts with three examples of each concept; one example of a highly typical item, one example from a mildly typical item, and one example of a less-typical item. For example, the concept of "cloth" was listed, followed by the three examples of "cotton", "satin", and "burlap." Participants were shown an example word and asked, "Is this a type of [concept name]?"

Experimenters immediately scored responses from participants, given in "yes-no" format. Results showed that those items deemed highly typical of a concept were identified and categorized faster than those items deemed less typical. Normally functioning older adults performed significantly faster than those adults with Alzheimer's disease, suggesting that the ability to categorize may be dramatically impacted by cognitive declines evident with Alzheimer's disease. However, it is not clear whether these declines are in the ability to classify typical items or the ability to remember previously learned natural categories. If the declines are related to memory, older adults may still be able to correctly learn new categories and demonstrate typicality within those categories. It is also unclear whether the declines demonstrated by the older adults with Alzheimer's disease are specific to typicality within categories or simply an overall slowing of mental processing. Johnson et al conclude that classification abilities may be a marker for progression of declines with age and propose that categorization abilities be used as a description of capabilities for those individuals with Alzheimer's disease.

These results suggest that older adults may perform differently than younger adults with categorization tasks, specifically with those tasks associated with typicality effects. The classifications used in the two studies mentioned previously use natural categories that participants had already formed prior to the experiment. These categories were most likely created before any age-related changes had begun to occur, and so it can be difficult to determine how much can be attributed to age declines as opposed to poor category formation initially. A solution to this problem is arbitrary class formation, such as the classes formed during stimulus equivalence procedures. By creating new categories, differences in ability may be controlled more so than those differences shown with natural categories. New categories also allow for the distinction between the ability to form categories and the ability to remember previously learned categories, thus distinguishing between declines in category formation abilities and declines in memory abilities.

Wilson and Milan (1995) compared older and younger adults using arbitrary stimuli and a match-to-sample procedure. Participants were initially taught three relationships, relating each of the printed Greek letters Λ, Ω, and Ξ (referred to as A stimuli) to a different abstract shape (B stimuli). These trials continued until participants could correctly complete 9 of 10 trials in a row. Participants were then given Greek letters as the samples and a separate set of Greek letters (Π, Γ, and Δ, referred to as C stimuli) as comparisons until 9 of 10 trials were accurate. Testing trials were then done to determine if participants could correctly demonstrate BC and CB relations. These trials were completed without feedback. If the 90% accuracy criterion had not been reached upon completion of one trial block of 120 trials, participants were given another block of equivalence trials. If the accuracy criterion had not been reached after two trials, the authors concluded that equivalence classes had not formed. Wilson and Milan found that older adults needed significantly more trials to reach the baseline criteria during training and responded significantly slower to all trials than did younger adults. Younger adults demonstrated equivalence class formation more often than did older adults. However, for those adults that did form the equivalence classes, they did so within the same number of trials blocks as the younger adults. Thus, while older adults were slower to respond during trials and less likely to form the classes as a whole, those older adults that did form classes were just as accurate as younger adults. These results seem to emphasize the idea that if given enough experience and time, older adults may be as likely to form classes as younger adults. Participants in this study were not given any training trials after completing the training structure of the experiment. This may have negatively impacted the results of the study, as participants may have forgotten the trained relations during the time between training and testing phases. Often, it is more beneficial to intermix training trials with novel trials during testing phases to ensure that participants are maintaining baseline levels of equivalence class formation. Without these measures in this study, it is impossible to know whether participants continued to maintain those classes taught during the training phase. It is possible that older adults would have performed just as well if reinforced training trials had been included in testing phases, or if more than two equivalence trial blocks had been given. However, the results of this study do show that some older adults are capable of forming basic equivalence classes just as well as younger adults are.

In a second study, Perez-Gonzales and Moreno-Sierra (1999) taught seven participants two equivalence classes using the match-to-sample procedure. Participants included a thirteen-year-old female, a twenty-one year-old female, a forty-four year-old female, a fifty-three year-old male, a sixty-six year-old female, a sixty-seven year-old male, a seventy year-old male, and a seventy-four year-old female. Thus, these experimenters compared many individuals of varying age, rather than two distinct groups of younger and older adults. Participants were taught initial relationships between shapes, which were labeled using the letter-number system described previously. Experimenters trained A1 to B1 and A2 to B2 in a series of trials. Participants were then taught B1 to C1 and B2 to C2 relations; upon mastering an accuracy criterion, participants were moved into a testing phase. Here, participants were given novel trials to test for symmetry, transitivity, and combined equivalence intermixed with previously trained baseline trials. The baseline trials were included during testing phases to ensure that participants maintained the accuracy levels achieved during the training phases. Results from the experiment showed that older adults (those over age fifty-one) made more errors in both the training and the testing phases than did younger adults (those under age fifty-one). Thus, it took more trials for the older adults to reach the baseline accuracy criterion than it did for younger adults to reach the same criterion, and older adults made more errors in the testing phase than did younger adults. These results seem to suggest that older adults require more training than younger adults initially, as well as being less able to actually form equivalence classes. This conflicts with the results found by Wilson and Milan, possibly pointing to a difference in the training structure that may impact the results. Perez-Gonzalez and Moreno-Sierra completed a second experiment with two older adults to address the question involving the number and structure of baseline training trials. In the second experiment, participants were given correction procedures to train specific relations more accurately; this baseline training structure did not significantly impact the accuracy of testing trials. The experimenters concluded that older adults are less able to form equivalence classes, possibly because of age-related declines in cognitive ability.

The results of these two experiments highlight the conflicting nature of equivalence class formation in older adults. While Wilson and Milan found that some older adults were not capable of learning new arbitrary classes within a certain number of trial blocks, Perez-Gonzalez and Moreno-Sierra found that more training might increase the likelihood that older adults correctly form arbitrary classes. Both studies conclude that older adults seem to have a much harder time forming the relations necessary to form new categories. Further, distinguishing between cognitive declines and physical declines is more difficult when using measures regarding reaction times. There is a need to draw these concepts together into a single experiment to determine just how class formation may break down or change as individuals age. Furthermore, little research has addressed the idea of transfer of function within arbitrary class formation, specifically in the area of older adult research. It is possible that older adults may be slower to form equivalence classes, but that once formed, older adults will show typicality effects and transfer of function just as well as younger adults. The need for research to answer these questions is apparent, and that is what this study attempted to achieve.

The first goal of this study was to observe differences between older and younger adults when forming equivalence classes; specifically, whether older adults would commit more errors during baseline training, whether overall response times would be longer for older adults than younger adults, and whether older adults could form complete classes from the training. The second goal was to determine whether older adults would show typicality effects within the created classes as well as younger adults have done in the past. The third and final goal of this study was to observe transformation of function differences between older and younger adults, and to determine whether those transformed relationships were linked to the typicality effects that were acquired during training and probe testing. It was hypothesized that both older and younger adults would form complete equivalence classes, though older adults would make more errors and require longer training sessions than younger adults. Further, older adults would show typicality effects and transformation of functions as well as younger adults, and the strength of those transferred functions would be related to the observed typicality effects for both age groups.

METHOD

Participants

Young adult participants were 11 undergraduate students from the University of North Carolina Wilmington that participated for credit to fulfill requirements for an introductory psychology class. The average age of this group was 18.6 years, ranging between 18.0 and 20.5 years of age. Four of the participants failed to meet the baseline training criteria and were subsequently dropped from the study, and so data from 7 younger adult participants are included in the analyses here. Individual younger adult participant ages were as follows: Y-1 was 18.2, Y-2 was 18.3, Y-3 was 19.4, Y-4 was 20.5, Y-5 was 18.1, Y-6 was 18.0, and Y-7 was 18.0.

Ten older adults participated in the experiment as volunteers, and were recruited in several ways. Four participants responded to a posted sign at a local senior center in the Wilmington area, and two more participants were recruited by word of mouth from these participants. The other four participants were recruited as part of a bridge group in Seaford, New York. The average age of these participants was 73.14 years, with a range between 67 and 87. In addition, adult participants also completed a Mini-Mental State Exam (Folstein, Folstein, and McHugh, 1975), which assesses mental awareness using several basic questions and activities that are scored out of 22 points. Questions include items such as the day's date and the participant's current location, and activities include things such as spelling the word 'HOUSE' backwards or remembering items mentioned earlier in the session. The questionnaire was administered at the beginning of the first session by the experimenter in a one-on-one setting. Scores were calculated by adding the number of points each participant received for the questions; scores lower than 17 are considered indicative of cognitive impairment. Participants in this study scored an average of 19.8, with scores ranging from 17 to 22; no participants were excluded from this study based on their scores on this questionnaire. Participants volunteered for a small reward averaging $25 in cost, and included things such as donations to a charity fund and gift certificates to local restaurants. Three of the adult participants failed to reach baseline training criteria in three sessions and further expressed an interest in no longer continuing the study; they were subsequently dropped from the analyses, leaving seven adult participants for analysis. Individual older adult participant ages were as follows: O-1 was 67.6, O-2 was 68.3, O-3 was 87.0, O-4 was 77.3, O-5 was 74.7, O-6 was 69.3, and O-7 was 67.9.

Apparatus and stimuli

The experiment was conducted using a Macintosh color computer and a match-to-sample program with software developed by Dube (1991).

During the initial training of relations, abstract stimuli served as comparisons with three nonsense trigrams used as samples (See Figure 1). Row A shows the sample trigrams, which were always printed in black against a white background. Comparison stimuli consisted of 8 features, four of which were critical to the formation of classes and four of which were irrelevant. The critical features (base shape, appendage shape, fill pattern, and insert shape) were shared within each class but not between classes.

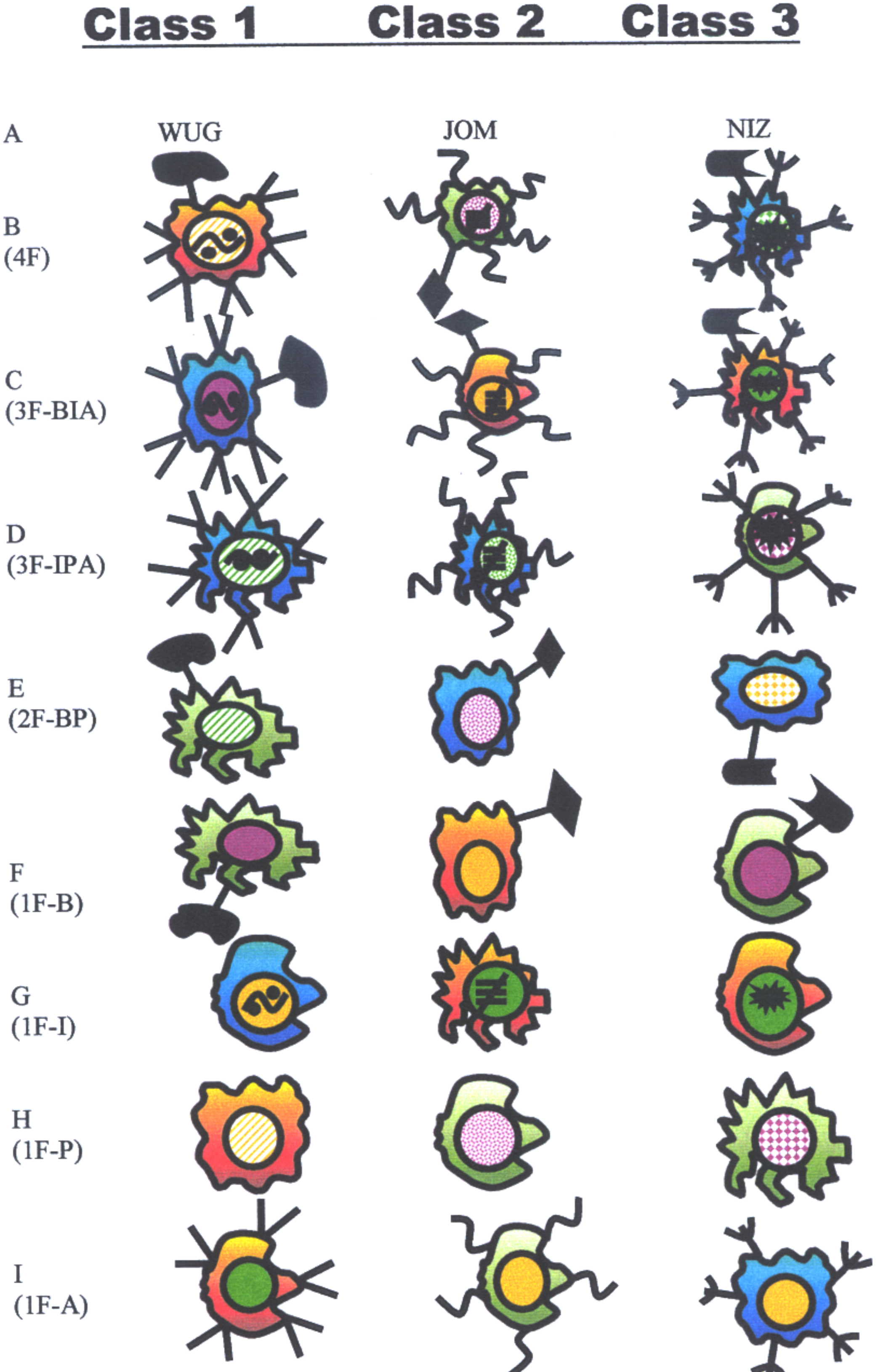


Figure 1. Abstract stimuli used in baseline training, symmetry, and equivalence probes

Irrelevant features (base location, stimulus shape, and stimulus color) were chosen for each stimulus and were distributed randomly across classes. Critical features appeared in combinations of one, two, three, or four relevant features (known as 1F, 2F, 3F, and 4F respectively) on any one stimulus; only some combinations were presented during training, to allow for novel combinations of features to use in equivalence testing. Row B of Figure 1 shows the stimuli with all four of the critical features present (known as the *prototype* stimuli). Rows C and D show the two three-feature combinations, which were used in baseline training. Row E shows the three stimuli used in baseline training with two critical features (fill pattern and insert). Rows F, G, H, and I show stimuli with one critical feature present (base, insert, fill pattern, and appendage respectively). Irrelevant features are also present in these stimuli and are equated across class. For example, the stimuli E1, D2, and C3 all have the same outside shape; stimuli C1, E2, and F3 all have the same base location. All abstract stimuli were presented in color against a white background, as comparisons during baseline training.

To observe the classification of unfamiliar stimuli into the arbitrary classes, novel stimuli were used in the testing phase. Two types of novel stimuli were used; some with equal numbers of features across all stimuli and those with unequal numbers of features across all stimuli. Unequal-feature stimuli were similar to those used in training, but used novel combinations of the critical features, plus new variants of the irrelevant features. Figure 2 gives examples of the abstract stimuli that were used in this phase of the testing. The equal-feature stimuli had new variants of the critical features, but these were distributed randomly across classes. For example, row W shows the equal-feature novel stimuli with the appendage as the class-consistent relevant feature. The other features

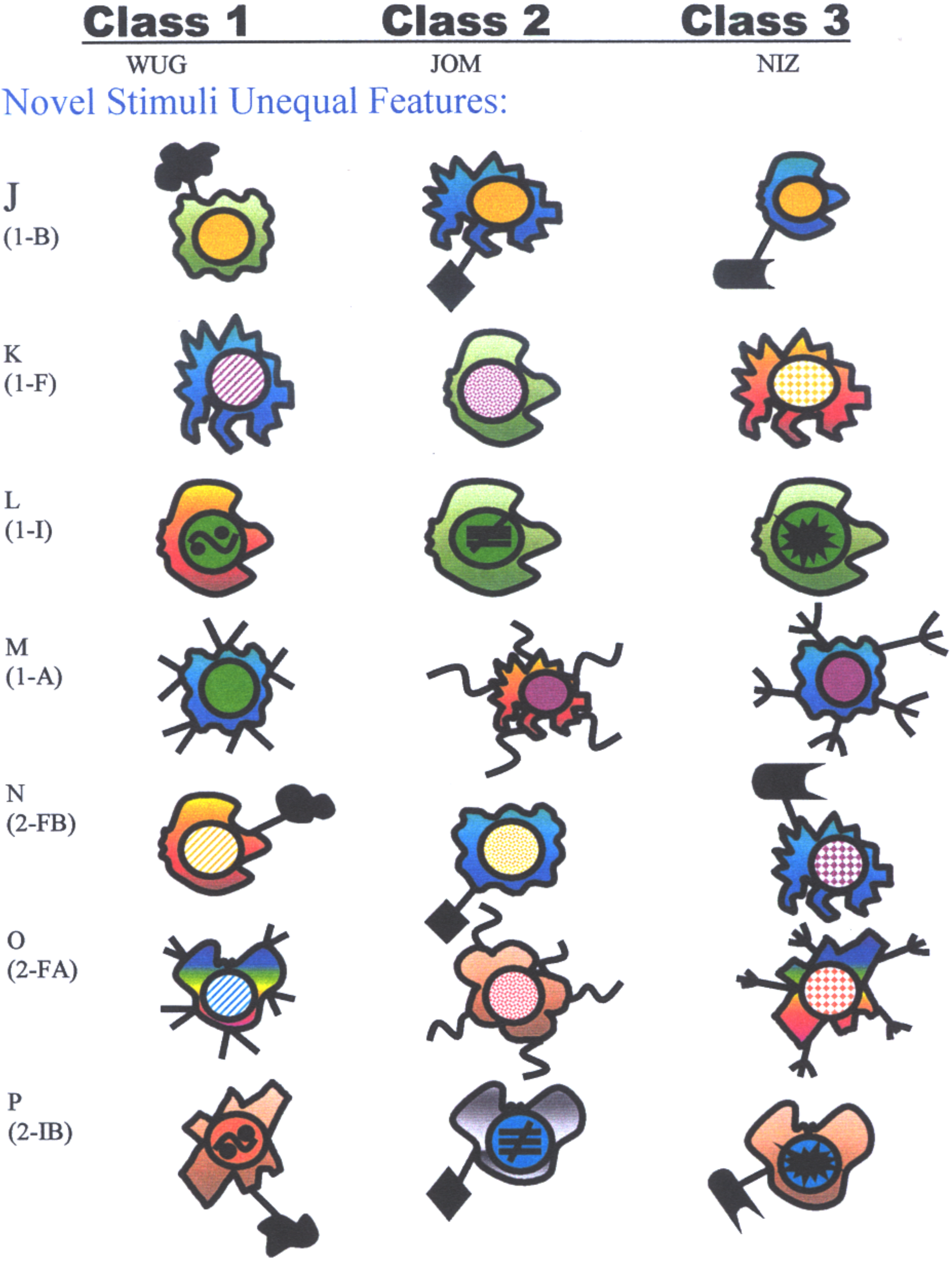
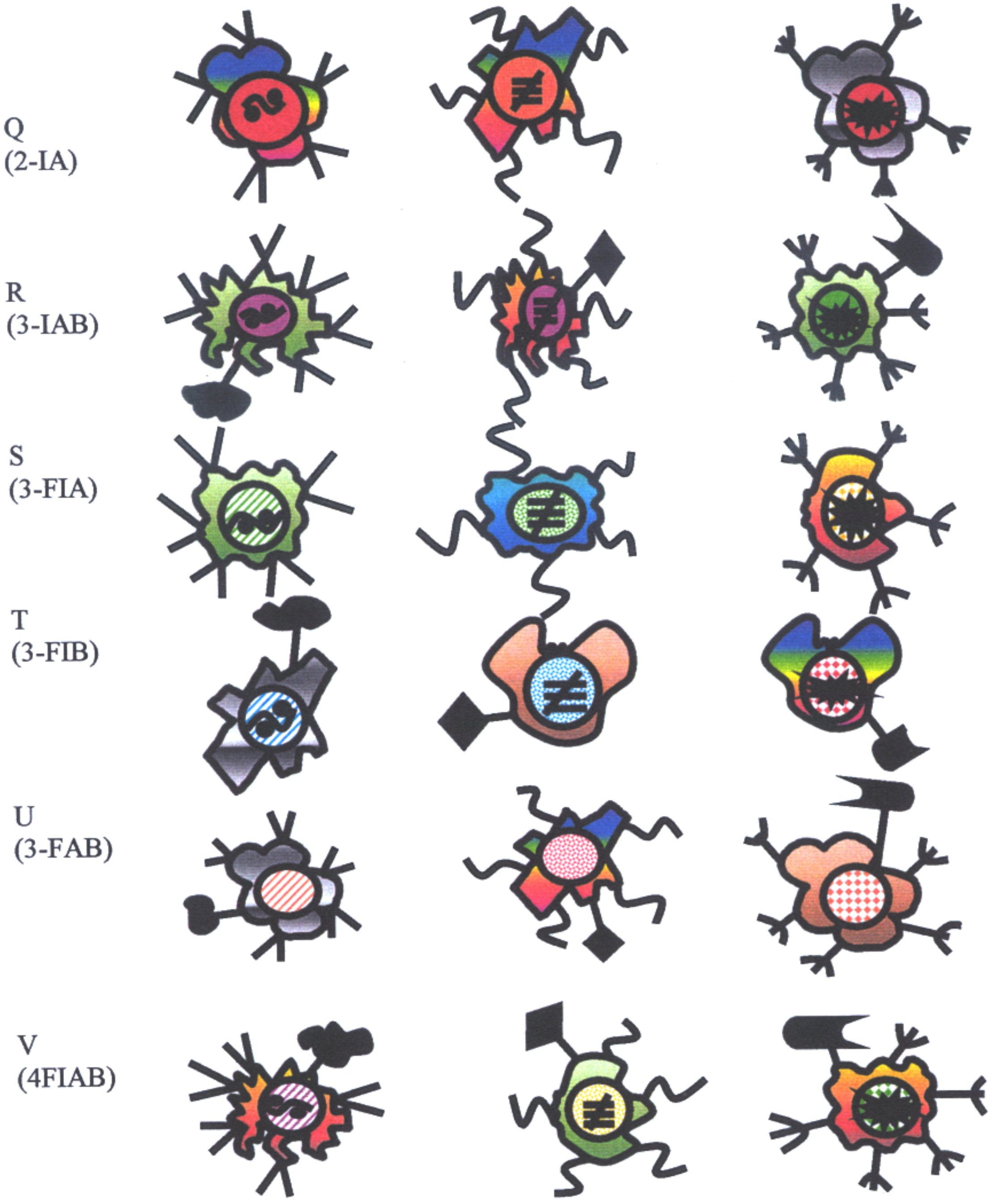


Figure 2. Unequal-feature and equal-feature stimuli used in novel testing probes.



Novel Stimuli Equal Features:

Figure 2. Unequal-feature and equal-feature stimuli used in novel testing probes.

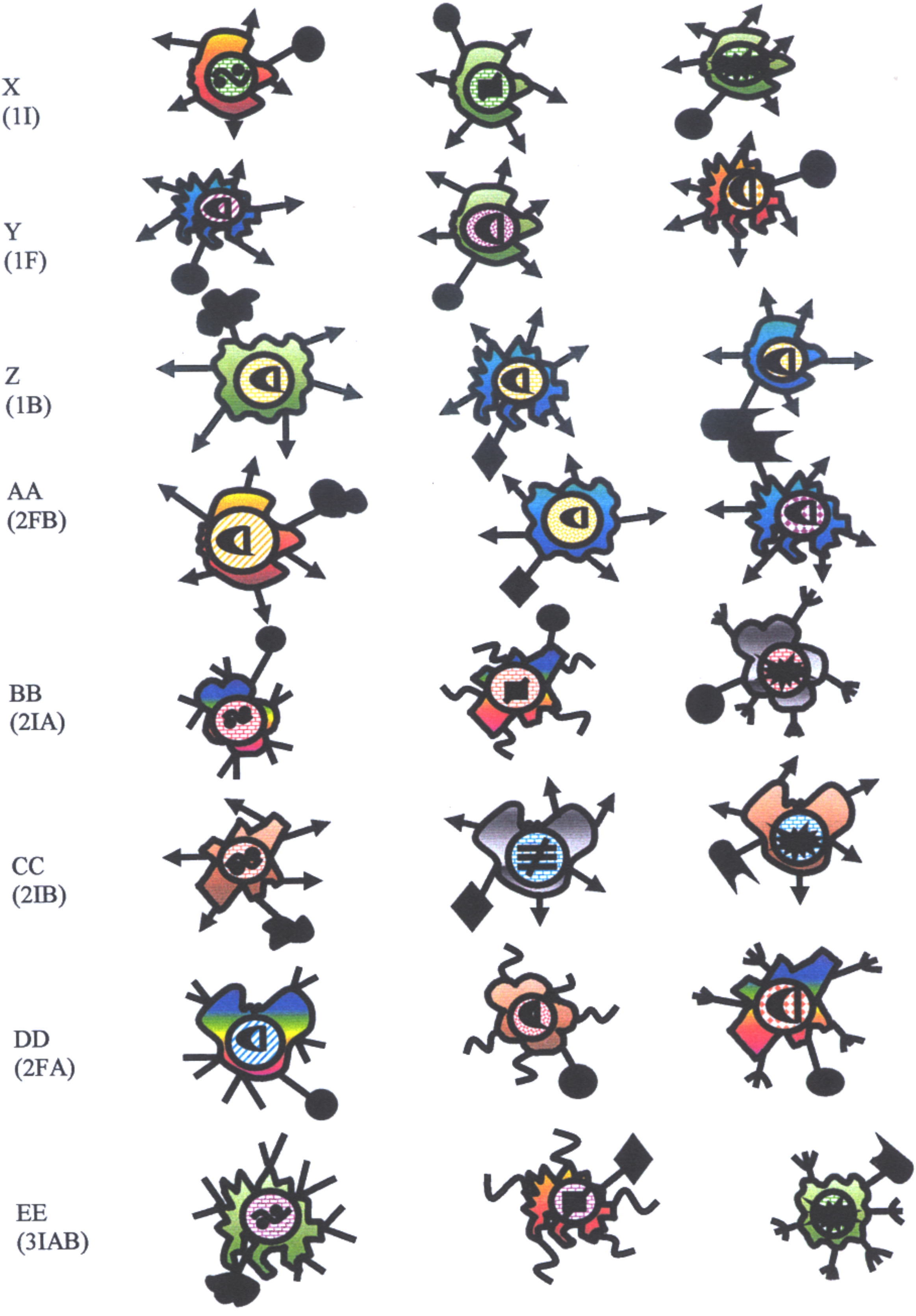


Figure 2. Unequal-feature and equal-feature stimuli used in novel testing probes.

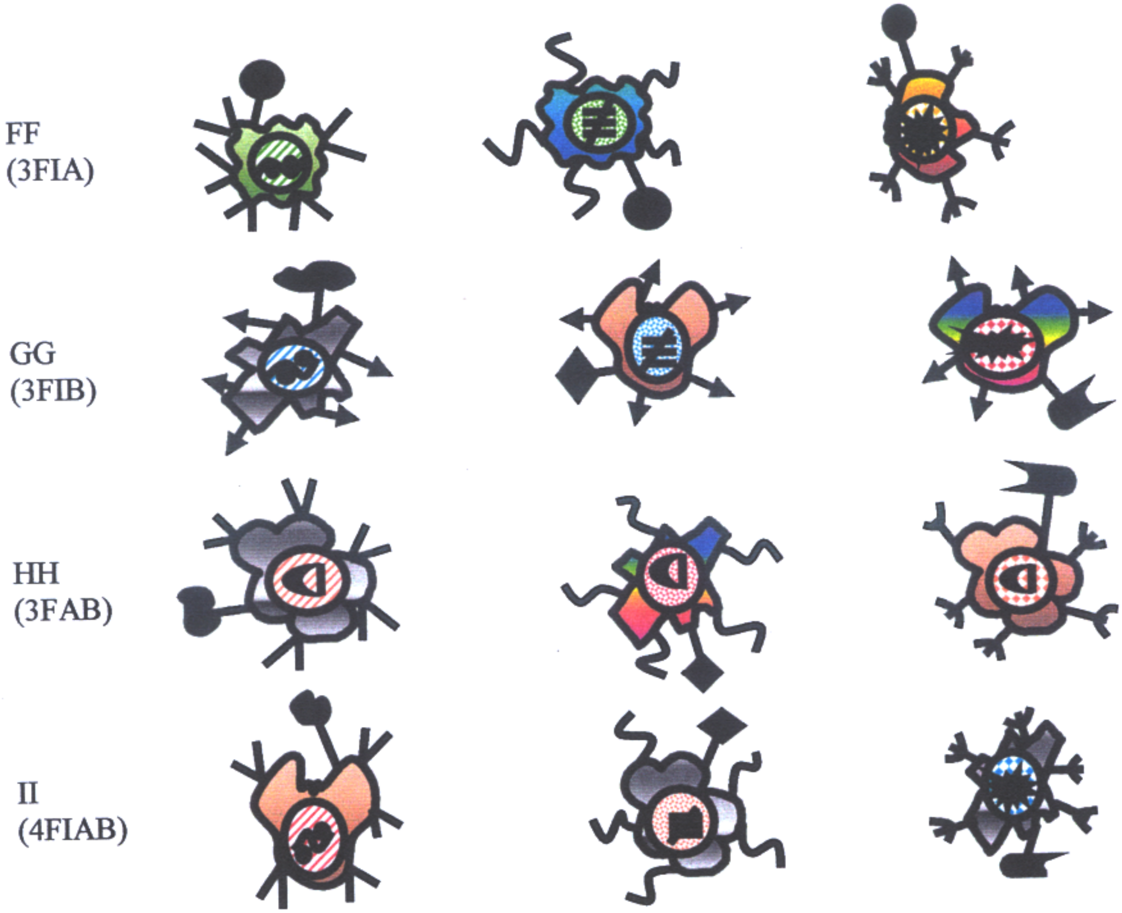


Figure 2. Unequal-feature and equal-feature stimuli used in novel testing probes.

that were relevant during training are still present, but are no longer relevant; base shape (circle), fill pattern (bricks), and insert shape (semi-circle) are the same for each class.

Procedure

Younger adults were tested for 3 one-hour sessions, all completed within one week of the first session. Sessions for older adults ranged from 4 to 8 in number and lasted between 40 and 75 minutes. Session frequency for the older adults was determined by interest level and scheduling between the experimenter and the participant, with at least one session per week. Older adults were told that daily sessions and overall experimental sessions could be terminated at any time, although the generally suggested session length was one hour. Termination was at the participant's request. Any extension over the one-hour suggestion was at the older adult participant's request.

The procedure for both groups consisted of five phases: a pre-sorting task, baseline training, novel probe testing, a transfer of function task, and a post-sorting task. Baseline training, novel probe testing, and the transfer of function task took place on the computer, while the pre-sort and post-sort tasks used laminated cards.

Pre-sort

Participants completed a pre-sort task using laminated cards with pictures of abstract stimuli printed on them and were asked to sort the cards into three groups and place them onto blank sheets of paper labeled "wug", "niz", or "jom" in whatever way they deemed appropriate. Stimuli used in the pre-sort were the abstract stimuli used in the baseline training phase of the procedure. There was no time length given for the pre-sort task.

Computerized task

Upon completion of the pre-sort task, participants were given the computerized training task, which consisted of an initial block of training trials followed by blocks of probe testing trials. Participants were read the following instructions:

> "The experiment you are about to participate in is a learning task; it is not a psychological test. We are interested in studying aspects of the learning process that are common to all people. More specifically, we are interested in the number of points that you can earn in completing the task. The way in which you earn points works like this. At the start of this experiment, a stimulus figure will appear in the center of the screen. The stimulus in the center of the screen is always the example stimulus. After you have looked at the example stimulus, use the mouse to position the cursor on it and click. Other figures will appear in the corners of the screen; these are your choices. Your goal is to pick the stimulus that goes with the stimulus in the center of the screen. Sometimes, after your choice, colored stars will appear on the screen, accompanied by music. Each time this happens, one point will be added to your total score. However other times your choice will be followed by a buzzer, and this will subtract one point from your total score. Also, sometimes there will be no feedback (no stars and no buzzer) to let you know if you have made a 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 the experimenter now."

Training was organized into blocks of 24 trials. Each trial began with the presentation of a nonsense trigram word as a sample. When the participant moved the cursor to the center of the computer screen and then clicked on it, the three comparison abstract stimuli appeared in three of the four corners of the screen (Figure 1). After either the music or the buzzer sounded the screen went blank for 1.5 seconds and the next trial began. Training blocks were composed of the shapes shown in Figure 1, with row A always as the sample stimulus. Comparison stimuli consisted of the shapes in rows B-I, with one row presented as the three comparison choices in any one trial. Trials were arranged so that every relevant feature was presented the same number of times, so that participants were not exposed to one relevant feature more than any other relevant feature. Presentation of the samples was evenly distributed across trials, with no more than three trials with the same sample presented consecutively. Similarly, presentation of the comparison stimuli was evenly distributed across trials, so that no row of comparisons appeared more than twice in a row. Further, correct comparisons were evenly placed in the corners of the screen, with no correct comparison appearing in the same corner more than three times in a row. Finally, trials were organized such that for every block of 24 trials, 50% (12) of the trials consisted of one-feature stimuli, 12.5% (3) consisted of two-feature and four-feature stimuli, and 25% (6) consisted of three-feature stimuli. This balance of trials was used throughout all of the blocks of testing trials.

Upon completion of two training blocks at 90% accuracy (22 out of 24 trials) with every trial receiving feedback (100% reinforcement), participants moved into a reduced reinforcement period of 75% reinforcement (6 unreinforced trials out of 24). This was done to lessen the likelihood that participants would be able to discriminate between training trial blocks and novel probe testing blocks, which were not reinforced. The unreinforced trials consisted of the same stimuli as the original baseline trials and were interspersed throughout a block of 24 such that each unreinforced trial was followed by a reinforced trial. Upon completion of two75% reinforcement training blocks at 90% accuracy followed by two 50% reinforcement training blocks at 90% accuracy, participants moved onto novel probe testing blocks. Trial blocks consisting of 50% reinforcement density (12 unreinforced trials for every block of 24) were also identical to the original baseline trials.

Probe tests consisted of four different types of trials, none of which were reinforcement during any testing phase. All probe trial blocks consist of probe trials intermixed with baseline trials, still at a reduced rate of reinforcement. This was done to ensure the accuracy levels of baseline trials are maintained during probe testing and to ensure that participants cannot discriminate the unreinforced probe trials from the unreinforced baseline trials. Symmetry trials used the same stimuli presented during the baseline training trials, but in reverse; sample stimuli were the abstract shapes in Rows B-I, and comparison stimuli were the nonsense trigrams from Row A of Figure 1. Symmetry trial blocks consisted of two blocks of 24 trials, twelve of which were symmetry probes and 12 of which were trials shown during baseline training, with reinforcement. Equivalence probes also use the same stimuli presented during the baseline training trials, but abstract stimuli from rows B-I in Figure 1 as both samples and comparisons. Sample stimuli and comparison stimuli were chosen such that they did not share any common relevant features; for example, if stimulus C2 served as the sample stimulus, comparison stimuli would be stimuli from row H. This was done to ensure that participants were not simply matching features on the sample and comparison stimuli, but were actually demonstrating equivalence class formation. Equivalence probes were also intermixed with baseline training trials, with 12 equivalence probes and 12 reinforced baseline trials in a block of 24 trials. Novel probe trials were those in which the sample stimulus was one of the nonsense trigram stimuli from Row A of Figure 1, and comparison stimuli were one of the rows presented in Figure 2. Novel probe tests were divided into unequal-feature comparison stimuli (rows J-V) and equal-feature comparison stimuli (rows W-II). Blocks of unequal-feature probe trials consisted of 45 trials, 21 of which were unequal-feature probes and 24 of which were baseline trials. Of the 24 baseline trials, 18 were reinforced and six were not. The same structure was used for the equal-feature probe trial blocks. Probe trial blocks were sequenced such that each participant completed a block of unequal-feature novel probes and a block of equal-feature novel probes twice, then completed two blocks of symmetry and two blocks of equivalence probes, followed by the unequal-feature and equal-feature probe blocks again.

Transfer of function

For the transfer of function task, participants were presented with one of the abstract stimuli shown during the baseline training trials (for example, B1) and given these instructions:

> The picture you see here is one that you've become familiar with during the course of this experiment. This picture is a germ, recently identified by doctors, which carries a disease that infects animals. Animals that contract the disease experience flu-like symptoms. However, only 50% of the animals that come into

> contact with this microbe actually contract the disease. Several other germ similar to this one have also been found recently, and you will see these shortly. Your job is to determine the percentage of animals that will get sick after coming into contact with each of these germs. Write the percentage that you feel is most appropriate on the sheet in front of you, making sure to match the number next to the germ with the number on the sheet. Pressing the space bar will bring up the next germ to observe, and there is no time limit on your decisions.

Participants were then presented with thirteen other abstract stimuli from the testing phase and asked to rate on a 1-100 scale how likely it was that animals that came into contact with that stimulus would get the same disease, based on the information given. The thirteen stimuli chosen included a 4F-stimulus and a 2F-stimulus from another category (for example, 2) as well as a 3F-stimulus and a 1F-stimulus from the third category (for example, 3). The other stimuli chosen were from the same category as the example stimulus and included various combinations of both the equal-featured stimuli and the unequal-featured stimuli such that each feature was included at least once alone and in combination with other features. These stimuli were chosen to allow for any transfer of function patterns that may be related to specific features, rather than to class members as a whole. The stimuli chosen for each of the transfer of function tasks are listed in Table 1 using the reference letters from Figure 2. Participants were then given a second stimulus from a different class (for example, H3) and a similar set of instructions involving plants instead of animals. Thirteen more testing stimuli from all three classes and with various numbers of features were presented, and the participant was asked to

Table 1.

Stimuli used to complete the transfer of function task.

	Scenario 1		Scenario 2	
Trained Stimulus	4F WUG	1F(f) NIZ	4F NIZ	1F(f) WUG
4F class-inconsistent	II 2	V2	V2	II2
3F class-inconsistent	U3	FF1	FF1	U3
2F class-inconsistent	AA2	DD2	DD2	AA2
1F class-inconsistent	J3	M1	M1	J3
4F class-consistent	V1	II3	II3	V1
3F class-consistent*	EE1	GG3	GG3	FF1
3F class-consistent	T1	R3	R3	R1
2F class-consistent*	O1	O3	O3	O1
2F class-consistent	AA1	P3	P3	BB1
1F class-consistent (i)	L1	X3	X3	L1
1F class-consistent (a)	W1	W3	W3	W1
1F class-consistent (f)*	Y1	Y3	Y3	Y1
1F class-consistent (b)	J1	J3	J3	J1

**-stimuli with the same relevant feature as the 1F target training stimulus.*

rate how likely it was that each of these stimuli would also make plants sick, on a 1-100 scale. Two scenarios were used in this phase of the study. In Scenario 1, the stimulus trained to animal infection was a 4F-stimulus from created class 1 (stimulus B1 in Figure 1) and the feature trained to plant infection was a 1F stimulus from created class 2 (stimulus H3 in Figure 1). In Scenario 2, the stimulus trained to animal infection was a 1F-stimulus from created class 1 (stimulus H1 in Figure 1), and the feature trained to plant infection was a 4F-stimulus from created class 2 (stimulus B3 in Figure 1). These scenarios were randomly given across participants such that half of the participants received Scenario 1, while the other half received Scenario 2. This counterbalancing was done to ensure that no one category was more prone to transfer of function than another, and to ensure that all categories formed equally well during the equivalence procedures.

Post-sort

After completing all training and novel probe testing sessions on the computer, participants were given a post-sort task. This task was very similar to the pre-sort task and used the cards with pictures of the abstract stimuli seen in training printed on them. Participants were asked to sort the cards into three groups in whatever way they deemed appropriate, and then to sort the cards by group from most representative of the category to least representative of the category.

RESULTS

Baseline trials

The number of training trials required for younger adult participants to reach the acquisition criterion ranged from 241 to 362, with a mean number of 302.86. For older adults, training trials ranged between 311 and 844, with a mean number of 493.86 (see

Table 2 for individual results for both younger and older adults). For the three younger adult participants that did not meet the training accuracy criterion, the mean number of completed trials was 1427.67 trials, with a range between 1289 and 1658 trials. For the three older adult participants that did not reach the training accuracy criterion, the number of trials ranged between 109 and 558, with a mean value of 394.33. Table 2 shows that the older adult participants consistently required more trials than younger adult participants ($t(12)=4.96$, $p<0.001$). Upon reaching an accuracy criterion of 88%, participants completed a series of training trials with reduced reinforcement (first at 75% and then at 50%) during which accuracy had to be maintained. Training trials were then also presented continuously during probe testing blocks, to ensure that participants maintained the accuracy achieved during training trials while completing probe trials. Further analyses were conducted by dividing the training trials into those trials occurring up until the final training criterion (completion of the training trial blocks at reduced reinforcement) was met (*pre-criterion* trials) and all subsequent baseline trials (*post-criterion* trials). Percent correct during pre-criterion and post-criterion trials were determined, as well as speed of responding during each of these phases.

For pre-criterion trials, the first block of training trials was excluded from both accuracy and speed analyses, as these trials were not based on any previous experience with the procedure and so were not relevant to any typicality analyses. Percent correct was calculated as the ratio of correct trials to total number of trials completed. Training trials were also broken down by number of features (1-4) present in comparison stimuli, and percent correct was determined for each of the four types of training trials. Speed scores were calculated by determining the reciprocal of the latency value between the

Table 2.
Number of trials required to meet criterion.

Younger Adults		Older Adults	
Participant	Number of Trials	Participant	Number of trials
Y-1	289	O-1	693
Y-2	265	O-2	844
Y-3	241	O-3	362
Y-4	362	O-4	450
Y-5	337	O-5	354
Y-6	337	O-6	311
Y-7	289	O-7	443
Mean	302.86	Mean	493.86

onset of the comparison stimuli and the selection of a comparison choice (correct trials only), and were also broken down by number of relevant features in addition to being calculated in total.

For the younger adult participants, the mean total percent correct for pre-criterion trials was 67.50%, with scores ranging between 60.83% and 82.70%. For older adults, scores ranged between 51.76% and 77.49% (with a mean score of 67.00%). Figure 3 shows the group average pre-criterion accuracy percentages presented by number of relevant features, which permits evaluation of typicality effects. Percent correct for trials with one-feature comparison stimuli were associated with the lowest level of accuracy, two-feature comparison stimuli were somewhat higher, and three-feature and four-feature comparison stimuli showed the highest levels of accuracy, demonstrating an effect of typicality ($F(3, 48) = 5.3$, $p=0.003$). There were no differences between the age groups ($F(1, 48)=0.1$, $p=0.99$) and no significant interaction ($F(3,48) = 40.02$, $p=0.84$). Table 3 shows the individual participant data for pre-criterion accuracy scores and reflects the typicality effect in all participants.

Speed scores were calculated by determining the reciprocal of the latency value between the onset of the comparison stimuli and the selection of a comparison choice (correct trials only), and were calculated together and also broken down by number of relevant features.

Figure 4 depicts the mean speed scores for the participants for pre-criterion trials, and shows a striking age difference. A 4 x 2 ANOVA (feature number x age group) demonstrates clearly that younger adults were significantly faster than older adult participants ($F(1,48)=122.3$, $p<0.001$). Note that there were no differences in speed

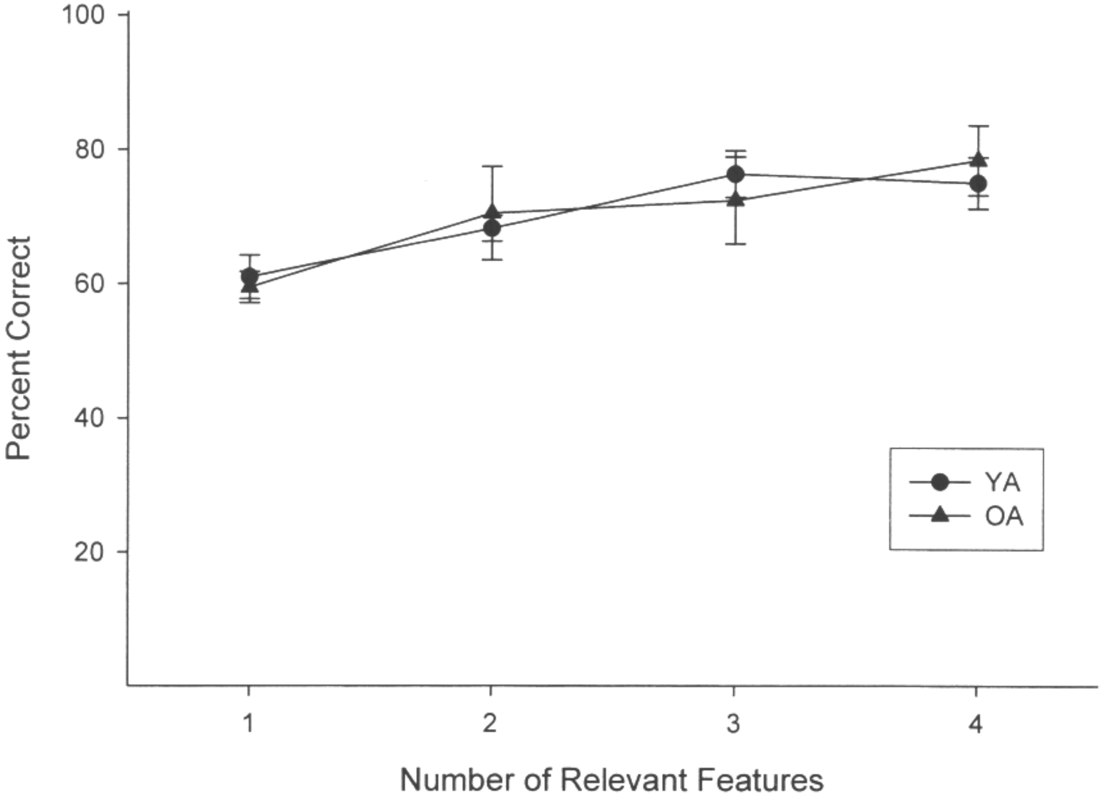


Figure 3. Mean percent correct for pre-criterion trials. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

Table 3.
Percent correct for pre-criterion trials by feature.

	Younger Adult						Older Adult				
Participant	1F	2F	3F	4F	Total	Participant	1F	2F	3F	4F	Total
Y-1	77.93	72.22	93.06	91.67	82.70	O-1	64.55	47.67	86.78	88.51	71.04
Y-2	57.89	66.67	75.76	72.73	65.28	O-2	53.08	84.83	42.86	91.00	63.39
Y-3	61.16	73.33	66.67	66.67	64.73	O-3	61.33	77.78	85.56	82.61	72.10
Y-4	61.88	66.67	82.42	84.44	70.44	O-4	56.64	51.79	75.89	75.00	63.11
Y-5	51.48	71.43	70.24	76.19	61.72	O-5	62.62	80.77	75.93	77.78	70.09
Y-6	54.44	69.05	69.05	61.90	60.83	O-6	67.31	94.87	85.71	84.62	77.49
Y-7	62.07	58.33	77.78	72.22	66.78	O-7	50.70	55.77	54.72	50.00	51.76
Mean	60.98	68.24	76.42	75.12	67.50	Mean	59.46	70.50	72.49	78.50	67.00

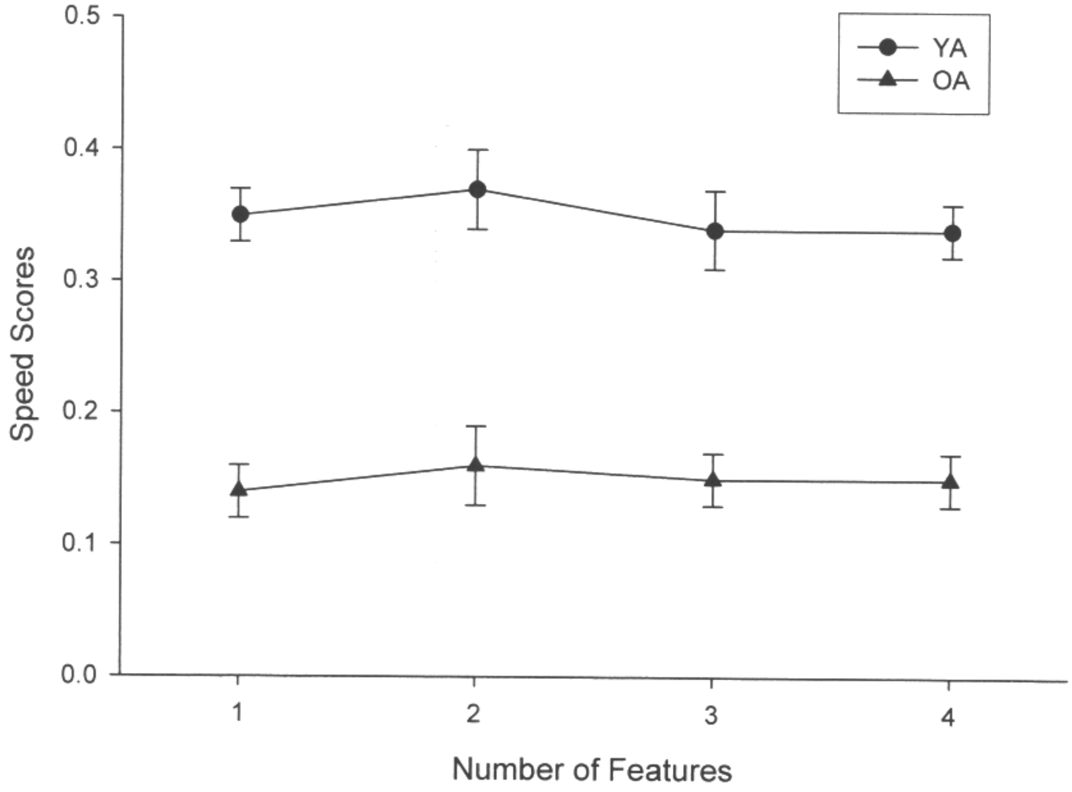


Figure 4. Mean speed scores for pre-criterion trials by number of features. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

as a function of the number of relevant features for either age group (*F*(3,48) = 0.02, *p*=0.88) and that there was no age by feature interaction (*F*(3,48) = 0.01, *p*=0.98).

Average speeds were very similar regardless of the number of features present on the comparison stimuli, averaging approximately 0.35 responses per second for younger adults and 0.15 responses per second for older adults. Individual younger adult data for pre-criterion baseline speed scores are shown in Table 4 and show very little variation in speed across either trial type or participant. Individual data for older adult participants (shown in Table 4) was much more varied, ranging between 0.06 responses per second (participant O-5) and 0.26 responses per second (O-7).

Post-criterion group accuracy percentage data are presented in Figure 5 and reflect very little variation and few errors regardless of trial type or participant age. The average accuracy was 97.96% for younger adults and 98.16% for older adults, and a within-subjects analysis of variance showed no significant difference between the number of relevant features presented on the comparison stimuli, *F*(3, 48)=0.8, *p*=0.48, of age, *F*(1,48)=0.2, *p*=0.66, or any interaction, *F*(3,48) = 1.5, *p*= 0.23. Table 5 shows the individual scores for younger adult post-criterion accuracy, in which only one participant (Y-4) scored lower than 95% on any type of trial. For older adult participants, only 2 of the 7 (O-1 and O-4) were below 95% on any type of trial, and 6 of the 7 participants scored 100% on trials with four-feature comparison stimuli (see Table 5).

Speed scores during post-criterion training trials are shown in Figure 6. Here younger adult participants responded with similar speeds for trials containing one-feature, two-feature, and three-feature stimuli (around 0.50 responses per second), but responded faster on trials containing four-feature stimuli (averaging 0.64 responses per second).

Table 4.

Percent correct for pre-criterion trials by feature.

	Younger Adult						Older Adult				
Participant	1F	2F	3F	4F	Total	Participant	1F	2F	3F	4F	Total
Y-1	77.93	72.22	93.06	91.67	82.70	O-1	64.55	47.67	86.78	88.51	71.04
Y-2	57.89	66.67	75.76	72.73	65.28	O-2	53.08	84.83	42.86	91.00	63.39
Y-3	61.16	73.33	66.67	66.67	64.73	O-3	61.33	77.78	85.56	82.61	72.10
Y-4	61.88	66.67	82.42	84.44	70.44	O-4	56.64	51.79	75.89	75.00	63.11
Y-5	51.48	71.43	70.24	76.19	61.72	O-5	62.62	80.77	75.93	77.78	70.09
Y-6	54.44	69.05	69.05	61.90	60.83	O-6	67.31	94.87	85.71	84.62	77.49
Y-7	62.07	58.33	77.78	72.22	66.78	O-7	50.70	55.77	54.72	50.00	51.76
Mean	60.98	68.24	76.42	75.12	67.50	Mean	59.46	70.50	72.49	78.50	67.00

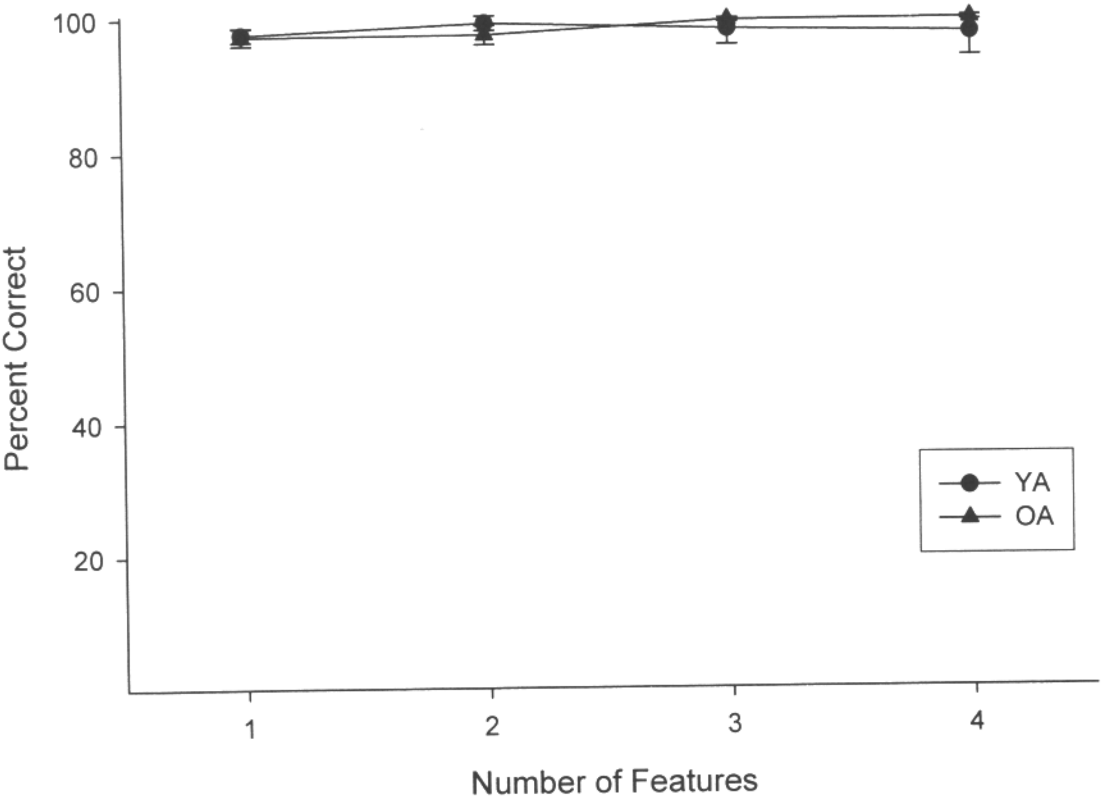


Figure 5. Mean percent correct for post-criterion trials Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

Table 5.

Percent correct for post-criterion baseline trials by feature.

	Younger Adults						Older Adults				
Participant	1F	2F	3F	4F	Total	Participant	1F	2F	3F	4F	Total
Y-1	98.81	100.00	100.00	100.00	99.40	O-1	96.59	90.91	100.00	97.73	96.88
Y-2	99.70	100.00	99.40	98.84	99.56	O-2	100.00	100.00	100.00	100.00	100.00
Y-3	99.72	100.00	98.81	98.89	99.42	O-3	97.22	100.00	100.00	100.00	98.61
Y-4	91.72	97.22	93.44	89.64	92.56	O-4	90.00	93.33	98.31	100.00	93.72
Y-5	99.08	100.00	98.75	98.81	99.08	O-5	100.00	100.00	100.00	100.00	100.00
Y-6	97.02	98.81	100.00	98.81	98.21	O-6	98.50	98.00	100.00	100.00	99.00
Y-7	97.18	98.57	97.10	98.61	97.51	O-7	98.57	100.00	98.57	100.00	98.92
Mean	97.60	99.23	98.21	97.66	97.96	Mean	97.27	97.46	99.55	99.68	98.16

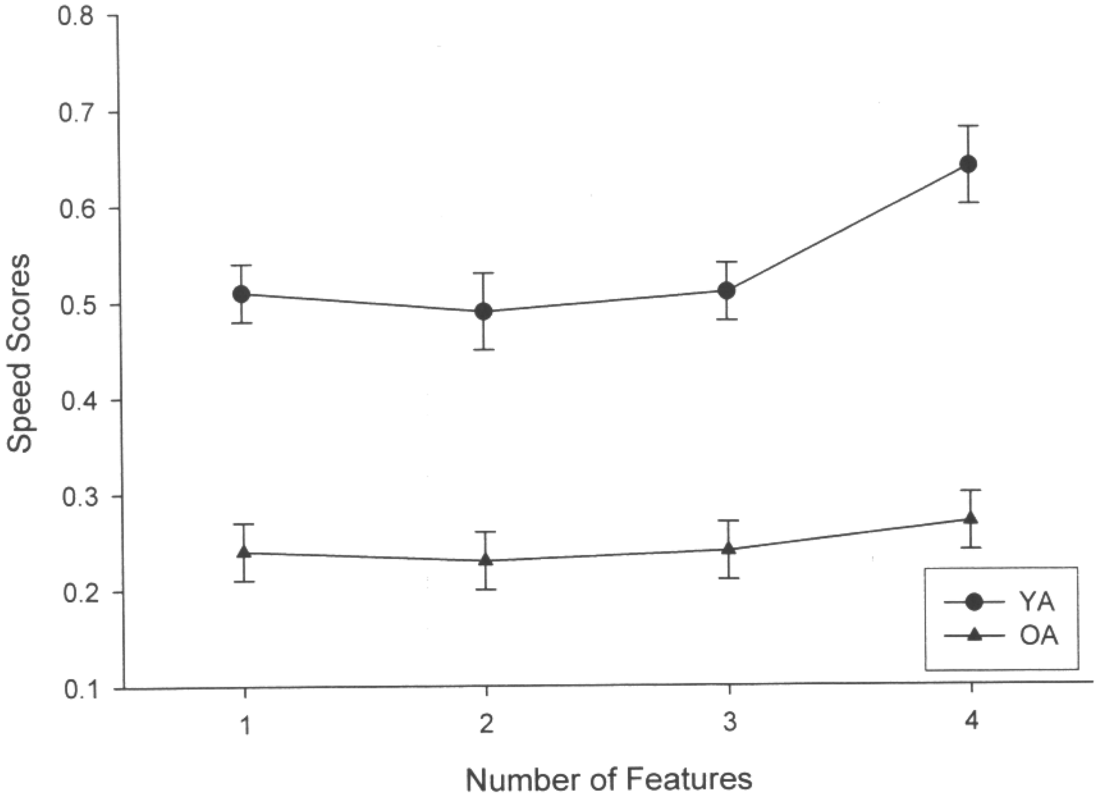


Figure 6. Mean speed scores for post-criterion trials by number of features. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

Older adult participants also showed a small increase in speed for four-feature stimuli, and overall responded more slowly and uniformly to all stimuli. A 4 x 2 ANOVA (feature by age) showed no significant interaction between age and the number of features present on the comparison stimuli ($F(3,48)=1.2$, $p=0.31$). However, there were significant main effects of both typicality ($F(3,48)=2.9$, $p=0.047$) and age ($F(1,48)=149.5$, $p<0.001$). Speed scores for post-criterion trials were also faster overall than pre-criterion trials, with average post-criterion speeds of 0.54 responses per second for younger adults and 0.25 responses per second for adults.

Individual participant data in Table 6 show speed scores by trial type for younger adults and reflect a typicality effect for four-feature comparison stimuli, with 6 of the 7 participants responding faster to those stimuli with four relevant features than to stimuli with one, two, or three relevant features. Older adult participants were less consistent, with only 4 of the 7 participants (O-1, O-3, O-4, and O-5) responding faster for those stimuli with more relevant features and slower for those with fewer relevant features, and the other 3 older adult participants responding with a reverse of this pattern.

Probe trials

Probe trials were broken up into those that determine the formation of equivalence classes (symmetry and equivalence probes) and those that demonstrate the categorization of novel stimuli (equal-feature and unequal-feature probes). Here correct trials were those in which the participant gave a response in which comparison stimuli chosen were in the same class as the sample stimulus for purposes of percent correct and speed score analyses. Figure 7 demonstrates the percent correct for all participants for each type of probe trial. Participants showed a consistently high level of performance across every

Table 6.
Speed scores by feature for post-criterion baseline trials.

	Younger Adults						Older Adults				
Participant	1F	2F	3F	4F	Mean	Participant	1F	2F	3F	4F	Mean
Y-1	0.58	0.56	0.61	0.73	0.66	O-1	0.26	0.24	0.28	0.31	0.27
Y-2	0.62	0.62	0.63	0.78	0.50	O-2	0.22	0.20	0.18	0.17	0.19
Y-3	0.50	0.48	0.47	0.53	0.44	O-3	0.21	0.22	0.20	0.32	0.24
Y-4	0.42	0.49	0.41	0.44	0.55	O-4	0.19	0.17	0.23	0.23	0.21
Y-5	0.50	0.53	0.46	0.70	0.52	O-5	0.25	0.22	0.24	0.30	0.25
Y-6	0.47	0.44	0.53	0.62	0.47	O-6	0.41	0.41	0.41	0.39	0.40
Y-7	0.46	0.33	0.46	0.65	0.62	O-7	0.16	0.18	0.13	0.14	0.15
Mean	0.51	0.49	0.51	0.64	0.51	Mean	0.24	0.23	0.24	0.27	0.25

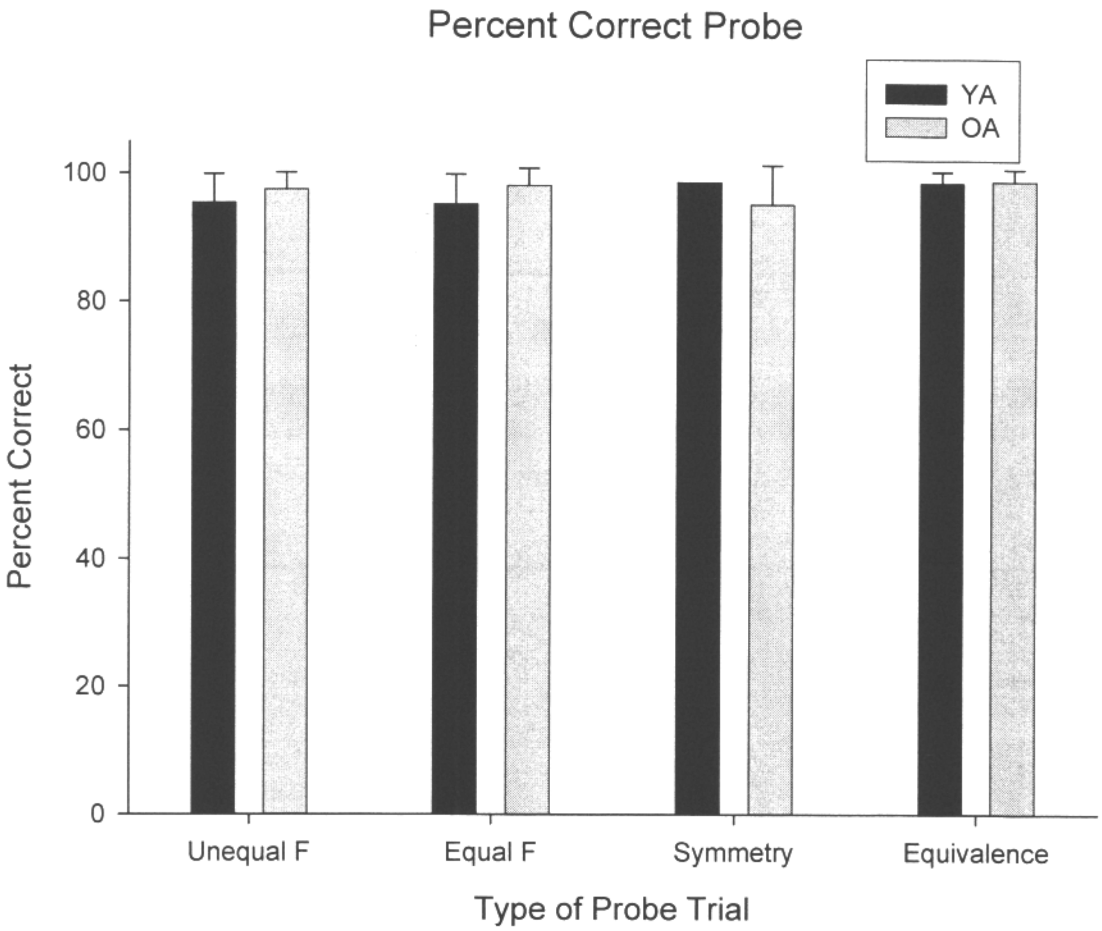


Figure 7. Mean percent correct for all probe trials. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by black columns, and data from older adults (OA) are represented by gray columns.

type of probe trial, with means ranging from 95.24% to 98.88% regardless of age (*F*(1,48)=0.1, *p*=0.72) or type of trial (*F*(3,48)=1.4, *p*=0.27). There was a significant interaction effect, in which older adult participants were more accurate on novel probe trials than younger adults, but younger adults were more accurate on symmetry trials (*F*(3,48) = 3.7, *p*=0.018). For younger adults, every participant also demonstrated equivalence-class formation, as evident in the high scores on both symmetry and equivalence probe trials (see Table 7). The mean accuracy for symmetry trials was 98.69%, with 4 of the 7 participants (Y-1, Y-2, Y-5, and Y-7) performing at 100%. The mean accuracy for equivalence trials was 98.61%, with 3 of 7 participants (Y-2, Y-3, and Y-5) performing perfectly. For older adults, 3 of the 7 participants (O-2, O-3, and O-5) performed perfectly and 6 of the 7 participants performed at or above 90% accuracy on both symmetry and equivalence trials, as shown in Table 7.

For symmetry trials, the mean percent correct was 95.24%, with 3 of the participants (O-2, O-6, and O-7) not missing a single trial. Only one participant (O-4) was noticeably less accurate on symmetry trials (83.3%); however, she reached 95.83% accuracy for equivalence probes, arguably still demonstrating class formation. For equivalence trials, the mean accuracy was 98.88, with all 7 of the participants scoring above 95%. Five of the 7 participants scored 100% on these trials.

Novel stimuli were presented either with an equal number of features or with an unequal number of features, to determine if stimulus complexity impacted categorization and reaction time of the participants. Calculations were done separately for each of these types of trials. For stimuli with an unequal number of features, younger adult participants categorized stimuli class-consistently an average of 95.41% of the time, with 2 of the 7

Table 7.

Percent correct for probe trials.

Younger Adults					Older Adults				
Subject	Unequal	Equal	Symmetry	Equivalence	Subject	Unequal	Equal	Symmetry	Equivalence
Y-1	100.0	99.5	100.00	97.90	O-1	95.06	95.00	91.67	96.30
Y-2	100.0	100.0	100.00	100.00	O-2	95.83	100.00	100.00	100.00
Y-3	89.0	89.4	97.20	100.00	O-3	97.56	100.00	95.83	100.00
Y-4	92.5	93.1	97.90	95.20	O-4	93.75	93.75	83.33	95.83
Y-5	91.0	94.7	100.00	100.00	O-5	100.00	100.00	95.83	100.00
Y-6	98.5	97.4	95.80	97.90	O-6	100.00	100.00	100.00	100.00
Y-7	95.8	88.5	100.00	97.90	O-7	100.00	98.08	100.00	100.00
Mean	95.26	94.66	98.70	98.41	Mean	97.46	98.12	95.24	98.88

participants (Y-1 and Y-2) categorizing comparison stimuli as class-consistent with the sample on every trial (see Table 7). Older adult participants responded correctly on 97.46% of the trials, with 3 of the 7 participants (O-5, O-6, and O-7) providing a class-consistent response on 100% of the trials. For stimuli with an equal number of features, younger adults responded accurately an average of 94.66% of the time, with scores ranging between 88.5% and 100%. For older adults, accuracy percentages ranged between 93.75% and 100.0%, with a mean of 98.12% and four of the participants (O-2, O-3, O-5, and O-7) scoring perfectly (Table 7).

Average group speed scores for the unequal-feature stimuli trials are shown in Figure 8, where older adult participants were consistently slower in responding than younger adult participants ($F(1,48) = 39.37, p<0.001$). There was no statistical evidence for a typicality effect ($F(3, 48)=0.65, p=0.62$) and no significant interaction between age and number of features ($F(3,48) = 2.94, p=0.162$). Individual speed scores for younger adult participants can be seen in Table 8 and show that only two of the participants (Y-2 and Y-6) show increasing speeds with increasing numbers of features consistently. Other participants reflect the unusually fast response speeds to those stimuli with only two of the class-consistent features. Three older adult participants (O-1, O-O-6, and O-7) demonstrated an impact of typicality in their speed in responding. Two others (O-4 and O-5) showed an increase in speed of responding when comparing one-feature stimuli trials to four-feature stimuli trials, although their responding was slower for those trials with two and three features.

For those trials using stimuli with an equal number of features there was a clear typicality effect for both groups ($F(3, 48)=11.0, p<0.001$). Younger adult participants

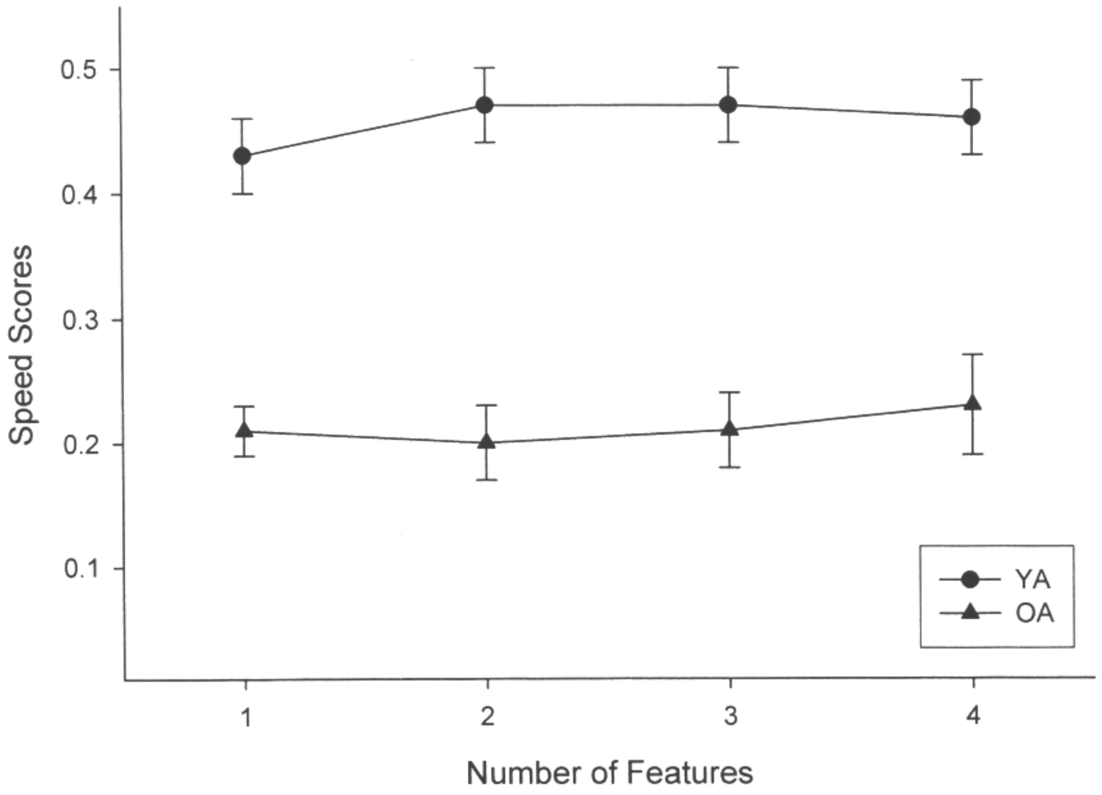


Figure 8. Mean speed scores for unequal-feature probe trials by number of features. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

Table 8.
Speed scores for unequal-feature probe trials by feature.

	Younger Adults						Older Adults				
Participant	1F	2F	3F	4F	Mean	Participant	1F	2F	3F	4F	Mean
Y-1	0.54	0.53	0.54	0.51	0.53	O-1	0.20	0.21	0.24	0.33	0.24
Y-2	0.57	0.55	0.55	0.55	0.56	O-2	0.19	0.15	0.15	0.18	0.17
Y-3	0.32	0.43	0.42	0.48	0.42	O-3	0.18	0.15	0.15	0.16	0.16
Y-4	0.40	0.44	0.36	0.36	0.39	O-4	0.19	0.18	0.18	0.20	0.19
Y-5	0.39	0.58	0.52	0.45	0.49	O-5	0.18	0.18	0.21	0.19	0.19
Y-6	0.38	0.41	0.45	0.49	0.43	O-6	0.36	0.37	0.38	0.41	0.38
Y-7	0.40	0.36	0.42	0.38	0.39	O-7	0.17	0.14	0.16	0.16	0.16
Mean	0.43	0.47	0.47	0.46	0.46	Mean	0.21	0.20	0.21	0.23	0.21

classified stimuli with four relevant features significantly faster (0.48 responses per second) than those with only one relevant feature (0.28 responses per second). Older adults showed the same pattern, with speeds of 0.24 responses per second and 0.14 responses per second, respectively (see Figure 9).

These differences reflect a significant main effect of age, $F(1,48) = 107.4$, $p<0.001$, and individual data in Table 9 also demonstrate the impact of typicality and age. There is no significant interaction between age and number of features ($F(3,48) = 1.5$, $p=0.24$).

Transfer of function

Recall that in the transfer of function task, participants saw either a four-feature target stimulus from one class or a one-feature target from another and were told that this stimulus was a "germ, which infects 50% of the people (or plants) that come into contact with it." Other stimuli, from all three of the trained classes and with varying numbers of relevant features, were then presented on the screen and participants were asked to rate the percentage of infection by contact with that stimulus. Participants' rankings for the compared stimuli were analyzed separately, according to target stimulus (either four-feature or one-feature).

Figure 10 shows the group means for ratings of class-consistent and class-inconsistent stimuli for both four-feature and one-feature target stimuli, and clearly illustrates higher ratings for those stimuli that were in the same class as the target, regardless of age. Participants consistently rated those stimuli that were in the target stimulus class as more likely to infect than those stimuli that were not class-consistent with the target stimulus ($F(1,24) =13.0$, $p=0.001$). The age of the participant did not

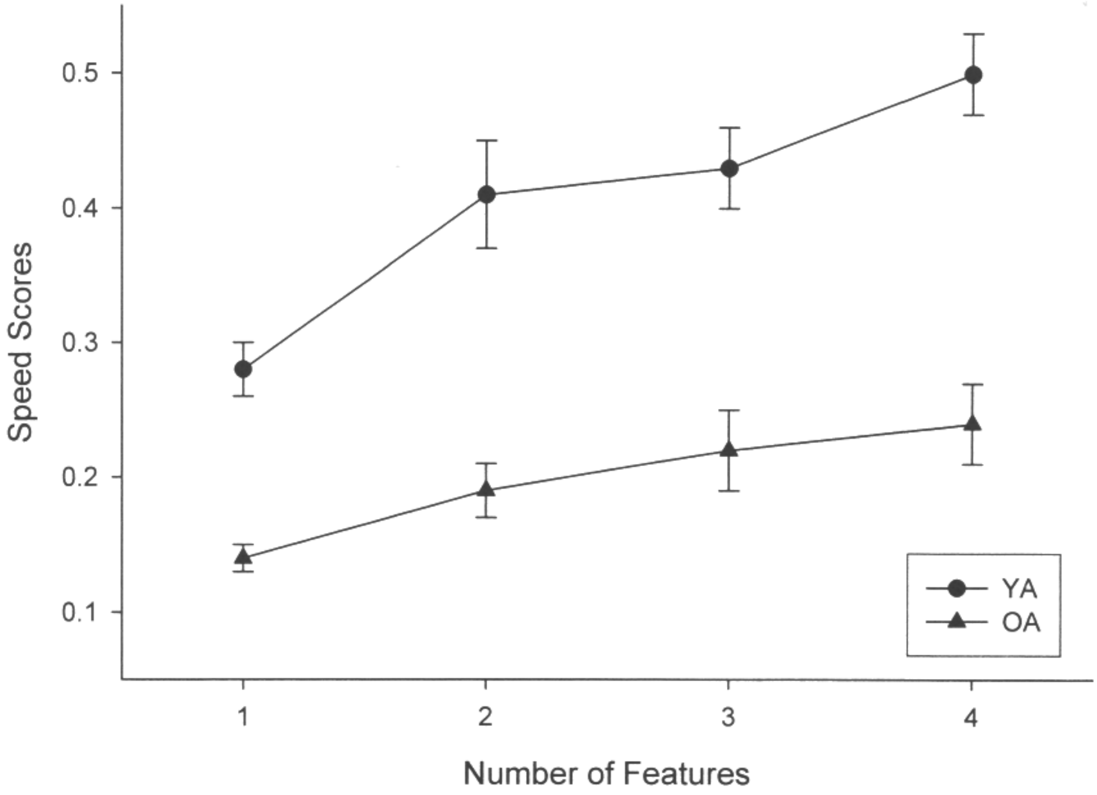


Figure 9. Mean speed scores for equal-feature probe trials by number of features. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

Table 9.
Speed scores for equal-feature probe trials by feature.

	Younger Adults						Older Adults				
Particpant	1F	2F	3F	4F	Mean	Participant	1F	2F	3F	4F	Mean
Y-1	0.36	0.50	0.52	0.53	0.48	O-1	0.14	0.20	0.21	0.26	0.20
Y-2	0.36	0.53	0.52	0.61	0.50	O-2	0.11	0.15	0.17	0.21	0.16
Y-3	0.25	0.38	0.37	0.38	0.34	O-3	0.10	0.17	0.23	0.23	0.19
Y-4	0.24	0.25	0.26	0.44	0.30	O-4	0.13	0.16	0.20	0.22	0.18
Y-5	0.27	0.46	0.44	0.54	0.43	O-5	0.16	0.22	0.26	0.22	0.21
Y-6	0.32	0.43	0.47	0.57	0.45	O-6	0.21	0.30	0.37	0.41	0.32
Y-7	0.19	0.35	0.45	0.46	0.36	O-7	0.12	0.12	0.12	0.13	0.12
Mean	0.28	0.41	0.43	0.50	0.41	Mean	0.14	0.19	0.22	0.24	0.20

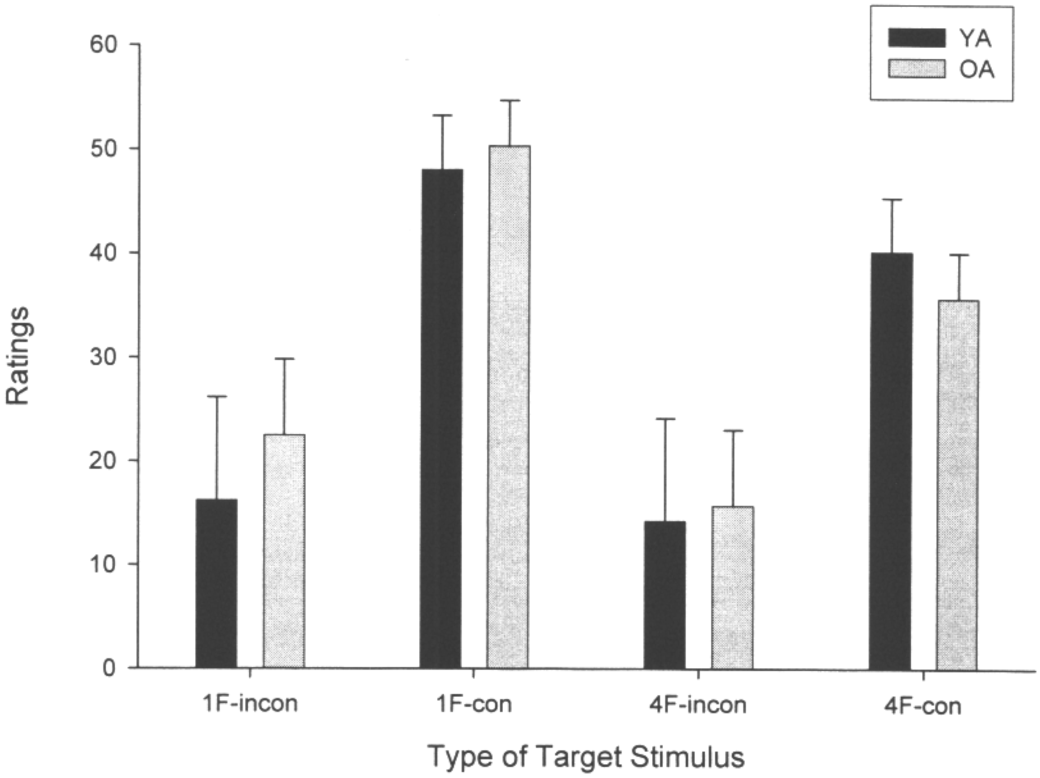


Figure 10. Mean ratings scores for target-consistent and target-inconsistent stimuli. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by black columns, and data from older adults (OA) are represented by gray columns.

target (*F*(1, 96) = 9.2, *p*=0.003) were observed within these ratings, but not of age significantly impact their rating of class-consistent versus class-inconsistent stimuli (*F*(1,24)=0.5, *p*=0.5), and there was no significant interaction between age and type of target stimulus (*F*(1,24) = 0.2, *p*=0.66). Table 10 shows the individual participant data for class-consistent and class-inconsistent responding for both younger adult and older adults and demonstrates the higher ratings for stimuli that were class-consistent with the training stimuli than for those that were class-inconsistent.

Ratings were also analyzed to determine whether the number of relevant features, age of the participant, or type of target stimulus impacted ratings for class-consistent comparison stimuli. A 2 x 2 x 4 ANOVA (age x target stimulus x number of features) was calculated to determine any main effects or interactions between these variables. Figures 11 and 12 show that for both the four-feature target stimulus and the one-feature target stimulus, group mean ratings increased as a function of the number of features present on the compared stimulus. Main effects of typicality (*F*(3, 96)=7.9, *p*<0.001) and type of (*F*(1,96)=0.1, *p*=0.71). There were also no significant interactions between either age and number of features (*F*(3,96) = 0.6, *p*=0.63), target and number of features (*F*(3,96) = 1.8, *p*=1.0, or age and target stimulus (*F*(1,96) = 1.8, *p*=0.19). The interaction between all three variables was significant (*F*(1,96) = 10.3, *p*=0.002) reflecting the somewhat different typicality functions for the two target stimuli and age groups. Figure 11 shows that the pattern of responding with higher ratings for stimuli with more relevant features and lower ratings for stimuli with fewer relevant features was the same for both groups given the four-feature target stimulus. For the one-feature target stimulus, Figure 12 shows that both younger adults and older adults rated stimuli with more relevant

Table 10.

Ratings scores for compared stimuli either class-consistent or class-inconsistent with the target stimulus.

	Younger Adult					Older Adult			
	4F Target		1F Target			4F Target		1F Target	
Subject	Con	Incon	Con	Incon	Subject	Con	Incon	Con	Incon
Y-1	33.75	0.0	50.00	10.0	O-1	35.42	37.5	25.00	50.0
Y-2	50.00	0.0	50.00	0.0	O-2	46.88	25.0	50.00	25.0
Y-3	23.13	7.5	23.75	3.8	O-3	50.00	37.5	62.81	52.5
Y-4	31.25	0.0	41.88	0.0	O-4	57.08	10.0	36.88	17.5
Y-5	50.00	0.0	46.88	0.0	O-5	48.29	0.0	81.25	0.0
Y-6	47.81	56.3	65.63	51.3	O-6	23.33	0.0	49.38	0.0
Y-7	54.38	50.0	78.75	65.0	O-7	29.17	0.0	70.63	12.5
Mean	41.47	16.3	50.98	18.6	Mean	41.45	15.7	53.71	22.5

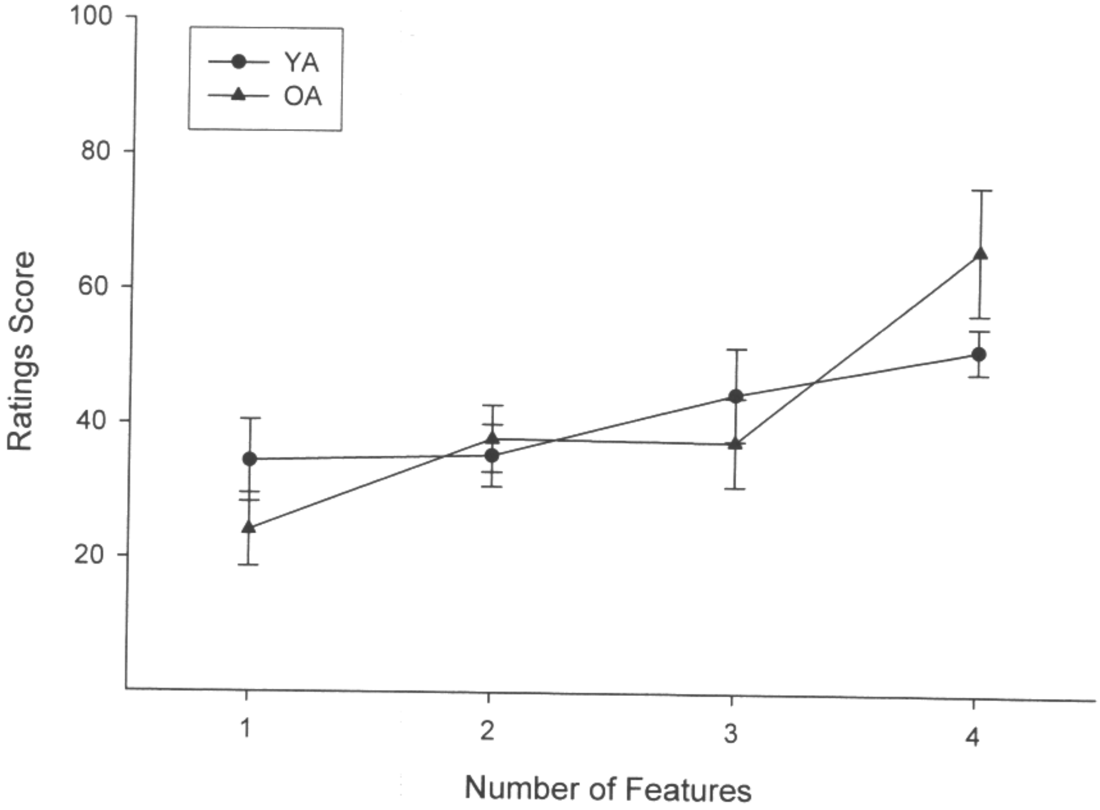


Figure 11. Mean ratings scores for class-consistent stimuli given a four-feature target stimulus. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

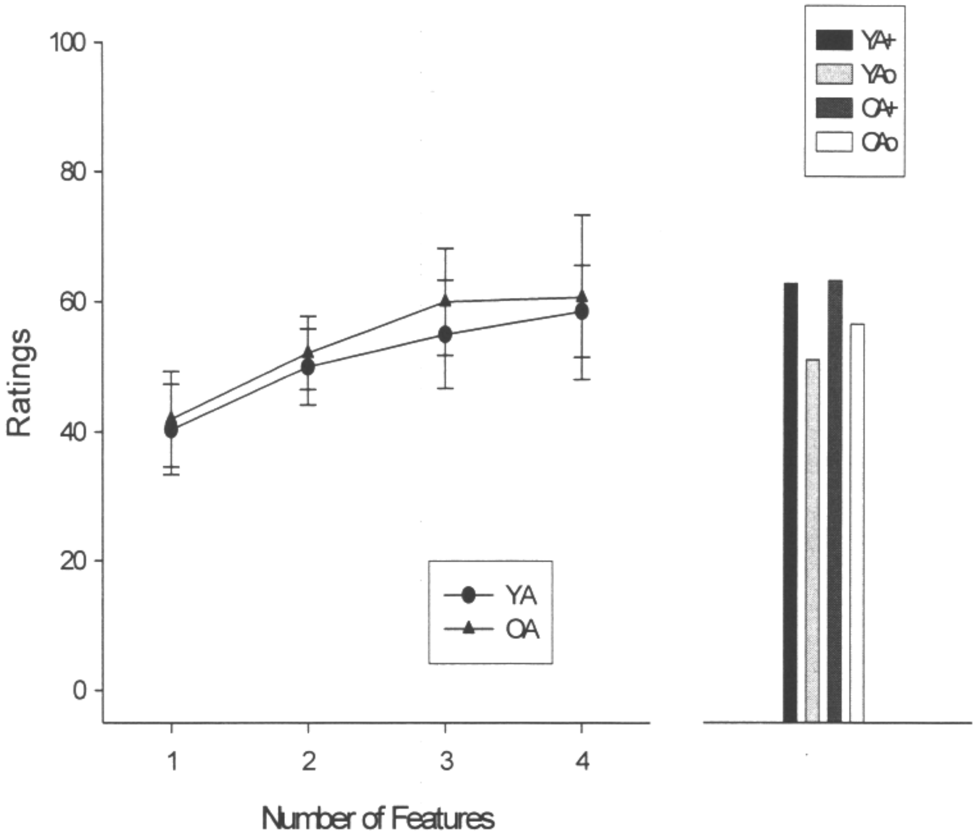


Figure 12. Mean ratings scores for class-consistent stimuli given a one-feature target stimulus and comparisons of ratings for stimuli with and without the target feature. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles. + indicates stimuli with the target feature, and o indicates stimuli without the target feature.

features as being more likely to transfer infection than stimuli with fewer relevant features.

Figures 13, 14, 15, and 16 show individual subject ratings for class-consistent (dotted lines) and class-inconsistent (solid lines) responses for both the four-feature target and the one-feature target scenarios, broken down by age and number of features present on the compared stimulus. Figures 13 (younger adults) and 14 (older adults) are from scenarios involving a four-feature target stimulus. These results can also be seen in Table 11. Figures 15 (younger adults) and 16 (older adults) are from scenarios involving a one-feature target stimulus, which can also be seen in Table 12.

Figure 13 shows that four of the 7 younger adult participants (Y-1, Y-2, Y-4, and Y-5) rated class-inconsistent stimuli as having a 0% infection rate after function training with the four-feature target. Three of these participants (Y-1, Y-2, and Y-5) showed no impact of typicality in their ratings of class-consistent stimuli either. Participant Y-4 showed a linear impact of typicality for class-consistent stimuli, with four-feature class-consistent stimuli receiving the same likelihood of infection as the target stimulus, three-feature class-consistent stimuli receiving 75% of the target stimulus infection rating, two-feature class-consistent stimuli receiving 50% of the target stimulus infection rating, and one-feature class-consistent stimuli receiving 25% of the target stimulus infection rating. Participants Y-3, Y-6, and Y-7 showed some transfer of function to stimuli that were not class-consistent with the target stimulus and also demonstrated some impact of typicality within those ratings. For these participants, it may have been that a characteristic of the stimuli other than equivalence-class membership that impacted the transfer of functions between stimuli.

Table 11.
Ratings scores for compared stimuli given a four-feature target stimulus.

Participant	1F	2F	3F	4F	Participant	1F	2F	3F	4F
Y-1	25.0	30.0	30.0	50.0	O-1	16.7	50.0	25.0	50.0
Y-2	50.0	50.0	50.0	50.0	O-2	50.0	37.5	50.0	50.0
Y-3	17.5	17.5	17.5	40.0	O-3	35.0	50.0	40.0	75.0
Y-4	12.5	25.0	37.5	50.0	O-4	23.3	35.0	70.0	100.0
Y-5	50.0	50.0	50.0	50.0	O-5	24.2	50.0	20.0	99.0
Y-6	48.8	40.0	52.5	50.0	O-6	8.3	17.5	27.5	40.0
Y-7	37.5	35.0	75.0	70.0	O-7	11.7	25.0	30.0	50.0
Mean	34.46	35.36	44.64	51.43	Mean	24.17	37.86	37.50	66.29

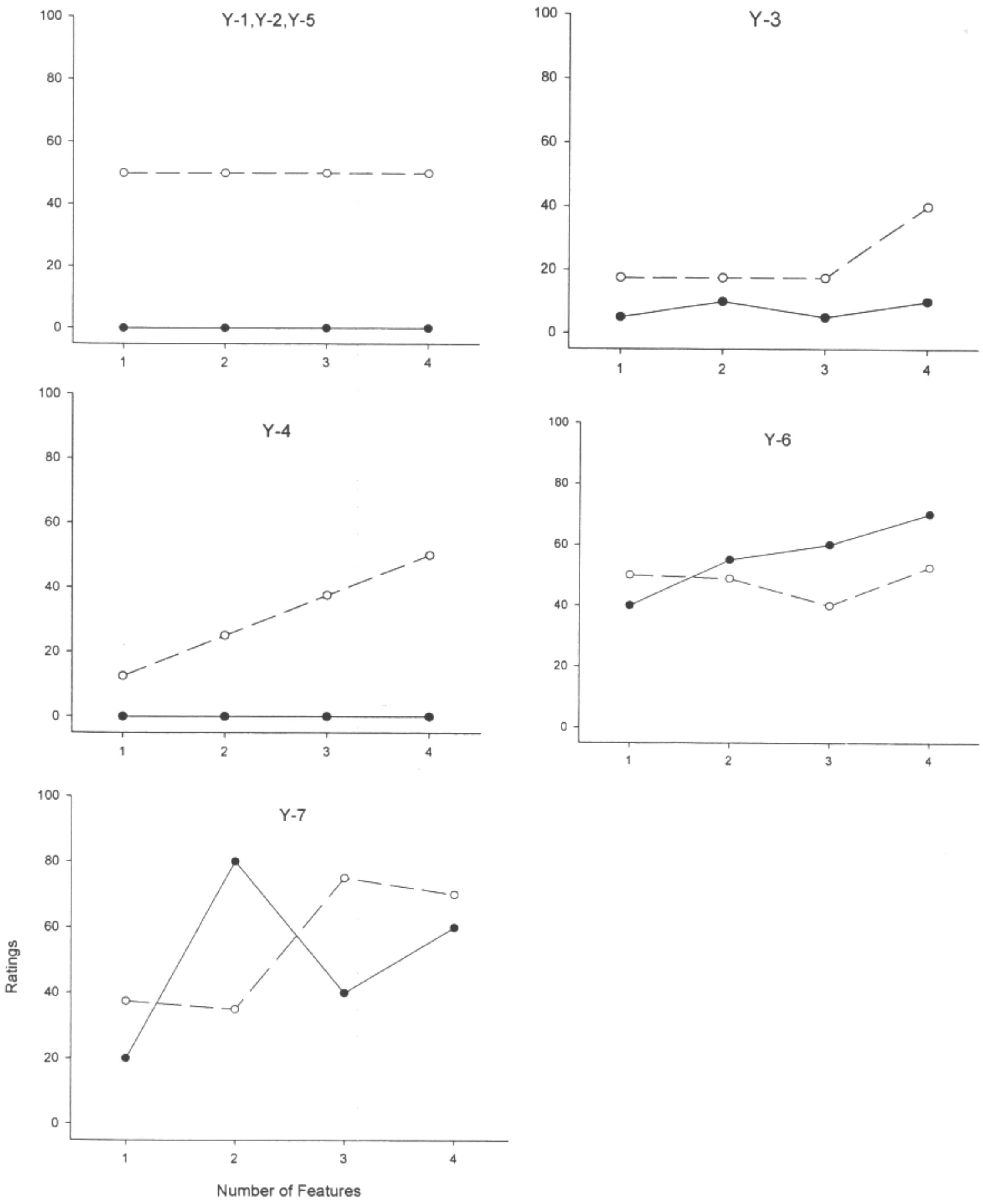


Figure 13. Individual younger adult graphs for class-consistent and class-inconsistent ratings given a four-feature target, by feature. Dotted lines and open circles indicate class-consistent responses, and solid lines and filled circles indicate class-inconsistent responses.

For the older adults, individual data for the four-feature target stimulus are presented in Figure 14. Again, three of the participants (O-5, O-6, and O-7) showed no transfer of function to stimuli inconsistent with the target stimulus. Participant O-4 also showed this pattern with the exception of a four-feature class-inconsistent stimulus, which received a 40% rating of likelihood of infection. Participants O-1, O-2, and O-3 all showed some transfer of function to class-inconsistent stimuli, although the majority of class-inconsistent stimuli were still rated as having a lower likelihood of infection than class-consistent stimuli. For class-consistent stimuli, participants O-3, O-4, O-6, and O-7 all showed a linear impact of typicality in their ratings of likelihood of infection. Participant O-5 also followed this pattern with the exception of three-feature stimuli, which were given a lower rating than either two-feature or four-feature compared stimuli. Participants O-1 and O-2 demonstrated mixed results for ratings of class-consistent stimuli, showing that perhaps number of relevant features was not the characteristic of the stimuli that determined the likelihood of infection given.

For the one-feature target, 5 of the 7 younger adults (Y-1, Y-2, Y-3, Y-4 and Y-5, shown in Figure 15) rated the stimuli not in the target stimulus class as having a 0% infection rate. Three of these participants (Y-1, Y-2, and Y-5) also showed no impact of typicality in their ratings of class-consistent stimuli, giving all class-consistent compared stimuli the same likelihood of infection as the target stimulus. Participants Y-3 and Y-4 demonstrated some impact of typicality in their ratings of class-consistent stimuli, with one-feature compared stimuli receiving a lower likelihood of infection rating than four-feature compared stimuli. Participants Y-6 and Y-7 showed transfer of function to all stimuli regardless of class, although class-inconsistent stimuli did receive lower ratings

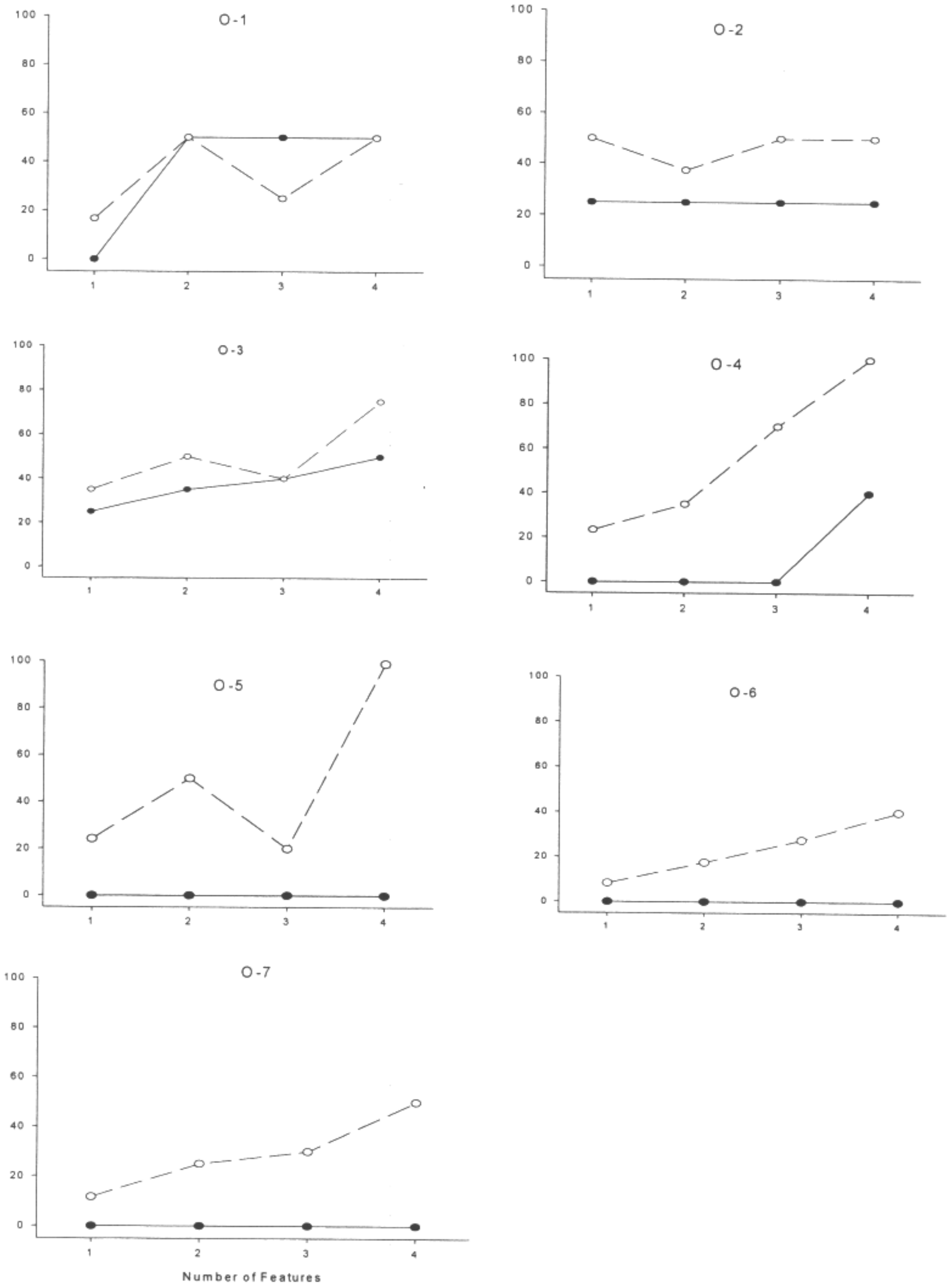


Figure 14. Individual older adult graphs for class-consistent and class-inconsistent ratings given a four-feature target, by feature. Dotted lines and open circles indicate class-consistent responses, and solid lines and filled circles indicate class-inconsistent responses.

Table 12.
Ratings scores for compared stimuli given a one-feature target stimulus.

Participant	1F	2F	3F	4F	Participant	1F	2F	3F	4F
Y-1	50.0	50.0	50.0	50.0	O-1	25.0	50.0	25.0	0.0
Y-2	50.0	50.0	50.0	50.0	O-2	50.0	50.0	50.0	50.0
Y-3	12.5	20.0	22.5	40.0	O-3	58.8	55.0	62.5	75.0
Y-4	20.0	47.5	50.0	50.0	O-4	22.5	40.0	45.0	40.0
Y-5	37.5	50.0	50.0	50.0	O-5	62.5	75.0	87.5	100.0
Y-6	47.5	67.5	67.5	80.0	O-6	17.5	30.0	70.0	80.0
Y-7	65.0	65.0	95.0	90.0	O-7	57.5	65.0	80.0	80.0
Mean	40.36	50.00	55.00	58.75	Mean	25.0	50.0	60.0	60.71

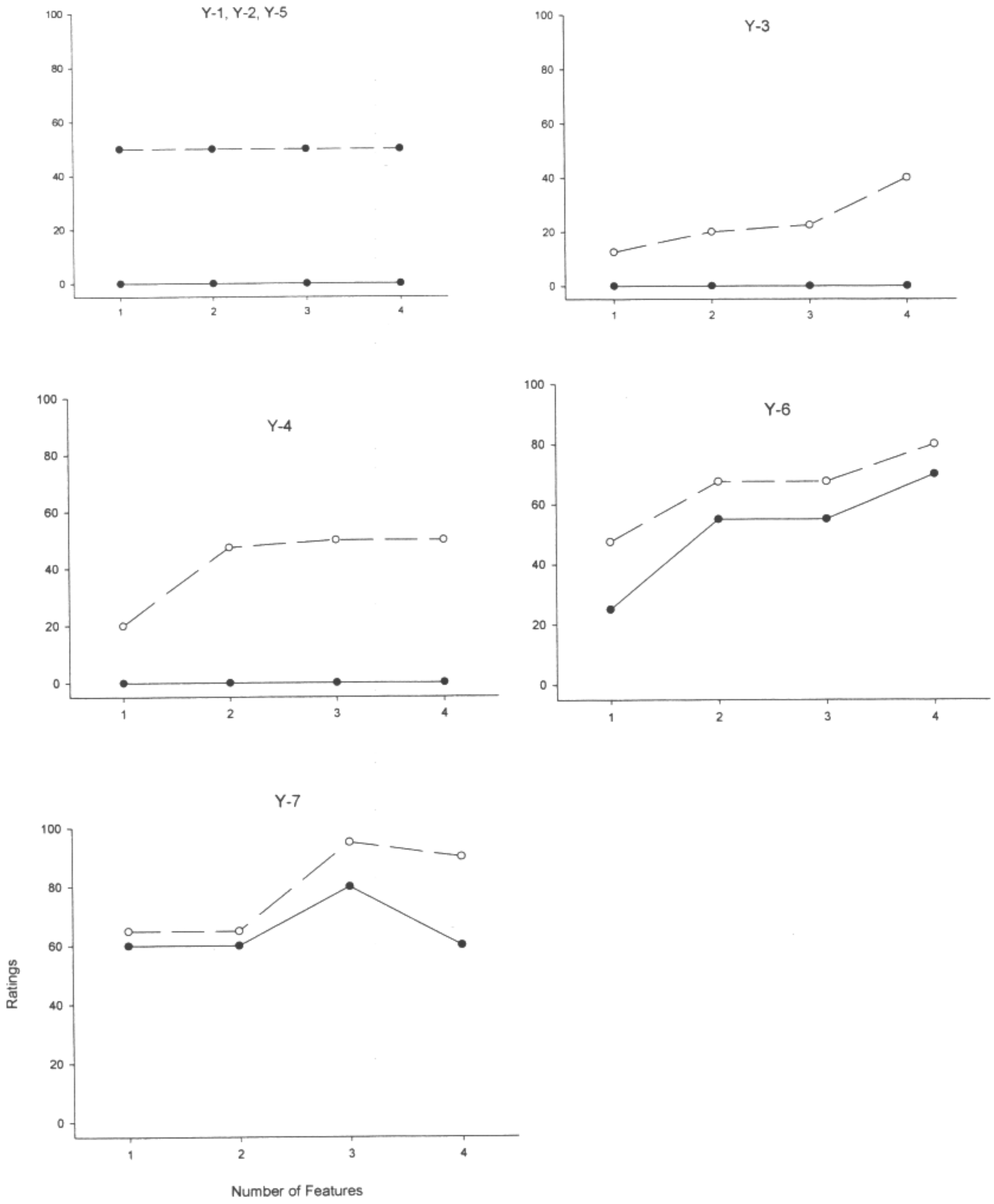


Figure 15. Individual younger adult graphs for class-consistent and class-inconsistent ratings given a one-feature target, by feature. Dotted lines and open circles indicate class-consistent responses, and solid lines and filled circles indicate class-inconsistent responses.

than class-consistent stimuli. These participants also demonstrated an impact of typicality in all ratings of stimuli, regardless of class-consistency.

Individual ratings for the older adult participants given the one-feature target (Figure 16) showed that 3 of the 7 older adult participants (O-5, O-6, and O-7) rated stimuli not in the target stimulus class as having a 0% likelihood of infection. Other older adult participants gave ratings that showed some transfer of function to the class-inconsistent stimuli, but less transfer than to class-consistent stimuli. Participant O-1 showed no impact of typicality in ratings of class-inconsistent stimuli, but rated stimuli that were class-consistent with the target as having a varying range of transfer of function. Participant O-2 showed similar results with class-inconsistent stimuli, but also showed no impact of typicality in ratings of class-consistent stimuli either, although all class-consistent stimuli were rated as having a higher likelihood of infection than class-inconsistent stimuli. Participants O-3 and O-4 showed both transfer of function and an impact of typicality in all ratings of compared stimuli, regardless of class.

Finally, the ratings scores for the one-feature target stimulus were analyzed in groups to determine whether stimuli that contained the specific feature found on the target (fill pattern), were rated higher than those stimuli that were class-consistent but did not contain the feature in the target stimulus. Although there was a trend for those stimuli containing the fill pattern feature to be rated higher than those stimuli not containing this feature, a 2 x 2 ANOVA comparison (age x target feature) showed no significant difference ($F(1,24))=1.1$, $p=0.30$) and no significant main effect of age ($F(1,24)=0.1$, $p=0.74$) or interaction ($F(1,24) = 0.1$, $p=0.77$). These results may have been impacted by the lack of variation in many of the participants' ratings of all class-

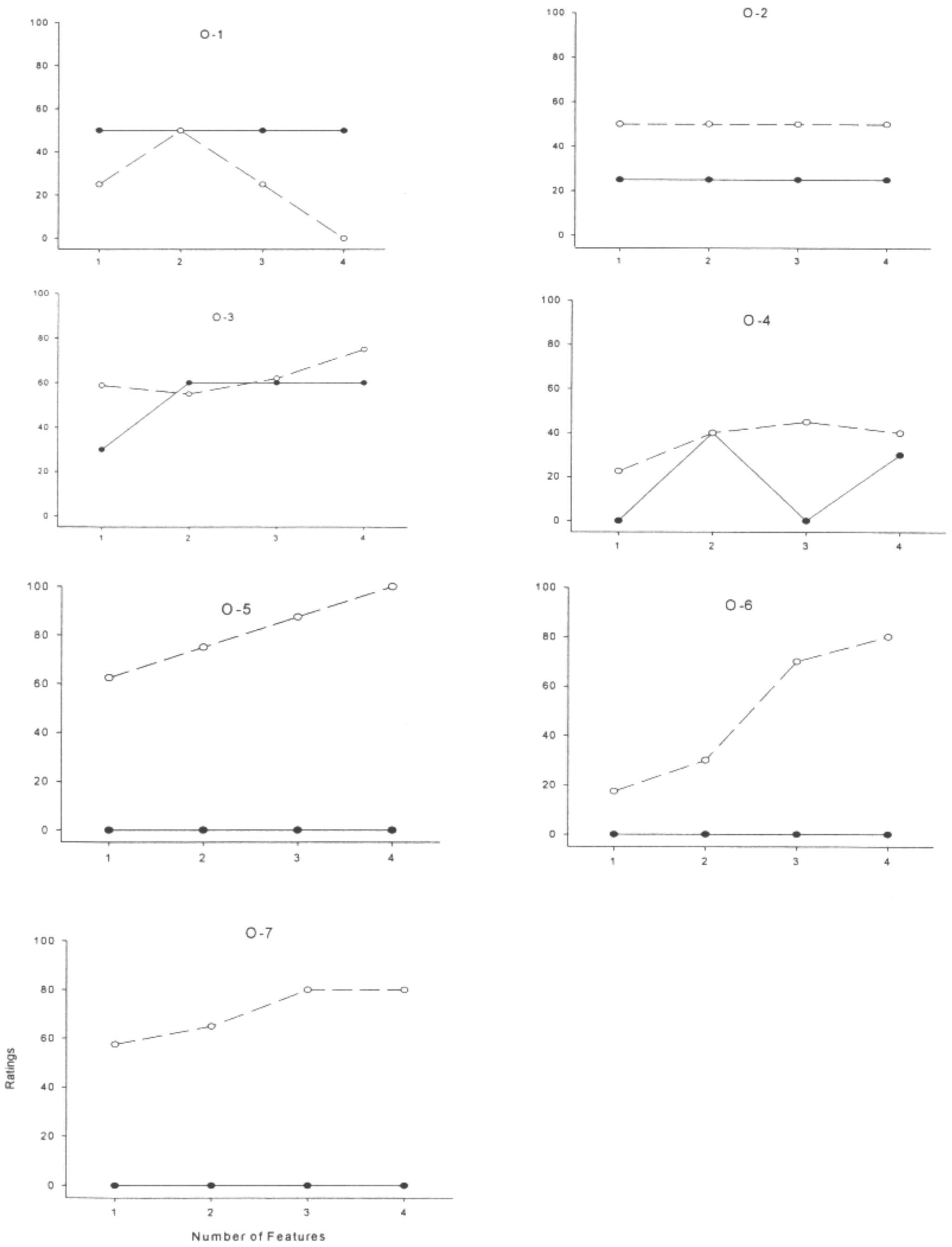


Figure 16. Individual older adult graphs for class-consistent and class-inconsistent ratings given a one-feature target, by feature. Dotted lines and open circles indicate class-consistent responses, and solid lines and filled circles indicate class-inconsistent responses.

consistent stimuli (see Table 12 and Figures 15 and 16 for individual results), but show that the typicality effects noted above were controlled by more than the target feature. Precautions were taken to ensure that the order of the transfer of function tasks did not impact the ratings of participants.

Pre-sorting and post-sorting task ratings

Participants were asked to group cards containing pictures of the training stimuli into three groups identified by the trigrams used in training for both the pre-sort task and the post-sort task. Comparing the pre-sort scores to post-sort scores showed a distinct difference between scores for both younger adults and older adults. Younger adults scored an average of 46.57% accuracy during pre-sort and a 98.81% accuracy during post-sort; older adults scored an average of 36.46% correct during pre-sort and a 99.05% correct during post-sort (see Table 13). A 2 x 2 (age group x test time) ANOVA identified a significant main effect of timing of sort ($F(1, 24) = 659.9$, $p<0.001$) and a main effect of age ($F(1,24) = 4.9$, $p=0.037$), as well as a significant interaction ($F(1,24) = 5.4$, $p=0.029$). Individual data are presented in Table 13 and show that only one participant in each group (Y-4 for the younger adult group and O-4 for the older adults) did not group every stimulus consistently with the classes trained during the computer task at the post-sort.

Additionally in the post-sort task, participants were also asked to assign a numerical value (1-8) to each card based on how representative of the class they rated the stimulus; those stimuli that were deemed “more representative” were assigned a higher value, and those stimuli that were deemed “less representative” were assigned a lower value.

Table 13.
Percentage of class-consistent items from pre- and post-sorting tasks.

Younger Adults								
	Pre-sort				Post-sort			
SN	W	N	J	Av	W	N	J	Av
Y-1	50.00	33.00	33.00	38.67	100.00	100.00	100.00	100.00
Y-2	50.00	100.00	33.00	61.00	100.00	100.00	100.00	100.00
Y-3	38.00	38.00	38.00	38.00	100.00	100.00	100.00	100.00
Y-4	38.00	38.00	38.00	38.00	100.00	87.50	87.50	91.67
Y-5	38.00	38.00	38.00	38.00	100.00	100.00	100.00	100.00
Y-6	66.00	44.00	44.00	51.33	100.00	100.00	100.00	100.00
Y-7	33.00	50.00	100.00	61.00	100.00	100.00	100.00	100.00
Mean	44.68	47.31	45.86	45.95	100.00	98.44	98.44	98.96

Older Adults								
	Pre-sort				Post-sort			
SN	W	N	J	Av	W	N	J	
O-1	36.36	42.86	33.33	37.52	100.00	100.00	100.00	100
O-2	36.36	33.33	33.33	34.34	100.00	100.00	100.00	100
O-3	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100
O-4	33.33	33.33	33.33	33.33	100.00	80.00	100.00	93
O-5	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100
O-6	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100
O-7	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100
Mean	36.58	37.07	35.71	36.46	100.00	97.14	100.00	99

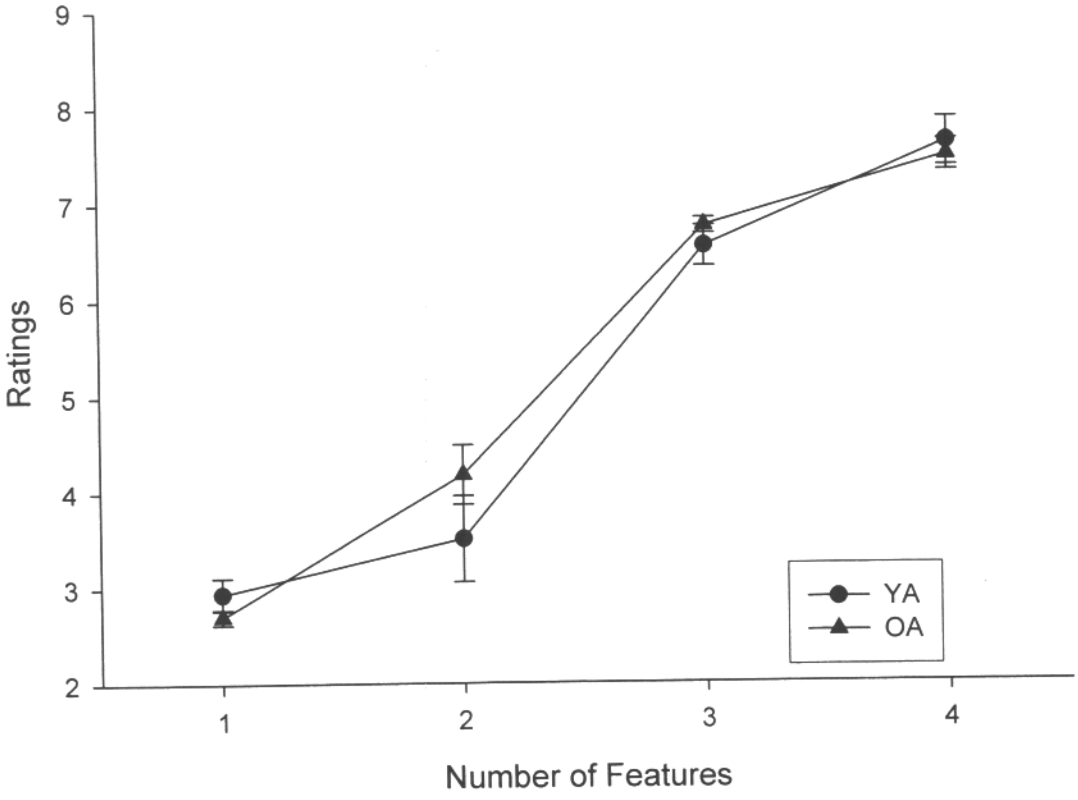
Figure 17 shows mean ratings for both groups and demonstrates that participants rated those stimuli with more relevant features as “more representative” than those stimuli with fewer relevant features, ($F(3, 48)=171.50$, $p<0.001$), with no real difference between younger adults and older adults ($F(1,48)=0.5$, $p=0.47$) and no significant interaction between age and number of features ($F(3,48) = 1.4$, $p=0.24$).

DISCUSSION

This experiment was conducted with several purposes. The first was to determine whether older adults would be able learn conditional discriminations and generalized equivalence classes as quickly and completely as younger adults. Along with this, a second goal was to determine whether older adults would show the same patterns of typicality effects within their generalized equivalence classes as younger adults have demonstrated in the past (Galizio et al, 2004). A third goal was to evaluate the transfer of functions within classes, for both older and younger adults.

Older adults did require more training trials to reach the baseline accuracy criterion of 88% than did younger adults. Although three of the ten older adults that began the experiment initially did not reach this criterion, there were a similar number of younger adults (five) who also did not reach criterion within three sessions. Once reaching the baseline training accuracy criterion, younger adults and older adults did not differ with respect to number of errors made on baseline trials or probe trials.

The most consistent difference between younger adults and older adults throughout the experiment was the speed of response; older adults responded significantly slower than younger adults on both pre-criterion and post-criterion training



<u>Figure 17.</u> Mean post-sort ratings scores by number of features. Bars represent standard deviations from the group means. Vertical bars represent standard deviation. Data from younger adults (YA) are represented by filled circles, and data from older adults (OA) are represented by filled triangles.

trials. Older adults were also significantly slower than younger adults in classifying novel stimuli (regardless of the number of features on the comparison stimuli).

It is important to note that in the pre-sort task, neither older adults nor younger adults grouped the cards into the categories determined by the experimenter. Thus, the stimuli within each experimenter-defined class were not rated as more similar to one another than to stimuli in other classes. The pre-sort grouping also demonstrates that the classes learned by the participants can be attributed to the training program alone. Both older adults and younger adults showed the change in sorting between the pre-sort task and the post-sort task, and once trained, both older adults and younger adults demonstrated robust equivalence-class formation, evidenced by strong performance on symmetry and combined equivalence probes. By using arbitrary shapes and ensuring that no sample and comparison shapes shared the same relevant features during equivalence testing, this experiment ensured that participants were not simply matching stimuli based on physical similarity, but based on the functional similarity of reinforcement learned during the training phase. It is this functional similarity that is important for the use of equivalence classes as a model for natural language categories, in which stimuli such as written words share no physically common characteristics with spoken words or with the objects identified, but are instead grouped together because of their function. Stimuli that are related by functional characteristics are the basis for equivalence classes as discussed by Sidman (1994).

In addition to demonstrating equivalence-class formation in both groups, the results of this study also demonstrate the generalization of those classes to novel stimuli. When given stimuli never presented during training, both older adult and younger adult participants were able to correctly classify the stimuli based on the features common to the class, regardless of the total number of features present on the stimulus. Thus, participants in this experiment demonstrated that classified stimuli did not have to be physically similar to be related, but that novel stimuli that were physically similar to class members could also be included in a class. Generalized equivalence is a phenomenon documented in previous stimulus-equivalence literature (Fields, Reeve, Adams, & Verhave, 1991) on the basis of primary stimulus generalization, in which physical similarity among class members allows for the inclusion of new class members. The results reported here support the results of Galizio et al. (2004) showing generalized equivalence among stimuli that, although not rated as physically similar, shared certain features, as do family resemblance categories.

The finding of generalized equivalence classes in older adults has implications for previous research. Wilson and Milan (1995) were not able to demonstrate equivalence-class formation consistently in older adults. Although some of their older participants completed the training with the same number of trials as the younger participants, all participants that did not meet the accuracy criterion within two blocks of 120 trials (eleven out of twenty older adults and four out of twenty younger adults) were dropped from the study with the conclusion that they were not able to form equivalence classes. The present experiment did demonstrate reliable class formation in most older adults (seven out of ten participants), but only after many more training trials had been presented. Older adults required an average of 493.86 trials in this experiment, more than twice as many as were presented to adults by Wilson and Milan (240 at most), who had much simpler baseline requirements. These differing results are therefore most likely due to the extended training exposure. Thus, the conclusion of Wilson and Milan that suggests that not all older adults are able to form equivalence classes is not the conclusion of the current study; older adults are very capable of forming classes if given sufficient training opportunity.

In their study of stimulus equivalence in older adults, Perez-Gonzales and Moreno-Sierra (1999) presented training trials to older adults until an accuracy criterion had been reached, just as the current experiment did. However, Perez-Gonzales and Moreno-Sierra found that even with training until accuracy, older adults were still more likely to make errors during probe testing blocks than younger adults were. Although this finding is similar to the finding of the current study that younger adults perform more accurately than older adults on symmetry trials, it does not match the current findings regarding the combined equivalence trials or generalization to novel stimuli. Both older and younger adults maintained at least a 95% accuracy for both symmetry and equivalence trials in this experiment. This difference may be due to the design of the experiment; participants in the Perez-Gonzales et al study were given exposure to the baseline trials only during training, and not during probe testing blocks. Thus, the argument could be made that the older adults may have failed to maintain their baseline relations during probe testing and therefore could not maintain probe-trial performance. The current experiment did include baseline trials throughout all types of probe testing trials to ensure accuracy maintenance. Both Perez-Gonzales et al (1999) and Wilson et al (1995) found that older adults made a significantly greater number of errors during training trials and required more trials to reach a baseline accuracy criterion. Results from the current experiment showed the same finding with regards to the number of baseline trials needed to reach the accuracy criterion, but found that with extended training, older adults became just as accurate as younger adult participants. Thus, the present experiment reliably demonstrated equivalence class-formation with older adults. Moreover, the present study showed that classification of novel stimuli in generalized equivalence classes was also similar in younger adults and older adults. Although there were differences in the rate of conditional discrimination learning and speed of responding in the present study, no differences in categorization between young and old subjects were noted.

A key property of categorization is the occurrence of typicality effects. In this regard, the present experiment replicated some of the previous findings with younger adults and demonstrated similar findings with older adults. There was a significant impact of typicality with respect to the number of errors made during pre-criterion training for both older adults and younger adults. Both younger adults and older adults made more errors with comparison stimuli with only one relevant feature than with the stimuli containing four relevant features. Although the groups are not identical in their demonstration of typicality effects, they do both show an impact of typicality in their likelihood of infection ratings. The typicality effect is not evident in pre-criterion speed scores or post-criterion accuracy scores (both of which showed older adults and younger adults responding very consistently on all trials), but are evident with speed scores for post-criterion trials. Younger adult participants were significantly faster in classifying stimuli with four relevant features than with stimuli containing less than four relevant features during post-criterion training trials. Older adults showed a similar pattern of responding, but to a lesser extent. This is different than the results presented in Galizio et al (2004), in which younger adults demonstrated typicality effects in baseline-training-trial speed scores for all sessions. This difference in findings may be due to the way in which the analyses were conducted. Galizio et al divided their analyses by session, and typicality effects had disappeared by the third session as participants reached a ceiling level of response speed. In this experiment, all baseline trials completed after the accuracy criterion was reached were grouped together to allow for comparisons between older and younger adults (who completed a varying number of sessions). Thus, it may be that typicality effects in post-criterion baseline trials were masked as participants reached a ceiling level of response speed.

Typicality effects were also evident in the categorization of novel stimuli for both groups. For trials involving novel stimuli with an equal number of features, both older and younger adults showed typicality in their speed of responding. Older adults as a group showed an almost linear progression of increased response speed as a function of increased numbers of features; younger adults were a bit more varied and showed distinctly slower response speeds for one-feature stimuli and distinctly faster response speeds for four-feature stimuli, but less difference between two and three feature stimuli. For trials involving stimuli with an unequal number of features, the impact of typicality was less evident. Younger adults did demonstrate a slower response speed to stimuli with only one relevant feature, but their response speeds were almost uniform for stimuli with two, three, or four relevant features. Older adults responded at almost the same speed across all trial types for stimuli with unequal features. This is somewhat similar to the findings of Galizio et al (2004), in which typicality effects were observed for both unequal-feature and equal-feature novel probe trials, but were much more pronounced for those trials involving stimuli with an equal number of features. The difference in the findings of Galizio et al (2004) and this experiment could be due to the smaller sample size and greater overall variability in response speed in the current study. Older adults were quite variable in their response speeds across all trials, making small trends in their data difficult to determine. Future research should speak to this problem, perhaps training adults to a steady rate of responding before beginning the training procedure.

Typicality effects were also demonstrated in the post-test rankings of stimuli once sorted into classes. Both older and younger adults rated stimuli with more relevant features as being "more representative" of the class than stimuli with a fewer number of relevant features. Thus, it seems that both older and younger adults demonstrate strong class formation given the training procedure and clear typicality effects in several different response measures. Again, the only major difference between the groups was in response speed.

The literature on categorization in older adults has suggested that they are less likely to demonstrate typicality effects than younger adults (Johnson et al, 1995; Hess and Slaughter, 1986). In their analysis of older adult categorization, Hess and Slaughter (1986) determined that categorization abilities in older adults may be less automatic than categorization abilities in younger adults, and that those abilities may be more impacted by age-related declines. They base these conclusions on the increased number of errors made by older adults as compared to younger adults in a categorization task and the inability of older adults to correctly name a rule for classification. Using categories created in the laboratory, the present study did not replicate those findings; older adults demonstrated typicality effects in pre-criterion trial accuracy and speed scores, post-criterion speed scores, novel probe trial speed scores, and post-testing sort scores. However, this study did not assess the ability of older adults to identify a classification rule; perhaps older adults are simply less able to put a name to a learned class than are younger adults, though they may be equally able to categorize stimuli. Older adults in this study did make more total errors in baseline training trials, just as the older adults in Hess and Slaughter's study; however, the older adults did reach the accuracy criterion and continue on with probe testing trials. These results are comparable to the younger adult performances and do not suggest that categorization abilities are impacted by age-related declines in cognitive functioning. Johnson, Hermann, and Bonilla (1995) found that adults with Alzheimer's disease were less able to correctly categorize stimuli, suggesting that declines in categorization ability may be a marker for early cognitive impairment. The results from the current experiment used only healthy older adults and could not speak to the declines experienced by adults diagnosed with Alzheimer's disease. However, results here may provide a base level of comparison for future experiments; it is now known that healthy older adults are able to categorize and demonstrate typicality effects, and so perhaps declines in these abilities may be suggestive of health problems only just beginning to show their impact. Performance of healthy older adults is essential to future comparisons of adults experiencing declines in health.

Several studies comparing older and younger adults found that older adults were slower to respond than younger adults (Baron and Journey, 1989; Baron, Menich, and Perone, 1983). This experiment also found that older adults were significantly slower to respond than younger adults in all types of trials, providing support for previous research. Baron, Menich, and Perone (1983) suggested that perhaps declines in physical abilities, rather than cognitive abilities, were the cause of slower older adult response times. Their study also showed that practice increased the response speed of older adults, bringing it close to the response speeds demonstrated by younger adults. Results from this experiment did not completely reflect the decreases in response speed differences between older and younger adults due to practice; older adults continued to show slower response speeds than younger adults well into the probe trial testing phases. However, older adults did show a significant increase in response speed between pre-criterion training trials and post-criterion training trials, suggesting that performance did in fact improve within the group. Thus, older adults did show an impact of age-related declines, but those declines were only evident in reaction time and not in class formation or the demonstration of typicality effects.

Finally, the present experiment was one of the first to assess transfer of function in generalized equivalence classes (see Belanich and Fields, 2003 for another example). Although the rating of likelihood of disease transmission is referred to here as transfer of function within stimulus classes, it should be noted that other literature might call this characteristic *transformation* of function. The difference lies in the theoretical implications of exactly what occurs as class members come to exhibit a function trained to only one member of the class. Literature involving Sidman's theory of equivalence classes and the naming theory of classification refer to the transfer of function, because all stimuli within the class will come to exhibit some amount of the same function trained to one class member (Hayes et al., 2001). Relational frame theorists prefer transformation of function, because functions may be altered across classes that are not equivalent and thus may not involve a simple transfer of a particular function. For example, Whelan and Barnes-Holmes (2004) taught participants classes of same and opposite and then established one class member as a punisher. Those class members that were in the same class exhibited the same function, but those class members in the opposite class exhibited an opposite function and became reinforcers. Thus, the function had not simply *transferred*, but had been *transformed*; some stimuli demonstrate a function that was never trained to any stimulus. The current study uses only equivalence classes and thus only demonstrates the transfer of functions between stimuli within a class. However, the argument could be made that this experiment does demonstrate transformation of function as well, as some participants rated compared stimuli as having a rate of infection not previously tied to any other class member.

Regarding the transfer of the function likelihood of infection within equivalence classes, this study allows for several conclusions. Both older and younger adults demonstrated transfer of function, as evidenced by the higher ratings of infection for those stimuli that were in the same class as the target stimulus than for those stimuli that were not. This supports the conclusions of Dougher, Augustson, Markham, Greenway, and Wulfert (1994), who found transfer of the function of fear in three learned classes of arbitrary shapes. Lowe, Horne, and Hughes (2005) have also demonstrated transfer of function with very young children; once the children had been taught to classify arbitrary stimuli into two groups, they demonstrated the transfer of a function trained to only one of the class members. Further, Whelan and Barnes-Holmes (2004) also demonstrated transfer of a punishment function within trained classes of arbitrary stimuli. Although the participants from these experiments were only younger adults and children (see Dymond & Rhefeld, 2000 for a complete review), the current study showed the same patterns of transfer of function with older adults as well. This finding also verifies the strength of classes formed by older adults, as the transfer of stimulus function within classes is a characteristic of natural categories as well. The results of this experiment also demonstrate the transfer of a new function (likelihood of infection), never before demonstrated in any previous literature. This function is one the experimenters feel has very real practicalities for the natural world; an example may be the shape of a hantavirus cell. Presented with a new cell with remarkably similar characteristics, doctors would be able to quickly identify some potential medicines that have been effective with other hantaviruses identified in the past. The ability of this function to transfer to not only stimuli within a learned class, but also to novel stimuli as yet unclassified is one that is very useful in the natural world, and thus one well worth studying in the laboratory.

In addition to demonstrating simple transfer of functions within classes, this study also aimed to determine whether the transfer would vary as a function of typicality effects. If typicality effects did impact the ratings of transfer to stimuli within the classes, one would expect to see those stimuli with fewer relevant features than the target stimulus to have a lower rating of infection than the target stimulus itself, and stimuli with more relevant features than the target stimulus to have a higher rating of infection than the target stimulus. This study approached this question by providing transfer of function tasks with both a four-feature target stimulus and a one-feature target stimulus. For the four-feature target stimulus, six of the younger adult participants rated stimuli that were class-consistent with the target stimulus as having a rate of infection of 50% (the rate given to the target stimulus) or lower. Four of the older adults demonstrated this pattern as well, reflecting an impact of typicality within ratings of transfer of function when given a prototype target stimulus. These results do demonstrate an impact of typicality given a prototype target stimulus; the majority of participants were more likely to rate stimuli with less than four relevant features as having a lower likelihood of infection than the target.

For the one-feature target scenario, results are a bit more mixed. The majority of both younger adults and older adult participants rated stimuli that were class-consistent with the target stimulus as having a 50% likelihood of infection (the same rate given to the target stimulus) or less. However, this is not necessarily the pattern of responding that would be predicted for this scenario; it is also possible that participants would rate comparison stimuli with more relevant features as having a higher rate of infection than the one-feature target stimulus. In fact, participants from both groups also demonstrated this type of responding. This finding does suggest some impact of typicality effect in the transfer of stimulus function within stimulus classes. One possible explanation of these findings could be that those comparison stimuli that had the same relevant feature as the target stimulus were rated as having a higher rate of infection than those stimuli without the feature present on the target stimulus, regardless of the number of other relevant features. However, further analyses showed this not to be the case for all participants. It may be that the two patterns of responding (rating the compared stimuli with more relevant features than the target stimulus as having a higher rate of infection versus rating the compared stimuli as having the same or lower likelihood of infection as the target) may have cancelled each other out when the participants were grouped together. Overall, participants in both groups did demonstrate typicality for both transfer scenarios, although the findings are more robust for the four-feature target scenario.

One explanation for this two-pattern finding may be evident in a closer examination of natural categories, where some features of a non-prototypical class member do not exist in more prototypical members of the class. Consider the natural class of birds, in which a robin may be a more typical member of the class and a penguin may be a more atypical member of the same class. Given the class "birds", one may expect that if a more prototypical member of the class has a feature (a beak), then most other members of the class would also have that same feature. However, it is not necessarily true that a feature of a less-typical member of the class (webbed feet) would be expectd to be seen in all members of the class. This phenomenon is discussed at length by Murphy (2002) in his review of category-based induction literature, in which he states that a characteristic of a prototype is much more likely to extend to an entire class than a characteristic of a less-typical member. The review includes a study by Rips (1975), which demonstrated category induction in a study using the natural category of birds. Rips' findings showed that when given a little-known fact about birds, participants were more likely to attribute the fact to other birds when the fact was tied to a prototypical bird (a robin) than when it was tied to a less typical bird (a duck). Murphy's review also includes a study by Osherson (1990) that identified *similarity* of the target and compared stimuli and *coverage* of the function across the class members as two important determinates of category induction. The results of the study showed that coverage of function to a class was more likely to stem from prototypical class members than from less typical class members. Murphy discusses other literature that supports these conclusions, as well as providing further elaboration in suggesting that prior knowledge of a characteristic (and how many members of a category contain that characteristic) may impact the induction of that characteristic to all members. These same phenomena are evident in the results found in this experiment. When told that a prototypical member of a class has a function, it is much more likely to exist in other members of the class than a function that exists on a less-prototypical member of the same class. Further, the similarity of the compared stimulus to the target stimulus impacted the similarity of the function transfer.

Overall, the findings in this section of the experiment did reliably demonstrate the transfer of stimulus function within the classes trained during the match-to-sample procedure. The impact of typicality within the ratings of transfer is somewhat mixed, with the typicality impact being stronger when the prototype is the target stimulus than when a one-feature stimulus is the target stimulus. These results held true for both younger adult participants and older adult participants, once again demonstrating that stimulus class properties are similar for both older and younger adults.

Future research may look further into older adult categorization abilities as a marker of age-related declines in cognitive functioning, as opposed to declines in physical functioning. As suggested by Johnson et al (1995), categorization abilities may be a marker of early Alzheimer's disease. If this is the case, than early intervention in this area may prove to be a successful means of retaining more abilities with age. More research in this area may investigate the categorization abilities of those adults in declining physical health but more stable mental health, as well as with adults showing declines in both physical and mental health areas. Investigating these ideas may shed more light on the area of aging and mental health more generally as well.

Future research may also further investigate the transfer of functions trained to non-prototypical members of stimulus classes. Although the results of this study are evident in more natural categories as well, it may be interesting to determine whether some stimulus functions may transfer from less typical to more typical members of a class, and which functions those might be. The transfer of stimulus function within generalized equivalence classes has become an area of interest more recently in this field, hopefully providing more evidence and understanding of the area as a whole.

The results of this experiment lend support to the idea of stimulus classes as a primary function of human behavior that does not decline with age. They also demonstrate evidence of transfer of stimulus function within arbitrary classes, showing that arbitrary classes do in fact demonstrate some of the characteristics evident in more natural categories. Thus, arbitrary classes do seem to be a valid way of studying more natural categories, in both older and younger adults.

REFERENCES

Belanich, J., & Fields, L. (2003). Generalized equivalence classes as response transfer networks. *Psychological Record, 53(3)*, 373-413.

Baron, A., & Journey, J. (1989). Age differences in manual versus vocal reaction times: Further evidence. *Psychological Sciences, 44(5)*, P157-159.

Baron, A., Menich, S.R., & Perone, M. (1983). Reaction times of younger and older men and temporal contingencies of reinforcement. *Journal of the Experimental Analysis of Behavior, 40*, 275-287.

Dougher, M.J., Augustson, E., Markham, M.R., Greenway, D.E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior, 62*, 331-351.

Dube, W.V (1991). Computer software for stimulus control research with Macintosh computers. *Experimental Analysis of Human Behavior Bulletin, 9*, 28-30.

Dymond, S., & Rehfeldt, R.A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst, 23*, 239-254.

Fields, L., Reeve, K.F., Adams, B.J., & Verhave, T. (1991). Stimulus generalization and equivalence classes: A model for natural categories. *Journal of the Experimental Analysis of Behavior, 55*, 305-312.

Fields, L., Reeve, K.F., Adams, B.J., Brown, J.L., & Verhave, T. (1996). Predicting the extension of equivalence classes from primary generalization gradients: The merger of equivalence classes and perceptual classes. *Journal of the Experimental Analysis of Behavior, 68*, 67-91.

Fields, L., Reeve, K.F., Rosen, D., Matneja, P., Varelas, A., Belanich, J., Fitzer, A., & Shamoun, K. (2002). The formation of a generalized categorization repertoire: Effect of training with multiple domains, samples, and comparisons. *Journal of the Experimental Analysis of Behavior, 78*, 291-313.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the linician. Journal of Psychiatric Research, 12, 189-198.

Galizio, M., Stewart, K.L., & Pilgrim, C. (2001). Clustering in artificial categories: An equivalence analysis. *Psychonomic Bulletin & Review, 8(3)*, 609-614.

Galizio, M., Stewart, K.L., & Pilgrim, C. (2004). Typicality effects in contingency-shaped generalized equivalence classes. *Journal of the Experimental Analysis of Behavior, 82*, 253-273.

Griffee, K., & Dougher, M.J. (2002). Contextual control of stimulus generalization and stimulus equivalence in hierarchical categorization. *Journal of the Experimental Analysis of Behavior, 78*, 433-447.

Hayes, S.C., & Bissett, R.T. (1998). Derived stimulus relations produce mediated and episodic priming. *The Psychological Record, 48*, 617-630.

Hayes, S.C., Fox, E., Gifford, E.V., Wilson, K.G., Barnes-Holmes, D., & Healy, O. (2001). Derived relational responding as learned behavior. In Hayes, S.C., Barnes-Holmes, D., & Roche, B. (Eds.). *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 21-49). New York: Kluwer Academic/Plenum Publishers.

Hess, T.M., & Slaughter, S.J. (1986). Aging effects on prototype abstraction and concept identification. *Journal of Gerontology, 41(2)*, 214-221.

Horne, P.J., and Lowe, C.F. (1996). On the origins of naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior, 68*, 271-296.

Johnson, M.K., Hermann, A.M., & Bonilla, J.L. (1995). Semantic relations and Alzheimer's disease: Typicality and direction of testing. *Neuropsychology, 9(4)*, 529-536.

Lowe, C.F., Horne, P.J., Harris, F.D.A., & Randle, V.R.L. (2002). Naming and categorization in young children: Vocal tact training. *Journal of the Experimental Analysis of Behavior, 78,* 527-549.

Lowe, C.F., Horne, P.J., & Hughes, J.C. (2005). Naming and categorization in young children: III. Vocal tact training and transfer of function. *Journal of the Experimental Analysis of Behavior, 83*, 47-65.

McDowd, J. M., & Shaw, R. J. (2001). Attention and aging: A functional perspective. In Craik, F. I. M., & Salthouse, T. A. (Eds.) *The Handbook of Aging and Cognition, Second Edition* (pp. 221-291). Mahwah, NJ: Lawrence Erlbaum Associate, Inc., Publishers.

Murphy, G. L. (2002). Typicality and the classical view of categories. In G.L. Murphy (Ed.) *The Big Book of Concepts*. (pp. 11-40), Cambridge, MA: MIT Press.

O'Donnell, J., & Saunders, K.J. (2003). Equivalence relations in individuals with language limitations and mental retardation. *Journal of the Experimental Analysis of Behavior, 80*, 131-157.

Perez-Gonzalez, L.A., & Moreno-Sierra, V. (1999). Equivalence class formation in elderly persons. *Psicothema, 11(2)*, 325-336.

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

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior, 74*, 127-146.

Sidman, M. (1994). *Equivalence Relations and Behavior: A Research Story.* Boston, MA: Authors Cooperative, Inc., Publishers.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*, 5-22.

Whelan, R., & Barnes-Holmes, D. (2004). The transformation of consequential functions in accordance with the relational frames of same and different. *Journal of the Experimental Analysis of Behavior, 82*, 177-195.

Wilson, K.M., & Milan, M.A. (1995). Age differences in the formation of equivalence classes. *Journal of Gerontology, 50B(4),* P212-P218.

Appendix

Source Tables for all Statistical Comparisons

Table A1.

Analysis of Variance for Pre-criterion Percent Correct Scores

Source	SS	df	MS	F	p
Feature	2269.71	3	756.57	5.3	0.0032
Age	0.03	1	0.03	0.1	0.9900
Feature*Age	120.07	3	40.02	0.3	0.8400
Error	6887.66	48	143.49		

Table A2.

Analysis of Variance for Pre-criterion Speed Scores

Source	SS	df	MS	F	p
Feature	0.01	3	0.01	0.2	0.8800
Age	0.55	1	0.55	122.3	0.0001
Feature*Age	0.01	3	0.01	0.1	0.9800
Error	0.22	48	0.01		

Table A3.

Analysis of Variance for Post-criterion Percent Correct Scores

Source	SS	df	MS	F	p
Feature	17.07	3	5.69	0.8	0.4800
Age	1.38	1	1.38	0.2	0.6600
Feature*Age	30.47	3	10.16	1.5	0.2300
Error	328.79	48	6.85		

Table A4.

Analysis of Variance for Post-criterion Speed Scores

Source	SS	df	MS	F	p
Feature	0.07	3	0.02	2.9	0.0470
Age	1.19	1	1.19	149.5	0.0001
Feature*Age	0.03	3	0.01	1.2	0.3100
Error	0.38	48	0.01		

Table A5.

Analysis of Variance for all Probe Trial Percent Correct Scores

Source	SS	df	MS	F	p
type	50.71	3	16.90	1.4	0.2700
age	1.62	1	1.62	0.1	0.7200
type*age	137.40	3	45.80	3.7	0.0180
Error	596.85	48	12.43		

Table A6.

Analysis of Variance for Unequal-feature Probe Trial Speed Scores

Source	SS	df	MS	F	p
feature	0.005	3	0.002	1.230	0.328
age	0.833	1	0.833	39.372	0.001
feature*age	0.007	3	0.002	1.692	0.204
Error	0.026	48	0.001		

Table A7.

Analysis of Variance for Equal-feature Probe Trial Speed Scores

Source	SS	df	MS	F	p
feature	0.19	3	0.06	11.0	0.0001
age	0.63	1	0.63	107.4	0.0001
feature*age	0.03	3	0.01	1.5	0.2400
Error	0.28	48	0.01		

Table A8.

Analysis of Variance for Consistent vs. Inconsistent Rating Scores

Source	SS	df	MS	F	p
class	4503.43	1	4503.43	13.0	0.0014
age	161.57	1	161.57	0.5	0.5000
class*age	67.64	1	67.64	0.2	0.6600
Error	8303.90	24	346.00		

Table A9.

Analysis of Variance Comparing Number of Features, Age, and Target Stimulus in Ratings Scores

Source	SS	df	MS	F	p
feature	8497.19	3	2832.40	7.9	0.0001
age	51.17	1	51.17	0.1	0.7100
target	3314.70	1	3314.70	9.2	0.0031
F*A	632.37	3	210.79	0.6	0.6300
F*T	-1955.31	3	-651.77	-1.8	1.0000
A*T	632.37	1	632.37	1.8	0.1900
F*A*T	3717.26	1	3717.26	10.3	0.0018
Error	34508.43	96	359.46		

Table A10.

Analysis of Variance for Stimuli With and Without the Target Feature

Source	SS	df	MS	F	p
feature	360.72	1	360.72	1.1	0.3000
age	37.72	1	37.72	0.1	0.7400
feature*age	27.01	1	27.01	0.1	0.7700
Error	7724.71	24	321.86		

Table A11.

Analysis of Variance for Pre-test and Post-test Sorting

Source	SS	df	MS	F	p
Sort	23075.38	1	23075.38	659.9	0.0001
Age	170.78	1	170.78	4.9	0.0370
Sort*Age	187.57	1	187.57	5.4	0.0290
Error	839.20	24	34.97		

Table A12.

Analysis of Variance for Post-test Class-Consistent Sorting

Source	SS	df	MS	F	p
Feature	211.51	3	70.50	171.5	0.0001
Age	0.22	1	0.22	0.5	0.4700
Feature*Age	1.77	3	0.59	1.4	0.2400
Error	19.73	48	0.41		