Clustering in artificial categories: An equivalence analysis

MARK GALIZIO, KATHERINE L. STEWART, and CAROL PILGRIM
University of North Carolina, Wilmington, North Carolina

Category clustering is a robust finding in the free recall of familiar category members, but has rarely been studied with artificial categories. In the present study, college students learned artificial categories via stimulus-equivalence methodology. Arbitrary match-to-sample training with nonsense syllables established three interrelated conditional discriminations, and, for most subjects, unreinforced test trials revealed the emergent stimulus-control relations considered to be evidence of equivalence classes. Free-recall tests revealed evidence of significant within-class clustering both before and after equivalence testing, but was more pronounced after the equivalence tests. These findings confirm that classic phenomena like clustering in free recall can be studied with stimulus-equivalence methodology, thus allowing for experimental control over relevant variables.

Perhaps no areas of modern psychology have been presented as pitted against each other so frequently as have cognitive psychology and behavior analysis. However, it is increasingly the case that these subdisciplines have overlapping interests and common methodological concerns. An important case in point is the study of natural language categories and stimulus equivalence.

Historically, the empirical study of categorization has been the province of cognitive psychology and psycholinguistics (e.g., Margolis & Laurence, 1999). For example, Bousfield's (1953) report of category clustering was one of the seminal events leading to a focus on organizational factors in memory. In that study, subjects were presented with words belonging to several semantic categories and, in a subsequent free-recall task, the words were recalled in clusters reflecting category membership regardless of the original presentation order. This classic finding has proven to be highly robust and has provided an important basis for interpretations of semantic organization in retrieval (e.g., Tulving & Donaldson, 1972). Still, many researchers have noted the difficulties inherent in studies involving natural language categories, the origins and character of which are beyond experimental control (e.g., Anderson & Bower, 1974; Rosch & Mervis, 1975). The extent to which uncontrolled factors play a role in the observation of clustering is unclear. Variables such as recency of exposure, familiarity, and category size and structure would seem to be of particular importance.

Recognizing these sorts of problems with the study of natural categories, Casey and Heath (1983) devised procedures

to investigate category clustering in artificial categories. Adolescents learned to categorize groups of "beings," labeled with trigrams, through stories about exploring an imaginary planet. The subjects showed clear evidence of category clustering when asked to recall the beings. The important conclusion reached by these authors was that classic findings with natural categories can be replicated with artificial ones, with the advantage of making possible the control over potentially critical dimensions of the categories being tested.

However, there remain limitations in procedures like those in the Casey and Heath (1983) study. For example, each of their category instances was directly instructed, unlike natural language categories in which novel exemplars may be readily categorized (see Margolis & Laurence, 1999). In addition, the use of stories required the subjects to have prior knowledge in order to establish the categories. Thus, much of the organizing of category members depended on subject experience that had occurred outside the laboratory, and was therefore uncontrolled.

The goal of the present study was to study category clustering under conditions in which relations between instances of the category were truly arbitrary and under experimental control. The stimulus-equivalence methodology developed by Sidman and colleagues (e.g., Sidman, 1994; Sidman & Tailby, 1982) represents an approach to the study of categorization that has attracted recent attention in the areas of behavior analysis and animal cognition (Zentall & Smeets, 1996). This methodology uses match-to-sample procedures to train arbitrary relations between unfamiliar stimuli, resulting in the emergence of novel relations indicative of stimulus class or category organization. It thus seems an ideal manner in which to establish artificial categories that are relatively free of extra-experimental factors.

Laboratory studies of stimulus equivalence typically begin with arbitrary match-to-sample training, in which

K. S. is now at the University of Kansas. This work was supported in part by NICHD Grant HD34265. The authors gratefully acknowledge the help of Abigail Newton, who assisted in data collection and analysis. Correspondence concerning this article should be addressed to Mark Galizio, Department of Psychology, University of North Carolina at Wilmington, Wilmington, NC 28403 (e-mail: galizio@uncwil.edu).

at least two interrelated conditional discriminations involving physically dissimilar stimuli are established. On any given trial, subjects are shown one of at least two possible sample stimuli and a group of at least two comparison stimuli. Choosing the correct comparison is reinforced, and the comparison designated as correct is conditional on the particular sample presented. For example, when shown sample stimulus A1 and comparison stimuli B1 and B2, the choice of comparison B1 is reinforced. When shown sample A2 and comparisons B1 and B2, the choice of B2 is reinforced. Similarly, subjects are trained with a new set of comparisons to choose C1 given an A1 sample, and C2 given an A2 sample.

After learning such a baseline, human subjects presented with novel trial types have reliably shown the emergence of stimulus relations that had not been explicitly reinforced. These novel performances have been designated as *stimulus reflexivity* (e.g., given A1 as sample, and A1 and A2 as comparisons, A1 is chosen; i.e., untrained relations are demonstrated between each stimulus and itself), *stimulus symmetry* (e.g., given B1 as sample, and A1 and A2 as comparisons, A1 is chosen; i.e., presenting stimuli in untrained roles reveals that the trained functions of sample and comparison stimuli are reversible), *stimulus transitivity* (e.g., given B1 as sample, and C1 and C2 as comparisons, C1 is chosen; i.e., relations are demonstrated between dissimilar stimuli that have never appeared together during training). Thus, after directly training a minimum of 4 stimulus relations (A1B1, A2B2, A1C1, A2C2), stimuli may become related to each other in ways that were not trained, and a total of 18 stimulus-control relations result (the 4 that were trained and A1A1, A2A2, B1B1, B2B2, C1C1, C2C2; B1A1, B2A2, C1A1, C2A2; B1C1, B2C2, C1B1, C2B2). If the stimuli become related to each other in this manner, they are termed *equivalence classes* (i.e., A1, B1, and C1 as one class; A2, B2, C2 as another), in that all elements are functionally substitutable within a given context (Sidman, 1994).

The equivalence phenomenon has received increasing levels of empirical attention over recent years for several important reasons. First, many theoretical interpretations of stimulus equivalence emphasize that equivalence classes formed in the laboratory provide a useful model of natural language categories (e.g., Hayes, Gifford, & Wilson, 1996; Pilgrim & Galizio, 1996; Sidman, 1994; Wilkinson, Dube, & McIlvane, 1998). These training and testing techniques have demonstrated an impressive generative capacity, the limits of which have yet to be revealed (e.g., Sidman, Kirk, & Willson-Morris, 1985). Indeed, applications of equivalence have had a major impact on language-skills training programs with special populations (Eikeseth & Smith, 1992; McIlvane, Dube, Kledaras, Iennaco, & Stoddard, 1990;). Finally, stimulus equivalence has been demonstrated in even very young children (e.g., Lipkens, Hayes, & Hayes, 1993; Pilgrim, Jackson, & Galizio, 2000) and in individuals with severe developmental disabilities (e.g., Sidman, 1994), and there is increasing interest in the possibility of equivalence

in nonhuman animals (e.g., Schusterman & Kastak, 1993; Urcuioli, 1996; Zentall, 1996). Thus, stimulus equivalence may permit the analysis of categories at their most fundamental level, avoiding the necessity of verbal interventions to create the classes (cf. Casey & Heath, 1983). In the present study, then, arbitrary match-to-sample procedures were used to establish equivalence classes, and free-recall tests assessed the extent to which class members would be recalled together in a cluster.

METHOD

Subjects

Two successive groups of introductory psychology students participated in two 1-h sessions, scheduled within at least 1 week of each other, in return for class credit. Group 1 included 36 students, and Group 2 included 30.

Apparatus

Stimuli were black nonsense trigrams (see Table 1) approximately 0.75×2.0 cm in size, presented on a white background on a Macintosh computer (30-cm diagonal screens), according to specialized match-to-sample programming (Dube, 1991). The sample trigram appeared in the center of the screen, and three comparison trigrams appeared in any three of the corners. The manipulation of a mouse moved a cursor on the screen, and the clicking of the left mouse button when the cursor was situated on or near a trigram registered a response.

Procedure

The procedure included three components: baseline training, equivalence testing, and free-recall testing. Each subject was tested individually. Minimal instructions specified that the subject was to "click on the center trigram to produce additional trigrams in the corners," that they should then "click on a corner trigram," and that "stars are worth one point, buzzers subtract one point, and the task is to earn as many points as you can." The subjects were further instructed that sometimes clicks would produce no feedback and that when this happened no points were earned or lost.

Baseline training. Each trial began with the presentation of a sample stimulus, and, following an observing response (i.e., mouse click), three comparison stimuli appeared immediately and remained on the screen until a selection was made. Following a response designated as correct, a brief fanfare sounded during which colored stars traversed the screen, and then the screen darkened for 1.5 sec before the next trial was presented. Following a response designated as incorrect, a buzzer sound was produced, and the screen immediately went blank for 1.5 sec. Baseline training was designed to establish three conditional discriminations (AB, AC, and CD) in a three-choice simultaneous arbitrary match-to-sample procedure. (The CD conditional discrimination was included to allow for exploration of larger equivalence classes.) Table 1 presents the specific trigrams used during training and their alpha-numeric designations, for ease in exposition. Half the subjects were exposed to Set 1 and half to Set 2.

Table 1
Nonsense Syllables and Alphanumeric Class Designations

Letter	Class					
	Set 1			Set 2		
	1	2	3	1	2	3
A	wug	niz	jom	yec	xil	sar
B	sar	leb	rup	rup	paf	keh
C	fid	paf	keh	niz	leb	tuj
D	yec	tuj	xil	jom	wug	fid

Initial trial blocks consisted of a baseline of nine trial types (i.e., A1: B1B2B3, A2: B2B3B1, A3: B3B1B2, A1: C1C2C3, A2: C2C3C1, A3: C3C1C2, C1: D1D2D3, C2: D2D3D1, C3: D3D1D2), in which the first alphanumeric represents the sample stimulus, the second represents the correct comparison stimulus, and the third and fourth represent incorrect comparison stimuli. Within a block, each trial type was presented an equal number of times in a quasi-random sequence, with the constraints that no sample appeared on more than three trials in succession and no comparison stimulus appeared in the same location for more than two trials in succession. Comparisons appeared in each location equally often. Upon meeting mastery criterion (i.e., 14 of 16 trials correct), to prepare subjects for the absence of feedback on test trials, the reinforcement rate was reduced to 75% and then to 50%, with unreinforced trials equally distributed across all trial types and the same mastery criteria in effect. Thus, to complete baseline training, subjects were required to meet the mastery criteria on three separate blocks of randomly ordered trial types.

Equivalence testing. Trial blocks in this phase included 6 test trials interspersed among 18 baseline trials (2 trials of each type). No 2 test trials ever occurred in succession, and no feedback was ever given on any test trial. The reinforcement percentage for baseline trials was increased to 67% during test blocks to keep the overall percentage at 50%. Trial blocks tested for either symmetry, or transitivity and combined symmetry/transitivity. Symmetry trial blocks included the following test trials: B1: A1A2A3, B2: A2A3A1, B3: A3A1A2, C1: A1A2A3, C2: A2A3A1, C3: A3A1A2, D1: C1C2C3, D2: C2C3C1, D3: C3C1C2. Transitivity and combined test trial blocks included B1: C1C2C3, B2: C2C3C1, B3: C3C1C2, A1: D1D2D3, A2: D2D3D1, A3: D3D1D2, B1: D1D2D3, B2: D2D3D1, and B3: D3D1D2.

Recall test. For each recall test, the subjects were given a sheet of paper that instructed them to write down as many of the trigrams as they could remember in the blank space below the instructions. The experimenter observed the order in which each trigram was

written on the data sheet and numbered them accordingly upon completion of the task. The subjects who produced fewer than five items on the recall test were dropped from the study.

Design and data analyses. For the subjects in the condition tested initially (Group 1), equivalence testing continued until the end of the initial 50-min session. The second session began with a baseline-review trial block at 50% reinforcement; equivalence testing blocks were re-introduced as soon as mastery criteria were met on the review. Equivalence testing was terminated after 45 min and was followed by the free-recall test. Because an active role for exposure to test trials has sometimes been postulated (e.g., Sidman, 1994), the subjects in Group 2 were given the free-recall test immediately upon meeting the mastery criterion for initial training blocks. The recall test ended Session 1. Session 2 was conducted as it had been for Group 1, ending with a second free-recall test after exposure to both training and probe trials.

Equivalence-test performances were analyzed for the final 24 test trials and were summarized in terms of the percentage of test trials on which the class-consistent comparison stimulus was selected. Thus, a score of 100% would reflect 12 class-consistent responses on symmetry, transitivity, and combined test trials. Because each trial had three comparisons, scores of 33% represented chance performance.

Clustering was characterized in terms of the number and length of runs, where the length of a run is defined as the number of trigrams recalled in succession for a given class. The degree of clustering was quantified using the adjusted ratio of clustering¹ (ARC; Roenker, Thompson, & Brown, 1971). For each recall list, scores of 0 represent chance clustering and scores of 1 represent perfect clustering (e.g., a list with the following sequence: A1B1C1D1A2B2C2D2A3B3C3D3). Two ARC scores were calculated for each subject's recall, one for each of the two training sets. This created one ARC score for runs consistent with the training actually received by the subject (i.e., a class-consistent ARC score) and a control score for runs consistent with the alternative stimulus set (i.e., a class-inconsistent ARC score).

Figure 1. Mean ARC scores for the groups trained with Set 1 (left bars) and Set 2 (right bars) relations. Solid bars show ARC scores computed on the basis of the classes that were consistent with the subject's training, and striped bars show class-inconsistent ARC scores. Vertical lines indicate standard errors of the mean.

Results

Group 1. Thirty-four of the 36 subjects in Group 1 met the acquisition and recall criteria. Figure 1 shows the mean ARC scores for the 17 subjects trained with Set 1 stimuli (leftmost bars) and for the 17 trained with Set 2 stimuli (rightmost bars). Class-consistent ARC scores (solid bars) were significantly higher than class-inconsistent ARC scores (striped bars) [$F(1,32) = 14.29$, $p < .01$]. Because the 2×2 analysis of variance (ANOVA) that provided this F ratio might be questioned on the basis of the independence of the within-group comparisons, t tests were also conducted. These showed that clustering with respect to a given stimulus set was significantly higher for the set that was consistent with the subject's training than for the set that was inconsistent with the subject's training (for Set 1, $t = 3.47$, $p < .01$; and for Set 2, $t = 2.78$, $p < .01$). In order to compare clustering to chance (an ARC score of zero), a binomial test was conducted in which class-consistent ARC scores of less than or equal to 0 were placed in one category and scores greater than 0 were placed in the other ($z = 2.23$, $p < .05$). Thus, ARC score analyses revealed strong evidence for clustering with respect to the experimentally determined classes.

Most subjects also showed evidence of equivalence categories, as indicated by accurate test performances. Total amount of training prior to presentation of the test trials was relatively limited, and mean accuracy across all test trials for these subjects was 77.2%. Eight subjects showed perfect scores, and 16 had scores greater than 90%. All but 1 subject responded at greater than chance levels (i.e., 33% class-consistent).

Group 2. Twenty-four of the 30 students who began the experiment met all criteria (12 trained with each set). Figure 2 shows the mean ARC scores for Test 1 (left panels) and Test 2 (right panels). On Test 1, class-consistent ARC scores (solid bars) were significantly higher than class-inconsistent ARC scores (striped bars) [$F(1,22) = 6.47$, $p < .05$]; t tests showed that clustering with respect to a given stimulus set was greater for the set consistent with the subject's training than for the set that was inconsistent, but the difference was significant only for the Set 1 stimuli [for Set 1, $t(22) = 2.61$, $p < .01$; for Set 2, $t(22) = 1.44$, $p > .05$], and the binomial test was not significant ($p > .05$).

For Test 2, there was stronger evidence of clustering, with class-consistent ARC scores again higher than class-inconsistent scores [$F(1,22) = 30.41$, $p < .01$]. Both

Figure 2. Mean ARC scores obtained from free-recall tests before equivalence testing (left graph) and after equivalence testing (right graph). The group trained with Set 1 is represented by the left bars in both graphs, and the group trained with Set 2 is represented by the right bars. Solid bars show ARC scores computed on the basis of the classes that were consistent with the subject's training, and striped bars show class-inconsistent ARC scores. Vertical lines indicate standard errors of the mean.

t tests and binomial tests confirmed significant clustering effects for both stimulus sets [for Set 1, $t(22) = 5.07$, $p < .01$; for Set 2, $t(22) = 4.02$, $p < .01$; $z = 4.69$, $p < .01$]. A t test comparing class-consistent ARC scores across the two tests revealed greater overall clustering in Test 2 than in Test 1 [$t(22) = 2.04$, $p < .05$]. Thus, ARC score analyses revealed some evidence for clustering both prior to and following exposure to test trials, although the extent of clustering was much greater in the second test.

Equivalence-test performances again demonstrated evidence of class formation for most subjects. The mean percentage of class-consistent responses on test trials across all subjects was 76.0%. All subjects showed greater than chance accuracy on the test trials, and 9 responded at 90% or better.

DISCUSSION

The analyses indicate that category clustering can occur after training with arbitrary stimulus relations. Although clustering was stronger after both training and equivalence testing, at least some subjects in Group 2 showed evidence of clustering on the first recall test, indicating that exposure to test trials was not essential for clustering to occur. Thus, classes evidenced on these free-recall tests can be seen as independent of many indirect sources of stimulus control (e.g., an instructive role of repeated testing).

The testing in the present study was relatively brief, and neither clustering nor equivalence-class formation was evident in all subjects. Equivalence performances have often been shown to emerge gradually upon extended testing (see Sidman, 1994). The greater degree of clustering observed in the second recall test for Group 2 suggests that repeated testing may have similar effects with this measure. More extended exposure to testing conditions might enhance the degree of both clustering and equivalence class formation.

Collectively, these data demonstrate the usefulness of stimulus-equivalence methodology for the study of category clustering. The robust nature of clustering is underscored by demonstrating its occurrence in purely arbitrary artificial categories (cf. Casey & Heath, 1983), the basis of which involved nonverbal training contingencies exclusively. The present results illustrate that such a contingency-based history is sufficient to provide for the structure of these categories; that is, the stimuli do not have to be tied to the subject's prior knowledge of the world or existing natural kinds (Heit, 1997; Keil, 1989; Murphy & Medin, 1985) for clustering to occur. This point seems particularly striking, given that the subjects here recalled together unfamiliar category members that had never been directly related during training. Stimulus equivalence is defined in terms of such emergent relations (i.e., symmetry and transitivity) and thus appears to hold much in common with natural language phenomena (Sidman, 1994). An additional advantage of

equivalence methodologies for the study of phenomena like category clustering is that relevant variables can be directly controlled and manipulated, and their analysis can thus be furthered.

REFERENCES

- ANDERSON, J. R. & BOWER, G. H. (1974). *Human associative memory*. Washington, DC: Wiley.
- BOUSFIELD, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, **49**, 229-240.
- CASEY, P. J., & HEATH, R. A. (1983). Categorization reaction time, category structure, and category size in semantic memory using artificial categories. *Memory & Cognition*, **11**, 228-236.
- DUBE, W. V. (1991). Computer software for stimulus control research with Macintosh computers. *Experimental Analysis of Human Behavior Bulletin*, **9**, 28-30.
- EIKESETH, S., & SMITH, T. (1992). The development of functional and equivalence classes in high-functioning autistic children: The role of naming. *Journal of the Experimental Analysis of Behavior*, **58**, 123-133.
- HAYES, S. C., GIFFORD, E. V., & WILSON, K. G. (1996). Stimulus classes and stimulus relations: Arbitrary applicable relational responding as an operant. In T. R. Zentall & P. M. Smeets, *Stimulus class formation in humans and animals*. The Netherlands: Elsevier.
- HEIT, E. (1997). Knowledge & concept learning. In K. Lamberts & D. Shanks (Eds.) *Knowledge, concepts and categories* (pp. 7-42). Cambridge, MA: MIT Press.
- KEIL, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- LIPKENS, R., HAYES, S. C., & HAYES, L. J. (1993). Longitudinal study of the development of derived relations in an infant. *Journal of Experimental Child Psychology*, **56**, 201-239.
- MARGOLIS, E., & LAURENCE, S. (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.
- MCILVANE, W. J., DUBE, W. V., KLEDARAS, J. B., IENNACO, F. M., STODDARD, L. T. (1990). Teaching relational discrimination to individuals with mental retardation: Some problems and possible solutions. *American Journal on Mental Retardation*, **95**, 283-296.
- MURPHY, G. L., & MEDIN, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316.
- PILGRIM, C., & GALIZIO, M. (1996). Stimulus equivalence: A class of correlations or a correlation of classes. In T. R. Zentall & P. M. Smeets, *Stimulus class formation in humans and animals*. The Netherlands: Elsevier.
- PILGRIM, C., JACKSON, J., & GALIZIO, M. (2000). Acquisition of arbitrary conditional discriminations by young normally developing children. *Journal of the Experimental Analysis of Behavior*, **73**, 177-193.
- ROENKER, D. L., THOMPSON, C. P., & BROWN, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, **76**, 45-48.
- ROSCH, E., & MERVIS, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.
- SCHUSTERMAN, R. J., & KASTAK, D. (1993). A California sea lion (*Zalophus californianus*) is capable of forming equivalence relations. *Psychological Record*, **43**, 823-839.
- SIDMAN, M. (1994). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.
- SIDMAN, M., KIRK, R., & WILLSON-MORRIS, M. (1985). Six-member stimulus classes generated by conditional-discrimination procedures. *Journal of the Experimental Analysis of Behavior*, **43**, 21-42.
- 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.
- TULVING, E., & DONALDSON, W. (1972). *Organization of memory*. New York: Academic Press.
- URCUIOLI, P. J. (1996). Acquired equivalences and mediated generalization in pigeon's matching-to-sample. In T. R. Zentall & P. M. Smeets,

Stimulus class formation in humans and animals. The Netherlands: Elsevier.

WILKINSON, K. M., DUBE, W. V., & MCILVANE, W. J. (1998). Fast mapping and exclusion (emergent matching) in developmental language, behavior analysis, and animal cognition research. *Psychological Record*, **48**, 407-422.

ZENTALL, T. R. (1996). An analysis of stimulus class formation in animals. In T. R. Zentall & P. M. Smeets, *Stimulus class formation in humans and animals*. The Netherlands: Elsevier Science.

ZENTALL, T. R., & SMEETS, P. M. (Eds.) (1996). *Stimulus class formation in humans and animals*. The Netherlands: Elsevier.

NOTE

1. The formula used to compute the adjusted ratio of clustering (ARC) score was based on Roenker, Thompson, and Brown (1971) and was as follows:

$$ARC = (R - E(R)) / (\max R - E(R)),$$

where

R = total number of observed class repetitions (i.e., the number of times a trigram followed another from the same class),

$\max R$ = maximum possible number of class repetitions—that is, total number of trigrams recalled minus the number of classes (3), and

$E(R)$ = expected (chance) number of class repetitions, or

$$E(R) = \sum_i n_i^2 / N - 1,$$

where

n_i = number of trigrams recalled from class i , and N is as above (total number of trigrams recalled).

(Manuscript received March 21, 2000;
revision accepted for publication October 18, 2000.)