JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2004, **82,** 253–273 NUMBER 3 (NOVEMBER)

TYPICALITY EFFECTS IN CONTINGENCY-SHAPED GENERALIZED EQUIVALENCE CLASSES

MARK GALIZIO, KATHERINE L. STEWART, AND CAROL PILGRIM

UNIVERSITY OF NORTH CAROLINA AT WILMINGTON

Two experiments were conducted using match-to-sample methodologies in an effort to model lexical classes, which include both arbitrary and perceptual relations between class members. Training in both experiments used a one-to-many mapping procedure with nonsense syllables as samples and eight sets of abstract stimuli as comparisons. These abstract stimuli differed along a number of dimensions, four of which were critical to the experimenter-defined class membership. Stimuli in some comparison sets included only one of the class-defining features, but stimuli in other sets included two, three, or all four of the critical features. After mastery of the baseline training, three types of probe tests were conducted: symmetry, transitivity/equivalence, and novel probe tests in which the training nonsense syllables served as samples, and comparisons were novel abstract stimuli that included one or more of the class-defining features. Symmetry and transitivity/equivalence probe tests showed that the stimuli used in training became members of equivalence classes. The novel stimuli also became class members on the basis of inclusion of any of the critical features. Thus these probe tests revealed the formation of open-ended generalized equivalence classes. In addition, typicality effects were observed such that comparison sets with more critical features were learned with fewer errors, responded to more rapidly, and judged to be better exemplars of the class. Contingency-shaped stimulus classes established through a match-to-sample procedure thus show several important behavioral similarities to natural lexical categories.

Key words: stimulus equivalence, categorization, concept learning, stimulus control, typicality effects, prototypes, adult humans

Sidman and Tailby's (1982) seminal analysis of stimulus equivalence showed that conditional discrimination contingencies could produce stimulus classes with language-like emergent properties. Although there is considerable theoretical debate on whether these phenomena are best viewed as relational frames (e.g., Hayes, Barnes-Holmes & Roche, 2001), the product of a special "naming" relation (e.g., Horne & Lowe, 1996), or as a fundamental process (e.g., Sidman, 1994, 2000), there *is* great agreement within the behavior-analytic community that these approaches are providing new insights into verbal behavior, particularly the control exerted by complex stimulus classes like natural language concepts or categories.

Unfortunately, behavior-analytic approaches do not appear to have had much impact on conceptualizations of categorization outside of behavior analysis. For example, a recent interdisciplinary collection of classic and current papers on concepts (Margolis & Laurence, 1999) presents over two dozen papers from philosophy, linguistics, and cognitive psychology. Concepts and categories are considered from a variety of theoretical perspectives including prototype theories, exemplar theories, classic definitional theories, theories that posit the innateness of categories (e.g., Fodor, 1990), and theories in which categories are defined in terms of "essence placeholders" (e.g., Medin & Ortony, 1989), but recent developments in behavior analysis are not represented. One explanation for this neglect may be that cognitive psychologists are simply unaware of these developments. The following quotation from the introduction of the book supports this: "Yet another alternative [to the view of concepts as "mental particulars," "structured mental representations," "abstract entities," or "essence placeholders"] is the view that concepts are . . . behavioral abilities. We take it that behavioral abilities are ruled out for the same reasons that argue against behaviorism in general (see Chomsky, 1959)" (p. 6).

Katherine Stewart is now at the Kansas University Parsons Research Center. This work was supported by NICHD grant HD34265. Experiment 1 formed part of the second author's master's thesis. The authors gratefully acknowledge the help of Tracy Peña and Robyn Stout in data collection and analysis. Helpful comments on the manuscript were provided by the UNCW stimulus control laboratory group, especially Maureen Aro.

Address correspondence to Mark Galizio, Department of Psychology, University of North Carolina at Wilmington, Wilmington, North Carolina 28403 (e-mail: galizio@uncw.edu).

Certainly if Chomsky's (1959) review of Skinner's (1957) *Verbal Behavior* is seen as the final word on behavioral approaches, a great deal has been missed. But another possibility is that the relevance of behavior-analytic work on topics related to categorization has not always been made clear to researchers outside of our field. For example, Harnad (1996) argued that equivalence methodologies lack relevance to problems of language and symbolization because the classes formed involve arbitrary relations, and in his view, an important aspect of natural language categories is that they involve perceptual relations—the abstraction of invariant properties of stimuli. The significance of this to Harnad is that once the category is learned through "honest toil" (i.e., through exposure to the contingencies), the subject is now able to categorize novel stimuli that possess the relevant properties, and to combine categories through what Harnad calls "symbolic theft." Thus a child who has learned to name horses can respond appropriately when exposed to a horse that she has never seen before. Moreover, if she has learned to categorize on the basis of stripes, she can identify her first zebra upon simply being told that it is a horse with stripes. As Harnad puts it: "that's symbolic/propositional theft; it can spare you an awful lot of honest toil; and it is the true power of language" (p. 264). Harnad sees equivalence research as irrelevant because it focuses on arbitrary classes, which lack much of this power.

Perhaps such confusions over issues of external validity follow from the behavior-analytic emphasis on experimental control. Behavior analysts have a long tradition of viewing concepts and categories in terms of basic behavioral principles. Keller and Schoenfeld (1950) defined concepts as involving generalization within, and discrimination between classes of stimuli. The procedures used to study stimulus equivalence represent one method of providing for discrimination between classes (or class partition; Sidman, 1994, 2000). In this instance, researchers have traditionally used arbitrary classes in order to provide a relatively pure experimental preparation—not for modeling purposes. Nothing in any of the behavior-analytic theories requires that stimulus class members must be physically unrelated. It is simply easier to identify the source of emergent relations when they are. It seems evident that equivalence classes would include not merely the trained stimuli, but through stimulus generalization would also include physically similar stimuli, and Fields and his associates have provided elegant experimental demonstrations of generalized equivalence classes (see Fields & Reeve, 2000 for a review). However, primary stimulus generalization is not sufficient to account for many examples of categorization. In many instances of categorizing, such as those described by Harnad (1996), control by abstracted stimulus properties, or combinations thereof, is required (see Herrnstein, 1990).

For example, many lexical classes have been described as "family resemblance" classes because class members may be characterized by one or more related features or properties, but no particular feature is necessarily common to all class members. Such classes are often characterized by typicality effects: that is, more typical class members—perhaps those with a greater number of family features—are responded to differently than less typical class members (Rosch & Mervis, 1975). In a number of experiments, stimuli rated as most representative of the category (prototypes) are more frequently produced in recall tests (Mervis, Catlin, & Rosch, 1976), are categorized more rapidly (Murphy & Brownell, 1985; Rips, Shoben, & Smith, 1973), and are more rapidly learned (Rosch, Simpson, & Miller, 1976) among other typicality effects. Indeed, such typicality effects have come to be considered a hallmark of lexical classes and have played a critical role in the development of cognitive theories of concepts and categories (see Murphy, 2002, for a review).

Stimulus equivalence methodology can be readily adapted to the study of family resemblance stimulus classes. In the present study we sought to determine whether generalized equivalence classes with family resemblance characteristics could be developed by arbitrary match-to-sample training. We wondered whether subjects would show symbolic theft in reaction to novel stimuli after such training, and whether parallels to the typicality effects noted with lexical categories would be observed. It is important for behavior analysts to develop experimental paradigms for the production and analysis of complex, naturally occurring phenomena, and the present study represents such an effort.

EXPERIMENT 1

METHOD

Subjects

A total of 19 undergraduate students from the University of North Carolina at Wilmington participated in the study. All subjects received credit to fulfill introductory psychology course requirements for their first experimental session. If eligible, subjects also received course credit for their second session, but if not, they were paid $5 for the second session. One subject who failed to meet the baseline training criteria was dropped from the study leaving 18 subjects who completed the experimental protocol.

Apparatus

Subjects were individually trained and tested on color Macintosh® computers with the use of match-to-sample software developed by Dube (1991). Stimuli were black nonsense trigrams approximately 7.5 mm by 20 mm in size, and black and white abstract drawings created with standard drawing software approximately 40 by 40 mm in size (see Figure 1). Subjects were tested individually in a small quiet room for two 1-hr sessions separated by no more than 1 week.

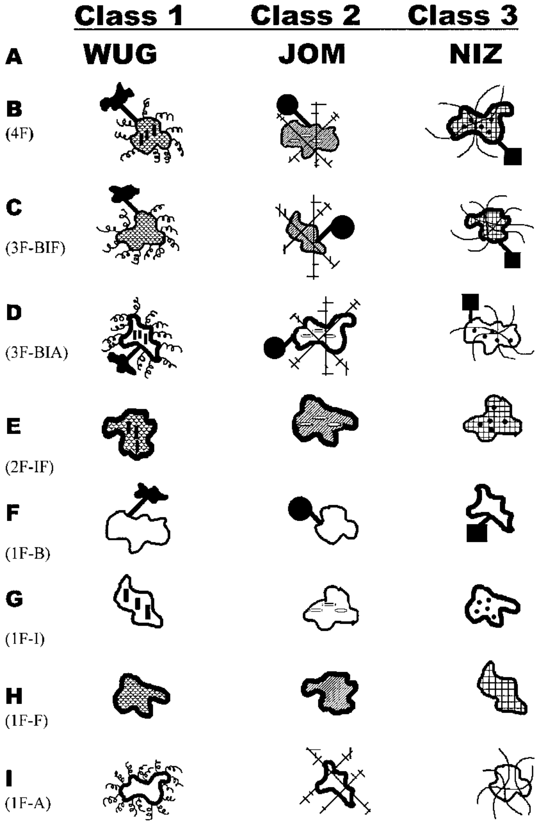

Fig. 1. Training stimuli used in Experiment 1. The A row of stimuli shows the trigrams that served as samples on training trials. Rows B through I show the stimuli that served as comparisons with each row revealing the comparison stimuli for a particular trial type. (1F, 2F, 3F, and 4F = one, two, three, and four critical features, respectively; B = base; I = inserts; F = fill; A = appendages.)

Procedure

The procedure included three phases: baseline training, equivalence and novel probe testing, and a sorting task. All subjects were instructed to look at the sample stimulus in the center of the screen and to "use the mouse to position the cursor on it and click" to produce other stimuli. Subjects also were instructed to pick the comparison stimulus that "goes with" the sample stimulus by clicking on it, that colored stars and music were worth one point, that a buzzer subtracted a point, and that they were to try to earn as many points as possible. The subjects were further instructed that sometimes responses would not produce feedback and that when this happened no points were added or subtracted from their score. Finally, subjects were told that the experimenter was interested in how rapidly they could make their choice and that "once you learn which objects go with each syllable, it is essential that you make your choices as quickly as you can."

Baseline training. Each trial began with the presentation of a sample stimulus in the center of the screen. Following the observing response (i.e., a mouse click on the sample), three comparison stimuli appeared immediately in the corners of the screen and remained on the screen until a selection was made. Following a response designated as correct, a brief fanfare sounded accompanied by the appearance of numerous colored stars on the screen. Following an incorrect response, a buzzer sounded. In either case, the screen went blank for a 1.5-s intertrial interval followed by the presentation of the next sample. Figure 1 illustrates the stimuli and also indicates the baseline trial types. One of the nonsense syllables (wug-A1, jom-A2, niz-A3) always served as the sample, and one of the eight stimulus sets (B through I) was presented as the three comparisons. Comparison stimuli differed along seven dimensions, four of which defined the contingencies of the conditional discrimination training. The critical features were fill (F), inserts (I), appendages (A), and base (B), seen in Figure 1; each could be presented as the only critical feature, or could appear in combinations of two, three, or four critical features. For example, when A1 (wug) was the sample, clicking on stimuli with scales for fill, bars for inserts, looped appendages, or an irregular base, alone or in any combination (stimuli B1 through I1), produced the display of stars (point gain). Responding to other comparison stimuli produced the buzzer (point loss). When A2 (jom) was the sample, the class-consistent features were a striped fill, elliptical inserts, straight appendages, and a circular base; and for A3 (niz), they were a checkered fill, dots for inserts, curved appendages, or a square base. Irrelevant features included the shape of the figure, thickness of outline, location of the base, and location and number of appendages and inserts. For example, to ensure that the shape of the stimuli would not provide a basis for correct responses, one set of eight arbitrary, abstract shapes was used. The eight different shapes were used equally often (three times) with each of the three trained classes. For example, stimuli B1, F2, and I3 all had the same shape (Figure 1). In addition, the shape outlines had two possible thickness values, and each was used in half of the stimuli of each trained class. Only 8 of the 15 possible arrangements of the four defining features were used. These eight arrangements ensured that each defining feature was presented the same number of times, to control for the amount of experience with each feature (i.e., each defining feature was represented in four of the comparison arrays). As Figure 1 shows, the B comparison array had all four of the class-consistent features, and thus could be described as prototypes. The C and D arrays included three of the four defining features (C: fill, appendages, and base; D: insert, appendages, and base). The E comparison array had a combination of two class-consistent features, fill and insert, and the F, G, H, and I comparison arrays included stimuli that had only one of the four defining features (F, base; G, inserts; H, fill; and I, appendages).

Training was arranged in blocks of 24 trials with each block including one each of the possible trial types illustrated in Figure 1 presented in random order (i.e., each comparison array was presented once with each of the three sample stimuli). The positions of the comparison stimuli were randomized throughout any given trial block with the constraints that for any two consecutive trials no one screen position was correlated with reinforcement, no one stimulus could occur in a given screen position, and no sample stimulus could be repeated. When subjects completed a block of trials, the next block began without interruption until a mastery criterion of two consecutive trial blocks with 22 of 24 trials correct (i.e., choices designated as correct were made on 22 of the 24 trials of that block) was met. To prepare subjects for the absence of feedback on probe trials, additional baseline training blocks were then programmed with reinforcement densities reduced to 75% and then to 50%, with no-reinforcement trials distributed randomly across trial types. Subjects had to complete one block at each reinforcement density with 22 of 24 trials correct to move to the next phase. Thus, to complete baseline training, subjects had to meet the mastery criteria on four separate training blocks.

Symmetry, equivalence, and novel stimulus probe tests. Following completion of baseline training, subjects were exposed to a series of probe-trial blocks that were designed to test for symmetry, equivalence, and generalized control by class-consistent features of novel stimuli. Each trial block contained only one of these three probe types intermixed with baseline trials, and subjects cycled through the three block types for the duration of the session. In order to minimize the effect of trial order effects, for each block type, trials were presented in one of three different randomized orders, called Orders A, B, and C. One third of the subjects were exposed to Order A first, whereas one third was exposed to Order B first and another third to Order C. After completing one set of probe blocks, subsequent blocks rotated through the three different orders of trial presentation. The first experimental session ended after 50 min or completion of two sets of all probe-trial blocks, whichever came first. In the second session, subjects were again required to meet criterion on baseline-only trial blocks before probe trials were presented, and completion of two sets of all probe-trial blocks or 40 min marked the end of computer testing. All subjects completed a minimum of two sets of probe-trial blocks and some completed as many as six probe-trial blocks.

Symmetry-trial blocks consisted of 48 probe trials in which each of the 24 stimuli that served as comparisons during training was presented as the sample with the trigrams wug, jom, and niz as comparisons. Thus all possible symmetry-trial types were represented twice in each trial block. Symmetry blocks included 72 baseline trials for a total of 120 trials. Feedback was never given for probe trials, and for symmetry blocks, feedback was also withheld for 12 of the baseline trials (distributed across baseline relations), which maintained a 50% overall block-reinforcement density.

Equivalence trial blocks consisted of 30 probe and 48 baseline trials. To maintain an overall block-reinforcement density of 50%, no feedback was provided for nine of the baseline trials. On equivalence probe trials, one of the baseline comparison stimuli was the sample (e.g., D1) with one of the baseline comparison sets as comparisons (e.g., H1, H2, H3). There were 56 possible transitive relations (i.e., 168 possible equivalence trials) given the baseline training. However, to keep these probes as similar as possible to those in traditional stimulus-equivalence testing (i.e., procedures involving arbitrary stimuli), only relations that involved stimuli without common class-relevant features were tested. For example, if the sample stimulus was H1, for which fill was the only class-consistent feature, then only the D, F, G, or I stimulus sets (which lack this feature) could serve as comparisons. The following 10 relations were tested: DH, EI, FE, FG, GC, GI, HF, HG, IF, and IH. This allowed each one- and two-feature stimulus to serve as both a sample and comparison.

A third type of probe trial of interest was made possible by the training structure. A group of novel stimuli with new combinations of the class-consistent and class-inconsistent features served as comparisons. Table 1 provides a description of each novel probe stimulus, and Figure 2 illustrates some representative novel probes (all stimuli are illustrated on the website, http://people.uncw.edu/galizio/galizio.htm—follow links to equivalence research page). There were four sets of novel stimuli that included just one class-consistent feature (Sets J through M in Table 1), one set for each feature (12 one-feature stimuli because each set included one stimulus from each class, wug, jom, and niz). Figure 2 illustrates one of the one-feature wug-class stimuli (upper left). Note that although this stimulus shares the insert feature with several of the wug-class training stimuli, its shape is different from any used in baseline (cf. Figure 1). There were six sets of novel stimuli with two class-consistent features (18 two-feature stimuli). These included one set of stimuli with fill and inserts, the two-feature combination used in baseline (Set N in Table 1), but with novel shapes, insert number, and insert placements. The remaining two-feature stimuli included combinations of class-consistent features not used in baseline, such as inserts and appendages (see Figure 2 and Sets O through S in Table 1). Similarly, four novel three-feature stimulus sets (12 stimuli) were used. Two included novel combinations of class-consistent features (fill, inserts, and base and fill, inserts, and appendages; Sets T and U in Table 1 and see Figure 2 for an example) and the other two involved combinations of class-consistent features used in baseline with novel irrelevant features (Sets V and W). Finally, three new four-feature stimulus sets (nine stimuli) were used to maintain relative balance with the number of stimuli presented with one-, two- and three-features. These included all four class-consistent features, but their shapes, outline thickness, position of inserts, and placement of appendages differed from the baseline four-feature stimuli (Sets X through Z in Table 1 and see Figure 2 for an example). Thus there were a total of 51 novel stimuli (17 from each class).

The format for a novel-stimulus probe trial was similar to a baseline-training trial except that feedback was never provided. The sample was always one of the nonsense syllables and comparisons were three-choice arrays of novel stimuli from one of the sets described

Table 1

Description of stimuli used in Experiment 1. Each stimulus is defined in terms of the four relevant features and two of the irrelevant features.

	Class 1					Class 2					Class 3				
Stimulus sets	Fill	Insert	Appendage	Base	Shape	Fill	Insert	Appendage	Base	Shape	Fill	Insert	Appendage	Base	Shape
Training stimuli															
A	—	—	—	—	Wug	—	—	—	—	Jom	—	—	—	—	Niz
B(4F)	Scale	Bars	Loops	Irr	A−	Stripe	Oval	Lines	Circ	C−	Check	Dots	Curve	Sq	B+
C(3F)	Scale	—	Loops	Irr	G−	Stripe	—	Lines	Circ	E−	Check	—	Curve	Sq	F+
D(3F)	—	Bars	Loops	Irr	H+	—	Oval	Lines	Circ	B+	—	Dots	Curve	Sq	C−
E(2F)	Scale	Bars	—	—	F+	Stripe	Oval	—	—	D+	Check	Dots	—	—	G−
F(1F)	—	—	—	Irr	C−	—	—	—	Circ	A−	—	—	—	Sq	H+
G(1F)	—	Bars	—	—	E−	—	Oval	—	—	G−	—	Dots	—	—	D+
H(1F)	Scale	—	—	—	D+	Stripe	—	—	—	F+	Check	—	—	—	E−
I(1f)	—	—	Loops	—	B+	—	—	Lines	—	H+	—	—	Curve	—	A−
Novel stimuli															
J(1F)	Scale	—	—	—	L+	Stripe	—	—	—	O+	Check	—	—	—	M−
K(1F)	—	Bars	—	—	P	—	Oval	—	—	R−	—	Dots	—	—	Q+
L(1F)	—	—	Loops	—	S+	—	—	Lines	—	U+	—	—	Curve	—	T−
M(1F)	—	—	—	Irr	T−	—	—	—	Circ	V−	—	—	—	Sq	S+
N(2F)	Scale	Bars	—	—	Q+	Stripe	Oval	—	—	W+	Check	Dots	—	—	P−
O(2F)	Scale	—	Loops	—	X−	Stripe	—	Lines	—	J+	Check	—	Curve	—	K+
P(2F)	Scale	—	—	Irr	W+	Stripe	—	—	Circ	Q+	Check	—	—	Sq	O+
Q(2F)	—	Bars	Loops	—	V−	—	Oval	Lines	—	T−	—	Dots	Curve	—	R−
R(2F)	—	Bars	—	Irr	U+	—	Oval	—	Circ	S+	—	Dots	—	Sq	V−
S(2F)	—	—	Loops	Irr	R−	—	—	Lines	Circ	P−	—	—	Curve	Sq	U+
T(3F)	Scale	Bars	Loops	—	O+	Stripe	Oval	Lines	—	L+	Check	Dots	Curve	—	W+
U(3F)	Scale	Bars	—	Irr	K+	Stripe	Oval	—	Circ	I−	Check	Dots	—	Sq	X−
V(3F)	Scale	—	Loops	Irr	M−	Stripe	—	Lines	Circ	X−	Check	—	Curve	Sq	L+
W(3F)	—	Bars	Loops	Irr	J+	—	Oval	Lines	Circ	M−	—	Dots	Curve	Sq	I−
X(4F)	Scale	Bars	Loops	Irr	I−	Stripe	Oval	Lines	Circ	K+	Check	Dots	Curve	Sq	J+
Y(4F)	Scale	Bars	Loops	Irr	Y−	Stripe	Oval	Lines	Circ	a+	Check	Dots	Curve	Sq	Z−
Z(4F)	Scale	Bars	Loops	Irr	b+	Stripe	Oval	Lines	Circ	d+	Check	Dots	Curve	Sq	c−

Note. Each letter designates a distinct shape (because more than 26 shapes were used, lowercase letters specify different shapes than uppercase letters). Plus and minus signs refer to whether the shape outline was thick or thin, respectively. Irr = irregular; Circ = circle; Check = checkered pattern; Sq = square; 1F, 2F, 3F, 4F = one feature, two features, three features, four features.

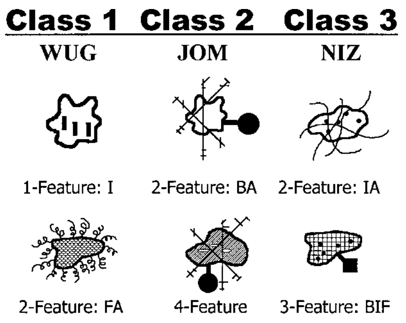

Fig. 2. Some examples of novel probe stimuli used in Experiment 1. The first row shows examples of stimuli that represented novel combinations of relevant features. The second row shows stimuli that represented novel shapes and other irrelevant features. Abbreviations as in Figure 1.

above and in Table 1 (e.g., J1, J2, and J3). Trial blocks consisted of a total of 126 trials; each one of the 54 novel stimulus tests was interspersed with 72 baseline trials (nine per baseline relation). To maintain a 50% reinforcement density for the block, nine baseline trials had feedback omitted.

Phase 3: Sorting and rating task. At the end of the computer testing of the final session, subjects were given a sorting and rating task. Each of the A-stimuli (i.e., wug, jom, and niz) was printed on a separate sheet of paper. The rest of the stimuli used for baseline training were individually printed and pasted on cardboard squares. After placing the sheets on the table in front of the subject, the experimenter gave the subject the cardboard squares

Table 2

Number of trial blocks to criterion, total errors, and percentage baseline errors per opportunity (E/O) for each feature number for the subjects of Experiment 1.

Subject	Criterion blocks	Errors	1F E/O	2F E/O	3F E/O	4F E/O
1A	14	109	35	31	31	26
2A	30	359	55	33	43	50
3A	9	58	19	26	22	19
4A	9	54	33	26	13	15
5A	12	80	37	39	13	11
6A	9	56	30	26	20	22
1B	9	59	37	30	11	19
2B	9	39	27	26	4	4
3B	19	210	49	58	37	40
4B	8	26	16	17	13	4
5B	12	37	19	8	6	8
6B	9	52	28	15	20	26
1C	13	111	40	36	22	33
2C	14	96	33	38	13	21
3C	11	85	35	39	29	24
4C	11	96	40	42	30	27
5C	8	33	17	50	8	4
6C	9	42	30	30	2	4

Note. 1F, 2F, 3F, 4F = one feature, two features, three features, four features.

with instructions as follows: "Now, I'd like you to sort these stimuli into the three categories." After the subject had placed each of cardboard squares on one of the three sheets, the experimenter removed two of the sets of sorted stimuli from the table and said, "Please arrange the stimuli from the most to the least representative of the category. So, arrange them from the best example on down." After recording the order of the first stimulus set, the experimenter removed the stimuli and the process was repeated with each of the other sets.

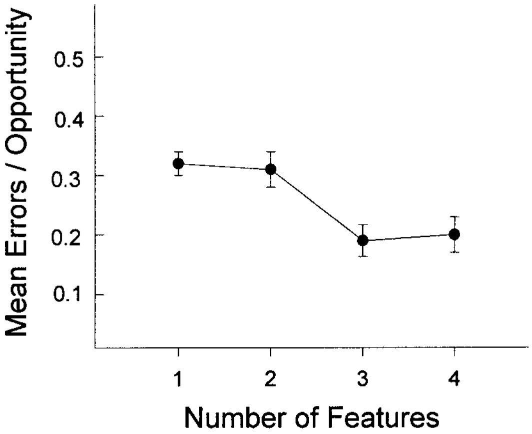

Fig. 3. Mean errors per opportunity as a function of feature number for Experiment 1. Vertical bars indicate standard error of the mean.

RESULTS

Acquisition

Number of trial blocks required to meet the initial acquisition criterion ranged from 8 to 30. Total number of trial blocks and errors during initial acquisition are shown in Table 2. The possibility of a typicality effect in acquisition was assessed by analyzing errors made in different trial types as a function of number of class-consistent features. First, the number of errors was summed independently for trials with comparisons including one, two, three, or four class-consistent features (omitting errors made in the initial trial block—these were omitted because we reasoned that there would be no basis for typicality effects until a history of differential reinforcement was established). These sums were then divided by the total number of trials for each feature number to produce error per opportunity scores for each subject. Table 2 shows these error scores at each feature number for individual subjects and overall means and standard errors are shown in Figure 3. A typicality effect would be indicated by an inverse relation between errors and number of class-consistent features, with the

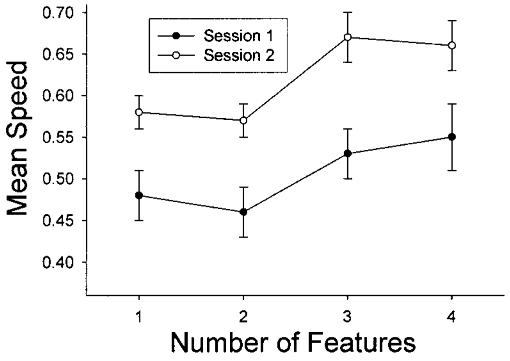

Fig. 4. Mean speed as a function of feature number on baseline trials in Session 1 (black circles) and Session 2 (white circles) of Experiment 1. Vertical bars indicate standard error of the mean.

greater number of errors made on trials involving the one- and two-feature stimuli relative to those with three- and four-feature stimuli as comparisons. A within-subjects analysis of variance confirmed the significance of this typicality effect with a main effect of feature number, $F(3, 51) = 17.38$, $p < .01$. Table 2 reveals that 16 out of the 18 subjects were individually consistent with the averaged results in Figure 3.

Baseline Reaction Times

Reaction times (latency from the presentation of the comparison stimuli to the subject's response) were summed independently for trials with comparisons of each feature number and divided by total number of trials in order to assess the possibility of typicality effects on response speed. As is typical in studies using reaction time measures, trials on which the subject made an incorrect response were excluded from the analyses. In addition, all reaction time data from the initial baseline-trial block of both sessions were excluded. Data from the first baseline trial block of the experiment were of little interest because this was the subject's first exposure to the conditional discriminations, and the first block of the second session was excluded to eliminate any "warm-up" effects. Because latency distributions are often skewed, a reciprocal transformation was used to convert latencies to speed scores (responses per second). Mean speeds for the group on Sessions 1 and 2 are shown in Figure 4. Reaction times in the second session were consistently faster than those obtained in the first session regardless of the number of defining features, $F(1, 17) = 8.22$, $p < .05$. A typicality effect was indicated by the slower speeds in responding to one- and two-feature stimuli relative to those for three- and four-feature stimuli, $F(3, 51) = 33.07$, $p < .01$. These overall means were generally representative of individual subjects, because typicality effects (faster responding to the three- and four-feature stimuli) were observed in 14 out of 18 subjects.

Table 3

Percentage class-consistent responses on probe trials for the subjects of Experiment 1.

Subject	Symmetry	Equivalence	Novel probes
1A	100	100	99
2A	94	90	98
3A	98	100	100
4A	100	100	100
5A	100	100	98
6A	98	100	100
1B	100	100	99
2B	100	98	98
3B	100	98	100
4B	96	95	100
5B	100	97	100
6B	100	100	100
1C	100	96	99
2C	100	90	100
3C	100	100	99
4C	100	100	99
5C	100	97	99
6C	96	97	99

Symmetry, Equivalence and Novel Probes

For all subjects, performance on symmetry and equivalence probes demonstrated the formation of equivalence classes. As summarized in Table 3, class-consistent choices on probe trials ranged from 94% to 100% for symmetry probes and from 90% to 100% for equivalence probes. As Table 3 also shows, choices on novel probes were remarkably class consistent, ranging from 98% to 100%. Thus the equivalence classes demonstrated by the symmetry and combined probes were not limited to those stimuli directly involved in baseline training, but also included novel stimuli with one or more of the class-consistent features.

Although each of the novel stimuli controlled class-appropriate responding based on the presence of class-consistent features, anal-

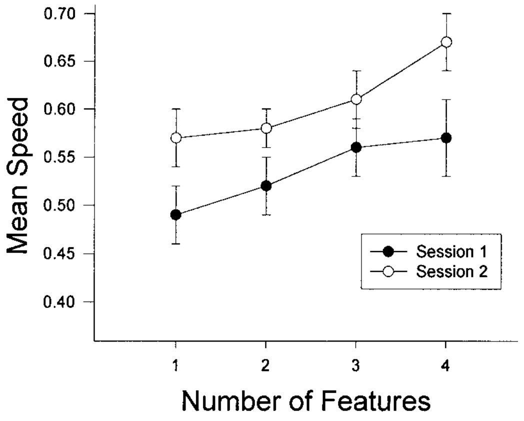

Fig. 5. Mean speed as a function of feature number on novel probe trials in Session 1 (black circles) and Session 2 (white circles) of Experiment 1. Vertical bars indicate standard error of the mean.

ysis of response speed for novel probes suggested that differential responding depended on the number of relevant features. Figure 5 shows the mean speed of responding on novel probe trials for Sessions 1 and 2 for each level of feature number. As was true on baseline trials, subjects reacted significantly more rapidly in Session 2 than in their initial session, $F(1, 17) = 7.21$, $p < .05$. Typicality effects were apparent in both sessions with speed directly related to the number of class-consistent features, $F(3, 51) = 26.68$, $p < .01$. Individual subject speeds presented in Figure 6 confirm the occurrence of typicality effects in a strong majority of the subjects.

Rating Task

Individual subjects' rankings of the stimuli for each class were assigned numbers from 1 to 8 (i.e., most typical = 8; least typical = 1) and pooled across classes at each level of feature number. For example, the stimulus the subject rated as most typical of Class 1 was assigned a value of 8, the second most typical, 7, and so on. The median rating was then determined for the three 4-feature stimuli, the six 3-feature stimuli, the three 2-feature stimuli and the twelve 1-feature stimuli. The subjects' mean of the median ratings and standard error at each level of feature inclusion are shown in Figure 7. There was a strong positive relation between number of features and typicality ratings, $F(3, 51) = 387.52$, $p < .01$.

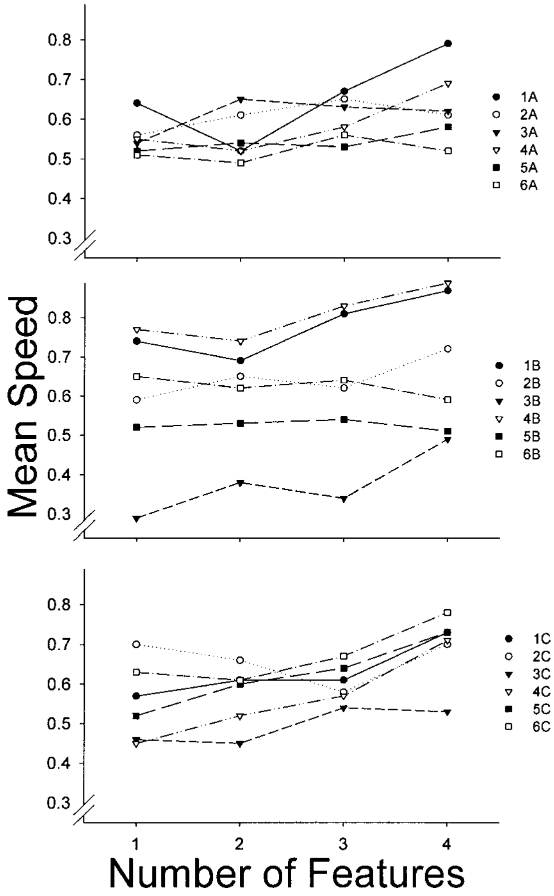

Fig. 6. Mean speed as a function of feature number on novel probe trials for each subject in Experiment 1.

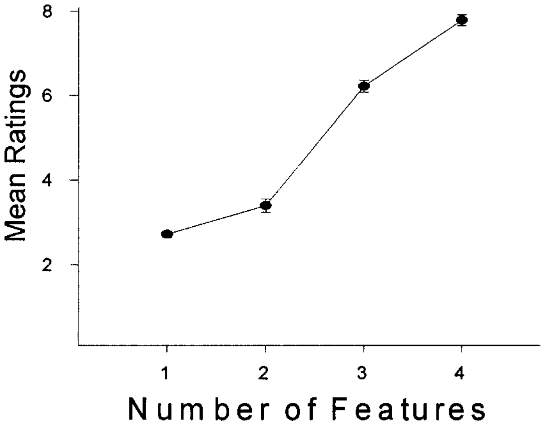

Fig. 7. Mean of the median typicality ratings produced following the stimulus sorting procedure of Experiment 1. Vertical bars indicate standard error of the mean (where no bar is visible, the *SEM* is smaller than the data symbol).

DISCUSSION

Subjects in the present study were trained on eight conditional discriminations that involved 24 directly trained relations. The symmetry and equivalence probe performances revealed the emergence of an additional 54 untrained relations consistent with the formation of three, 8-member classes. However, accurate performances on the novel probes indicated that another 51 stimuli were class members. It would appear that the inclusion of one or more class-consistent features was sufficient to result in class membership of a novel stimulus. In principle, then, these classes were ''open-ended''; that is, any number of stimuli could be included in the classes without additional training if they possessed one or more of the relevant features. Thus subjects' classification of the stimuli illustrated the important generative properties that help to define natural language categories (Harnad, 1996; Markman, 1989).

The typicality effects identified in the present results are also consistent with findings from natural language categories. There were four different measures in which it was possible to observe typicality effects. These measures included baseline training errors, reaction time to the baseline training stimuli, reaction time to novel-probe stimuli, and verbal rating of stimuli following the sorting task. For each measure, typicality effects were present—indexed by an inverse relation between number of class-consistent features and errors and a direct relation between number of class-consistent features and response speed and ratings of typicality.

Certain features of Experiment 1 may complicate interpretation. For example, it is assumed that the differential reinforcement provided during baseline training was critical to the formation of the complex classes that were observed. The relation between each of the trigrams and the stimuli that became class members was completely arbitrary, but the need for differential reinforcement to produce control by the class-consistent features might be questioned. Some class members were physically similar to one another, and stimulus generalization might be invoked to explain common responding to these stimuli. For example, although stimuli B1 and C1 (Figure 1) were shaped differently, they appear similar in many other respects and perhaps the shape differences were insufficient to make these stimuli discriminable. This sort of argument appears to break down when applied to the one-feature stimuli (e. g., H1 appears more similar to E2 than to any of the Class 1 stimuli), but it is possible that some aspects of stimulus control in Experiment 1 might not have required the reinforcement history provided. One of the purposes of Experiment 2 was to replicate Experiment 1 under conditions in which the likelihood of control by class-consistent features without differential reinforcement could be assessed.

To evaluate the factors governing stimulus classification without training, a sorting task like that used in Experiment 1 was administered prior to discrimination training in Experiment 2. In addition, a new stimulus set was created for Experiment 2. The Experiment 2 stimuli had the same four class-consistent features (appendages, base, fill, and inserts) used in Experiment 1 but new variants of each feature were developed for Experiment 2 and two new class-inconsistent features were used (color of an inner circle and color outside the circle), along with two of the class-inconsistent features used in Experiment 1 (shape and position of base). In addition, three of the irrelevant features (shape and the two colors) had only three variants just as the class-consistent features did, ruling out number of variants as a factor that could bias classification.

Another methodological change in Experiment 2 was a control for stimulus complexity of the novel-probe stimuli. In Experiment 1, reaction times were a function of number of class-consistent features, and although this may be interpreted as a typicality effect, number of relevant features also was confounded with stimulus complexity. Although it may seem counterintuitive, perhaps the overall complexity of the stimuli (i.e., the number of features per se), rather than the number of class-consistent features, produced the reaction time effects. In support of this possibility, lexical decision studies have found that response speeds to concrete nouns are directly related to the number of features that subjects associate with their referents (Pexman, Holyk, & Monfils, 2003; Pexman, Lupker, & Hino, 2002). Although these ''number of features'' effects involve somewhat different procedures than those used here, it may be that the present results were due to the overall number of features present rather than to the number of relevant features. To control for this possibility, novel probe stimuli that included an equal number of features for each stimulus were developed for Experiment 2. To accomplish this, variants of the class-consistent features were developed that were not used during baseline training, but appeared for the first time among the novel probes. These new variants were not consistently associated with any of the classes, but were used to equate stimuli for complexity—they were, in effect, irrelevant versions of features that were always class-consistent during baseline training. So, for example, a one-feature novel probe might possess one of the class-consistent inserts, but would also have class-irrelevant appendages, base and fill. Thus the number of class-consistent features in this novel probe set could be varied while holding the overall complexity of the stimuli constant.

EXPERIMENT 2

Subjects and Apparatus

Subjects in Experiment 2 participated for three, 1-hr sessions. No more than two sessions were scheduled per day and all three sessions were completed within a 7-day period. Apparatus and subject selection procedures were as described in Experiment 1. Twelve experimentally naive subjects served in Experiment 2.

Procedure

Preliminary sorting task. Experimental procedures generally followed those of Experiment 1 except that the sorting task was conducted both at the outset of the experiment and again after equivalence testing was complete. The preliminary sorting task was conducted to determine how subjects would categorize the stimuli to be used in baseline training of Experiment 2 (described below) before the conditional discrimination training history. Instructions and procedures were the same as in Experiment 1 except that after placing each of the stimulus cards on one of the blank sheets of paper, the task ended without requiring subjects to rank order stimuli with respect to typicality.

Baseline training. As in Experiment 1, eight conditional discriminations were simultaneously trained using a one-to-many structure with one of the nonsense syllables (wug-A1, jom-A2, niz-A3) serving as the sample and one of the eight stimulus sets (B through I) providing the three comparisons. (See Appendix for a description of the stimuli and Figure 8 for illustration of a subset. All baseline and probe stimuli are illustrated on the website, http://people.uncw.edu/galizio/Galizio.htm—follow links to equivalence research page.) These stimuli differed with respect to eight possible features, and as in Experiment 1, four of them (fill, inserts, appendages, and bases) were related to the contingencies in a class-consistent fashion; choice of each of the three variants was reinforced following wug, jom, or niz samples, respectively. For example, Class 1 stimuli (those whose selection was reinforced when the sam-

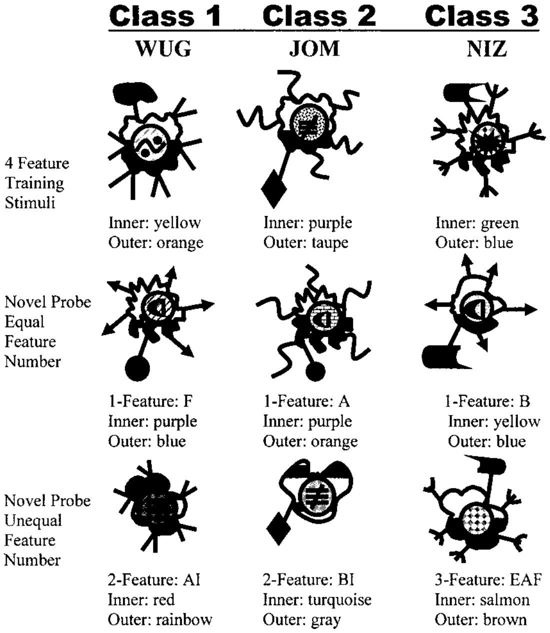

Fig. 8. Some examples of training and novel probe stimuli used in Experiment 2. The first row shows the three, 4-feature (prototype) stimuli used in baseline training that illustrate the values of all four of the relevant features (appendages [A], base [B], insert [I], and fill [F]). The second row shows examples of the novel probe stimuli equated for number of features. For example, the leftmost stimulus shows a stimulus with one relevant feature (fill) and illustrates the class-inconsistent variants of the other critical features. The bottom row shows examples of novel probe stimuli with unequal numbers of features.

ple was ''wug'') had a striped fill in a circle situated in the center of each figure, a curved line with two dots as a insert (Yin), pairs of lines resembling a ''V'' as appendages (Twin), and an irregularly shaped base (Irr; see Appendix and Figure 8 for more details). There were four class-inconsistent features and each of these also had three variants that were inconsistently associated with wug, jom, or niz samples: the shape of the figure (see Figure 8), the color of a circle situated in the center of each figure (green, yellow, or purple–inside), the color that filled the shape outside the circle (blue, taupe, or orange–outside) and clock position of base (2:00, 7:00, or 11:00). The same eight combinations of the four defining features used in Experiment 1 made up the baseline training. Again, these eight combinations were arranged such that each defining feature was presented an equal number of times to control for the amount of experience with each feature. Trial block arrangement and criteria required to move to the probe-testing phase of the experiment were the same as in Experiment 1.

Symmetry, equivalence, and novel probes. The main departure from the procedures of Experiment 1 was that two types of novel probe stimuli were developed. The first type closely followed the strategy used before of representing each combination of class-consistent features (see Appendix and Figure 8). For example, one-feature stimuli included three or four of the irrelevant features (shape, color of inner circle, color of outside circle, and position of the base, if one was present) and one of the four class-consistent features. As in Experiment 1, each block of trials included three exposures to each of the four 1-feature comparison sets and each of the four possible three-feature combination sets (one with each of the three samples). Only four of the six possible two-feature combinations were used and one novel four-feature set was developed. Thus there were a total of 13 comparison sets composed of 39 novel stimuli.

A second group of novel-probe stimuli was designed to control for unequal number of total features. Each stimulus of this type had some variant of all eight possible features. So, for example, a one-feature comparison from the ''equal feature'' set included only one class-consistent feature (e.g., fill), but also had a base, appendages and an insert that were not consistently associated with any of the three classes (see Figure 8 and Appendix). In this way, number of class-consistent stimuli was varied while holding overall stimulus complexity constant. The equal-feature novel probe sets were constructed in the same way as the unequal feature sets, so there were 13 equal-feature comparison sets with 39 novel stimuli. In both equal and unequal feature sets, irrelevant features for some stimuli used novel combinations of the values used in training, but in order to make the novelty of these stimuli more distinctive, other stimuli used shapes and colors not seen in training (see Appendix). Trial blocks for both novel probe types consisted of the 39 probe trials intermixed with 48 baseline trials and 36 of them were reinforced, producing an overall reinforcement rate of 41%.

On any given session, subjects received a block of one of the novel probe types immediately upon mastering the baseline training steps (6 received the unequal feature number block first, and 6 received the equal feature number block first). Completion of the first novel probe block was followed by a symmetry probe block, an equivalence probe block, and then novel probe blocks of both types were alternated for the duration of the session. Because of the additional novel probe set used in Experiment 2, subjects were tested for three 50-min sessions. Each session began with baseline trial blocks and, after criterion was met, probe-trial blocks were presented in the sequence described above until 50 min had elapsed in the first two sessions, and until 40 min had elapsed for the third session, after which the sorting and rating task was administered.

Posttraining sorting and rating task. Procedures were the same as those of Experiment 1.

RESULTS AND DISCUSSION

Acquisition

Total number of trial blocks and errors during baseline acquisition are shown in Table 4, and mean errors per opportunity are presented as a function of feature number in Figure 9. All 12 subjects met the acquisition criterion at a rate comparable to those of Experiment 1 (trial blocks to criterion ranged from 5 to 23). Figure 9 shows that, as in Experiment 1, errors decreased inversely with

Table 4

Number of trial blocks to criterion, total errors and baseline percentage errors per opportunity (E/O) for each feature number for the subjects of Experiment 2.

Subject	Criterion blocks	Errors	1F E/O	2F E/O	3F E/O	4F E/O
1	8	51	71	71	58	67
2	5	16	38	37	36	33
3	22	133	47	54	40	20
4	23	141	49	61	18	07
5	13	86	65	37	62	36
6	8	53	62	65	34	55
7	11	65	31	57	42	07
8	9	52	48	54	28	37
9	6	28	33	31	15	38
10	5	17	41	20	20	37
11	16	130	58	50	58	40
12	6	25	44	17	42	16

Note. 1F, 2F, 3F, 4F = one feature, two features, three features, four features.

the number of class-consistent features in the comparison stimuli, $F(3, 33) = 4.14$, $p < .05$. This typicality effect was clearly evident at the individual subject level with all subjects making fewer errors to four- and three-feature stimuli than to one- and two-feature stimuli (see Table 4).

Baseline Reaction Times

Figure 10 presents mean response speeds for baseline trials across the three sessions of Experiment 2. There was a significant effect of feature number with increased speeds associated with more class-consistent features, $F(3, 33) = 18.72$, $p < .01$. Response speeds also increased across sessions, $F(2, 22) = 13.24$, $p < .01$, but as Figure 10 shows, in the first session responding was slowest with the one- and two-feature stimuli and increased in linear fashion to three- and four-feature stimuli. On the second and third sessions, the effect was largely based on slower responding to the one-feature stimuli. A significant Session X feature number interaction, $F(6, 66) = 8.45$, $p < .01$, was consistent with this interpretation.

Symmetry, Equivalence and Novel Probes

All subjects showed high levels of accuracy (class-consistent responding) on symmetry, equivalence, and both types of novel probes (Table 5). Figure 11 shows response speeds as a function of number of class-consistent features for both novel probe types averaged across subjects and sessions (few subjects completed both equal- and unequal-feature probe types within a session, so comparisons across sessions were not meaningful. Howev-

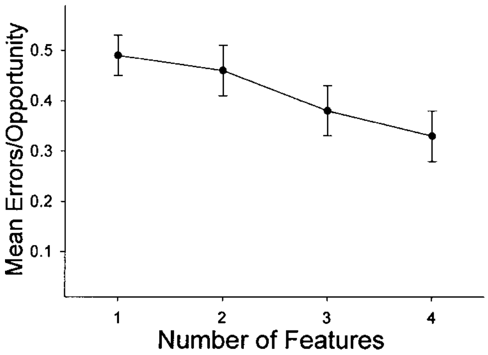

Fig. 9. Mean errors per opportunity as a function of feature number for Experiment 1. Vertical bars indicate standard error of the mean.

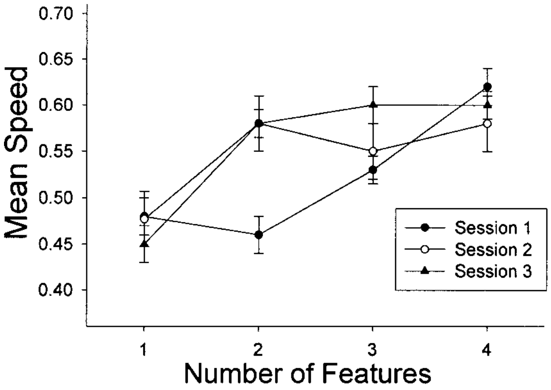

Fig. 10. Mean speed as a function of feature number on baseline trials in Session 1 (black circles), Session 2 (white circles), and Session 3 (triangles) of Experiment 2. Vertical bars indicate standard error of the mean.

Table 5

Percentage class-consistent responses on probe trials for Experiment 2.

Subject	Symmetry	Equivalence	Novel-equal	Novel-unequal
1	97	98	98	99
2	100	100	96	96
3	90	92	85	83
4	88	88	78	78
5	100	90	94	92
6	98	100	100	100
7	100	100	99	99
8	100	100	98	100
9	100	100	100	100
10	99	99	97	99
11	98	94	90	90
12	100	100	99	99

er, each subject completed at least two blocks with each probe type). The top panel of Figure 11 shows speeds for unequal-feature probe types and provides some evidence of typicality effects with fastest responding evident for four-feature comparisons and slowest for one-feature stimuli, $F(3, 33) = 10.57$, $p < .01$. Performances on the equal-feature probe types, new to Experiment 2, are shown in the bottom panel of Figure 11, and also revealed strong evidence of typicality effects, $F(3, 33) = 46.52$, $p < .01$. In fact, results from the equal-feature probe trials provided the most linear typicality function of any of the speed analyses, largely due to slower responding for the one-feature stimuli than in the analyses involving comparisons with unequal features. Individual subject speeds for novel probes (collapsed across probe type) are presented in Figure 12 and evidence of typicality effects is present for every subject.

Pre- and Posttraining Sorting and Rating Tasks

Prior to training, none of the subjects sorted the stimuli on the basis of the features that were to be designated as class-consistent when training began. Many subjects appeared to sort the stimuli on the basis of one of the colors, a few used one or more of the class-consistent features as a basis for sorting, and in other cases it was not possible to determine any particular pattern. In order to analyze the subject's pretraining sorts, each stack was examined to determine which stimulus class had the highest number of exemplars and

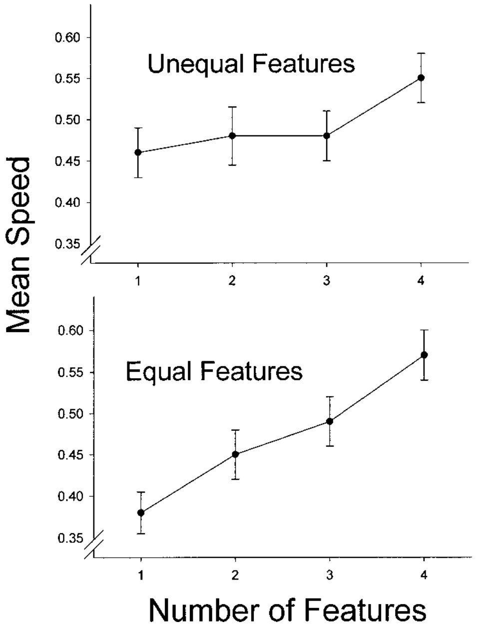

Fig. 11. Mean speed as a function of feature number on novel probe trials with unequal total features (top panel) and equal total features (bottom panel) for Experiment 2. Vertical bars indicate standard error of the mean.

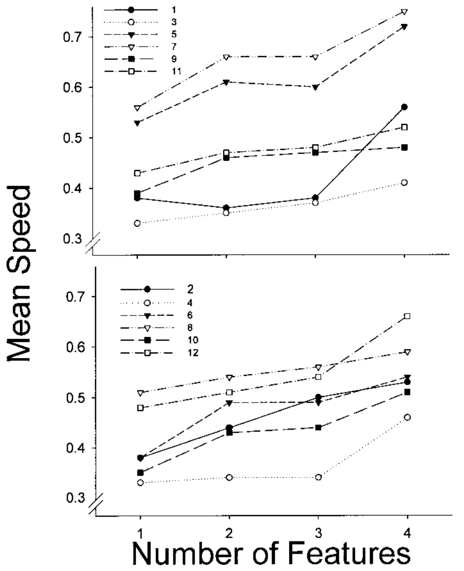

Fig. 12. Mean speed as a function of feature number on novel probe trials (equal and unequal probe types combined) for each subject in Experiment 2.

Table 6

Percentage and number (in parenthesis) of class-consistent items of items sorted in each class per- and posttraining.

Subject	Pre-WUG	Pre-JOM	Pre-NIZ	Pre-avg	Post-Wug	Post-JOM	Post-NIZ	Post-avg
1	33.33	33.33	33.33	33.33	100.00	100.00	100.00	100.00
	(2)	(2)	(4)	(2.7)	(8)	(8)	(8)	(8.0)
2	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100.00
	(3)	(3)	(3)	(3.0)	(8)	(8)	(8)	(8.0)
3	37.50	37.50	37.50	37.50	87.50	87.50	100.00	91.67
	(3)	(3)	(3)	(3.0)	(7)	(8)	(7)	(7.3)
4	55.56	50.00	44.44	50.00	62.50	77.78	71.43	70.57
	(3)	(4)	(5)	(4.0)	(5)	(5)	(7)	(5.7)
5	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100.00
	(3)	(3)	(3)	(3.0)	(8)	(8)	(8)	(8.0)
6	28.57	50.00	55.56	44.71	100.00	100.00	100.00	100.00
	(2)	(5)	(4)	(3.7)	(8)	(8)	(8)	(8.0)
7	33.33	33.33	33.33	33.33	100.00	100.00	100.00	100.00
	(4)	(2)	(2)	(2.7)	(8)	(8)	(8)	(8.0)
8	50.00	100.00	100.00	83.33	100.00	100.00	100.00	100.00
	(8)	(4)	(4)	(5.3)	(8)	(8)	(8)	(8.0)
9	33.33	33.33	33.33	33.33	100.00	100.00	100.00	100.00
	(4)	(2)	(2)	(2.7)	(8)	(8)	(8)	(8.0)
10	71.43	66.67	36.36	58.15	100.00	100.00	100.00	100.00
	(4)	(5)	(4)	(4.3)	(8)	(8)	(8)	(8.0)
11	37.50	37.50	37.50	37.50	100.00	100.00	100.00	100.00
	(3)	(3)	(3)	(3.0)	(8)	(8)	(8)	(8.0)
12	33.33	33.33	33.33	33.33	100.00	100.00	100.00	100.00
	(4)	(2)	(2)	(2.7)	(8)	(8)	(8)	(8.0)
Mean	40.74	45.83	43.31	43.29	95.83	97.11	97.62	96.85

that class was designated as "correct" for that stack. Thus random responding with respect to class-consistent features yielded 33.3% correct (the lowest possible score given the analysis) and class-consistent sorting produced 100% correct. As Table 6 shows, 8 of the 12 subjects' sorts resulted in accuracy levels between 33% and 38%. Only 1 subject (Subject 8) showed a bias toward class-consistent responding prior to training (83.3% correct). Subject 8 placed each of the stimuli with identical bases in stacks, but put all 12 remaining stimuli in one of the three stacks, so even in this case high percentage was not indicative of control by any of the class-consistent features except for the base.

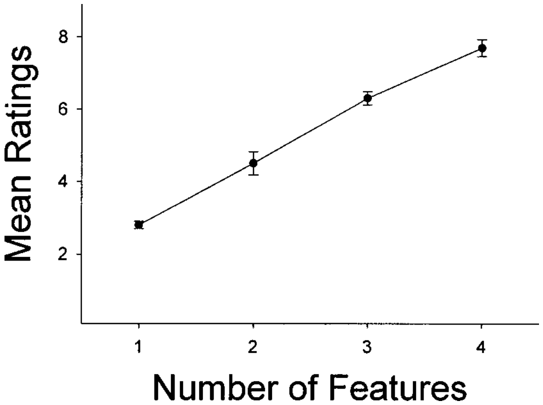

Fig. 13. Mean of the median typicality ratings produced following the stimulus sorting procedure of Experiment 2. Vertical bars indicate standard error of the mean (where no bar is visible, the *SEM* is smaller than the data symbol).

In contrast, 11 of the 12 subjects showed strong evidence of class-consistent sorting after training, and 10 of the 12 had 100% correct after training. In each of these cases, all eight of the stimuli for each class were correctly placed together (although it should be noted that 1 subject was quite inaccurate at 70.6% and was excluded from the subsequent typicality rating analysis). These data provide evidence that the stimulus classes defined by the conditional discrimination training did not generally control responding until after that training had been provided. Finally, Figure 13 shows the results of the typicality ratings and reveals a strong linear relation between number of class-consistent features and judgments of typicality, $F(3, 33) = 66.66$, $p < .01$. In summary, the results of Experiment 2 closely paralleled those of Experiment 1 and showed that the effects were replicable with a different stimulus set in which the relevant and irrelevant stimulus features were balanced.

GENERAL DISCUSSION

The stimulus classes that came to control responding in the present studies were of interest, in part, because they included both stimuli that were perceptually related (different abstract shapes with common features) and stimuli that were arbitrarily related (nonsense syllables and shapes). These classes thus shared some of the relational properties of natural lexical or family resemblance categories (Rosch & Mervis, 1975; Wittgenstein, 1953/1999). For example, like word referents, the abstract class members studied here often had one feature (and sometimes two, three, or four features) in common with one another, but no features were common to all items. In the present study, such family resemblance stimulus classes were shaped through the contingencies of a one-to-many matching-to-sample procedure.

Importantly, not all of the observed stimulus relations were directly trained. For example, symmetry probes revealed that subjects chose the class-consistent nonsense syllable when one of the abstract stimuli was presented as a sample (although the nonsense syllables never served as comparisons nor the abstract stimuli as samples in training). Equivalence probes also showed that when one of the abstract stimuli was presented as a sample, it occasioned selection of a class-consistent stimulus, even when the two had no common features (in all equivalence probe trials, stimuli were programmed such that the sample and class-consistent comparison shared no relevant features). Thus the three stimulus classes that emerged in both studies showed the defining properties of equivalence classes (Sidman, 1994; Sidman & Tailby, 1982). Although the relations between the nonsense syllables and the shapes were arbitrary, the family resemblances among the shapes in each class led to the emergence of numerous additional untrained relations observed on novel probe tests. In both experiments, novel stimuli with at least one of the relevant features were reliably selected when the class-consistent nonsense syllable was the sample. As noted above, the open-ended feature of these classes provides a model of the generative capacity of lexical classes by permitting the subject to behave adaptively (in accord with past contingencies) in response to a novel stimulus—a simple type of ‘‘symbolic theft’’ or category-based induction (see Harnad, 1996; Markman, 1989; Murphy, 2002).

Fields and his colleagues (e.g., Fields, Reeve, Adams, Brown, & Verhave, 1997; Fields, Reeve, Adams, & Verhave, 1991; see Fields & Reeve, 2000, for a review) previously have demonstrated that novel stimuli closely related on a physical dimension to equivalence-class members may be included in the class through primary stimulus generalization. Like the classes observed in the present study, these generalized equivalence classes also may be described as open-ended, but an important difference is that class formation in the present study cannot be accounted for completely in terms of primary stimulus generalization. Because class members did not always share physical features (see Figures 1, 2, and 8), they were often quite dissimilar. The pretraining sort procedure used in Experiment 2 showed that subjects did not classify the stimuli according to the experimenter-defined classes before conditional discrimination contingencies were applied, and this provided direct evidence that class membership did not involve primary stimulus generalization. Pretraining sorts were controlled by overall stimulus similarity in which shape and color were critical determinants, but in post-training sorts as well as matching-to-sample performances the presence of even a single relevant feature was sufficient to override stimulus similarity on all other dimensions. Rather, particular features came to control responding after a history of differential reinforcement training, as has been described for feature classes in nonhumans (e.g., Herrnstein, 1990; Lea, 1984). During that training, features that were distributed across classes lost whatever initial control they had over responding (i.e., relative to the presort), and stimulus control developed for the features that were class consistent (abstraction). As Herrnstein (1984) put it: ‘‘besides shaping response topographies, reinforcement appears also to shape perceptual features. From the available stimulus dimensions, differential reinforcement selects those that differentiate positive and negative instances'' (p. 254). Thus reinforcement selects particular stimulus control topographies (see McIlvane, Serna, Dube, & Stromer, 2000). The stimulus classes in the present study might be described as a merger between feature and equivalence classes established simultaneously through discrimination training. Although discrimination training was necessary to bring selection under the control of the relevant features, the classes demonstrated here still might be referred to as generalized equivalence classes in keeping with Keller and Schoenfeld's (1950) definition of stimulus classes/concepts that included relations between class members based on acquired similarity or equivalence as well as primary generalization.

Few studies have investigated classes with these mixed properties. Astley and Wasserman (1996) conducted experiments in which mergers of perceptually distinct feature classes were demonstrated in children. Children were exposed to four sets of perceptually related items and learned to make one response to two of the sets and another response to the other two. Novel stimuli from the same perceptual classes were responded to in class-appropriate ways. The present results were consistent with those of Astley and Wasserman and extended their findings by adding equivalence tests and by including stimuli that did not share features and that varied in the number of relevant features. This last aspect of the present study allowed the analysis of typicality effects.

Typicality effects are commonly observed in lexical and artificial categories and involve relations in which speed of learning, number of errors, speed of responding (reaction time), and judgments of exemplariness depend on number of category-relevant features or properties (e.g., Murphy, 2002; Rosch, 1978). The present studies demonstrated these same typicality effects in contingency-shaped generalized equivalence classes. For example, in both studies the number of errors was inversely related to the number of relevant features during conditional discrimination training. Response speeds were directly related to the number of relevant features both for the training stimuli and for novel probe stimuli. Further, in Experiment 2 the relation was demonstrated for novel probes even when stimulus complexity was equated for stimuli varying in the number of class-defining features. Finally, subjects' judgments of exemplariness showed the same direct relation. In sum, all major dependent variables that traditionally have been considered as measures of typicality were related to number of relevant features in the classes studied here.

It should be noted that the determinants of typicality effects in the present study may well be different than those in lexical classes. In fact, there are certainly multiple determinants of typicality ratings in language categories, because not all typicality effects are correlated with number of relevant features (e.g., the number ''7'' is judged to be a better exemplar of the category of odd numbers than the number ''91,'' Armstrong, Gleitman & Gleitman, 1983). The three measures of typicality obtained in the present study may have different sources as well. For example, error rates may simply reflect differential acquisition of particular features. Once one feature (e.g., ''inserts'') has come to control responding, then four-, three-, and two-feature training comparisons with inserts may be correctly classified even though the other relevant features have yet to acquire control over responding. Alternatively, the reaction time results might reflect visual search such that one of the relevant features is more quickly identified in stimuli with multiple critical features. Neither of the above accounts appears likely to explain the rating data. An advantage of the present procedures is that they could be altered to provide an experimental evaluation of explanations like those offered above. Thus typicality effects may be viewed as products of a particular training history rather than as properties of a mental structure.

The present experiments add to a growing body of literature that shows that effects observed in natural lexical categories can be observed and studied in the laboratory using the various methodologies that have been developed in the behavioral analysis of stimulus relations and stimulus equivalence. For example, phenomena such as semantic priming

(Hayes & Bisset, 1998), category clustering in free recall (Galizio, Stewart, & Pilgrim, 2001), fast lexical mapping (Wilkinson, Dube, & McIlvane, 1996), transformation of function (Barnes & Keenan, 1993; Dougher, Augustson, Markham, Greenway, & Wulfert, 1994; Green, Sigurdardottir, & Saunders, 1991), and class-consistent sorting (Pilgrim & Galizio, 1996), as well as the generative emergence of new stimulus relations (see Sidman, 1994, for a review) have all been demonstrated with equivalence-class methodologies. The data presented here add the emergence of novel relations based on abstracted stimulus features and observation of typicality effects to the list of similarities that support the utility of these approaches to the analysis of language and categorization. These studies are best understood as preliminary approaches to explore the value of a contingency-based account of effects such as category-based induction and typicality, and it remains to be seen whether the account developed can be extended to interpret the various empirical and theoretical problems associated with these complex phenomena (Harnad, 1996; Margolis & Laurence, 1999; Murphy, 2002). Although the type of account developed here is preliminary, it seems promising and illustrates a functional alternative to more traditional structural accounts of concepts.

REFERENCES

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13,* 263–308.

Astley, S. L., & Wasserman, E. A. (1996). Mediating associations, essentialism and nonsimilarity-based categorization. In T. R. Zentall & P. M. Smeets (Eds.), *Stimulus class formation in humans and animals* (pp. 111–133). Amsterdam: Elsevier.

Barnes, D., & Keenan, M. (1993). A transfer of functions through derived arbitrary and nonarbitrary stimulus relations. *Journal of the Experimental Analysis of Behavior, 59,* 61–81.

Chomsky, N. (1959). Review of Skinner's *Verbal Behavior. Language, 35,* 26–58.

Dougher, M. J., Augustson, E. M., 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.

Fields, L., & Reeve, F. K. (2000). Synthesizing equivalence classes and natural categories from perceptual and relational classes. In J. C. Leslie & D. Blackman (Eds.), *Experimental and applied analysis of human behavior* (pp. 59–84). Reno, NV: Context Press.

Fields, L., Reeve, F. K., Adams, B. J., Brown, J. L., & Verhave, T. (1997). Predicting the base 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, F. K., 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.

Fodor, J. A. (1990). Information and representation. In P. Hanson (Ed.), *Information, language and cognition* (pp. 175–190). Vancouver: University of British Columbia Press.

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

Green, G., Sigurdardottir, Z. G., & Saunders, R. R. (1991). The role of instructions in the transfer of ordinal functions through equivalence classes. *Journal of the Experimental Analysis of Behavior, 55,* 287–304.

Harnad, S. (1996). Experimental analysis of naming behavior cannot explain naming capacity. *Journal of the Experimental Analysis of Behavior, 65,* 262–264.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post- Skinnerian account of human language and cognition.* New York: Kluwer Academic/Plenum.

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

Herrnstein, R. J. (1984). Objects, categories and discriminative stimuli. In H. L. Roitblat, T. G. Bever, & H. S. Terrace (Eds.), *Animal cognition* (pp. 233–261). Hillsdale, NJ: Erlbaum.

Herrnstein, R. J. (1990). Levels of stimulus control: A functional approach. *Cognition, 37,* 133–166.

Horne, P. J., & Lowe, C. F. (1996). On the origins of naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior, 65,* 185–241.

Keller, F. S., & Schoenfeld, N. (1950). *The principles of psychology.* New York: Appleton-Century-Crofts.

Lea, S. E. G. (1984). In what sense do pigeons learn concepts? In H. L. Roitblat, T. G. Bever, & H. S. Terrace (Eds.), *Animal cognition* (pp. 263–276). Hillsdale, NJ: Erlbaum.

Margolis, E., & Laurence S. (Eds.). (1999). *Concepts: Core readings.* Cambridge, MA: MIT Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

McIlvane, W. J., Serna, R. W., Dube, W. V., & Stromer, R. (2000). Stimulus control topography coherence and stimulus equivalence: Reconciling test outcomes with theory. In J. C. Leslie & D. Blackman (Eds.), *Experimental and applied analysis of human behavior* (pp. 85–110). Reno, NV: Context Press.

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). New York: Cambridge University Press.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society, 7,* 283–284.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11,* 70–84.

Pexman, P. M., Holyk, G. G., & Monfils, M. H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition, 31,* 842–855.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9,* 542–549.

Pilgrim, C., & Galizio, M. (1996). Stimulus equivalence: A class of correlations or a correlation of classes? In T. R. Zentall & P. M. Smeets (Eds.), *Stimulus class formation in humans and animals* (pp. 173–195). Amsterdam: Elsevier.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12,* 1–20.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 491–502.

Sidman, M. (1994). *Equivalence relations and behavior: A research story.* Boston: Authors Cooperative.

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

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.

Skinner, B. F. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.

Wilkinson, K. M., Dube, W. V., & McIlvane, W. J. (1996). A crossdisciplinary perspective on studies of rapid word mapping in psycholinguistics and behavior analysis. *Developmental Review, 16,* 125–148.

Wittgenstein, L. (1999). Philosophical investigations, Sections 65–78. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 171–174). Cambridge, MA: MIT Press. (Original work published 1953)

Received January 27, 2004
Final acceptance September 12, 2004

APPENDIX

Description of stimuli used in Experiment 2.

Stimuli		Critical features				Irrelevant features			
Array	Class	Fill	Insert	Appendages	Base	Shape	Color (O)	Color (I)	Base position
Training stimuli									
B(4F)	WUG	Stripe	Yin	Twin	Irr	A	Orange	Yellow	11
	JOM	Dots	Bars	Wavy	Diam	A	Taupe	Purple	7
	NIZ	Check	Star	Fork	Ears	B	Blue	Green	11
C(3F)	WUG	—	Yin	Twin	Irr	A	Blue	Purple	2
	JOM	—	Bars	Wavy	Diam	C	Orange	Yellow	11
	NIZ	—	Star	Fork	Ears	B	Orange	Green	11
D(3F)	WUG	Stripe	Yin	Twin	—	B	Blue	Green	—
	JOM	Dots	Bars	Wavy	—	B	Blue	Green	—
	NIZ	Check	Star	Fork	—	C	Taupe	Purple	—
E(2F)	WUG	Stripe	—	—	Irr	B	Taupe	Green	11
	JOM	Dots	—	—	Diam	A	Blue	Purple	2
	NIZ	Check	—	—	Ears	A	Blue	Yellow	7
F(1F)	WUG	—	—	—	Irr	B	Taupe	Purple	7
	JOM	—	—	—	Diam	A	Orange	Yellow	2
	NIZ	—	—	—	Ears	C	Taupe	Purple	2
G(1F)	WUG	—	—	—	Irr	C	Blue	Yellow	—
	JOM	—	—	—	Diam	B	Orange	Green	—
	NIZ	—	—	—	Ears	C	Orange	Green	—
H(1F)	WUG	Stripe	—	—	—	A	Orange	Yellow	—
	JOM	Dots	—	—	—	C	Taupe	Purple	—
	NIZ	Check	—	—	—	B	Taupe	Purple	—
I(1F)	WUG	—	—	Twin	—	C	Orange	Green	—
	JOM	—	—	Wavy	—	C	Taupe	Yellow	—
	NIZ	—	—	Fork	—	A	Blue	Yellow	—
Novel stimuli: unequal features									
J(1F)	WUG	—	—	—	Irr	A	Taupe	Yellow	11
	JOM	—	—	—	Diam	B	Blue	Yellow	7
	NIZ	—	—	—	Ears	C	Blue	Yellow	7
K(1F)	WUG	Stripe	—	—	—	B	Blue	Purple	—
	JOM	Dots	—	—	—	C	Taupe	Purple	—
	NIZ	Check	—	—	—	B	Orange	Yellow	—
L(1F)	WUG	—	Yin	—	—	C	Orange	Green	—
	JOM	—	Bars	—	—	C	Taupe	Green	—
	NIZ	—	Star	—	—	C	Taupe	Green	—
M(1F)	WUG	—	—	Twin	—	A	Blue	Green	—
	JOM	—	—	Wavy	—	B	Orange	Purple	—
	NIZ	—	—	Fork	—	A	Blue	Purple	—
N(2F)	WUG	Stripe	—	—	Irr	C	Orange	Yellow	11
	JOM	Dots	—	—	Diam	A	Blue	Yellow	7
	NIZ	Check	—	—	Ears	B	Blue	Purple	11
O(2F)	WUG	Stripe	—	Twin	—	D	Multi	Blue	—
	JOM	Dots	—	Wavy	—	E	Brown	Red	—
	NIZ	Check	—	Fork	—	F	Multi	Orange	—
P(2F)	WUG	—	Yin	—	Irr	F	Brown	Orange	11
	JOM	—	Bars	—	Diam	D	Gray	Blue	2
	NIZ	—	Star	—	Ears	D	Brown	Blue	2
Q(2F)	WUG	—	Yin	Twin	—	E	Multi	Red	—
	JOM	—	Bars	Wavy	—	F	Multi	Orange	—
	NIZ	—	Star	Fork	—	E	Gray	Red	—
R(3F)	WUG	—	Yin	Twin	Irr	B	Taupe	Purple	7
	JOM	—	Bars	Wavy	Diam	B	Orange	Purple	2
	NIZ	—	Star	Fork	Ears	B	Taupe	Green	2
S(3F)	WUG	Stripe	Yin	Twin	—	A	Taupe	Green	—
	JOM	Dots	Bars	Wavy	—	A	Blue	Green	—
	NIZ	Check	Star	Fork	—	C	Orange	Yellow	—

APPENDIX

(*Continued*)

Stimuli		Critical features				Irrelevant features			
Array	Class	Fill	Insert	Appendages	Base	Shape	Color (O)	Color (I)	Base position
T(3F)	WUG	Stripe	Yin	—	Irr	F	Gray	Blue	7
	JOM	Dots	Bars	—	Diam	D	Brown	Blue	2
	NIZ	Check	Star	—	Ears	D	Multi	Red	11
U(3F)	WUG	Stripe	—	Twin	Irr	E	Gray	Red	2
	JOM	Dots	—	Wavy	Diam	F	Multi	Red	11
	NIZ	Check	—	Fork	Ears	E	Brown	Orange	11
V(4F)	WUG	Stripe	Yin	Twin	Irr	H	Orange	Purple	8
	JOM	Dots	Bars	Wavy	Diam	I	Taupe	Yellow	5
	NIZ	Check	Star	Fork	Ears	G	Orange	Green	5
Novel stimuli: equal features									
W(1F)	WUG	Bric	Cres	Twin	Circ	G	Blue	Green	5
	JOM	Bric	Cres	Wavy	Circ	H	Orange	Purple	1
	NIZ	Bric	Cres	Fork	Circ	G	Blue	Purple	8
X(1F)	WUG	Bric	Yin	Arr	Circ	I	Orange	Green	8
	JOM	Bric	Bars	Arr	Circ	I	Taupe	Green	5
	NIZ	Bric	Star	Arr	Circ	I	Blue	Purple	8
Y(1F)	WUG	Stripe	Cres	Arr	Circ	H	Blue	Purple	1
	JOM	Dots	Cres	Arr	Circ	I	Taupe	Purple	1
	NIZ	Check	Cres	Arr	Circ	H	Orange	Yellow	8
Z(1F)	WUG	Bric	Cres	Arr	Irr	G	Taupe	Yellow	5
	JOM	Bric	Cres	Arr	Diam	H	Blue	Yellow	1
	NIZ	Bric	Cres	Arr	Ears	I	Blue	Yellow	5
AA(2F)	WUG	Stripe	Cres	Arr	Irr	I	Orange	Yellow	8
	JOM	Dots	Cres	Arr	Diam	G	Blue	Yellow	1
	NIZ	Check	Cres	Arr	Ears	H	Blue	Purple	5
BB(2F)	WUG	Bric	Yin	Twin	Circ	E	Multi	Red	1
	JOM	Bric	Bars	Wavy	Circ	F	Multi	Orange	8
	NIZ	Bric	Star	Fork	Circ	E	Gray	Red	8
CC(2F)	WUG	Bric	Yin	Arr	Irr	F	Brown	Orange	5
	JOM	Bric	Bars	Arr	Diam	D	Gray	Blue	1
	NIZ	Bric	Star	Arr	Ears	D	Brown	Blue	8
DD(2F)	WUG	Strip	Cres	Twin	Circ	D	Multi	Blue	5
	JOM	Dots	Cres	Wavy	Circ	E	Brown	Red	5
	NIZ	Check	Cres	Fork	Circ	F	Multi	Orange	5
EE(3F)	WUG	Bric	Yin	Twin	Irr	H	Taupe	Purple	1
	JOM	Bric	Bars	Wavy	Diam	H	Orange	Purple	8
	NIZ	Bric	Star	Fork	Ears	G	Taupe	Green	8
FF(3F)	WUG	Stripe	Yin	Twin	Circ	G	Taupe	Green	5
	JOM	Dots	Bars	Wavy	Circ	G	Blue	Green	1
	NIZ	Check	Star	Fork	Circ	I	Orange	Yellow	5
GG(3F)	WUG	Stripe	Yin	Arr	Irr	F	Gray	Blue	1
	JOM	Dots	Bars	Arr	Diam	D	Brown	Blue	1
	NIZ	Check	Star	Arr	Ears	D	Multi	Red	5
HH(3F)	WUG	Stripe	Cres	Twin	Irr	E	Gray	Orange	8
	JOM	Dots	Cres	Wavy	Diam	F	Multi	Red	5
	NIZ	Check	Cres	Fork	Ears	E	Brown	Orange	1
II(4F)	WUG	Stripe	Yin	Twin	Irr	D	Brown	Red	8
	JOM	Dots	Bars	Wavy	Diam	E	Gray	Orange	8
	NIZ	Check	Star	Fork	Ears	F	Gray	Blue	5

Note. Abbreviations: Check (checkerboard pattern), Stripe (striped pattern), Dots (small dots in pattern color), Yin (yin-yang symbol), Bard (three horizontal bars with a fourth bar crossing through them), Star (starburst image), Twin (v-shaped appendages), Wavy (curved lines), Fork (lines ending in three branches), Irr (irregularly shaped base), Diam (diamond-shaped base), Ears (base with two ear-like points), Bric (brick pattern), Arr (arrow-shaped appendages), Cres (crescent shaped insert), Circ (circle base). Letters A-H in shape column refer to the eight different shape outlines used. Color (O) refers to the color of the shape outside the center circle, and Color (I) refers to the color inside the circle. Base position labels the point at which the base was attached to the shape using the hands of the clock as a reference. 1F, 2F, 3F, 4F = one feature, two features, three features, four features.