VERBAL SELF-REPORTS OF EMERGENT RELATIONS IN A STIMULUS EQUIVALENCE PROCEDURE

SCOTT D. LANE AND THOMAS S. CRITCHFIELD

AUBURN UNIVERSITY

During a preliminary training phase, college students were taught to categorize each of their responses accurately in a conditional discrimination task as either correct or incorrect. Next, in the absence of self-reports, subjects acquired conditional discriminations (involving novel stimuli) prerequisite to the formation of two four-member equivalence classes. The self-report procedure was reinstated during probe sessions that tested for untrained relations indicative of equivalence class formation. Interspersed trials involving trained relations provided a positive control, and trials with no class-consistent comparison provided a negative control. Eight of 10 subjects demonstrated equivalence class formation; all accurately reported their performance on trained relations and on trials with no class-consistent comparison. Subjects also reported their performance on most untrained (emergent) relations accurately, but in several instances self-reports indicated failure or uncertainty despite nearly perfect emergent-relations performance. These inconsistencies add to a growing body of literature that suggests there are differences between individual types of emergent relations. We suggest that the present procedure may be helpful in understanding these differences and other equivalence-related effects.

Key words: equivalence class, matching to sample, differential discrimination, verbal self-reports, joystick response, college students

Match-to-sample (MTS) conditional discrimination training in humans often engenders performances that suggest that all arbitrarily related stimuli come to function equally as both sample and comparison (DeGrandpre & Bickel, 1993; Lazar, 1977; Lazar, Davis-Lang, & Sanchez, 1984; R. Saunders, Wachter, & Spradlin, 1988; Sidman, 1995; Sidman & Cresson, 1973; Sidman & Tailby, 1982). For example, following training on the relations A1-B1 and B1-C1, a subject will then select A1 when A1 is the sample (reflexivity), A1 when B1 is the sample (symmetry), C1 when A1 is the sample (transitivity), and A1 when C1 is the sample (combined symmetry and transitivity, or equivalence). The emergence of equivalent stimulus classes typically is defined by tests for these formal properties of reflexivity, symmetry, transitivity, and equivalence (Fields & Verhave, 1987; Sidman & Tailby, 1982). Describing members of a stimulus class as equivalent thus implies a generalized interchange-

ability that can be demonstrated in transfer-of-function tests. If one stimulus in an equivalence class is endowed with a particular function (e.g., clapping hands if stimulus B1 is present), other members of the class spontaneously come to serve that function as well (e.g., clapping to A1 and C1, see Barnes & Keenan, 1993; Hayes, Kohlenberg, & Hayes, 1991; Wulfert & Hayes, 1988). A variety of functions have been found to propagate across the members of equivalence classes in this way.

Despite the robust findings of transfer-of-function studies, there is reason to believe that under some conditions the individual relations that define the equivalence class may differ in important ways. That is, even when a class is demonstrated to be equivalent by the conventional tests, differences between specific stimulus relations may remain. Some of these differences are not apparent in accuracy scores on traditional MTS test procedures (Fields, Adams, Verhave, & Newman, 1993; Mackay, 1991). Traditional test procedures may, for example, fail to detect spuriously high accuracy scores resulting from control by negative comparisons rather than the programmed contingencies between sample and positive comparison (Carrigan & Sidman, 1992; Johnson & Sidman, 1993); performances on tests of reflexivity that are con-

Portions of this paper were presented at the 1994 meeting of the Association for Behavior Analysis in Atlanta, Georgia. We thank Jennifer Willis and David Martin for invaluable assistance in scheduling subjects and collecting and analyzing data.

Reprint requests may be sent to Scott Lane, Department of Psychiatry and Behavioral Sciences, Texas Health Science Center at Houston, 1300 Moursund St., Houston, Texas 77054 (E-mail: GrabberSL@aol.com).

trolled by the perceptual similarity between sample and comparison rather than the general equivalence relation shown by symmetry and transitivity performances (see R. Saunders & Green, 1992); differential acquisition of transitive test performances as a function of the nodal distance between stimuli (Fields, Adams, Verhave, & Newman, 1990); and dissociations between symmetrical and transitive equivalence relations when baseline contingencies are reversed (Pilgrim & Galizio, 1990, 1995).

The examples just noted highlight the potential limitations of MTS accuracy as a solitary measure of equivalence class formation. It seems unlikely, in fact, that any single dependent measure will prove to be sensitive enough to capture the complexity of all the factors involved. This places a premium on the development of measures that can provide supplemental information about behavior during studies of stimulus equivalence. Supplemental measures might, for example, be informative when subjects fail to form equivalence classes; when equivalence classes form gradually over the course of the equivalence tests; when reversing baseline conditional discriminations produces dissociations between individual relations; and when investigators seek to understand equivalence-related effects such as contextual control or transfer of function.

Several studies have employed supplemental dependent measures of equivalence performances. Latency to MTS choice, for example, has been used in several studies of equivalence class formation. Reaction times have been demonstrated to be significantly different for trained and symmetrical versus transitive relations (Wulfert & Hayes, 1988); for probe trials differing in complexity of relatedness (Steele & Hayes, 1991); and for relations with stimuli separated by zero nodes versus those separated by one node (Bentall, Dickins, & Fox, 1993; Bentall, Dickins, & Jones, 1994¹).

To date, perhaps the most common supplemental measure used in studies of equivalence has been subject self-reports about

some aspect of their equivalence performance. In many cases, the self-reports have been collected via postexperimental interviews or questionnaires. For example, postexperimental verbal reports have been used to explore the role of naming in equivalence class formation (Sidman & Tailby, 1982; Sidman, Willson-Morris, & Kirk, 1986), and to provide post hoc corroboration of the apparent finding that unexpected equivalence test outcomes were due to control by negative comparison stimuli during conditional discrimination training (Dube, Green, & Serna, 1993).

Another approach to procuring self-report data has been to ask subjects to "think aloud" during equivalence training, testing, or both, and then to analyze the correspondence between subjects' verbal behavior and equivalence test performances (Wulfert, Dougher, & Greenway, 1991; Wulfert, Greenway, & Dougher, 1994). The resulting protocol analysis (Ericsson & Simon, 1984) requires potentially complex narrative comments to be coded into categories to facilitate a search for relations between a subject's verbal behavior and concurrent performance. For example, Wulfert and colleagues placed subjects into one of two categories, depending on whether their think-aloud comments tended to describe the relations between stimuli or to describe the stimuli as sample-comparison compounds. Subjects who made relational comments often formed equivalence classes consistent with the conditional discrimination training, whereas subjects who described the stimuli as compounds tended to show equivalence failures (Wulfert et al., 1991).

The present study explored a different way to generate self-reports during the course of an equivalence experiment. Our procedure differed from postexperimental approaches in that we obtained self-reports following every trial of the equivalence tests. Our procedure differed from the protocol analysis approach used by Wulfert and colleagues in two ways. First, our subjects reported on their conditional discrimination performance rather than on thoughts or other covert responses that were taking place concurrently. Second, to facilitate the analysis of correspondence between performance and self-reports, our forced-choice report procedure constrained the range of possible report topographies.

¹ Bentall, R., Dickins, D., & Jones, R. (1994, April). *Beyond the ABC of equivalence: Going the distance*. Paper presented at the Experimental Analysis of Behaviour Group annual conference, London.

Subjects first completed a preliminary training phase that familiarized them with the conditional discrimination trial format and with a self-report procedure in which they categorized the response on each conditional discrimination trial as correct or incorrect and then made a confidence rating about how sure they were of their MTS response. Following the preliminary training, subjects were given the conditional discrimination training necessary for establishing two new four-member stimulus classes. Next, during tests for equivalence class formation, the self-report procedure used during preliminary training was reintroduced. This self-report procedure is based on one used in previous studies to collect subjects' self-reports about delayed identity MTS selections (Critchfield, 1993, 1994; Critchfield & Perone, 1990). In those studies, self-report accuracy generally was high, but deviations between performance and self-reports were systematically related to characteristics of the match-to-sample task. We reasoned that if self-reports in the Critchfield and Perone studies were sensitive to task parameters, then a similar self-report procedure might provide a useful supplemental measure of equivalence-related outcomes.

METHOD

Subjects and Apparatus

Twelve undergraduates volunteered to participate in an experiment described in the informed consent agreement as focusing on "reasoning in symbolic logic." None had previously participated in an experiment involving conditional discriminations. Two subjects (S1 and S5) were dismissed prior to completing the study after failing to attend scheduled sessions. No data are reported for those subjects. Subjects received extra credit in psychology courses based on the amount of time they spent participating in the research. Subjects participated for a total of 8 to 10 hr distributed over a period of 4 to 12 days. Each visit to the laboratory lasted about 1.5 to 2 hr.

Subjects worked alone in a small room containing a table, a chair, a video monitor, a joystick, one pair of stereo headphones, and a fluorescent lamp. They performed the MTS conditional discrimination task by using the joystick to move a cursor to appropriate lo-

cations on the video screen. An IBM®-compatible microcomputer located in an adjacent room was used to control the experimental procedures and collect the data. Headphones permitted the delivery of auditory messages with a SoundBlaster® sound card.

Each subject completed three phases. A preliminary training phase was designed to familiarize subjects with the MTS and self-report procedures and to establish self-reports as a discriminated operant. In the main experiment, a conditional discrimination training phase established the stimulus-stimulus relations necessary to the formation of equivalence classes. No self-reports took place during this phase. Finally, an equivalence test phase assessed performance on the expected emergent relations. No feedback was presented during this phase, and subjects made self-reports about the accuracy of the MTS choices immediately after each trial.

Preliminary Training

Instructions. The informed consent agreement indicated that, "For the purposes of understanding the extent to which logical conclusions are unconscious or 'automatic' we are interested in people's self-evaluations of these logical conclusions as they first emerge." Just before starting the experiment subjects read printed instructions. The first paragraph instructed subjects to calibrate the joystick (without this precaution, *x* and *y* coordinates of the cursor position on the screen tended to drift across sessions). The rest of the instructions are reproduced below:

The selections that you make will be worth seconds. These seconds will add up and constitute the extra credit that you earn in your psychology course. There may be times during the experiment in which there is no correct selection available. If at any time during the experiment it is unclear what to do, or things appear to have changed, please do your best to figure out how to proceed and continue working on your own. There will be a short break between each session, and a longer intermission occurring about every 30 minutes. You will be notified on the monitor screen during each of these breaks. Please do not leave your work space unless there is an emergency (i.e., you are ill, etc.) or you are prompted to do so by the experimenter. Each correct selection and guess about that selection is worth time. **THE EXTRA CREDIT YOU EARN**

WILL BE BASED ENTIRELY ON THE TIME YOU ACCUMULATE DURING THE EXPERIMENT. Your goal is to earn as many seconds, minutes, and hours (i.e., extra credit time) as possible.

The last section of the instructions was intended to establish "seconds" as reinforcers. The experimental protocol actually required that subjects receive credit for all time spent in the study, regardless of their performance, and to be immediately debriefed should seconds accumulated at the end of the study fall short of the amount of time spent participating. All subjects earned the maximum possible extra credit, however, and thus were debriefed about the deception and the nature of the experiment at the end of the academic quarter.

On the first day of the experiment, subjects completed a series of 21-trial sessions (range, 9 to 14, determined by each subject's rate of completing sessions), each involving conditional discrimination trials using stimuli different from those of the main experiment. After each MTS trial, subjects made a self-report about the accuracy of their MTS selection. Immediately following each self-report, subjects rated their confidence in the accuracy of the preceding MTS selection. Following the confidence rating, feedback messages described the accuracy of the MTS selection, the accuracy of the self-report, and the number of seconds earned or lost on that trial (described below). After each session, subjects were given feedback about the number of correct matches, number of accurate self-reports, and number of seconds earned. A written total of seconds earned was given at the end of each visit to the laboratory.

MTS trials. Each trial began with the presentation of four boxes drawn in white against a black background on the video screen. One box, used to display sample stimuli, was centered in the upper part of the screen, and the three remaining boxes, used to display three comparison stimuli, were distributed horizontally across the bottom of the screen, as shown in Figure 1. A small red rectangular cursor appeared in the sample box just below the sample stimulus, accompanied by the printed message, "press the top orange joystick button to continue." A press cleared the message and activated the joystick, allowing subjects to move the cursor

Fig. 1. Graphic display of the computer screen, including the stimuli used, during each segment of a typical test trial. The top of the figure shows the eight stimuli used and the classes in which they were trained. The next panel shows the MTS boxes and sample and comparison stimuli. The third panel shows the self-report display. The bottom panel shows the confidence rating display.

around the screen. A new message then prompted subjects to "Use the joystick to move the cursor into one of the lower boxes and push the top orange joystick button to make a selection." After the first five trials of the experiment, the instructions did not appear again. A response inside one of the three comparison boxes cleared the MTS screen and immediately initiated the self-report procedure. The basic MTS trial format remained unchanged through the rest of the experiment.

Table 1
Stimuli used during the preliminary training phase.

Set number	Stimuli in group	Color
1 (A1, A2, A3) (B1, B2, B3)	411 41 41 41 41 41	All white All white
2 (B1, B2, B3) (C1, C2, C3)	41 41 41 41 41 41	All white All white
3 (A1, A2, A3) (B1, B2, B3)	411 41 41 41 41 41	(Blue, green, cyan) (Red, magenta, brown)
4 (B1, B2, B3) (C1, C2, C3)	41 41 41 41 41 41	(Red, magenta, brown) (White, gray, light blue)
5 (xA1, xA2, xA3) (xB1, xB2, xB3)	41 41 41 41 41 41	All white All white
6 (xB1, xB2, xB3) (xC1, xC2, xC3)	41 41 41 41 41 41	All white All white
7 (xA1, xA2, xA3) (xB1, xB2, xB3)	41 41 41 41 41 41	(Blue, green, cyan) (Red, magenta, brown)
8 (xB1, xB2, xB3) (xC1, xC2, xC3)	41 41 41 41 41 41	(Red, magenta, brown) (White, gray, light blue)

Stimuli used in preliminary training were created by combining two ASCII character codes from the range 179 to 218 and printing the resulting shapes in varying colors (see Table 1). Trials involving two groups of three stimuli were intermingled in each training session. These two groups of three stimuli were arranged as three two-member classes per set. As shown in Table 1, there were eight of these stimulus sets. Normal adult subjects tend to acquire conditional discriminations rapidly in traditional MTS training formats (K. Saunders, Saunders, Williams, & Spradlin, 1993). Therefore, errors were induced by introducing new sets of stimuli at the start of every odd-numbered preliminary training session (1, 3, 5, 7, etc.), for two reasons. It was important that subjects make both correct and incorrect MTS selections (a) to contact contingencies designed to promote accurate self-reports (described below) and (b) to allow an analysis of the extent to which self-reports and confidence ratings varied as a function of MTS accuracy.

Table 2 shows the arrangement of stimulus presentations used during the preliminary training. Note that S8 through S12 also received interspersed sets of identity matching trials (described below). For each stimulus set in the preliminary training, correct matches were defined by experimentally designated sample-comparison stimulus pairs (e.g., A1-B1, A2-B2, and A3-B3). On each trial (except "catch" trials), stimuli from the other two

classes served as negative comparisons (A1-B1, with B2 and B3 as S-; see Table 2).

Preliminary training sessions were comprised of 18 standard conditional discrimination trials and three catch trials (described below). Excluding catch trials, each sample stimulus appeared six times per session and each comparison stimulus appeared 18 times per session. Each comparison stimulus occurred an equal number of times in each comparison location; no configuration of comparison stimuli occurred more than twice consecutively; and no correct comparison appeared in the same box more than twice consecutively. On catch trials, no class-consistent comparison was presented (e.g., Sample A1 with Comparisons A2, B2, and B3). Each sample stimulus (e.g., A1, A2, and A3) was presented once per session with no class-consistent comparison stimulus available. These trials were included to provide a measure of sensitivity of self-reports. If self-reports reliably distinguished correct from incorrect MTS responses, then responses on catch trials would always be categorized as incorrect.

The 10 subjects who completed the experiment were treated identically except that S8 through S12 completed a block of 18 identity matching trials after every two sessions of arbitrary matching during the preliminary training. The stimuli were from the same set presented in the preceding MTS session. Stimulus arrangements for identity matching blocks are shown in Table 2. The identity

Table 2

Stimulus arrangements used during the preliminary training phase. For ease of visual inspection, choices that were reinforced are underlined. In each session comparison stimuli (CO) appeared an equal number of times in each comparison location, with the exception of catch trials. Each subject was exposed to each set arrangement twice before the next set was introduced (i.e., 1, 1, 2, 2, 3, 3, etc.).

Set	Sample	CO 1	CO 2	CO 3	Times presented
1 (white S)	A1	B1	B2	B3	6
	A2	<u>B2</u>	B1	B3	6
	A3	<u>B3</u>	B1	B2	6
	A1	<u>B2</u>	B3	A2	1 ^a
	A2	B1	B3	A3	1 ^a
2 (white S)	A3	B1	B2	A1	1 ^a
	B1	C1	C2	C3	6
	B2	<u>C2</u>	C1	C3	6
	B3	<u>C3</u>	C1	C2	6
	B1	<u>C2</u>	C3	B2	1 ^a
3 (colored S)	B2	C1	C3	B3	1 ^a
	B3	C1	C2	B1	1 ^a
	A1	B2	B1	B3	6
	A2	<u>B3</u>	B1	B2	6
	A3	<u>B1</u>	B2	B3	6
4 (colored S)	A1	<u>B2</u>	B3	A2	1 ^a
	A2	B1	B3	A3	1 ^a
	A3	B1	B2	A1	1 ^a
	B1	C2	C1	C3	6
	B2	<u>C3</u>	C1	C2	6
5 (white S)	B3	<u>C1</u>	C2	C3	6
	B1	<u>C2</u>	C3	C2	1 ^a
	B2	C1	C3	C3	1 ^a
	B3	C1	C2	C1	1 ^a
	xA1	<u>xB1</u>	xB2	xB3	6
6 (white S)	xA2	<u>xB2</u>	xB1	xB3	6
	xA3	<u>xB3</u>	xB1	xB2	6
	xA1	<u>xB2</u>	xB3	xA2	1 ^a
	xA2	xB1	xB3	xA3	1 ^a
	xA3	xB1	xB2	xA1	1 ^a
7 (colored S)	xB1	<u>xC1</u>	xC2	xC3	6
	xB2	<u>xC2</u>	xC1	xC3	6
	xB3	<u>xC3</u>	xC1	xC2	6
	xB1	<u>xC2</u>	xC3	xB2	1 ^a
	xB2	xC1	xC3	xB3	1 ^a
8 (colored S)	xB3	xC1	xC2	xB1	1 ^a
	xA1	xB2	xB1	xB3	6
	xA2	<u>xB3</u>	xB1	xB2	6
	xA3	<u>xB1</u>	xB2	xB3	6
	xA1	<u>xB2</u>	xB3	xA2	1 ^a
	xA2	xB1	xB3	xA3	1 ^a
	xA3	xB1	xB2	xA1	1 ^a
	xB1	<u>xC2</u>	xC1	xC3	6
	xB2	<u>xC3</u>	xC1	xC2	6
	xB3	<u>xC1</u>	xC2	xC3	6
	xB1	<u>xC2</u>	xC3	xC2	1 ^a
	xB2	xC1	xC3	xC3	1 ^a
	xB3	xC1	xC2	xC1	1 ^a

Table 2

(Continued)

Set	Sample	CO 1	CO 2	CO 3	Times presented
Identity MTS ^b					
1 (white S)	A1	A1	B2	B3	6
	A2	<u>A2</u>	B1	B3	6
	A3	<u>A3</u>	B1	B2	6
2 (white S)	B1	<u>B1</u>	C2	C3	6
	B2	<u>B2</u>	C1	C3	6
	B3	<u>B3</u>	C1	C2	6
3 (colored S)	xA1	<u>xA1</u>	xB2	xB3	6
	xA2	<u>xA2</u>	xB1	xB3	6
	xA3	<u>xA3</u>	xB1	xB2	6
4 (colored S)	xB1	<u>xB1</u>	xC2	xC3	6
	xB2	<u>xB2</u>	xC1	xC3	6
	xB3	<u>xB3</u>	xC1	xC2	6

^a Catch trials in which no correct comparison was available.

^b Subjects 8 through 12 were exposed to each identity set following the presentation of two arbitrary MTS sessions with a given set (i.e., Arbitrary 1, Arbitrary 1, Identity 1, Arbitrary 2, Arbitrary 2, Identity 2).

matching blocks were included after some of the first 5 subjects produced consistently inaccurate self-reports on reflexivity test trials. Identity matching during preliminary training did not appear to bear on the results in the main study, however, and will be addressed hereafter only briefly in the Discussion.

Self-reports and confidence ratings. Immediately following each MTS selection, the screen cleared and two boxes, outlined in white, appeared in the upper corners of the screen, with the cursor in the center of the screen. "Correct" was printed in the center of the left box, and "incorrect" was printed in the center of the right box (Figure 1). On the first five trials, a vocal query² presented through the headphones asked, "Was your selection correct or incorrect?" Subjects moved the cursor into one of the two boxes and pushed the joystick button to make a selection. The selected box became highlighted by the addition of two larger boxes surrounding it. Throughout the rest of the preliminary training, the vocal query occurred randomly

² Headphones were used to pilot auditory stimuli for use in studies with children too young to read printed messages. The recorded messages are not essential to the present procedures.

on 25% of the trials. On trials with no vocal query, the self-report boxes simply appeared on the screen until subjects made a selection.

A confidence rating took place following a selection of the "correct" or "incorrect" self-report box. The cursor automatically recentered and two columns of four yellow boxes appeared, one column on each side of the screen (Figure 1). The boxes contained values ranging from "0 seconds" to "21 seconds," in increments of 3 s. The word "sure" appeared in yellow to the right of the 21 seconds box and "unsure" appeared to the left of the 0 seconds box. On the first five trials, the recorded vocal query, "How confident are you about your selection?" prompted the selection of one of the yellow boxes. Throughout the remainder of the preliminary training, the query occurred randomly on 25% of the trials. Subjects moved the cursor into one of the eight boxes and pushed the joystick button to make a selection. As a result, points were gained or lost depending on the magnitude of the rating and the accuracy of the MTS selection.

Feedback and consequences. On each preliminary training trial, a feedback screen immediately followed the confidence rating. At the top of the screen, printed in large letters, was either "correct" or "incorrect," consistent with the just-preceding MTS selection. Three printed messages appeared underneath. The first stated either "Your selection was correct = +25 seconds" or "Your selection was incorrect = +0 seconds." On the first five trials a recorded vocal message stating either "correct selection" or "incorrect selection," consistent with the subject's MTS response, accompanied this message. Throughout the remainder of the preliminary training, vocal feedback messages occurred randomly on 25% of the trials.

The second printed message, appearing just below the first, stated, "your report was correct = $+x$ seconds" or "your report was incorrect = $-x$ seconds." The value of x was based on the confidence rating made by the subject on that trial. Following accurate self-reports, the subject was awarded the number of seconds "wagered" in the confidence rating (0 to 21 in increments of 3). Following inaccurate self-reports, the subject lost seconds based on the magnitude of the confidence rating. When subjects falsely reported

that they had been "incorrect," the deduction equaled the magnitude of the rating. When subjects falsely reported that they had been "correct," the deduction was double the magnitude of the confidence rating. For example, following an incorrect MTS response, a self-report of "correct," accompanied by a confidence rating of 21 seconds, would produce the feedback message, "your report was incorrect = -42 seconds ($\times 2$)." The " $(\times 2)$ " portion of the message was printed in red flashing characters. The purpose of this extra penalty was to discourage overconfidence in self-reports (e.g., overreporting success), a pattern that had been observed both in pilot research and in previous studies using similar self-report procedures (e.g., Critchfield, 1993, 1994).

The third feedback message summarized the point consequences for the trial, reading, "total time earned = $\pm x$ seconds." The value of x was the sum of earnings and losses contingent on both MTS selections and self-reports and confidence ratings. On each trial a subject could earn up to 46 seconds (25 for a correct MTS selection plus 21 for an accurate self-report) or lose a maximum of 42 seconds (0 for an incorrect MTS selection minus as much as 42 for falsely reporting MTS success). Preliminary training took place on the 1st day of the experiment and lasted until the end of the 2-hr visit for that day.

Main Experiment

Conditional discrimination training procedures. The basic trial format was similar to that at the conclusion of the preliminary training except that (a) no self-reports occurred, (b) the stimuli and stimulus arrangements were different, and (c) no catch trials occurred. Conditional relations consistent with two four-member stimulus classes were trained using the procedures summarized at the top of Table 3. The stimuli were red and had no intended topographic similarity to each other or to the stimuli used in the preliminary training. No self-reports or confidence ratings took place during this phase until just before the equivalence tests. Until that point, subjects earned either 25 or 0 seconds (based solely on MTS performance) following each trial.

Conditional discrimination training employed a "sample-as-node" procedure, in

Table 3

Stimulus arrangement used during each phase of the study. For ease of visual inspection, choices that were reinforced during training are underlined. During conditional discrimination training, each comparison stimulus (CO) occurred an equal number of times in each comparison location during each session. During equivalence testing, each comparison stimulus occurred an equal number of times in each comparison location across three sessions. For probe trials, experimentally defined class-consistent selections are listed under CO 1.

Relation type	Sample	CO 1	CO 2	CO 3	Times presented per session
AB training	A1	B1	B2	C2	12 (trn), 6 (mix), 2 (probe)
	A2	<u>B2</u>	B1	C1	12 (trn), 6 (mix), 2 (probe)
AC training	A1	<u>C1</u>	C2	D2	12 (trn), 6 (mix), 2 (probe)
	A2	<u>C2</u>	C1	D1	12 (trn), 6 (mix), 2 (probe)
AD training	A1	<u>D1</u>	D2	B2	12 (trn), 6 (mix), 2 (probe)
	A2	<u>D2</u>	D1	B1	12 (trn), 6 (mix), 2 (probe)
Catch trials	A1	<u>B2</u>	C2	D2	1
	A2	B1	C1	D1	1
	B1	A2	C2	D2	1
	B2	A1	C1	D1	1
	C1	A2	B2	D2	1
	C2	A1	B1	D1	1
	D1	A2	B2	C2	1
	D2	A1	B1	C1	1
Reflexivity	A1	A1	A2	C2	1
	A2	A2	A1	C1	1
	B1	B1	B2	D2	1
	B2	B2	B1	D1	1
	C1	C1	C2	A2	1
	C2	C2	C1	A1	1
	D1	D1	D2	B2	1
	D2	D2	D1	B1	1
Symmetry	B1	A1	A2	C2	3
	B2	A2	A1	C1	3
	C1	A1	A2	D2	3
	C2	A2	A1	D1	3
	D1	A1	A2	B2	3
	D2	A2	A1	B1	3
Combined (symmetry and transitivity)	B1	C1	C2	A2	2
	B2	C2	C1	A1	2
	B1	D1	D2	A2	2
	B2	D2	D1	A1	2
	C1	D1	D2	A2	2
	C2	D2	D1	A1	2
	C1	B1	B2	A2	2
	C2	B2	B1	A1	2
	D1	B1	B2	A2	2
	D2	B2	B1	A1	2
	D1	C1	C2	A2	2
	D2	C2	C1	A1	2

which a single sample stimulus (A) served for each of three comparison stimuli (B, C, and D). Thus, six relations were trained (A1-B1, A1-C1, A1-D1, A2-B2, A2-C2, A2-D2) in order to establish the prerequisites necessary for two four-member classes. Trials were presented in random order, except that no trial type occurred on more than three consecutive trials. The positive comparison stimulus never occurred in the same box on more than three consecutive trials. Placement of correct

and incorrect comparison stimuli was counterbalanced so that each stimulus occurred in each of the three comparison boxes an equal number of times during a single session. The stimuli used in the main experiment are reproduced above the top panel of Figure 1. Table 3 shows how the A-B, A-C, and A-D relations were trained separately in consecutive blocks of 24-trial sessions, with A1 and A2 each presented as sample stimulus on 12 trials per session (e.g., 12 trials A1-B1 and 12

Table 4

Summary of experimental manipulations for each of the different experimental conditions.

Condition	Relation(s) trained or tested	Self-reports and confidence wagers?	<i>p</i> of auditory queries or feedback	<i>p</i> of vi- sual feed- back	Catch trials?	Tri- als per ses- sion	Mastery criterion
Preliminary training (Day 1)	One set of three two-member relations	Yes	1.0	1.0	Only on trials 16-21	21	None
Conditional dis- crimination training (Days 2 and 3)	A-B, A-C, A-D	No	.25	1.0	No	24	3 of 4 sessions at 100%, 4th > 95%
Same as above	All types mixed	No	.25	1.0	No	36	2 of 3 sessions at 100%, 3rd > 92%
Same as above	All types mixed	No	.25	.75	No	36	Same as above
Same as above	All types mixed	No	.25	.35	No	36	Same as above
Same as above	All types mixed	On last 12 trials	.25	.10	No	36	Same as above
Equivalence re- lations tests (Days 3-5)	Reflexive, sym- metrical com- bined (symm/ trans)	Yes	.25 (queries only)	0	Yes	70	None

trials A2-B2). Feedback followed each trial. Mastery of each relation was defined as 100% correct on three consecutive sessions or on three of four consecutive sessions with the fourth at higher than 95% (see last column in Table 4). The AB, AC, and AD relations were each trained, in that order, until mastery was achieved.

After all relations were mastered individually, they were intermingled in 36-trial mixed sessions, in which each of the six relations was presented six times under the same counter-balanced conditions as the 24-trial sessions (see Table 3). Probability of feedback on these sessions was 1.0 until performance reached 100% on two consecutive sessions, or on two of three consecutive sessions with the third above 92%. The probability of feedback was then reduced to .75 with the same mastery criterion, then .35, then .10. Following trials for which no feedback was given, the screen cleared and the printed words "information withheld" appeared in the center of the screen for 1 s.

When the feedback probability reached .10, the self-report and confidence rating procedures used in preliminary training were re-introduced on the last 12 trials of each 36-trial session. Thus, subjects made at least 24

self-reports and confidence ratings immediately prior to equivalence tests (12 self-reports in each of the minimum two sessions required to meet the mastery criterion). There were no catch trials during any conditional discrimination sessions prior to the equivalence tests. Sessions with .10 feedback probabilities always occurred on the same day as the introduction of the equivalence tests.

Following the final training sessions, the experimenter asked the subjects if they remained familiar with the self-report and confidence rating procedures by asking, "Did you remember the ratings from the first day of the experiment?" All subjects answered affirmatively. Subjects were then told by the experimenter, "This situation will now be in place for the remainder of the experiment. You have the same opportunity to gain or lose extra seconds as you did on the first day of the experiment."

Equivalence probes. Table 3 shows the emergent relations test trial arrangements. Equivalence test sessions consisted of 70 trials of all trained and potential emergent relations derivable from the conditional discrimination training. Each session included eight reflexivity test trials and 18 symmetry test trials. With the sample-as-node training procedure

employed in this investigation, the remaining tests are considered to be a combined test for symmetry and transitivity (R. Saunders & Green, 1992). Thus, for example, the relation B-C constitutes the same relation type as C-B. We will refer to these relations as *combined* tests. Each emergent relations test session included 24 combined test trials. Interspersed among the emergent relations test trials were 12 training trials and eight catch trials in which no class-consistent MTS selection was available (all comparisons were from the class to which the sample stimulus did not belong; see Table 3).

Trials were presented in random order, with the exception that no trial type or test type occurred on more than three consecutive trials. The correct comparison stimulus never occurred in the same location on more than three consecutive trials. Across three consecutive sessions, the positive comparison stimulus was presented an equal number of times in each of the three possible locations (the four-member class arrangements did not permit counterbalancing of the positive comparison stimulus location within a single session for all relation types).

Following each of the trials in emergent relations test sessions, subjects made self-reports and confidence ratings in the same format as during the preliminary training (including vocal queries on 25% of the trials) except that no feedback occurred on any of the trials or at the end of a session. Instead, the "information withheld" message appeared on the screen at the end of each trial and at the end of the session.

Emergent relations test sessions were presented until both MTS performance and self-report confidence ratings appeared to be stable upon visual inspection (9 to 11 sessions per subject). Subjects received a written record of their total seconds accumulated at the end of each visit to the laboratory.

RESULTS

Preliminary Training

One purpose of the preliminary training was to insure that prior to the start of the main experiment self-reports were sensitive to actual MTS performance—that is, that subjects tended to report differently following

Fig. 2. Results of the preliminary training procedures. The top panel shows the mean confidence rating about correct (filled circles) and incorrect (open circles) MTS responses for all 10 subjects. The manner in which the confidence rating data were derived is described in the text. Identity matching sessions are removed from the data (subjects typically made no incorrect responses on identity matching sessions). The bottom two rows of panels show confidence ratings for correct and incorrect MTS responses for individual subjects on the first two and last two sessions of preliminary training.

correct MTS responses than following incorrect ones. The introduction of new stimulus sets every two sessions produced a variety of levels of MTS accuracy during preliminary training sessions (mean MTS accuracy across all subjects = 55%, range, 6% to 100%). Thus, all subjects presumably made enough errors to contact the self-report contingencies, and Figure 2 suggests that, for 9 of 10 subjects, the preliminary training had its intended effect on self-reports. For economy of presentation, self-reports and confidence ratings were integrated by assigning a positive value to the confidence rating when subjects reported that their MTS response was correct and a negative value when subjects reported that their MTS response was incorrect. The resulting scale ranged from +21 (very sure of a correct MTS response) to -21 (very sure

of an incorrect MTS response). Mean confidence ratings were derived separately for trials involving correct and incorrect MTS responses. This integrated measure will be referred to simply as the confidence rating.

The top panel of Figure 2 shows mean confidence ratings for all 10 subjects across the first nine sessions (the maximum number completed by all 10 subjects) of preliminary training. Identity matching trials are excluded for S8 through S12. Each smaller panel shows an individual subject's confidence ratings on the first and last two sessions of preliminary training. Separate functions summarize confidence ratings on correct and incorrect MTS responses. In general, confidence ratings about correct and incorrect MTS responses were similar in the early sessions, but tended to diverge during the later sessions. An analysis of variance (ANOVA) corroborated the visually apparent main effect of MTS success on confidence rating, $F(1, 9) = 38.1, p = .0002$, and the interaction between confidence rating and sessions, $F(8, 9) = 4.1, p = .0005$. The lower panels of Figure 2 show that this general pattern held for all subjects but S4. The most important feature of these data is that 9 of 10 subjects entered the main study reliably discriminating correct from incorrect MTS responses.

Conditional Discrimination Training

Table 5 shows the number of sessions required by each subject to meet the mastery criteria at each stage of the conditional discrimination training and the percentage of correct MTS responses in each session. Data are collapsed across trial types (e.g., A1-B1 combined with A2-B2). Subjects typically met the mastery criterion within two to five sessions per stage. Thus, the prerequisite relations for equivalence class formation were in place by the beginning of the equivalence test sessions.

Self-Reports During Equivalence Tests

Table 6 summarizes properties of self-reports during the equivalence testing phase. For trained relations, self-reports were considered to be accurate if they were consistent with the status of the preceding MTS selection as defined by the contingencies during conditional discrimination training (e.g., reporting "correct" after choosing the correct

comparison). For catch trials, self-reports of "incorrect" were considered to be accurate and self-reports of "correct" were considered to be inaccurate. For untrained relations, self-reports were considered to be accurate if they were consistent with the status of the MTS selection, as defined by the expected emergent relations (e.g., reporting correct after selecting C1 in response to B1).

Table 6 shows that self-report accuracy on trained and catch trials was near 100% for all subjects. Accuracy also was generally high for self-reports about the emergent relations (24 of 31 cases higher than 90%). Table 6 also shows mean confidence ratings following class-consistent (correct) and class-inconsistent (incorrect) conditional discrimination responses on all emergent relations trials during all sessions of equivalence testing. Self-reports and confidence ratings were integrated to produce the bipolar scale described previously for preliminary training data (+21 = very sure of a correct MTS response; -21 = very sure of an incorrect MTS response). Because subjects tended to make relatively few conditional discrimination errors during the equivalence tests, it was not possible to conduct this analysis separately for each type of emergent relation. Thus, data were summed across reflexive, symmetrical, and combined test trials before analysis. In general, just as at the conclusion of preliminary training, self-reports reliably distinguished correct from incorrect conditional discrimination responses, as evidenced by the generally positive ratings following correct MTS responses and generally negative ratings following incorrect MTS responses.

Table 6 also shows a measure of self-report bias, the difference between the percentage of trials on which subjects reported correct conditional discrimination responses and the percentage of trials on which correct responses actually occurred. A positive score indicates an overestimation of conditional discrimination success, and a negative score indicates an underestimate. There was no consistent pattern of bias across subjects. All deviations greater than 10% reflected underestimates, an outcome consistent with the self-report contingencies.

Figure 3 shows the relationship between MTS performance and confidence ratings across each of the sessions of equivalence test-

Table 5
Summary of conditional discrimination training results as percentage correct for each subject, each relation trained, and each feedback probability.

Subject	Condition trained	Feedback proba- bility	Sessions								
			1	2	3	4	5	6	7	8	9
S2	AB	1.0	.41	.83	1.0	1.0	1.0				
	AC	1.0	.71	.92	1.0	1.0	1.0				
	AD	1.0	.83	1.0	1.0	.95	1.0				
	Mixed	1.0	.97	1.0	1.0						
	Mixed	.75	1.0	1.0							
	Mixed	.35	1.0	1.0							
	Mixed	.10	1.0	.97	1.0						
S3	AB	1.0	.79	1.0	1.0	1.0					
	AC	1.0	.71	1.0	1.0	1.0	1.0				
	AD	1.0	.88	1.0	.96	1.0					
	Mixed	1.0	.94	.97	1.0	1.0					
	Mixed	.75	1.0	1.0	1.0						
	Mixed	.35	1.0	1.0							
	Mixed	.10	.97	1.0							
S4	AB	1.0	.41	.88	1.0	1.0	1.0				
	AC	1.0	.71	1.0	.96	1.0	1.0				
	AD	1.0	.88	1.0	1.0	1.0					
	Mixed	1.0	.69	.88	1.0	.97	1.0				
	Mixed	.75	1.0	.97	1.0						
	Mixed	.35	.97	1.0	1.0						
	Mixed	.10	1.0	1.0							
S6	AB	1.0	.17	.88	1.0	.96	.92	.92	.96	1.0	1.0
	AC	1.0	.88	1.0	1.0	1.0					
	AD	1.0	1.0	1.0	1.0						
	Mixed	1.0	1.0	1.0							
	Mixed	.75	.97	1.0	1.0						
	Mixed	.35	1.0	1.0							
	Mixed	.10	.97	1.0	1.0						
S7	AB	1.0	.88	1.0	1.0	1.0					
	AC	1.0	.88	1.0	1.0	1.0					
	AD	1.0	.83	1.0	1.0	1.0					
	Mixed	1.0	1.0	1.0							
	Mixed	.75	1.0	1.0							
	Mixed	.35	1.0	1.0							
	Mixed	.10	1.0	1.0							
S8	AB	1.0	.71	.96	1.0	1.0	1.0				
	AC	1.0	.83	.96	.92	1.0	1.0	1.0			
	AD	1.0	.96	.83	1.0	1.0	.96	1.0			
	Mixed	1.0	.97	.81	.89	1.0	.97	.97	1.0	1.0	
	Mixed	.75	1.0	.97	1.0	1.0					
	Mixed	.35	1.0	1.0							
	Mixed	.10	.94	1.0	1.0						
S9	AB	1.0	.47	.88	.92	.96	1.0	.96	1.0	1.0	
	AC	1.0	.79	1.0	1.0	1.0					
	AD	1.0	1.0	1.0	1.0						
	Mixed	1.0	1.0	1.0							
	Mixed	.75	1.0	1.0							
	Mixed	.35	.97	1.0	1.0						
	Mixed	.10	1.0	1.0							
S10	AB	1.0	.71	1.0	1.0	1.0					
	AC	1.0	.38	1.0	1.0	1.0					
	AD	1.0	.96	1.0	1.0	.98	1.0				
	Mixed	1.0	1.0	1.0							
	Mixed	.75	1.0	1.0							
	Mixed	.35	.97	1.0	1.0						
	Mixed	.10	1.0	1.0							

Table 5
(Continued)

Subject	Condition trained	Feedback probability	Sessions								
			1	2	3	4	5	6	7	8	9
S11	AB	1.0	.58	1.0	1.0	1.0					
	AC	1.0	.74	1.0	1.0	1.0					
	AD	1.0	.63	1.0	1.0	1.0					
	Mixed	1.0	.97	1.0	1.0						
	Mixed	.75	1.0	1.0							
	Mixed	.35	1.0	1.0							
	Mixed	.10	1.0	1.0							
S12	AB	1.0	.54	1.0	1.0	1.0					
	AC	1.0	1.0	1.0	1.0						
	AD	1.0	.95	1.0	1.0	1.0					
	Mixed	1.0	1.0	1.0							
	Mixed	.75	1.0	1.0							
	Mixed	.35	1.0	1.0							
	Mixed	.10	1.0	1.0							

ing. In general, MTS response accuracy was high on the trained relations for all subjects. Eight of 10 subjects also demonstrated performances at or above 95% on all three emergent relations (symmetry, reflexivity, and combined equivalence). In the only exceptions (S10's combined relations and S4's reflexive and combined relations), performance never exceeded chance (33% with three comparisons). Confidence ratings following correct MTS responses are shown as connected filled circles. Incorrect MTS responses occurred infrequently, so the data points for confidence ratings following them

are left unconnected. Comparison of the filled circles and open triangles confirms the outcome shown in broad terms in Table 5. Self-reports reliably distinguished correct from incorrect MTS responses during the equivalence tests. That is, in sessions containing both correct and incorrect MTS responses, confidence ratings tended to be more positive following correct responses. The only consistent exception to this pattern occurred in the three cases in which conditional discrimination performance failed to exceed chance (S10's combined relations and S4's reflexive and combined relations). In

Table 6
Characteristics of self-reports during equivalence testing sessions.

Subject	% Accurate self-reports of MTS class consistency						Confidence rating when MTS response is		Over- or underestimate of class-consistent MTS responses					
	Trained	Catch	Re-flex-ive	Sym-metri-cal	Com-bined	Emer-gent	Class consistent (SE)	Class inconsistent (SE)	Trained	Catch	Re-flex-ive	Sym-metri-cal	Com-bined	All emergent
S2	100	100	66	100	100	94	14.9 (2.3)	— ^a	0.0	0.0	-33.4	0.0	-0.4	-5.6
S3	100	100	100	100	100	100	12.4 (1.1)	— ^a	0.0	0.0	0.0	0.0	0.0	0.0
S4	100	100	65	100	68	79	-6.5 (3.6)	-21.0 (0.0)	0.0	0.0	0.0	0.0	-35.0	-21.2
S6	100	100	96	98	98	98	15.8 (1.4)	-5.6 (2.4)	0.0	0.0	4.2	1.9	1.9	2.2
S7	100	100	3	99	100	84	7.4 (3.5)	— ^a	0.0	0.0	1.3	0.0	0.4	0.5
S8	99	100	100	99	89	94	17.1 (1.0)	-10.0 (3.8)	0.7	0.0	0.0	0.0	1.9	0.9
S9	99	100	93	98	99	98	7.8 (0.6)	-5.3 (1.6)	-0.8	0.0	-5.1	0.6	0.0	-0.6
S10	100	100	99	99	58	79	7.0 (3.6)	-20.8 (0.2)	0.0	0.0	1.3	-0.6	-42.5	-20.4
S11	97	96	100	96	100	98	20.1 (0.2)	-21.0 (0.0)	0.9	4.2	0.0	4.3	0.0	1.6
S12	99	100	86	99	100	97	14.7 (2.0)	10.5 (6.1)	0.9	0.0	-11.1	1.2	0.0	-1.3

Note. MTS, match to sample; SE, standard error of the measurement. Mean confidence ratings represent all three types of emergent relations tests (reflexive, symmetrical, and combined) for all test sessions.

^a Fewer than five incorrect match-to-sample responses.

Fig. 3. Percentage of accurate MTS responses (open bars, left y axis) and confidence ratings following correct MTS responses (filled circles, right y axis) and incorrect MTS responses (open triangles, right y axis) across all equivalence test sessions for all relation types. The manner in which the confidence ratings were derived is described in the text. Missing triangles indicate either 100% accuracy or identical (overlapping) mean ratings for correct and incorrect MTS.

these cases, confidence ratings were routinely and strongly negative, regardless of conditional discrimination performance.

Figure 3 also shows performance and confidence ratings for trained relations. For all subjects, MTS performance on trained relations was maintained during the equivalence tests, although S6 exhibited transient disruption of trained performances during the first two equivalence test sessions. For all subjects, confidence ratings on trained relations were correspondingly high (near +21), although S6 showed evidence of acquisition effects in parallel with changes in conditional discrimination performance. Figure 3 also shows that confidence ratings on catch-trial responses (by definition, always incorrect)

were uniformly negative (near -21), although S6 and S8 showed evidence of a gradual acquisition of this pattern. This contrast between confidence ratings on trained relations and catch trials lends confidence to the view of self-reports as generally being sensitive to changes in conditional discrimination performance.

Figure 4 categorizes patterns shown in Figure 3, summarizing correspondence between MTS performance and self-reports for the three untrained trial types (symmetry, reflexivity, and combined). Conditional discrimination and confidence-rating functions were coded by two naive independent raters using procedures described in the Appendix. Too few conditional discrimination er-

Fig. 3. Continued.

Fig. 4. Summary of conditional discrimination and confidence rating patterns during equivalence testing.

rors occurred to facilitate coding of patterns of confidence ratings following incorrect MTS responses. Each cell of Figure 4 shows the conjunction of one conditional discrimination pattern and one confidence-rating pattern. Shown within each cell is the total number of panels from Figure 3 that were judged to fit the relevant categories, plus a breakdown of the relations for each subject that was categorized within that cell.

The three shaded cells, encompassing 25 of 30 relations across the 10 subjects, represent cases of qualitative correspondence between conditional discrimination and confidence-rating functions. In 18 of these cases, immediately accurate conditional discrimination performance was routinely and immediately accompanied by positive confidence ratings. In four cases, gradual acquisition of conditional discrimination performance was accompanied in the early sessions of equivalence testing by gradual increases in confidence ratings. In three cases,

chance conditional discrimination performance was accompanied by routinely negative and immediately low confidence ratings.

The five remaining cases show apparent incongruities between performance and confidence ratings. In all five cases, untrained reflexive relations emerged immediately at near 100% accuracy, but confidence ratings sometimes indicated otherwise. In three cases (S2, S3, and S8), ratings tended to be either negative or near zero (a point of uncertainty) during initial equivalence testing sessions, before consistently indicating conditional discrimination success thereafter. In a fourth case (S7), ratings were near zero during the first testing sessions, and thereafter responses were consistently rated as incorrect, whereas in the fifth (S12), ratings initially indicated conditional discrimination success but thereafter were essentially variable or near zero, again indicating a general uncertainty.

DISCUSSION

Following conditional discrimination training, 8 of 10 subjects produced MTS performances indicative of the formation of two four-member equivalence classes. Subjects had learned during preliminary training to make self-reports about their MTS selections, and when the self-report procedure was reinstated during tests for the emergence of untrained relations, self-reports generally corresponded to MTS performance (Table 6). For example, when MTS selections on test trials were consistent with the emergent relations expected based on the conditional discrimination training contingencies, subjects usually reported their selections as correct and expressed strong confidence in the accuracy of their selections. Five of 10 subjects, however, showed either transient or persistent noncorrespondence between MTS performance and self-descriptions of that performance. In each of these cases, MTS performances on reflexive relations were near 100% consistent with the defined equivalence relation, but self-reports described some lower level of accuracy. Interestingly, 3 subjects (S2, S3, and S7) showed this pattern despite performing virtually without error on any of the equivalence tests, and 1 subject (S12) showed a pattern of uncertainty about reflexive relations despite a preliminary training history with identity matching.

Although there are many reasons to treat the present data as preliminary, it is probably significant that self-report patterns were not perfectly predictable from MTS performance. At a broad level, this outcome supports the assertion that percentage correct on emergent relations tests constitutes a limited measure of behavior during equivalence class formation (Mackay, 1991; Pilgrim & Galizio, 1995; Sidman, 1980, 1995), and provides cause for optimism about the value of further developing self-report methods for use in stimulus equivalence studies. In many previous studies, self-report data have been viewed as useful because of their apparent correspondence to MTS performance (e.g., Dube et al., 1993; Steele & Hayes, 1991; Wulfert et al., 1991, 1994; see also Hayes, 1986). The present data may be noteworthy in their implication that self-reports can also reveal ef-

fects not evident in the MTS performance of standard equivalence tests.

Among the plausible explanations for the noncorrespondence between self-reports and reflexive MTS performance observed in some subjects is that reflexive performance was differentially discriminable compared to the other relations. Such an outcome could be related to the observation by R. Saunders and Green (1992) that, of the three properties defining an equivalence class (reflexivity, symmetry, and transitivity), reflexivity is unique in that the typical conditional discrimination training provides experimental experience with arbitrary but not with identity matching. With this observation in mind, we provided identity matching experience during the preliminary training for 5 subjects (S8 through S12), but 1 of these subjects (S12) still failed to report accurate reflexive performance, and S8 did not do so immediately. It is possible that we provided too little identity matching experience or provided it too far in advance of the emergent relations tests, but in the absence of additional data our results appear to leave open the possibility that performances on different types of emergent relations were differentially discriminable for other reasons. A key point of curiosity in this regard is why the noncorrespondence on reflexive trials occurred for some subjects but not for others.

If borne out by additional research, the finding of differential self-reporting on reflexive trials would add to a growing body of evidence of discrepancies across emergent stimulus relations despite the "equivalence" of stimuli within the class as measured by traditional MTS tests. For example, Fields and Verhave (1987) have noted that transitive test performances with high levels of accuracy could exist at different levels of associative strength as a result of nodal distance (the number of links in a chain of stimulus-stimulus relations). Recent studies have provided some support for this idea, indicating that the control by sample stimuli over comparison-stimulus selection (Fields et al., 1990, 1993) and latency of accurate MTS responses (Bentall et al., 1994) both are related to nodal distance. If, as in the present study, relations of identical nodal distance may be differentially discriminable, then it seems reasonable to wonder whether the discriminability of emer-

gent relations performance would vary as a function of nodal distance. The self-report procedures of the present study seem well suited for finding out.

Another series of studies has demonstrated that, following the emergence of equivalence classes, the reversal of baseline contingencies can have differential effects on class membership, as measured through different types of emergent relations. For example, Pilgrim and Galizio (1990, 1995) found that when class membership was reversed for one stimulus in each of two four-member classes (e.g., A1-D1 becomes A1-D2 and A2-D2 becomes A2-D1), adult subjects performed consistent with the reversed contingencies on symmetry probes but consistent with the originally trained contingencies on transitive equivalence probes. Pilgrim and Galizio (1995) suggested that their data "raise questions about the nature of equivalence as an underlying, fundamental, and integrated process" (p. 237). An interesting application of the present procedures would be to determine whether baseline reversal manipulations also differentially affect the discriminability of different emergent relations.

The present study might best be viewed as an exercise in the development of method. Its purpose was to employ a simple, structured self-report procedure to examine the extent to which subjects could describe the accuracy of their responses on emergent relations tests. Although many studies have employed self-report measurement to shed light on the variables that control performance during equivalence tests, to our knowledge this is the first investigation of its kind to establish contingencies to promote accurate self-reporting and to provide general markers by which to compare self-reports and emergent relations performances. This is a qualitatively different approach than that employed when think-aloud or postexperimental reports are assumed to be accurate or are interpreted based on broad, and sometimes untested, assumptions about the conditions that mediate their validity (e.g., Ericsson & Simon, 1984; Hayes, 1986).

A possible objection to the present procedure is that it focuses only on public responses and therefore ignores issues of interest to many stimulus equivalence researchers. If the present procedures have any bearing

on the private correlates of equivalence performance, it may be in the same sense that drug-discrimination responses have bearing on the binding of drugs at neurochemical receptor sites in the brain. Drug discriminations presumably reflect neurochemical activity and associated private sensations, but drug-discrimination test procedures and results can be described entirely in terms of public events, and provide useful information at this level of description. The typical procedure (see Overton, 1984) is structurally similar to the procedures of the present study. Reinforcement contingencies are used to promote a discriminated "reporting" repertoire based on the administration of a placebo or drug of known chemical properties. Establishing and maintaining the baseline discrimination are critical to the interpretation of subsequent unreinforced generalization probes, during which new drugs are introduced and subjects "report" which of the training stimuli to which the novel drug bears similarity. A robust finding is that drugs with similar discriminative-stimulus properties also tend to share other publicly evident functions, such as reinforcing and clinical efficacy (Branch, 1991; DeGrandpre & Bickel, 1993). The investigation of these relationships is aided by, but not dependent on, neuropharmacological assays that provide direct evidence of the relevant receptor action. It is worth noting, for example, that drug-discrimination studies in the past sometimes guided inferences about receptor action when the relevant receptors had not yet been identified (Colpaert, 1986; Herling & Woods, 1981).

Private events such as covert naming may facilitate equivalence performance, but no behavioral analogue or receptor-binding assays exist to guide the evaluation of this assumption. A currently popular argument holds that, limitations notwithstanding, it makes more sense to collect and interpret introspective self-reports than to abandon the potentially profitable study of private events (e.g., Lieberman, 1975). Based on the example of drug-discrimination research, it seems equally reasonable to argue that the discriminative-stimulus properties of equivalence-related responses merit study for their own sake. Self-reports about these responses may well provide insight into the structure and function of equivalence classes without

reference to private events. Moreover, if two variables are closely related, then the study of one necessarily informs the study of the other. Much work lies ahead in mapping out the discriminative-stimulus properties of equivalence behavior. Methods and data resulting from this enterprise should be useful for their own sake and, if public performances really are related to inner events, useful as well to those interested in the study of private correlates.

REFERENCES

- 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.
- Bentall, R., Dickins, D., & Fox, S.R.A. (1993). Naming and equivalence: Response latencies for emergent relations. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology*, 46B, 187–214.
- Branch, M. N. (1991). Behavioral pharmacology. In I. H. Iversen & K. A. Lattal (Eds.), *Experimental analysis of human behavior* (Part 2, pp. 21–77). Amsterdam: Elsevier.
- Carrigan, P. F., Jr., & Sidman, M. (1992). Conditional discrimination and equivalence relations: A theoretical analysis of control by negative stimuli. *Journal of the Experimental Analysis of Behavior*, 58, 459–504.
- Colpaert, F. C. (1986). Drug discrimination: Behavioral, pharmacological, and molecular mechanisms of discriminative drug effects. In S. R. Goldberg & I. P. Stolerman (Eds.), *Behavior analysis of drug dependence* (pp. 161–194). Orlando, FL: Academic Press.
- Critchfield, T. S. (1993). Signal-detection properties of verbal self-reports. *Journal of the Experimental Analysis of Behavior*, 60, 495–514.
- Critchfield, T. S. (1994). Bias in self-evaluation: Signal-probability effects. *Journal of the Experimental Analysis of Behavior*, 62, 235–250.
- Critchfield, T. S., & Perone, M. (1990). Verbal self-reports of delayed matching to sample by humans. *Journal of the Experimental Analysis of Behavior*, 53, 321–344.
- DeGrandpre, R. J., & Bickel, W. K. (1993). Stimulus control and drug dependence. *The Psychological Record*, 43, 651–666.
- Dube, W. V., Green, G., & Serna, R. W. (1993). Auditory successive conditional discrimination and auditory stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior*, 59, 103–114.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Fields, L., Adams, B. J., Verhave, S., & Newman, S. (1990). The effects of nodality on the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior*, 53, 345–358.
- Fields, L., Adams, B. J., Verhave, S., & Newman, S. (1993). Are stimuli in equivalence classes equally related to each other? *The Psychological Record*, 43, 85–105.
- Fields, L., & Verhave, T. (1987). The structure of equivalence classes. *Journal of the Experimental Analysis of Behavior*, 48, 317–332.
- Hayes, S. C. (1986). The case of the silent dog—verbal reports and the analysis of rules: A review of Ericsson and Simon's *Protocol Analysis: Verbal Reports as Data*. *Journal of the Experimental Analysis of Behavior*, 45, 351–363.
- Hayes, S. C., Kohlenberg, B. S., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence relations. *Journal of the Experimental Analysis of Behavior*, 56, 119–137.
- Herling, S., & Woods, J. H. (1981). Discriminative stimulus effects of narcotics: Evidence for multiple receptor-mediated actions. *Life Sciences*, 28, 1571–1584.
- Johnson, C., & Sidman, M. (1993). Conditional discrimination and equivalence relations: Control by negative stimuli. *Journal of the Experimental Analysis of Behavior*, 59, 333–347.
- Lazar, R. (1977). Extending sequence-class membership with matching to sample. *Journal of the Experimental Analysis of Behavior*, 27, 381–392.
- Lazar, R. M., Davis-Lang, D., & Sanchez, L. (1984). The formation of visual stimulus equivalences in children. *Journal of the Experimental Analysis of Behavior*, 41, 251–256.
- Lieberman, D. A. (1975). Behaviorism and the mind: A (limited) call for a return to introspection. *American Psychologist*, 34, 319–333.
- Mackay, H. A. (1991). Conditional stimulus control. In I. H. Iversen & K. A. Lattal (Eds.), *Experimental analysis of human behavior* (Part 1, pp. 301–350). Amsterdam: Elsevier.
- Overton, D. A. (1984). State dependent learning and drug discriminations. In L. L. Iversen, S. D. Iversen, & S. H. Snyder (Eds.), *Handbook of psychopharmacology* (Vol. 18, pp. 59–127). New York: Plenum.
- Pilgrim, C., & Galizio, M. (1990). Relations between baseline contingencies and equivalence probe performances. *Journal of the Experimental Analysis of Behavior*, 54, 213–224.
- Pilgrim, C., & Galizio, M. (1995). Reversal of baseline relations and stimulus equivalence: I. Adults. *Journal of the Experimental Analysis of Behavior*, 63, 225–238.
- Saunders, K. J., Saunders, R. R., Williams, D. C., & Spradlin, J. E. (1993). An interaction of instructions and training design on stimulus class formation: Extending the analysis of equivalence. *The Psychological Record*, 43, 725–744.
- Saunders, R. R., & Green, G. (1992). The nonequivalence of behavioral and mathematical equivalence. *Journal of the Experimental Analysis of Behavior*, 57, 227–241.
- Saunders, R. R., Wachter, J., & Spradlin, J. E. (1988). Establishing auditory stimulus control over an eight-member class via conditional discrimination procedures. *Journal of the Experimental Analysis of Behavior*, 49, 95–115.
- Sidman, M. (1980). A note on the measurement of conditional discrimination. *Journal of the Experimental Analysis of Behavior*, 33, 285–289.
- Sidman, M. (1995). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.
- Sidman, M., & Cresson, O. (1973). Reading and cross-modal transfer of stimulus equivalences in severe re-

- tardation. *American Journal of Mental Deficiency*, 77, 515-523.
- Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An extension of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, 37, 5-22.
- Sidman, M., Willson-Morris, M., & Kirk, B. (1986). Matching-to-sample procedures and the development of equivalence relations: The role of naming. *Analysis and Intervention in Developmental Disabilities*, 6, 1-29.
- Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrary applicable relations responding. *Journal of the Experimental Analysis of Behavior*, 56, 519-555.
- Wulfert, E., Dougher, M. J., & Greenway, D. E. (1991). Protocol analysis of the correspondence of verbal behavior and equivalence class formation. *Journal of the Experimental Analysis of Behavior*, 56, 489-504.
- Wulfert, E., Greenway, D. E., & Dougher, M. J. (1994). Third-order equivalence classes. *Psychological Record*, 44, 411-439.
- Wulfert, E., & Hayes, S. C. (1988). Transfer of a conditional ordering response through conditional equivalence classes. *Journal of the Experimental Analysis of Behavior*, 50, 125-144.

Received October 4, 1994

Final acceptance November 9, 1995

APPENDIX

The following categories of conditional discrimination performance and confidence ratings were created by the investigators based on visual inspection of the data and were then refined based on the comments of two naive raters who made informal preliminary ratings. Two new naive raters, one a doctoral-level clinical psychologist and the other a clinical psychology doctoral student, then independently categorized patterns of conditional discrimination performance and self-reports during the equivalence probe sessions to produce Figure 4. Raters were given copies of Figure 3 with the panels numbered from 1 to 40 and subject information, axis labels, and relation-type labels removed. Written instructions stated:

Please use the patterns described below to categorize the data in the attached graphs. Write your answers on the attached coding form. Each graph contains data from two variables. Values for both variables are plotted as a func-

tion of sessions on the abscissa. Variable X (bars) is scaled on the left ordinate (values from 0 to 1.0). Variable Y (black lines) is scaled on the right ordinate (values from -21 to +21). For each graph, circle one code only for Variable X and one code only for Variable Y. If you choose the "other" category (D or H), please write a description of the pattern.

Raters agreed on 77 of 80 patterns. The three disagreements were resolved by retaining the rating more suggestive of correspondence between conditional discrimination performance and confidence ratings. We viewed this approach as conservative in that it is biased toward supporting the conventional wisdom of self-report data-collection methods and certain verbal mediation hypotheses that saying and doing should correspond. Thus, in Figure 4, (a) S6's confidence rating functions on trained and symmetrical relations are listed as "gradual increase" instead of "other," and (b) S8's conditional discrimination function for reflexivity trials is listed as "immediately high" instead of "other."

Conditional Discrimination Patterns

- A. Immediately high: Consistently around .85 or higher from the first sessions. No evidence of trend in early sessions.
- B. Gradual increase to high: Relatively lower in the early sessions; increase to high steady state.
- C. Chance: Typically around .33, with or without trend in early sessions.
- D. Other: Doesn't fit into any of the other three categories.

Confidence Rating Patterns

- E. Immediately positive: Consistently around +7 or higher from the first sessions; no evidence of trend in early sessions.
- F. Gradual increase to positive: Relatively less positive in early sessions; increase to positive steady state of around +7 or higher.
- G. Negative: Typically around -7 or lower, with or without trend in early sessions.
- H. Other: Doesn't fit into any of the other three categories.