JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2011, **44,** 435–450 NUMBER 3 (FALL 2011)

TRANSPORTABILITY OF EQUIVALENCE-BASED PROGRAMMED INSTRUCTION: EFFICACY AND EFFICIENCY IN A COLLEGE CLASSROOM

DANIEL M. FIENUP AND THOMAS S. CRITCHFIELD

ILLINOIS STATE UNIVERSITY

College students in a psychology research-methods course learned concepts related to inferential statistics and hypothesis decision making. One group received equivalence-based instruction on conditional discriminations that were expected to promote the emergence of many untaught, academically useful abilities (i.e., stimulus equivalence group). A negative control group received no instruction, and a positive (complete instruction) control group received instruction on all possible relations (those taught to, and emerging untaught in, the stimulus equivalence group). On posttests, the stimulus equivalence group performed as well as the positive control group (and both outperformed the negative control group), but those in the equivalence-based instruction condition achieved this outcome with significantly less training, thereby demonstrating the efficiency of equivalence-based instruction. Social validity measures indicated that participants found the instruction to be beneficial and as enjoyable as traditional teaching methods.

Key words: college students, effectiveness, efficiency, statistics, stimulus equivalence

Stimulus equivalence involves relating two stimuli that have never been formally paired but share a relation with a common stimulus (Critchfield & Fienup, 2008). The potential academic benefits of stimulus equivalence were recognized in the earliest investigations. For example, Sidman and Cresson (1973) used an equivalence framework to teach individuals with disabilities to match written words and pictures that had never been experienced together. To accomplish this, the individuals were taught to select pictures in response to a spoken word, and to select written words in response to spoken words (spoken word → picture, spoken word → written word). After these two relations were taught, four relations emerged without any direct training. Participants demonstrated untaught symmetrical relations by naming pictures and words (spoken word → picture, written word → picture) and untaught transitive and equivalence relations by relating pictures to written words and written words to pictures, respectively. The procedure was effective for all directly trained relations and was also efficient in that more relations were mastered than were directly taught. Because of these findings, Stromer, Mackay, and Stoddard (1992; see also Critchfield & Fienup, 2008) proposed that efficiency is a hallmark of instruction based on the principles of stimulus equivalence. For economy of expression, we will call this approach *equivalence-based instruction* (EBI).

Although most EBI research has addressed the needs of persons with disabilities, recent studies show that EBI also may be useful in instruction with advanced learners. For example, recent EBI studies have taught college students algebra and trigonometry concepts (Ninness et al., 2006, 2009), statistical concepts such as interaction effects in factorial research

This research was conducted at Illinois State University as part of the first author's requirements for a doctoral degree in school psychology. We thank Daniel Covey for data-collection assistance, and Larry Alferink, Dawn McBride, Gary Cates, James Dougan, Kimberley Reyes, and Jeffery Hamelin for constructive comments about the study. In developing this study, we benefited substantially from the unpublished doctoral dissertation of Tracy Zinn (2002); we thank her for sharing this work with us.

Address correspondence to Daniel M. Fienup, who is now at the Department of Psychology, Queens College, Flushing, New York, 11367 (e-mail: daniel.fienup@qc.cuny.edu).

doi: 10.1901/jaba.2011.44-435

(Fields et al., 2009), and brain–behavior relations (Fienup, Covey, & Critchfield, 2010). In these studies and in our recent work teaching concepts of inferential statistics and hypothesis decision making (Critchfield & Fienup, 2010, in press; Fienup & Critchfield, 2010; Fienup, Critchfield, & Covey, 2009), students consistently mastered more relations than were directly trained. However, no published study of EBI has directly documented the instructional efficiency, compared to traditional instruction, that Stromer et al. (1992) identified as EBI's defining feature.

One limitation of the literature on EBI is that prior studies took place in highly controlled, laboratory-like environments rather than in the typical instructional milieu (Rehfeldt, 2011). That is, the instruction occurred in a relatively distraction-free environment and was not part of a formal program of academic study. Thus, although existing research shows that EBI *can* promote mastery of college-level concepts, it does not document that EBI *will* produce these effects during instruction in college courses. This issue is not unique to EBI and is paralleled in the literature on evidence-based practices. A great deal of attention is focused on the transportability of interventions, that is, on determining whether interventions that show promise in controlled research will yield substantial benefits in less controlled field settings (Chorpita, 2003). Because of practical constraints in field settings, quite often these interventions do not perform as well as in published controlled research studies (Shoenwald & Hoagwood, 2001).

A second limitation of the EBI literature concerns what is known about relative effectiveness and efficiency compared to traditional instructional practice. In most cases, EBI has been compared to an uninstructed baseline condition, which demonstrates that EBI is better than no instruction at all. However, the effectiveness and efficiency of an instructional intervention are best evaluated by comparing its outcomes (e.g., postintervention performance, time to accomplish those effects) to those of typical practice. For example, Taylor and O'Reilly (2000) compared EBI to non-EBI instruction with adults with mild intellectual disabilities. Although the two approaches were equally effective in establishing the targeted skills (i.e., grocery shopping), EBI training required less time to promote mastery. Thus, both instructional interventions were effective, but EBI was more efficient.

The present investigation was designed to extend our ongoing line of research on EBI for college-level statistics concepts by evaluating its relative efficiency in a college course of instruction. As in our past studies, certain statistical relations were taught, and related concepts were tested but not directly taught (e.g., Critchfield & Fienup, 2010, in press; Fienup & Critchfield, 2010; Fienup et al., 2009). Students first learned about stimuli related to statistical significance and nonsignificance. Next, they learned how results that do, and do not, match the scientific hypothesis influence decisions regarding the null and scientific hypotheses. Finally, students learned a conditional skill in which statistical information was combined with hypothesis statements and research results to make decisions about null and scientific hypotheses.

In the current study, participants were college students completing a psychology research methods course in a group setting with all of its attendant distractions. Unlike in the typical EBI study, students' efforts earned them points that directly influenced course grades. Course-grade contingencies might produce better effort than artificial contingencies usually associated with laboratory research, but the converse also could be true. As Critchfield and Fienup (2010) noted, "we have sometimes seen students work energetically as research volunteers to gain bonus credit even while neglecting course assignments that could make bonus credit unnecessary" (p. 774). Thus, it was of interest

to see how well the instructional tasks would be completed when they were presented as part of academic contingencies rather than as an additional research experience. This study employed an approximation of a randomized controlled trial, a group-comparison experimental design that is regarded as the gold standard of efficacy evidence (Evans, 2003). We compared a stimulus equivalence (SE) group, in which students completed EBI lessons identical to those in our prior research (Fienup & Critchfield, 2010), a negative control (CTL) group that received no formal instruction, and a positive control complete instruction (CI) group, in which students were directly taught all of the targeted relations. If SE students mastered the same relations with less time and effort than CI students, then relative efficiency of the EBI lessons would be shown.

METHOD

Participants, Settings, and Materials

Participants and Group Assignment

Participants were students enrolled in a sophomore-level college course in research methods. Not all enrolled individuals participated, so the term *student* will refer to all those enrolled in the course and *participant* will refer to individuals who were part of the final analysis. The research-methods course was required of all psychology majors and minors and included two 1-hr lectures and one 2-hr small-group meeting per week. Students enrolled in one of four small-group sections (each section contained about 20 students) via standard university registration procedures prior to the semester. Thus, assignment to small-group sections was not randomized. Students who attended small-group sections during the 2nd week of the term received an informed consent agreement to sign, if they wished, and a paper retest (described in the materials section below) to complete. Sixty-eight students provided informed consent, and those who scored 70% or higher on the pretest (nine students) were excluded from the experimental design, leaving 59 students eligible to participate.

Participants were assigned to conditions based on their small-group section assignment with the instructional and control groups counterbalanced by day and instructor. Two small-group instructors each hosted one small-group meeting on Wednesday and one on Friday. Students in Instructor A's Wednesday section were eligible for the instructional (SE or CI) groups, and those in Instructor A's Friday section were eligible for the control group. The opposite was true of Instructor B's students (i.e., Wednesday section was the control group, Friday section contained the instructional groups). The small-group instructors were not involved in the design or conduct of the experiment. They were told that their students would complete activities related to the unit on inferential statistics, but were not informed about the purpose of the experiment, group assignment, or the specific activities completed by the groups.

To constitute the final groups, triads of eligible student participants were matched on paper pretest scores. A triad included one student eligible for the control group and two students eligible for instructional groups. The two participants eligible for the instructional groups were assigned to either SE or CI using random numbers generated in Microsoft Excel. This process resulted in 16 triads of students whose scores differed by no more than two correct responses on the 40-item pretest. Eleven eligible students could not be matched to a triad, based on pretest scores, and were excluded from the analyses. Due to experimenter error, two participants were exposed to the wrong computer lessons and were subsequently dropped from the study along with the other members of their triads. Analyses were based on the remaining 14 triads of participants (a total of 42, including 8 men and 34 women). See Table 1 for a summary of group demographics (data obtained using demographic question-

Table 1
Demographic Information

	CTL *M* (*SD*)	SE *M* (*SD*)	CI *M* (*SD*)	Statistic
GPA	3.02 (0.56)	3.15 (0.53)	3.10 (0.50)	$F(2, 38) = 0.21, p = .81$
ACT	24.50 (3.78)	24.00 (2.42)	24.77 (3.72)	$F(2, 38) = 0.18, p = .83$
Age	20.36 (1.34)	20.71 (1.86)	20.14 (1.56)	$F(2, 39) = .046, p = .64$
Years in college	2.50 (0.85)	2.79 (1.86)	2.57 (0.65)	$\chi^2(2, N = 42) = 0.70, p = .71$
Previous PSY138	0.36 (0.50)	0.29 (0.47)	0.36 (0.50)	$\chi^2(2, N = 42) = 0.21, p = .90$
Current PSY138	0.36 (0.50)	0.14 (0.36)	0.35 (0.50)	$\chi^2(2, N = 42) = 2.05, p = .36$
Other statistics courses	0.71 (0.91)	1.00 (0.88)	0.43 (0.51)	$F(2, 39) = 1.84, p = .17$

Note. GPA = college grade point average on a four-point scale. ACT = score on the ACT standardized college entrance examination (maximum score = 36). Years in college represents whole years including the one during which the study took place. PSY138 (scored as 1 or 0) was an introductory course in statistical concepts; previous refers to students who had completed the course, and current refers to students who were currently taking that course. Other statistics courses were entered as the total number of classes taken.

naire from Fienup & Critchfield, 2010). There were no significant differences between groups on grade point average (GPA), ACT college admissions test score, age, and number of self-reported previously completed statistics and math courses (tested by one-way ANOVAs). Chi square tests confirmed no differences in the groups in terms of year in college and whether participants had taken or were currently taking the psychology department statistics course.

Settings and Materials

The experiment was conducted in three settings. First, the weekly small-group sections took place in a pair of classrooms that seated up to 25 students. Second, all computerized teaching and testing occurred in a computer classroom that contained 30 individual workstations with Windows-compatible computers equipped with stereo headphones. An instructor station and a ceiling-mounted projector were located at the front of the room. Third, students who missed class experienced individual remedial appointments in an office (2.7 m by 3.7 m) that contained a desk and one Windows-compatible computer.

Computer program. The lessons were conducted using a custom written Visual Basic computer program (Dixon & MacLin, 2003) that operated on Windows-compatible computers (see additional programming details in Fienup & Critchfield, 2010). The participant's computer displayed a series of screens with a sample stimulus and three comparison stimuli (S+, S−, and a blank box that was considered an S−). During training and testing, the order of sample stimuli and placement of comparison stimuli were randomized. During training, a box in the upper right corner of the screen displayed the number of consecutive correct responses needed to master a unit (i.e., 12) and the current number of consecutive correct responses. Immediately following the selection of a comparison stimulus on each trial, this counter changed to reflect the current consecutive correct count. An ascending sound (chime) indicated a correct response, and a descending sound (chord) indicated an incorrect response. No accuracy feedback was provided during testing.

Computer stimuli. The learning stimuli (Table 2) were identical to those used by Fienup and Critchfield (2010) and were presented in boxes (7.6 cm by 7.6 cm) on the screen. Lesson 1 stimuli (ABC) were related to statistical significance, Lesson 2 stimuli (DEF) were related to hypothesis decisions, and Lesson 3 stimuli paired descriptions of the direction of research results (D stimuli) with statistical information.

Paper pretest and posttest. Students completed a paper-and-pencil multiple-choice test (used

Table 2
Stimuli and Notation

Notation	Set 1	Set 2
A	Low p value	High p value
B	Statistically significant	Not statistically significant
C	$p \leq .05$	$p > .05$
↑ D	Scientific hypothesis: The IV will increase the DV Results: The DV increased	Scientific hypothesis: The IV will increase the DV Results: The DV did not increase
↓ D	Scientific hypothesis: The IV will decrease the DV Results: The DV decreased	Scientific hypothesis: The IV will decrease the DV Results: The DV did not decrease
↕D	Scientific hypothesis: The IV will change the DV Results: The DV changed	Scientific hypothesis: The IV will change the DV Results: The DV did not change
E	Consistent with scientific hypothesis	Not consistent with scientific hypothesis
F	Reject null hypothesis	Fail to reject null hypothesis

Note. Stimuli within a set were associated with each other during the study. Lesson 1 used the A, B, and C stimuli. Lesson 2 used the D, E, and F stimuli. In Lesson 3, students were required to attend to A stimuli to make decisions about how D stimuli were related to the E and F stimuli. Note that in Lessons 2 and 3, there were three separate versions of the D stimuli, representing different types of predictions about changes in a dependent variable (↑ = increases, ↓ = decreases, ↕ = changes).

previously by Critchfield & Fienup, 2010) to evaluate the target relations for the computer lessons. The purpose of the test was to assess responding using a format typically used in a classroom setting. Participants did not receive feedback on their answers. Each question was followed by two potential answers to remain consistent with the format of match-to-sample training in the computerized lessons. For instance, a question asking which p value would indicate that results were statistically significant would have options (a) $p \leq .05$ and (b) $p > .05$. The multiple-choice test was divided into two parts. Part 1 included 24 questions related to the stimuli from Lessons 1 and 2. During Lesson 2, the posed questions indicated that a dependent variable was hypothesized to either increase or decrease. To keep the test battery brief, we asked only Lesson 2 questions in which a dependent variable was hypothesized to increase (designated as ↑ D in Table 2). Part 2 involved 16 total questions related to the stimuli from Lesson 3. Questions involved both results and statistical information, and students had to select answers relevant to either statistical information (B or C stimuli) or hypothesis information (E or F).

Satisfaction survey. After the study concluded, participants in the SE and CI groups completed the satisfaction survey in Table 3. The response scale for all items ranged from 1 (*strong disagreement*) to 7 (*strong agreement*).

Contingencies for Performance

Contingencies for performance varied across activities. Performance on the computer lessons directly influenced course grades for students in the SE and CI, groups with 50 of 1,050 course points (4.8% of the total) contingent on computerized posttest performance. Students in the CTL group earned up to 50 points through small-group activities unrelated to the study or the target topic. Scores on the paper pretest and posttest did not influence course points for both practical and ethical reasons (e.g., the pretest occurred prior to instruction for all groups and the CTL group did not receive any instruction prior to the paper posttest). Course bonus credit (5 points) was tied to pretest scores of 50% or higher, and participants in the SE and CI groups could avoid completing the computer lessons with a score of 80% or higher. Each correct response on the paper posttest counted as one entry in a

Table 3
Satisfaction Survey Questions and Ratings

	SE group ($N = 13$, except Item 6)			CI group ($N = 14$)		
	Distribution 1-2-3-4-5-6-7	*M*	*SD*	Distribution 1-2-3-4-5-6-7	*M*	*SD*
1. The computer lessons helped me to master information about inferential statistics and hypothesis decision. I was better prepared on this topic *after* the modules than I was *before* them.	0-0-0-0-3-7-3	6.00	0.71	0-0-0-1-2-4-7	6.21	0.97
2. What I learned in the computer lessons helped me on the classroom examination on inferential statistics.	0-0-0-2-2-6-3	5.77	1.01	0-0-1-0-3-7-3	5.79	1.05
3. In the lab section, there were practice exercises on inferential statistics and hypotheses. What I learned in the computer lessons helped me to do these exercises.	0-0-1-2-3-6-1	5.31	1.11	0-0-0-2-2-8-2	5.71	0.91
4. I drew on what I learned in the computer lessons to explain ideas about statistics and hypotheses to other students (e.g., in practice exercises or in my project team).	0-0-0-2-3-7-1	5.54	0.88	0-1-1-3-5-3-1	4.79	1.31
5. I would have done just as well on the course examination if I hadn't worked on the computer lessons.	3-4-1-0-3-2-0	3.15	1.95	0-9-1-2-0-2-0	2.93	1.49
6. I enjoyed completing the computer lessons.	2-0-1-5-2-2-0	3.92	1.62	0-4-2-4-1-2-1	3.86	1.66
7. The computer lessons were more boring than other class activities on the same topic (e.g., lectures, practice exercises).	1-2-1-2-4-3-0	4.15	1.68	3-2-3-4-0-1-1	3.21	1.81
8. Psychology 231 students next semester would benefit from the computer lessons.	0-0-0-1-3-5-4	5.92	0.95	0-0-1-0-3-3-7	6.07	1.21

Note. This table displays the frequency of scores, mean, and standard deviation for each question by both the SE and CI groups on the satisfaction survey. No differences were found between groups on ratings ($p > .05$). Rating of 1 indicates *strong disagreement* with the statement whereas rating of 7 indicates *strong agreement* with the statement. One participant in the SE group, who completed all training and testing, was no longer enrolled in the course during Week 7 when the survey was administered. In addition, one SE participant failed to complete Item 6.

prize drawing, with prizes including cash amounts (ranging from $20 to $150), course bonus credit, and the option of being excused from the final exam. Note that these incentives were included to promote careful attention to the pretest and posttest, and would not have been needed if the relevant skills were evaluated in regularly scheduled course examinations.

Procedure

To minimize the likelihood that participant gains were related to aspects of the course instruction besides the experimental instruction, the study began and ended before regular instruction on inferential statistics and hypothesis decision making occurred in the course.

Table 4 lists specific weeks during the academic term in which study events occurred. Note that students who missed study activities due to a small-group meeting absence completed the missed activities during an individual, remedial appointment prior to the next scheduled small-group meeting (i.e., within 1 week).

Week 1: Overview of Project and Points (All Students)

The small-group instructors discussed the general procedures of the study (including performance contingencies; see above) with students during the first meeting of the semester. As part of this discussion, students in the instruction condition sections also learned that they would complete computerized lessons relevant to inferential statistics during the small-group meetings. All students were told that although the instructional activities (e.g., paper tests, computer lessons) were mandatory in the course, individual information would be used for research purposes only after informed consent was given. They were

Table 4
Timeline of Study

Week	SE and CI groups	CTL group
1	Overview of project and points	Overview of project and points
2	Informed consent	Informed consent
	Paper pretest (Parts 1 and 2)	Paper pretest (Parts 1 and 2)
3	Lesson 1 pretest	Demographic questionnaire
	Lesson 1 training	
	Lesson 1 posttest	
	Demographic questionnaire	
4	Lesson 2 pretest	Paper posttest (Part 1)
	Lesson 2 training	
	Lesson 2 posttest	
	Paper posttest (Part 1)	
5	Lesson 3 pretest	
	Lesson 3 training	
	Lesson 3 posttest	
6	Paper posttest (Part 2)	Paper posttest (Part 2)
7–10	Instruction as usual and exam	Instruction as usual and exam
11	Satisfaction survey	

further told that the decision to participate or not had no bearing on their standing in the course, and that the instructor would not learn who provided consent until after final course grades were recorded.

Week 2: Informed Consent and Pretest (All Students)

Students read and were given the opportunity to sign the informed consent agreement. Next, students were asked to complete the paper pretest (Parts 1 and 2), and, as noted above, were told that their answers could earn bonus credit and excuse them from the computer lessons.

Week 3: Lesson 1: A-B-C Relations (SE and CI Groups, Except as Noted)

Participants in the CTL group completed the demographics questionnaire during the Week 3 small-group section. Participants in the SE and CI groups made their first visits to the computer classroom for orientation to the computer learning environment, followed by testing and training for the A, B, and C stimuli (see Table 2). The experimenter (first author) provided a 2- to 5-min presentation explaining how the students' efforts that week were related to the mission and contingencies of the course. The experimenter indicated that participants would complete the first of three computerized lessons with three parts (i.e., pretest, training, and posttest) and that a posttest score of 90% correct was required for completion. The experimenter also reminded students that computer posttest performance was related to course points. Next, the participants completed the pretest followed by instruction in how the following stimuli relate: *p*-value descriptors, statistical significance, and *p*-value ranges. Next, they completed the posttest followed by the demographic questionnaire and were excused for the day.

Preliminary training and testing. Participants learned about the match-to-sample format and feedback delivery during a 5-min automated tutorial program (described by Fienup & Critchfield, 2010). Next, they completed the Lesson 1 computer pretest (identical to that of Fienup & Critchfield, 2010), which included 48 trials with four of each of the combinations of relations of the A-B-C stimuli (see Table 2). Next, participants completed a brief training program (Fienup & Critchfield, 2010) that verified that all participants understood inequality notation ($>$ and $\leq$) used in the C

stimuli. Participants in the SE group met a criterion of 12 consecutive correct responses in a mean of 15.9 (SD = 5.9) trials requiring a mean of 74.4 s (SD = 39.2) to complete. Participants in the CI group met the criterion in a mean of 12.7 (SD = 1.73) trials requiring 57.2 s (SD = 21.8) to complete. According to a t test for unpaired scores, groups did not differ significantly in terms of either trials to criterion, $t(26) = 1.97$, $p = .06$, or training time, $t(26) = 1.44$, $p = .16$.

Lesson 1 training. Participants in the SE group completed the same computerized Lesson 1 that was used in Fienup and Critchfield (2010), which taught two relations to mastery (i.e., 12 consecutive correct responses per relation) in the following order: A→B, C→A. Participants in the CI group completed training on all six of the relations in the following order: A→B, B→A, C→A, A→C, B→C, C→B.

Lesson 1 computer posttest. This test was identical to the Lesson 1 computer pretest except that students continued in the program until they scored 90% correct or higher. A score of less than 90% correct initiated remediation in the form of repeating the Lesson 1 training, followed by another attempt at the Lesson 1 computer posttest. This process repeated until the student met the posttest mastery criterion.

Week 4: Lesson 2: D-E-F Relations (SE and CI Groups, Except as Noted)

Lesson 2 training and testing focused on the D, E, and F stimuli (see Table 2). Participants learned how the following ideas related: hypothesis paired with research results, decisions regarding the scientific hypothesis, and decisions regarding the null hypothesis. The prelesson overview indicated that participants were completing the second of three computerized lessons, reiterated that each correct response on the paper posttest was worth one entry in a prize drawing, and described the prizes that would be awarded.

Lesson 2 computer pretest and posttest. The test (same as in Fienup & Critchfield, 2010) involved 52 total trials, two of each of the D-E-F relations (see Table 2). Students who scored 90% correct or higher on the posttest exited the computer program, and those who scored less than 90% correct received remediation training until they met the posttest mastery criterion.

Lesson 2 training. Participants in the SE group completed the computerized Lesson 2 described by Fienup and Critchfield (2010), which taught the six relations involving change in a variable in the following order: ↑D→E, ↑D→F, ↓D→E, ↓D→F, ↕D→E, ↕D→F. Participants in the CI group completed computerized training on all 14 of the relations tested in the D-E-F pretest. This involved learning the following relations (see Table 2) to mastery criterion in the following order: ↑D→E, E→D↑, ↑D→F, F→D↑, ↓D→E, E→D↓, ↓D→F, F→D↓, ↕D→E, E→D↕, ↕D→F, F→D↕, E→F, F→E.

Paper posttest Part 1. After completing the computerized activities, each participant was given Part 1 of the paper posttest. The experimenter reminded participants that each correct response was worth one entry into the prize drawing and handed them the test. After completing the test, SE and CI participants were excused for the day, and participants in the CTL group continued with their scheduled small-group activities.

Week 5: Lesson 3: Contextual Relations (SE and CI groups)

Training and testing focused on the participants learning the conditional relation between D-E-F (i.e., hypothesis decision) relations and statistical significance. The prelesson overview was similar to that of Weeks 3 and 4 except that participants were told that they were completing the third of three lessons.

Lesson 3 computer pretest and posttest. Participants in both the SE and CI groups completed the test used by Fienup and Critchfield (2010), which involved four trials of each of the 12 relations that were taught during Lesson 3 training. A student exited the computer pro-

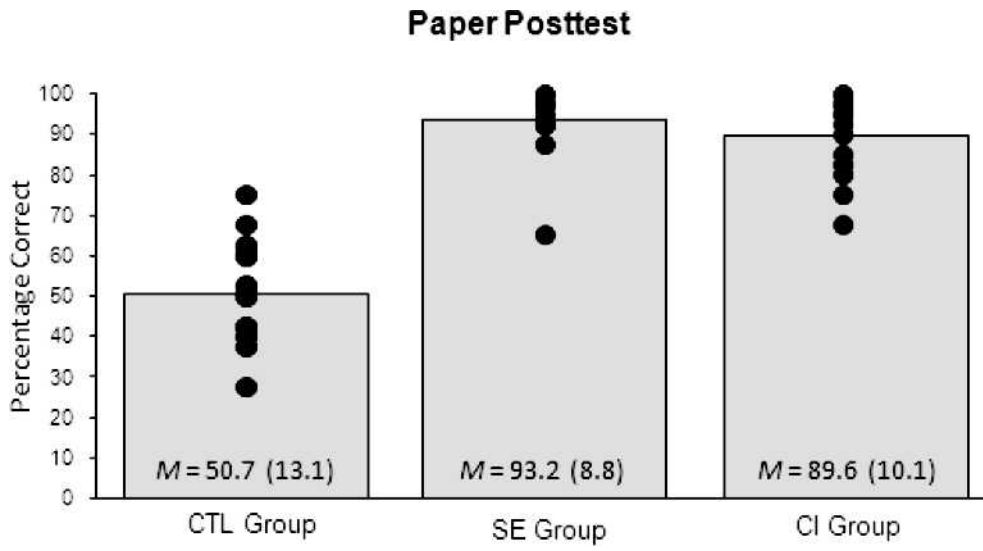

Figure 1. Paper posttest outcomes. Bars show group means, and dots show individual scores.

gram after scoring 90% correct or higher on the posttest or received remediation training until achieving this posttest mastery criterion.

Lesson 3 training. Participants in both the SE and CI groups completed the same Lesson 3 training, which involved learning the 12 relations (see Table 2) in the following order: ↑ D/A1→E, ↑ D/A2→E, ↑ D/A1→F, ↑ D/A2→F, ↓ D/A1→E, ↓ D/A2→E, ↓ D/A1→F, ↓ D/A2→F, ↕D/A1→E, ↕D/A2→E, ↕D/A1→F, and ↕D/A2→F (see Fienup & Critchfield, 2010, for a full description).

Week 6: Paper Posttest Part 2 (All Students)

One week after the participants completed all computerized training and testing, all students completed Part 2 of the paper posttest in small-group meetings. The test was distributed by the small-group instructors following the completion of other class activities, and students were reminded that each correct response was worth one entry into the prize drawing.

Week 7: Prize Drawing (All Students)

The prize drawing was conducted and prizes were distributed to the winning students in a lecture meeting.

Week 11: Satisfaction Survey (SE and CI Groups)

During Weeks 7, 8, and 10, regular course routines resumed, with all students receiving instruction as usual (i.e., nonstudy activities) on various topics. The university's spring break occurred during Week 9. During Week 11, participants in the SE and CI groups were asked, at the end of small-group meetings, to complete the satisfaction survey that asked about their experience with the computer training and how helpful this was in terms of knowledge acquisition.

RESULTS

Instructional Efficacy

The mean percentage accuracy on the paper pretest was virtually identical for the CTL group (M = 49.1% correct, SD = 10.7), the SE group (M = 49.3% correct, SD = 10.5), and the CI group (M = 49.6% correct, SD = 10.6). Paper posttest scores are shown in Figure 1. Instruction produced similar increases in accuracy for the two instruction groups, and no change was noted for the CTL group scores. A one-way ANOVA revealed a statistically significant overall group effect, $F(2, 39) = 66.38$, $p < .001$, and post hoc tests using a Bonferroni

correction technique confirmed the visually apparent similarities and differences among groups. Specifically, the CTL group scored significantly lower than both the SE group ($p < .001$) and the CI group ($p < .001$), and the SE and CI groups did not differ significantly ($p = \sim 1.0$).

In the evidence-based practice literature, it is customary to accompany statistical evaluations of whether an effect occurred with an estimate of effect size (Lipsey &Wilson, 2001), which evaluates the magnitude of an effect. This is accomplished by comparing means between two groups in relation to the variance of the sample data. Standardized mean difference effect sizes are estimated as follows:

$$\mathrm{ES}_{\mathrm{sm}} = \frac{M_2 - M_1}{s_{\mathrm{pooled}}} \quad (1)$$

with M_1 and M_2 representing means of two groups and s representing standard deviation. The s_{pooled} term is defined as

$$s_{\mathrm{pooled}} = \sqrt{\frac{(N_1 - 1){s_1}^2 + (N_2 - 1){s_2}^2}{N_1 + N_2 - 2}} \quad (2)$$

with N_1 and N_2 as the sample size and s_1 and s_2 representing the standard deviations of the groups. Typically, effect size is considered to be small at 0.2 or less; medium if at least 0.5; and large if at least 0.8 (Lipsey & Wilson, 2001). Effects sizes were 3.8 for the SE versus CTL comparison and 3.3 for the CI versus CTL comparison. Thus, the two interventions produced very large effect sizes.

The paper posttest can be broken down into relevant clusters of questions, such as ABC, DEF, and contextual relations. These clusters were analyzed separately to determine if the different lessons produced similar acquisition and whether any differences were observed for directly trained versus emergent relations. In all cases, the SE and CI groups performed better than the CTL group ($p < .002$), and no differences were found between the SE and CI groups ($p > .05$). This means that for all relevant clusters of questions, instruction resulted in higher scores than having no instruction, but the two different types of instruction resulted in similar outcomes.

Figure 2 displays the scores on computerized tests for the two instruction groups at individual and group levels. Each bar depicts the group average for pretest and posttest scores for Lesson 1, Lesson 2, and Lesson 3, and the dots represent individual performances. Computerized pretest means were near the 50% level of accuracy that would be expected with chance responding for Lessons 1 and 2. Although a few individuals scored highly on the computerized pretest (this was possible because only the paper pretest was considered in determining participant inclusion), all participants scored ≥90% on the computerized posttests. Scores of ≥90% were required to exit the posttest and all participants in both groups successfully exited on their first attempt at each posttest ($M \geq 96\%$ correct in all cases). Because some of the material in Lesson 3 overlapped with content from prior lessons, participants who mastered all aspects of Lessons 1 and 2 would be expected to score 75% correct (for a detailed explanation, see Fienup & Critchfield, 2010). Lesson 3 pretest means were near 75% correct, with no significant differences between the SE and CI groups (t tests for unpaired scores, $p > .05$ in all cases). In addition, there were no significant differences between the SE and CI groups on any computerized posttest ($p > .05$ in all cases). In summary, the SE group performed as well on all relations (i.e., trained and untrained) as the CI group, whose members were taught all of the relations.

Instructional Efficiency

Figure 3 (left) shows the average (bar) and individual (dots) number of training trials to criterion (i.e., 12 consecutive correct) for participants in the SE and CI groups to proceed to the computerized posttest. During Lessons 1 and 2, the SE participants practiced only selected relations and CI participants practiced all relations; perhaps not surprisingly, the SE participants completed training in fewer trials than CI participants did (t tests for unpaired

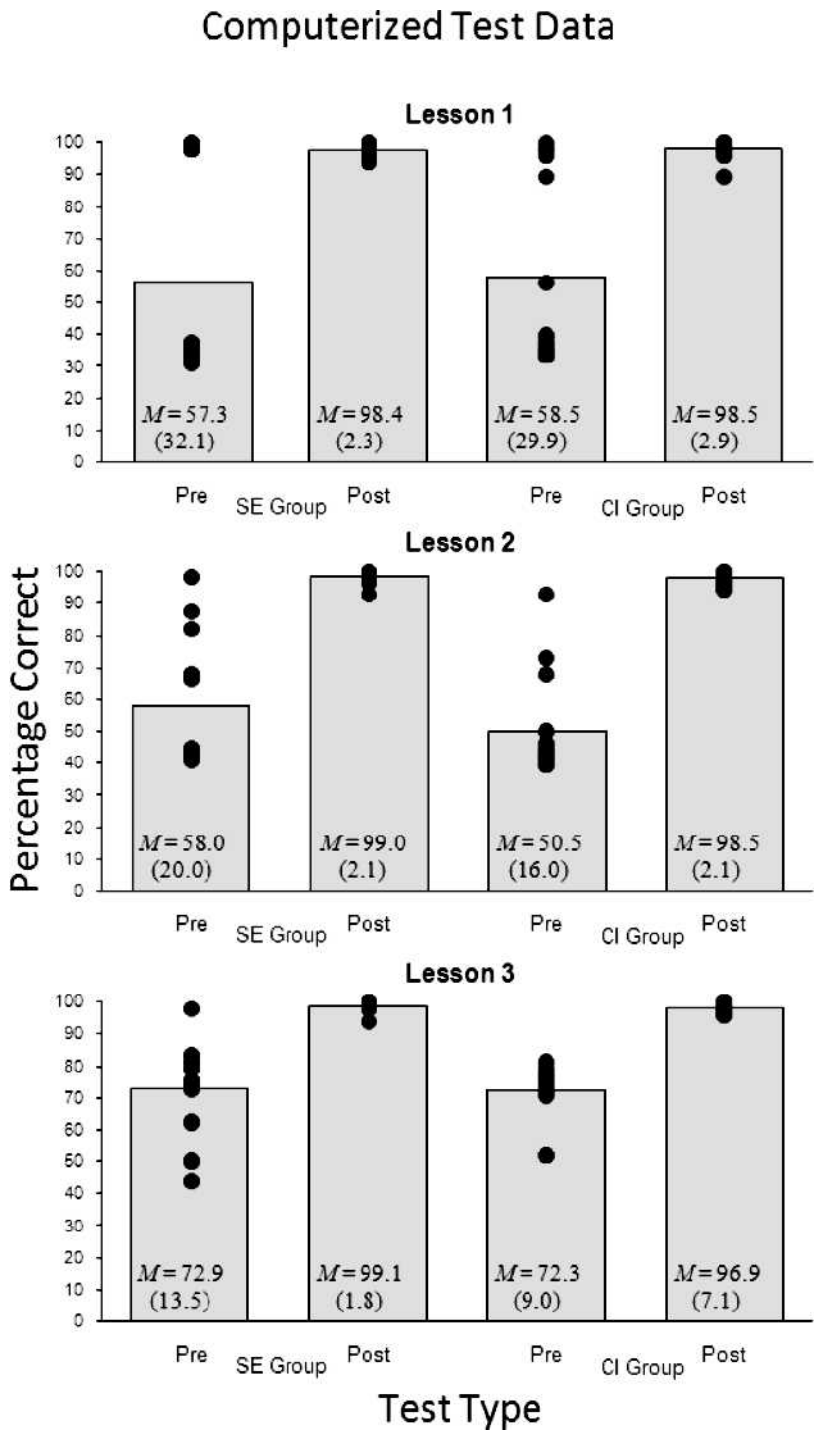

Figure 2. Computerized pretest and posttest scores for the stimulus equivalence (SE) and complete instruction (CI) groups for each of the three lessons. Bars show group means, and dots show individual scores.

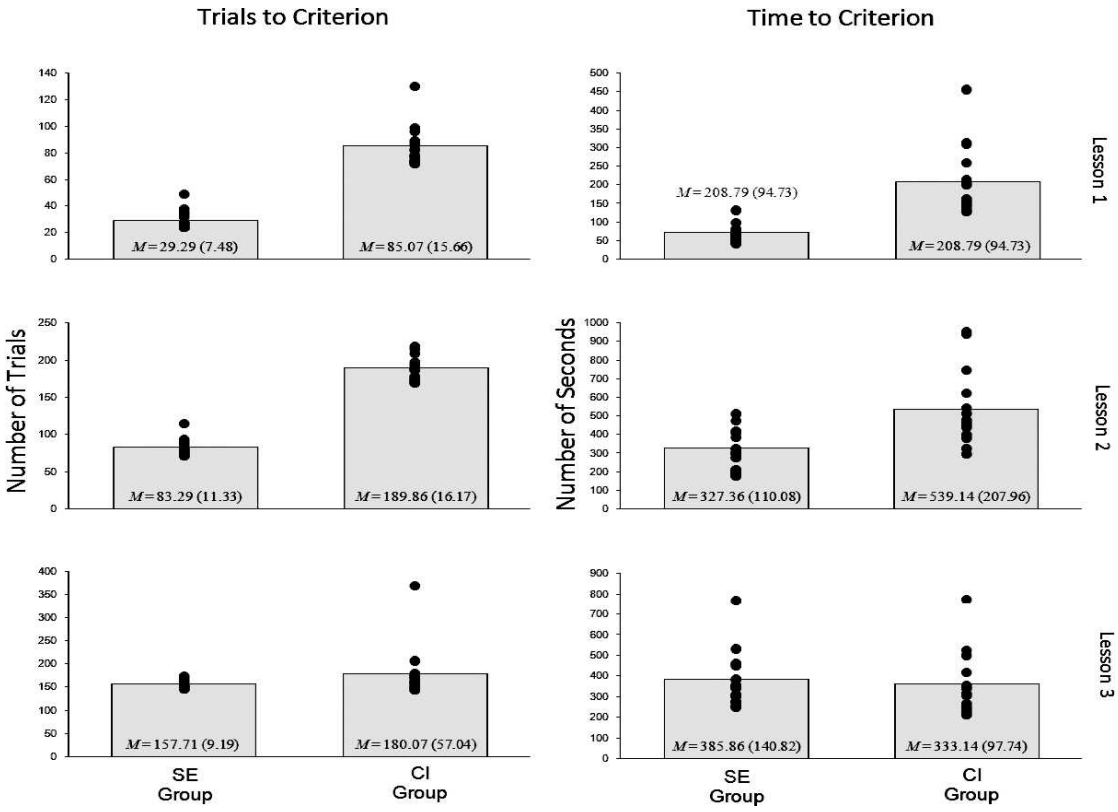

Figure 3. Number of trials (left) and time (right) required to meet the mastery criteria during training for participants in the two intervention groups across the three lessons. Bars show group means, and dots show individual scores.

scores, $p \leq .001$). In Lesson 1, CI participants practiced three times as many relations as SE participants and required 2.9 times as many trials to complete training. In Lesson 2, CI participants practiced 2.3 times as many relations and required 2.3 times as many trials to complete training than SE participants. Lesson 3 employed identical training for the SE and CI groups, with no significant difference in the number of trials required to complete training ($p = .16$). Figure 3 (right) shows the average (bar) and individual (dots) training time required to meet the mastery criterion. In addition to requiring more trials, the CI participants who practiced all relations required significantly more time to complete training in Lessons 1 and 2 (t tests for unpaired scores, $p \leq .002$). Lesson 1 required three times more training time for the CI group, and Lesson 2 required 1.6 times more training time. Training was identical in Lesson 3, and there was no significant difference in the number of seconds required to complete training ($p = .26$).

Satisfaction

The results of the satisfaction survey are presented in Table 3 as group means and standard deviations accompanied by frequency distributions of individual responses. There were no significant differences between the two instructional groups for any survey item (t tests for unpaired scores, $p > .05$ in all cases). Participants tended to report that the lessons helped them to master information about inferential statistics and hypothesis decision making; perform well on class activities that

were not part of the study; and discuss relevant concepts with other students during small-group meetings. Participants also described the lessons as neither very enjoyable nor very unenjoyable, and as no more or less boring than other types of academic assignments. All participants but one (in the CI group) indicated that it would be beneficial for future students to complete the lessons.

DISCUSSION

The present results join those of other published studies in demonstrating that instruction based on principles of stimulus equivalence, or EBI, can promote the emergence of academically useful, college-level concepts (e.g., Fields et al., 2009; Fienup et al., 2010; Ninness et al., 2005, 2006, 2009). Our previous reports have shown that concepts related to inferential statistics and hypothesis decision making could be taught to college students under laboratory conditions (e.g., Critchfield & Fienup, 2010; Fienup & Critchfield, 2010; Fienup et al., 2009). The current study replicates our previous findings in a more naturalistic classroom environment in which programmed instruction was delivered as a class assignment in a group setting, and student gains were evaluated using a typical college assessment (i.e., multiple-choice exam). The current study demonstrated that training a targeted subset of relations was as effective as training all relations, but it was significantly more efficient. Both instructional procedures were more effective than no instruction on traditional paper-and-pencil measures, which is not unexpected given that skills of inferential statistics and hypothesis decision making tend not to exist in lay repertoires and require special instruction to master (e.g., Kranzler, 2007).

Students in the two instructional conditions achieved similar levels of mastery although the training investment differed substantially. The CI group practiced all 20 target relations (i.e., six in Lesson 1, 14 in Lesson 2) and required significantly more trials and time for mastery than the SE group, which only practiced eight of these relations (i.e., two in Lesson 1, six in Lesson 2). In computerized tests, the SE group demonstrated proficiency on all 20 relations, including the 12 that were not directly taught, thereby demonstrating the relative efficiency of EBI compared to the CI approach. This proficiency carried over to other experiences, in that participants in the two groups profited equally from a subsequent common training experience in Lesson 3 and performed equally well on the paper posttest.

The finding that the SE group required fewer training trials may seem intuitive given that these participants practiced fewer relations, but this outcome was not a given. The less extensive training of the SE group could have yielded more failures on computerized posttests with untrained relations, which would have necessitated a repeat of the training experiences. No previous studies have directly compared EBI to other instructional procedures with college students, so these findings provide the first clear evidence of superior relative efficiency of EBI with advanced academic material and advanced learners. Thus, our study joins with that of Taylor and O'Reilly in providing direct evidence for the efficiency of learning that previous writers have claimed is the hallmark of EBI (e.g., Critchfield & Fienup, 2008; Stromer et al., 1992).

Because many types of control groups can be useful in empirical evaluations of practice, it is worth reviewing the logic that led to the present experimental design. First, although the recipient of *any* kind of intervention is likely to outperform a no-treatment control group, we included a no-instruction group to rule out the possibility that student skills improved for reasons unrelated to the experiment (e.g., instruction from sources outside the experiment). Such history effects are a special risk in experiments of considerable duration (Christensen, 2001); however, the present CTL group

showed no gains from pretest to posttest. Second, it might be argued that the present CI group was not an ecologically valid comparison for the SE group because college instructors are unlikely to "teach everything" directly. We elected to train all relations because there are no relevant published data indicating the degree to which the sum of all traditional college instructional activities (e.g., lecture, reading textbooks, practice exercises, and examinations) teaches every desired relation. In addition, some kinds of instruction, such as programmed instruction of the sort that Skinner (1968) pioneered, *do* tend toward "teaching everything" (see Holland & Skinner, 1961).

Although these findings are promising, future research should include more rigorous evaluations of both relative efficacy and relative efficiency. It would be especially valuable to compare EBI to instructional procedures based on other theories of instruction and to other behavioral instructional approaches (e.g., direct instruction, precision teaching). It would also be useful to examine other aspects of relative efficiency by altering the control procedures. For example, Zinn (2002) included a positive control group that practiced all relations, as ours did, but Zinn matched the total amount of practice to the stimulus equivalence group (i.e., only as many practice trials as the EBI group needed to master their subset). Using this approach, Zinn found that participants in the control condition mastered fewer relations than did participants in the EBI condition.

The contemporary evidence-based practice movement relies heavily on group-design randomized trials to evaluate population-wide efficacy (Chorpita, 2003; Christensen, 2001; Evans, 2003). In addressing large-scale implementation and public policy decisions, studies with large number of subjects are valued for providing excellent estimates of general effectiveness and efficacy for the general consumer. The current study was designed as an approximation of the randomized controlled trial that has proven to be so influential in other areas of efficacy evaluation (Evans, 2003); however, the randomization procedure was affected by practical matters such as the student-driven enrollment process at the university. Although small-group sections were assigned to conditions via counterbalancing of times and instructors and students were matched via pretests, student entry into a given section was potentially affected by various personal variables. However, this approach of using a quasirandom assignment is not uncommon in applied outcome research (e.g., Lovaas, 1987).

Laboratory-like studies often constitute the first step in validating an intervention by determining whether the independent variable produces the intended effect under controlled and favorable conditions (Chorpita, 2003). Transportability research must confront additional variables that occur in natural settings that could alter the impact of an intervention in everyday implementation in nonlaboratory settings. In this sense, transportability research may be considered a more rigorous test of an intervention. Our previous studies illustrated that using EBI to teach statistical concepts can produce excellent acquisition in a laboratory (e.g., Critchfield & Fienup, 2010; Fienup & Critchfield, 2010; Fienup et al., 2009). The current findings provide preliminary evidence that those same positive outcomes hold true in situations that are a closer approximation to the traditional college classroom milieu. To be clear, some of the present procedures were atypical of everyday instruction (e.g., prize drawings to increase effort on tests, individual make-up sessions when classes were missed). However, these artificial aspects of our procedure were designed to support good measurement of student accomplishment. Many other aspects were similar to what students would encounter in other courses. For instance, the EBI lessons took place under a point-for-performance contingency similar to those employed in many college courses, and the group computer lab was used as

the primary instructional setting in other psychology courses.

Citing the empirical outcomes for an intervention does not ensure that potential adopters will be convinced to use the intervention (Rogers, 2003). Among the issues that drive dissemination is social validity (Wolf, 1978) or the degree to which consumers like a given intervention. Because our EBI lessons were not typical college-classroom experiences, our participants might have disliked them, which could affect student evaluations of instruction and instructor job security in a typical college environment. However, the social validity assessment indicated that our participants appreciated that the lessons helped them to learn about a technical subject matter that is among the most difficult of instructional challenges in undergraduate social science curricula (Kranzler, 2007). Consistent with this finding, prior to completing the social validity questionnaire, several participants spontaneously indicated to the first author that they wished their other courses included similar forms of instruction. Such evidence suggests that student consumers would not present an impediment to disseminating EBI in college classrooms.

A more substantial hurdle to the dissemination of EBI might be instructor familiarity with stimulus equivalence principles. Although the equivalence framework has been applied to instructional problems for over 30 years, it remains conceptually daunting to many with a strong background in behavior analysis. The complexities of the equivalence framework and the specific procedural arrangements of EBI may fall outside the comfort zone of the typical college instructor. A proximal strategy might be to begin to disseminate the general conceptual framework of stimulus equivalence through mainstream publications that will reach college teachers. A more distal strategy might involve developing user-friendly EBI technologies that can be implemented by instructors with limited expertise in the science of stimulus relations. A similar approach has been taken in human services to guide individuals through the functional assessment and treatment process for problem behavior (e.g., O'Neill et al., 1996). At this point, however, no best practices guidelines exist for EBI, and no automated delivery products have been designed for easy implementation of EBI by nonexperts. A product akin to the Headsprout Early Reading online instructional modules for literacy instruction (Clairfield & Stoner, 2005) might be useful for this purpose. Thus, there is currently a sizable body of research that documents the efficacy of EBI and an emerging and important evidence of efficiency and transportability of EBI technologies, suggesting that it may be time to begin larger scale dissemination efforts.

REFERENCES

Chorpita, B. F. (2003). The frontier of evidence-based practice. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 42–59). New York: Guilford.

Christensen, L. B. (2001). *Experimental methodology* (8th ed.). Boston: Allyn and Bacon.

Clairfield, J., & Stoner, G. (2005). The effects of computerized reading instruction on the academic performance of students identified with ADHD. *School Psychology Review*, *34*, 246–254.

Critchfield, T. S., & Fienup, D. M. (2008). Stimulus equivalence. In S. F. Davis & W. F. Buskist (Eds.), *21st century psychology* (pp. 360–372). Thousand Oaks, CA: Sage.

Critchfield, T. S., & Fienup, D. M. (2010). Using stimulus equivalence technology to teach about statistical inference in a group setting. *Journal of Applied Behavior Analysis*, *43*, 437–462.

Critchfield, T. S., & Fienup, D. M. (in press). A "happy hour" effect in translational stimulus relations research. *Experimental Analysis of Human Behavior Bulletin*.

Dixon, M. R., & MacLin, O. H. (2003). *Visual basic for behavioral psychologists*. Reno, NV: Context Press.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*, 77–84.

Fields, L., Travis, R., Roy, D., Yadlovker, E., de Aguiar-Rocha, L., & Sturmey, P. (2009). Equivalence class formation: A method for teaching statistical interactions. *Journal of Applied Behavior Analysis*, *42*, 575–593.

Fienup, D. M., Covey, D. P., & Critchfield, T. S. (2010). Teaching brain–behavior relations economically with stimulus equivalence technology. *Journal of Applied Behavior Analysis*, *43*, 19–34.

Fienup, D. M., & Critchfield, T. S. (2010). Efficiently establishing concepts of inferential statistics and hypothesis decision making through contextually controlled equivalence classes. *Journal of Applied Behavior Analysis*, *43*, 437–462.

Fienup, D. M., Critchfield, T. S., & Covey, D. P. (2009). Building contextually controlled equivalence classes to teach about inferential statistics: A preliminary demonstration. *Experimental Analysis of Human Behavior Bulletin*, *30*, 1–10.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction*. New York: McGraw-Hill.

Kranzler, J. H. (2007). *Statistics for the terrified* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. London: Sage.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, *55*, 3–9.

Ninness, C., Barnes-Holmes, D., Rumph, R., McCullen, G., Ford, A. M., Payne, R., et al. (2006). Transformation of mathematical and stimulus functions. *Journal of Applied Behavior Analysis*, *39*, 299–321.

Ninness, C., Dixon, M., Barnes-Holmes, D., Rehfeldt, R. A., Rumph, R., McCuller, G., et al. (2009). Constructing and deriving reciprocal trigonometric relations: A functional analytic approach. *Journal of Applied Behavior Analysis*, *42*, 191–208.

Ninness, C., Rumph, R., McCullen, G., Harrison, C., Ford, A. M., & Ninness, S. K. (2005). Functional analytic approach to computer-interactive mathematics. *Journal of Applied Behavior Analysis*, *38*, 1–22.

O'Neill, R. E., Horner, R. H., Albin, R. W., Sprague, J. R., Storey, K., & Newton, J. S. (1996). *Functional assessment and program development for problem behavior: A practical handbook*. Pacific Grove, CA: Brooks/Cole.

Rehfeldt, R. A. (2011). Toward a technology of derived stimulus relations: An analysis of articles published in the *Journal of Applied Behavior Analysis* 1992–2009. *Journal of Applied Behavior Analysis*, *44*, 109–119.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Shoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services*, *52*, 1190–1197.

Sidman, M., & Cresson, O. (1973). Reading and crossmodal transfer of stimulus equivalences in severe retardation. *American Journal of Mental Deficiency*, *77*, 515–523.

Skinner, B. F. (1968). *The technology of teaching*. New York: Appleton-Century-Crofts.

Stromer, R., Mackay, H. A., & Stoddard, L. T. (1992). Classroom applications of stimulus equivalence technology. *Journal of Behavioral Education*, *2*, 225–256.

Taylor, I., & O'Reilly, M. F. (2000). Generalization of supermarket shopping skills for individuals with mild intellectual disabilities using stimulus equivalence training. *The Psychological Record*, *50*, 49–62.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.

Zinn, T. E. (2002). Using stimulus equivalence to teach drug names: A component analysis and drug name classification procedure. *Dissertation Abstracts International*, *63*, 3051.

Received October 29, 2009
Final acceptance January 11, 2011
Action Editor, Linda LeBlanc