EQUIVALENCE CLASS FORMATION: A METHOD FOR TEACHING STATISTICAL INTERACTIONS

LANNY FIELDS

THE GRADUATE CENTER OF THE CITY UNIVERSITY OF NEW YORK
QUEENS COLLEGE OF THE CITY UNIVERSITY OF NEW YORK

ROBERT TRAVIS

THE GRADUATE CENTER OF THE CITY UNIVERSITY OF NEW YORK

DEBORAH ROY

UNIVERSITY OF ULSTER, COLERAINE

EYTAN YADLOVKER AND LILIANE DE AGUIAR-ROCHA

THE GRADUATE CENTER OF THE CITY UNIVERSITY OF NEW YORK

AND

PETER STURMEY

THE GRADUATE CENTER OF THE CITY UNIVERSITY OF NEW YORK
QUEENS COLLEGE OF THE CITY UNIVERSITY OF NEW YORK

Many students struggle with statistical concepts such as interaction. In an experimental group, participants took a paper-and-pencil test and then were given training to establish equivalent classes containing four different statistical interactions. All participants formed the equivalence classes and showed maintenance when probes contained novel negative exemplars. Thereafter, participants took a second paper-and-pencil test. Participants in the control group received two versions of the paper-and-pencil test without equivalence-based instruction. All participants in the experimental group showed increased paper-and-pencil test scores after forming the interaction-indicative equivalence classes. Class-indicative responding also generalized to novel exemplars and the novel question format used in the paper-and-pencil test. Test scores did not change with repetition for control group participants. Implications for behavioral diagnostics and teaching technology are discussed.

DESCRIPTORS: college students, computer-based training, equivalence classes, generalization to novel exemplars

The ability to manipulate, interpret, and describe data are key skills needed to evaluate published empirical work, plan experimental research, and function effectively in the natural and social sciences (Mulhern & Wylie, 2004; Ward & Kaflowitz, 1986). In addition, these skills can enhance a person's ability to understand the complex information encountered in everyday settings in our increasingly sophisticated world. For example, health and longevity can be influenced in complex ways by variables such as genetic background, exercise, diet, years of marriage, and so on. The enhancement of longevity and health then might depend on an ability to understand what it means for these factors to interact and how those interactions might inform the implementation of beneficial changes in lifestyle.

For many individuals, notions of interaction are introduced in college courses in statistics.

Address correspondence to Lanny Fields, Department of Psychology, Queens College/CUNY, 65-30 Kissena Boulevard, Flushing, New York 11367 (e-mail: Lanny.Fields@QC.cuny.edu).

doi: 10.1901/jaba.2009.42-575

Therefore, the concepts imparted in a statistics course could have a beneficial influence on an individual's quality of life. Many college students, however, find it difficult to master the content of a statistics course (Rosenthal, 1992; Simon & Bruce, 1991). Explanations of these difficulties include interference with performance by affective variables such as anxiety (Nasser, 1999), deficiencies of the primarily lecture-based instructional methods used to teach concepts in statistics (Christopher & Marek, 2002; Peden, 2001), and deficiencies in mathematical skills (Mulhern & Wylie, 2004). A cooperative learning approach to teaching statistics that combines in-class group activities with conceptual material provided during lectures appears to improve performance in (Hinde & Kovac, 2001) and student ratings of (Davidson & Kroll, 1991) statistics courses. These studies, however, did not operationalize how the teaching factors influenced the learning of the statistical concepts. In another study, although students in a traditionally taught statistics course learned to manipulate definitions and algorithms, often they were unable to apply these concepts to real-world problems (Bradstreet, 1996). Finally, Seipel and Apigian (2005) noted that a better understanding of the "behavioral weaknesses" of students might lead to new instructional modes designed to correct these deficits. The present experiment sough to address these shortcomings by the application of an equivalence class analysis to a difficult topic in statistics: interaction.

Equivalence classes. Three or more physically disparate stimuli are equivalent when the presentation of any stimulus from the set evokes selection of any other stimulus in the same set (Fields & Reeve, 2000; Sidman, 1971). The procedural variables that lead to the formation of equivalence classes in laboratory settings have been well documented (Fields, Reeve, Adams, & Verhave, 1991; Fields & Verhave, 1987; Fields, Verhave, & Fath, 1984; Sidman, Kirk, & Willson-Morris, 1985; Sidman & Tailby, 1982; Smeets & Barnes-Holmes, 2005) and have been used in applied settings to establish equivalence classes indicative of reading repertoires by individuals with developmental disabilities (Connell & Witt, 2004; de Rose, de Souza, & Hanna, 1996; Sidman, 1971; Sidman & Cresson, 1973), facial recognition in adults with brain damage (Cowley, Green, & Braunling-McMorrow, 1992; Guercio, Podolska-Schroeder, & Rehfeldt, 2004), geographic relations in children with autism (LeBlanc, Miguel, Cummings, Goldsmith, & Carr, 2003), and fraction-decimal relations in children (Lynch & Cuvo, 1995). Thus, similar procedures might be effective for teaching relations among the complex stimuli typically encountered by college students in statistics.

Statistical interaction. Personal observation and those of many colleagues who have taught courses in statistics and experimental psychology indicate that many college students have difficulties recognizing representations of the combined effects of two independent variables on some dependent variable (i.e., statistical interaction). Specifically, when two independent variables are simultaneously manipulated, two possible outcomes can occur. First, an alteration in the value of one independent variable can produce a constant change in the effects of a second independent variable on a dependent variable. In this case, the effects of the two independent variables are said to be additive (i.e., the effect of the second variable on the first is constant across manipulations). Second, an alteration in the value of one independent variable can modulate the effect of a second independent variable on a dependent variable. In this case, the effect of one variable on a dependent variable is determined by the value of the other variable. When the manipulations of independent variables produce such an outcome, the effect is referred to as an *interaction*. In addition, a change in the value of one variable can reverse, enhance, or diminish the effects of a second independent variable.

Finally, each type of interaction can be depicted in many ways (e.g., as a graph, a textual description, a definition, and a name).

The representations of a statistical interaction can be viewed as four different stimuli that are presented to a student during instruction. *Comprehending* a particular type of statistical interaction can be operationally defined as selecting any stimulus from a given set of representations when presented with any other stimulus from the same four-member interaction set. This goal can be achieved by the establishment of interaction-indicative equivalence classes. To illustrate, assume that the stimuli for a type of statistical interaction are a graph (A), a written description of the data in the graph (B), the label of the type of interaction (C), and its definition (D). Matching-to-sample training can be used to establish the relations for each class of four stimuli representing a particular interaction: A-B, B-C, and C-D. *Grasping* a statistical interaction can be inferred when a student responds in a class-consistent manner to the trained and untrained relations among the stimuli in the set. Specifically, a student must select the correct description (B) when given its graph (A-B), the correct graph when given the description (B-A), the correct label (C) when presented with the corresponding description (B-C), the description (B) when presented with the correct label (C-B), the correct label when given the correct graph (A-C), and vice versa (C-A). Further, a student should be able to select the correct definition when given its corresponding graph (A-D), description (B-D), or label (C-D) and vice versa (D-A, D-B, and D-C). Thus, the emergence of the three symmetrical (B-A, C-B, D-C), three transitive (A-C, A-D, B-D), and three equivalence (C-A, D-A, D-B) relations would indicate the formation of a four-member equivalence class after the training of only three baseline conditional discriminations (A-B, B-C, and C-D).

To be of practical value, the selection of any stimulus in a class that represents an interaction would also have to generalize to new variations of each member of that class. For example, presenting some novel graphic or textual variant of an A or a B stimulus as a sample should occasion selection of the stimuli in the class that had been used as comparisons during training and vice versa. A graphic variant (A) would be an interaction graph that contained functions with slopes and intercepts that differed from those used in training and also had different independent and dependent variables. A variant of a descriptive variable (B) would be text that paraphrased the trained descriptions. Further, presentation of any of these novel stimulus variants as samples should also occasion selection of any novel stimulus variant as a comparison. Such an outcome would demonstrate that the perceptually distinct exemplars of a given class along with their variants were functioning as members of a generalized equivalence class (Fields & Reeve, 2001). Finally, these performances would indicate generalization among stimuli within a class and discrimination between stimuli in different classes, the behavior-analytic definition of concept formation (Keller & Schoenfeld, 1950). These data, then would operationally define the establishment of the concept of interaction.

The present study addressed four questions. First, can computer-based procedures that are known to form equivalence classes with arbitrary stimuli also be used to establish classes of stimuli that represent four types of statistical interaction in which each class contains different depictions of the designated type of interaction? Second, would the trained and derived relations in the equivalence classes be maintained when tested in the context of novel negative exemplars, a form of generalization across contexts? Third, would the trained and derived relations in the equivalence classes generalize to novel representations of statistical interactions in a novel paper-and-pencil testing format that contained more choices than those used during class formation? Fourth, would students have a preference for the procedure used to establish the interaction-based equivalence classes (i.e., social validity)? These questions were answered in a two-group pretest–posttest design. An experimental group received a paper-and-pencil pretest on statistical interactions, computer-based equivalence class formation training, and then a paper-and-pencil posttest. A control group received only the pretest and posttest alone. Outcomes were determined by comparing the scores obtained from the pretests and posttests for both groups.

METHOD

Participants

Twenty-one students, enrolled in a class in introductory psychology, satisfied one of the course requirements by participation in the present experiment. To participate, a student first signed an informed consent statement for the 3- to 3.5-hr experiment. All participants received the same credit toward satisfaction of the course requirement.

Apparatus

Setting and hardware. All computer training phases took place in cubicles (1.8 m by 1.5 m) that contained an IBM computer, a keyboard, a dot matrix printer, and a desk and chair. All stimuli were presented on the computer monitor, and all responses to the stimuli involved pressing specific keys on the computer keyboard.

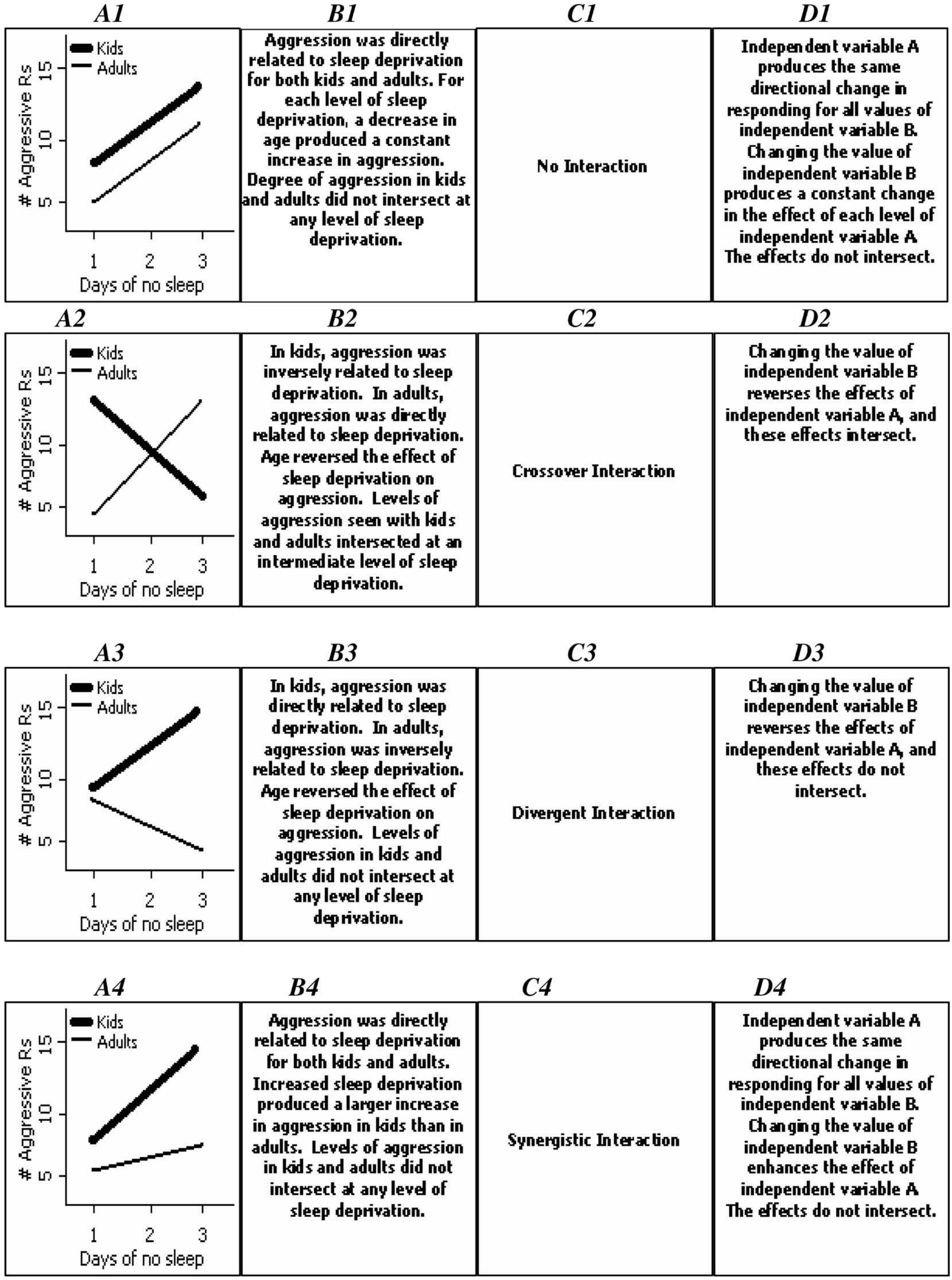
Software. A customized DOS-based program written in Visual Basic controlled all aspects of computer-based training, testing, and recording of the relations presented for training and testing, the choices made by the participant, reaction times, and the feedback provided on every trial. All stimuli measured 5 cm by 5 cm and were presented on a 380-mm SVGA computer monitor.

Stimuli used in equivalence class formation. The four members of each statistical stimulus class used during computer-based equivalence training are shown in Figure 1. Each stimulus class contained four different stimulus types that were assigned a letter designation. The A stimuli were line graphs depicting four types of statistical interactions. The B stimuli were textual descriptions of the interactions depicted in each graph. The C stimuli were labels of each interaction or no interaction. The D stimuli were textual definitions of each type of interaction. Each stimulus class was also numbered (1 = no interaction, 2 = crossover interaction, 3 = divergent interaction, and 4 = synergistic interaction). For example, the A1 stimulus was a line graph from the no-interaction class, and the D3 stimulus was a definition from the divergent class.

Procedure

Experimental design. The experiment was a pretest–posttest design with control and experimental groups. Participants in the control group received two versions of the paper-and-pencil test without intervening establishment of equivalence classes. Participants in the experimental group received a paper-and-pencil test, computer training to form four four-member equivalence classes, and then a second version of the paper-and-pencil test. Across groups, participants were matched on pretest scores and then randomly assigned to the experimental or control group by the flip of a coin to reduce between-groups variability by ensuring that participants in both conditions performed essentially equally before the intervention. Because 1 participant dropped out of the experimental group after group assignment, an uneven number of participants were in the two conditions. The dependent variable was performance on the paper-and-pencil test. Finally, all participants completed a social validity questionnaire to evaluate four aspects of the experiment.

Paper-and-pencil pretest. The paper-and-pencil tests contained 24 multiple-choice questions about statistical interactions with four options



B1

Aggression was directly related to sleep deprivation for both kids and adults. For each level of sleep deprivation, a decrease in age produced a constant increase in aggression. Degree of aggression in kids and adults did not intersect at any level of sleep deprivation.

C1

No Interaction

D1

Independent variable A produces the same directional change in responding for all values of independent variable B. Changing the value of independent variable B produces a constant change in the effect of each level of independent variable A. The effects do not intersect.

B2

In kids, aggression was inversely related to sleep deprivation. In adults, aggression was directly related to sleep deprivation. Age reversed the effect of sleep deprivation on aggression. Levels of aggression seen with kids and adults intersected at an intermediate level of sleep deprivation.

C2

Crossover Interaction

D2

Changing the value of independent variable B reverses the effects of independent variable A, and these effects intersect.

B3

In kids, aggression was directly related to sleep deprivation. In adults, aggression was inversely related to sleep deprivation. Age reversed the effect of sleep deprivation on aggression. Levels of aggression in kids and adults did not intersect at any level of sleep deprivation.

C3

Divergent Interaction

D3

Changing the value of independent variable B reverses the effects of independent variable A, and these effects do not intersect.

B4

Aggression was directly related to sleep deprivation for both kids and adults. Increased sleep deprivation produced a larger increase in aggression in kids than in adults. Levels of aggression in kids and adults did not intersect at any level of sleep deprivation.

C4

Synergistic Interaction

D4

Independent variable A produces the same directional change in responding for all values of independent variable B. Changing the value of independent variable B enhances the effect of independent variable A. The effects do not intersect.

Figure 1. An example of the four members of each class of stimuli used during the equivalence training.

as answers (a, b, c, and d). A participant answered the questions by entering the letter corresponding to the correct answer on a standard Scantron sheet that was scored electronically. Of the 24 questions, two were included from each possible stimulus relation (A-B, B-A, B-C, C-B, A-C, C-A, C-D, D-C, A-D, D-A, B-D, D-B), with six questions from each stimulus class. The information in each question in the paper-and-pencil tests contained statements and graphs that differed in textual and graphic content from those used as stimuli for the computer-induced equivalence classes. Thus, a B1-A1 question contained a description of a graph that was similar to but differed from the description of the B1 stimulus depicted in Figure 1. In addition, the answer options consisted of four graphs that were similar to but differed from those used as the A1 through A4 stimuli depicted in Figure 1.

These distinctions are illustrated in Figure 2. Whereas all of the B stimuli used in the computer training depicted the effects of age and sleep deprivation on aggressive responses, the B2 stimulus used in the B2-A2 question in the paper-and-pencil test described the effects of light exposure and water intake on plant growth. Whereas the B2 stimulus used in training included the phrase "intersected at an intermediate level of sleep deprivation," the B2 stimulus in the paper-and-pencil test contained the phrase "did intersect." Similarly, the four A graphs used in the paper-and-pencil test were the same format as those used for computer-based training; the graphs used in the paper-and-pencil test contained functions with slopes and intercepts that were different from those used in the A stimuli in Figure 1. Three faculty members in the Department of Psychology at Queens College/CUNY, each recognized as an expert teacher of statistics, assessed the validity of the test and concluded that it would measure knowledge of each type of statistical interaction accurately.

Although unlikely, it is possible that the answers to the questions in the two versions of the test could be determined by listing the questions in the same order or by listing the answers to each question in the same order. To obviate such a source of control, the two versions of the test listed the same questions in different orders and listed the answers to the same question in a different order. (The tests can be obtained from the first author.)

All participants in both experimental and control groups were randomly assigned to receive A or B versions of the paper pretest in alternating orders. The test was conducted in a classroom and given to all participants at the same time in a group format. Instructions for completing the test were dictated from a typed sheet. All participants were given a maximum of 45 min to complete the test. After completion, experimental participants were led to cubicles and began computer-based training, and control group participants were given a 1.5-hr break before returning to take the posttest.

The two versions of the test were randomly assigned as pretest and posttest with the constraint that each was used equally in each test. The sequence of test administrations was nearly balanced across both groups; because of the odd number of participants, the A then B test sequence was presented one more time than the B then A test sequence. Thus, differences in scores on the pretest and posttest could not be attributed to the particular version of the test.

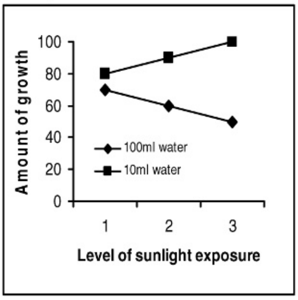
Computer-based procedure. Equivalence classes were established with trials presented in matching-to-sample format. Three stimuli were presented on the computer screen in an equilateral triangular array with the sample at the top of the triangle and the two comparisons at the bottom left and right of the triangle. A trial began by pressing the ENTER key, which produced the sample stimulus. Pressing the space bar then produced the two comparison stimuli. All three stimuli remained on the screen until the participant selected the comparison on the left by pressing the 1 key or the comparison

Sample Test Questions

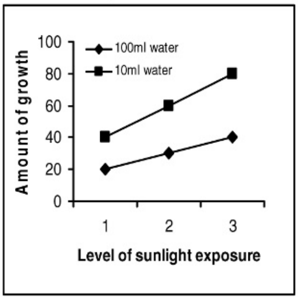
B-A relation from Class 2

1) Which graph depicts the following relation? An increase in the amount of water received by a plant reversed the effect of sunlight exposure on plant growth: In addition, the effects did intersect.

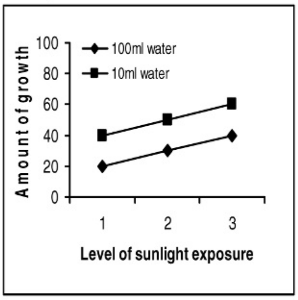
A)

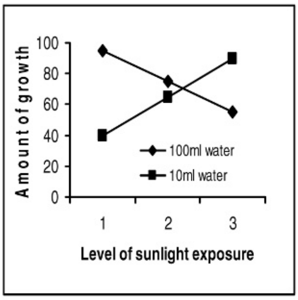


B)



C) D)





D-C relation from Class 4

2) Independent variable A produced the same directional change in responding for all values of independent variable B. Changing the value of independent variable B enhanced the effect of independent variable A. The effects do not intersect. Which label below best suits this definition?

A) Divergent interaction
B) No interaction
C) Synergistic interaction
D) Crossover interaction

Figure 2. Two examples of questions on the paper pretest and posttest. The first question tests a B-A relation from Class 2, and the second question tests a D-C relation from Class 4.

on the right by pressing the 2 key. Immediately thereafter, the stimuli disappeared and were replaced with one of two informative feedback messages or a noninformative feedback message. Correct and incorrect choices produced the words ‘‘right’’ and ‘‘wrong,’’ respectively. If a trial was scheduled for noninformative feedback, the letter E appeared in the screen. The feedback messages remained on the screen until the participant pressed the R key in the presence of right, the W key in the presence of wrong, and the E key in the presence of E.

All training and testing were conducted in blocks of trials presented in a randomized order without replacement. For training, a block was repeated until performances reached a mastery criterion, and trials in a block were conducted in conditions that either produced (a) informative feedback on 100% of the trials; (b) informative feedback on 75%, 25%, or 0% of the trials; or (c) noninformative feedback. During blocks that tracked the emergence of derived relations, all trials produced noninformative feedback.

Keyboard familiarization. Training began a procedure to teach participants the sequence of responses needed to negotiate the matching-to-sample trials used throughout the experiment (Fields et al., 1997). The stimuli were two sets of three words semantically related to each other. Each trial in the block consisted of a sample and positive comparison that was from the same semantically related set and a negative comparison that was from the other set. In addition, the response keys were indicated with onscreen prompts. If the performance criterion of 100% accuracy was achieved in a block of trials, the next block contained fewer prompts, which were faded in four steps. Familiarization training was complete once a block of trials produced the mastery level of responding in the presence of trials that did not contain any prompts.

Equivalence class formation. At the completion of keyboard familiarization training, participants in the experimental group were exposed to computer-based protocol to induce four four-member interaction-indicative equivalence classes (Class 1 = no interaction, Class 2 = crossover interaction, Class 3 = divergent interaction, and Class 4 = synergistic interaction). Trials were presented in the same matching-to-sample format used during keyboard familiarization, but with no prompts. The sequence of training and testing blocks followed the simple-to-complex protocol (Adams, Fields, & Verhave, 1993; Imam, 2006). Because the participants were university students, it was assumed that a generalized identity-matching repertoire was present already (i.e., if given an A1 stimulus, participants would be able to select the A1 comparison because it was identical to the sample stimulus); therefore, identity relations were not tested.

During all training and testing phases, unless otherwise noted, stimuli from one class were locked with stimuli from a specific corresponding class as negative comparisons. Stimuli from Class 1 served as negative comparisons for Class 4 members, and Class 2 stimuli served as negative comparisons for Class 3 members, and vice versa. For example, when training A1 to B1, the negative comparisons for the Class 1 stimuli consisted of B members from Class 4 (B4). Trials used to train or test for each relation are listed in Table 1.

Training for baseline conditional discriminations and testing for the emergence of derived relations began with establishing the baseline A-B relations, using a block that contained 16 trials: four presentations of each trial listed in the A-B section of Table 1. Training continued with 100% feedback until the mastery criterion was achieved. Thereafter, feedback in successive blocks was systematically reduced from 100% to 75% to 25% and then to 0% of trials as long as performance was maintained at the mastery level of responding. These blocks contained only eight A-B trials. This method established the baseline conditional discriminations using 100% feedback and assessed the maintenance of these relations with the reduction of feedback.

The maintenance of the A-B relations was followed with tests for the emergence of the symmetrical properties of the A-B relation with B-A probes. This B-A test block contained eight B-A trials: two presentations of each trial listed in the B-A section of Table 1. These trials were presented with no informative feedback. The block was repeated up to three times or until a participant responded in a class-indicative manner on all trials (the mastery criterion of 100% accuracy). After passing the B-A test, the

Table 1
Symbolic Representation of Samples (Sa), Positive Comparisons (Co+), and Negative Comparisons (Co−) Used During Equivalence Class Formation

Three-member classes					Four-member classes				
Rel	Type	Sa	Co+	Co−	Rel	Type	Sa	Co+	Co−
A-B	BL	A1	B1	B4			A3	C3	C2
		A4	B4	B1					
		A2	B2	B3	C-A	EQV	C1	A1	A4
		A3	B3	B2			C4	A4	A1
							C2	A2	A3
B-A	SYM	B1	A1	A4			C3	A3	A2
		B4	A4	A1					
		B2	A2	A3	C-D	BL	C1	D1	D4
		B3	A3	A2			C4	D4	D1
							C2	D2	D3
B-C	BL	B1	C1	C4			C3	D3	D2
		B4	C4	C1					
		B2	C2	C3	D-C	SYM	D1	C1	C4
		B3	C3	C2			D4	C4	C1
							D2	C2	C3
C-B	SYM	C1	B1	B4			D3	C3	C2
		C4	B4	B1					
		C2	B2	B3	B-D	TTY	B1	D1	D4
		C3	B3	B2			B4	D4	D1
							B2	D2	D3
A-C	TTY	A1	C1	C4			B3	D3	D2
		A4	C4	C1					
		A2	C2	C3	A-D	TTY	A1	D1	D4
		A4	D4	D1					
		A2	D2	D3					
		A3	D3	D2					
D-B	EQV	D1	B1	B4					
		D4	B4	B1					
		D2	B2	B3					
		D3	B3	B2					
D-A	EQV	D1	A1	A4					
		D4	A4	A1					
		D2	A2	A3					
		D3	A3	A2					

Note. Entries in the Rel column indicate the stimulus–stimulus pairs in the equivalence classes. Entries in the Type column indicate the kind of relation served by each stimulus–stimulus pair. BL indicates the trials used to train the baseline relations, and SYM indicates the symmetry probe trials. TTY indicates transitivity probe trials. EQV indicates equivalence probe trials. Each Sa/Co+/Co− trial was presented two times per block, once each with the positive comparison presented on the left and the right, and vice versa for the negative comparisons.

B-C relations were trained in the same manner as A-B relations. The block used for training with 100% feedback contained 16 B-C trials: four presentations of each trial listed in the B-C section of Table 1. Maintenance of B-C relations during feedback reduction used a block that contained eight B-C trials: two presentations of each trial listed in the B-C section of Table 1. This was followed by a test for C-B symmetry that was conducted in the same manner as the B-A test. The C-B testing block contained two presentations of each trial listed in the C-B section of Table 1. After passing the C-B test, a maintenance test of both symmetrical relations was conducted by presenting the B-A and C-B relations together in the same block of 16 that contained two presentations of each trial listed in the B-A and C-B sections of Table 1. This was followed by a test for transitivity with a block that contained the eight trials listed in the A-C section of Table 1. Finally, the emergence of the equivalence relations was assessed with a test block that contained eight C-A trials (C-A section of Table 1). The occurrence of class-consistent responding on all training and probe blocks would indicate the formation of four three-member equivalence classes. For each derived relations test, a block was repeated up to three times or until a participant responded in a class-indicative manner on all trials within a block (the mastery criterion of 100% accuracy).

The next phase was a three-mix probe test that involved the presentation of A-B, B-A, B-C, C-B, A-C, and C-A trials in one test block. Each relation was presented eight times, all with no informative feedback. This test was presented in three blocks, each of which contained 16 trials. The presentation of each relation was balanced across the three blocks, and each class appeared an equal number of times within and across these three blocks. Class-consistent performances on these blocks would indicate the maintenance of the four three-member interaction classes when all baseline relations and derived relations were presented together.

Once maintenance of the three-member classes was established, the class membership was expanded by training C-D relations for all four equivalence classes in blocks of descending feedback. After C-D training, participants were presented with a four-mix test that included all possible relations, A-B, B-A, B-C, C-B, A-C, C-A, C-D, D-C, A-D, D-A, B-D, and D-B. Each relation was assessed with the presentation of eight trials, as listed in Table 1. This test consisted of 96 trials that represented all possible stimulus relation in the four classes presented in four separate blocks containing 24 trials each to avoid participant fatigue. Progress through each testing block was not dependent on performance.

In all previous training and testing blocks, a sample stimulus on a trial was presented with a positive comparison from the same class and a negative comparison that was drawn from one specific class (i.e., the locked class: Class 1 with Class 4 and Class 2 with Class 3). Because the positive and negative comparisons were from invariant classes, it was possible that the four classes would remain intact only in the context of the stimuli used as negative comparisons. Alternatively, the classes might have remained intact regardless of the stimuli used as negative comparisons. These possibilities were evaluated with the next battery of probes, called a four-mix-plus test.

The four-mix-plus test involved the presentation of trials that contained examples of all of the relations used in the four-mix test with the following extension. Each sample and positive comparison (Co+) from the same class was now presented with negative comparisons (Co−) that were drawn from the two classes that had not been used during class formation. For samples and positive comparisons drawn from Classes 1 and 4, the negative comparisons were drawn from Classes 2 and 3, and for samples and positive comparisons drawn from Classes 2 and 3, the negative comparisons were drawn from Classes 1 and 4. In addition, the new Co–s were used on different trials. For example, in the four-mix test, the A1 stimulus would be presented with B1 as the positive comparison and B4 as the only negative comparison in every trial. By contrast, a trial in the four-mix-plus test that contained A1 and B1 as the sample and positive comparison would now be presented with the novel negative comparisons B2 and B3 in two separate trials, but not with B4.

To avoid participant fatigue, the 192 trials were presented in 16 blocks that contained 12 trials each presented once each and in the same order for all participants. The correct comparison appeared with equal probability in the left and right positions in each block. In addition, each stimulus relation contained three questions in each block. Because there were three questions per relation, the number of questions from each stimulus class could not be balanced per block given this uneven number. Nevertheless, if one block contained fewer questions from a certain class, the following block would correct the imbalance by presenting more questions from that class. This would create an imbalance within another block that was again corrected in the subsequent block. Thus, the questions from each relation were balanced within each block, but the number of questions drawn from each class was balanced over the entire 16 blocks.

Paper-and-pencil retest. The second version of the paper-and-pencil test was administered after completion of computer-based class induction for participants in the experimental group, and about 90 min after the administration of the first paper-and-pencil tests for participants in the control group.

Social validity questionnaire. A social validity questionnaire assessed the goals, methods, and outcome of the experiment. Participants answered four questions by assignment of scores from 1 to 7 on a Likert scale, with 1 and 7 being the lowest and highest rankings, respectively. Once completed, participants were debriefed, given the opportunity to ask questions, provided with a means to contact experimenters in the future, and issued course credit.

RESULTS

Time spent in the experiment. The participants in the class formation group spent from 2.8 to 3.5 hr in the experiment; about 1.5 hr was spent in the formation of the interaction-indicative equivalence classes. The participants in the control group were given a 1.5-hr delay between the completion of the first paper-and-pencil test and the presentation of the second paper-and-pencil test. Thus, the time between test administrations was equivalent for participants in both conditions.

Formation of three-member equivalence classes. All participants in the experimental group formed four four-member interaction-indicative equivalence classes. Therefore, equivalence class formation was depicted using group means for each phase of training and testing (Figure 3). A minimum of four blocks were needed to establish each baseline relation. The A-B and B-C relations were acquired rapidly, in means of 5.5 and 5.6 blocks, respectively. The narrowness of the standard error bars indicates the similarity in performances across participants. With few exceptions, all emergent relations tests (B-A, C-B, A-C, and C-A) produced mastery levels of responding in the first block of a test. The few participants who needed to repeat test blocks were able to meet mastery criterion on the second presentation of the block. Along with the mastery levels of responding produced by the baseline relations (A-B and B-C), these performances documented the formation of four three-member equivalence classes.

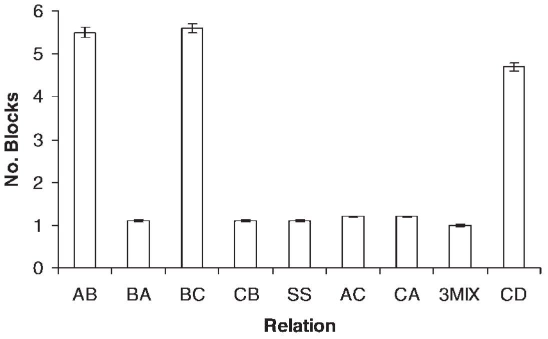


Figure 3. The mean number of blocks needed for all experimental group participants to achieve mastery criterion during the computer-based equivalence training. Each training and testing phase of equivalence class formation appears as a separate bar, and the left to right position of each bar corresponds to the order in which each relation was trained or tested. The height of each bar indicates the mean number of blocks needed to form a baseline relation or to pass an emergent relations test.

Maintenance of the three-member equivalence classes. In all cases, these probes produced mastery levels of responding during the first presentation of the test block when all relations were mixed together, rather than being presented on an individual basis in separate test blocks. These performances, obtained with all 10 participants, demonstrated the maintenance of all four three-member classes. Thus, the performances produced by all of the emergent relations were not compromised by their presentation in a single test block.

Table 2
Scores in Test Blocks on the Two Posttraining Computer Tests

	Four-mix blocks				Four-mix-plus blocks															
Participant	1	2	3	4	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
3271	100	100	100	100	92	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3287	100	100	94	100	100	92	100	100	100	100	100	100	92	100	100	100	100	100	100	100
3270	94	100	100	94	100	100	92	100	100	92	100	100	100	100	100	100	92	100	100	100
3293	100	83	100	100	100	92	100	92	100	100	100	100	100	100	100	100	100	100	100	100
3268	100	100	100	100	100	100	100	92	92	100	92	100	92	92	100	100	100	100	100	100
3314	100	100	100	100	100	92	100	100	100	100	100	100	100	100	92	100	100	100	83	92
3316	100	83	100	83	92	92	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3273	100	83	83	89	92	92	100	100	83	92	92	83	100	92	92	100	83	100	100	83
3309	89	94	94	94	100	100	100	83	100	100	100	100	92	100	92	83	92	75	83	92
3311	100	100	100	100	100	100	100	67	50	100	100	67	100	100	100	100	100	50	42	33

Expansion to four-member equivalence classes. The C-D baseline relations were acquired in a mean of 4.7 blocks (Table 2). The criterion used to define the formation of an equivalence class was the experimenter-selected score of at least 90% class-consistent comparison selections when averaged across all four blocks of the four-mix test. Nine of the 10 participants met this criterion, which demonstrated the formation of the four-member interaction-indicative equivalence classes. The stability of accuracy scores across the four blocks of the four-mix test also demonstrated the immediate emergence of all four interaction-indicative equivalence classes. One participant (3273) just missed the criterion level of responding needed to demonstrate class formation (i.e., 89% correct).

Four-member classes with novel negative comparisons. The emergence of the three- and four-member classes could have been contextually limited to the particular negative comparisons used for training and testing. The four-mix-plus test evaluated that possibility by presenting all trials with negative comparisons from all classes (Table 2). In the first three blocks of the four-mix-plus test, performances were typically 100% accurate for all 10 participants. The maintenance of criterion levels of responding with the introduction of the four-mix-plus test demonstrated that the relations among the stimuli in the four interaction-indicative equivalence classes were maintained in the presence of new Co–s in the baseline and emergent relations test trials. Notably, these class-indicative performances were maintained even with the sudden substitution of trials that contained new comparisons. These performances then demonstrated one level of generalization of the four interaction-indicative equivalence classes.

With a continuation of four-mix-plus testing, different patterns of responding emerged for different participants. Six of the 10 participants responded at the mastery level for the entirety of the four-mix-plus test, which demonstrated the maintenance of the classes with extensive testing under conditions of uninformative feedback. Two of the 4 remaining participants (3311 and 3309) showed some minor performance breakdowns in some of the test blocks (shaded cells in Table 2). For Participants 3311 and 3309, the performance breakdowns were more precipitous and occurred with increased frequency in the later test blocks (shaded cells). For them, the classes did not remain intact. Additional research will be needed to identify factors that are responsible for the maintenance of equivalence relations with continued testing and with novel negative comparisons.

Overall effects: Paper-and-pencil test scores. Figure 4 depicts the overall effects of the two independent variables by plotting the mean scores on the paper-and-pencil tests for the participants in the experimental and control

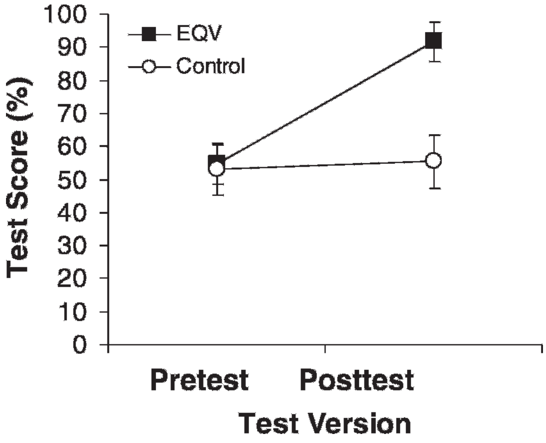


Figure 4. The mean pretest–posttest scores for both experimental (filled squares) and control (open circles) groups. The I beams that bracket each data point indicated ± 1 *SE.*

groups on the first and second administrations of the test. By design, the pretest scores for both groups were very similar to each other. Thus, any post-class-formation differences could not be attributed to participant-based variables. In the control group, the mean posttest score was only 2% greater than the pretest score. The overlap in standard errors showed that the difference was not significant. In the experimental group, the mean posttest score was 37% higher than the mean pretest score. When the posttest scores were compared, the participants in the experimental group had paper-and-pencil posttest scores that were 35% greater than the corresponding scores for the participants in the control group. The difference between groups on the posttest score was significant after

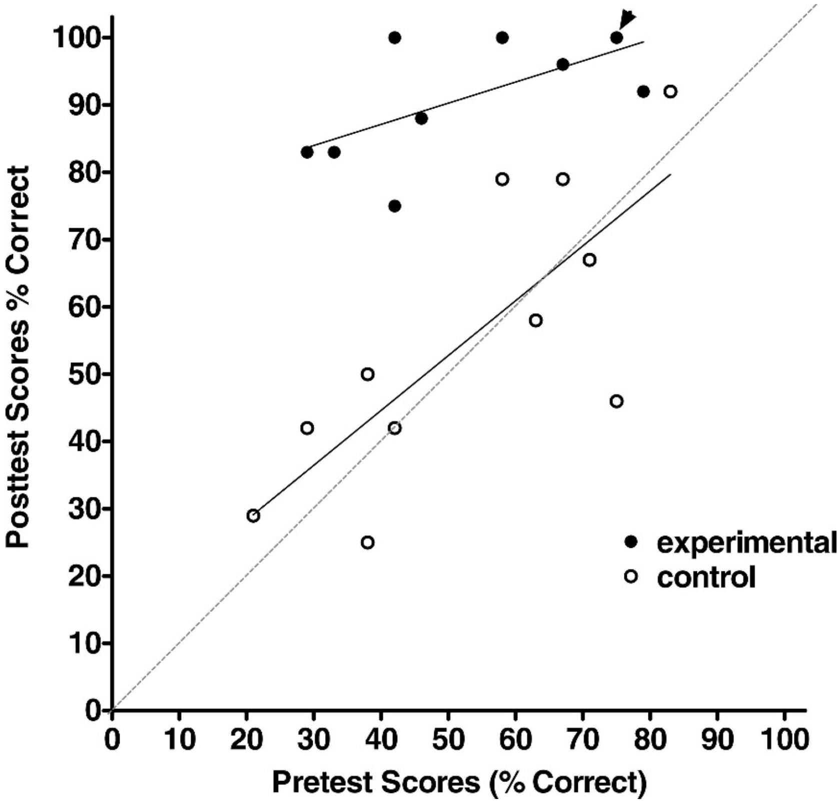


Figure 5. A scattergram showing posttest scores plotted as function of pretest scores for each participant in the experimental (filled circles) and control (open circles) groups. Two participants in the experimental group produced identical pretest and posttest scores, indicated by the arrow. Separate regression lines are also shown for the data obtained from participants in the experimental and control groups.

controlling for any potential differences between the groups on pretest scores (ANCOVA, *df* 19, $n = 21$, $F = 42.56$, $p < .000004$). In addition, $r^2 = 0.775$ indicated that more than 77% of the variance in the values of the dependent variable was accounted for by the experimental intervention. Finally, effect size was $d = 2.3$ (Cohen, 1992). Because effect sizes that are greater than 0.8 are considered to be large, the obtained effect size obtained in the present experiment is exceptionally large.

Performances by matched participants. The data in Figure 4 did not permit a comparison of individual participants who were matched in terms of initial knowledge of interaction. That information is presented in Figure 5, which plots posttest scores as a function of pretest scores for each participant. The diagonal line with a slope of 1 that began at the origin indicated one-to-one correspondences of pre- and posttest scores. The scores on the pretest varied from 29% to 83% for participants in both conditions. For participants in the control condition, the posttest scores were quite similar to the pretest scores. These scores straddled the diagonal line, thereby indicating a nearly one-to-one correspondence of pretest and posttest scores.

For participants in the experimental condition, posttest scores were reliably higher than the scores produced by matching participants in the control condition. Although the posttest scores were similar to each other, there was a small increase in posttest score that was directly correlated with pretest score. This was indicated by the shallow positive slope of the regression line that was fitted to the data obtained from participants in the experimental condition. The weakness of the correlation was documented by the fact that only 42% of the variance in the posttest scores was accounted for by the pretest scores. For these participants, the magnitude of the increment in posttest score over pretest score became smaller with increases in pretest scores. This ceiling effect was inevitable because high pretest scores precluded large increases in posttest scores.

The data presented in Figure 5 can also be viewed in terms of traditional letter grades earned on a typical classroom quiz. Test scores of at least 80% correct correspond to letter grades of A and B. Test scores no greater than 69% correct correspond to letter grades of D and F. As can be seen in the posttest data in Figure 5, 10 of the 11 participants with grades in the A and B range were in the experimental group and one was in the control group. The 1 participant in the control group who did obtain a high grade had already scored a passing grade in the pretest. By contrast, all 8 of the participants with grades in the D and F range were in the control group, and none were in the experimental group. These differences could have occurred by chance with an exact probability of .0001 (Fisher's exact test). If grades on an examination can be used to assess social validity in an academic setting, this analysis indexed the high level of social validity that can be ascribed to equivalence-based instruction.

Social validity. The four questions on the social validity questionnaire provided the following mean ratings. "Please rate your current understanding of statistical interactions" produced a mean rating of 6 (*SE* = 0.5) for participants in the experimental group and 3 (*SE* = 0.5) for the participants in the control group. "Are you happy with the methods used in this study?" produced a mean rating of 6 (*SE* = 1.0) for participants in the experimental group and 3 (*SE* = 0.33) for the participants in the control group. "Are the methods used in this study acceptable?" produced a mean rating of 6 (*SE* = 0.66) for participants in the experimental group and 3 (*SE* = 0.66) for the participants in the control group. Thus, participants in the experimental group reported that the computer training was acceptable and effective. "Is it a good goal to use effective teaching methods to teach the concept of statistical interactions to students?" produced a mean rating of 6 (*SE* = 0.5) for participants in the experimental group and 6 (*SE* = 0.66) for participants in the control group. Finally, during a postexperiment debriefing, the participants in the experimental group reported feeling more confident in their understanding of statistical interactions, and several students reported that they would like to see a similar teaching format used for other difficult concepts in statistics.

DISCUSSION

Formation of interaction-indicative equivalence classes. Knowledge of statistical interaction was evaluated with a paper-and-pencil test that determined whether an individual could match four different representations of interactions with each other for four different types of interaction. Before intervention, the population of college students enrolled in a course in introductory psychology provided correct answers to about 54% of the questions on the pretest, indicating a low level of knowledge regarding the interchangeability of representations of each type of interaction. Participants in an experimental condition were exposed to a computer-based program that resulted in the formation of four interaction-indicative equivalence classes. In this part of the experiment, after training three stimulus–stimulus relations in each of the four classes, 12 new relations among the stimuli in each class emerged immediately and without benefit of direct training. Further, the paper-and-pencil test administered after class formation yielded scores that were on average 37% higher than pretest scores. Thus, equivalence class induction procedures established knowledge of the interchangeability of perceptually distinct representations of four different forms of statistical interaction. Because the representations used in the paper-and-pencil tests differed from those used as members of the trained equivalence classes, the participants generalized the knowledge learned during training to novel exemplars. By implication, those participants should also be able to apply what had been learned to new examples encountered in real-world settings.

Prior research has shown that test repetition can increase scores on a test without any explicit intervention (Lievens, Buyse, & Sackett, 2005; Wing, 1980). Therefore, the increment in the paper-and-pencil test scores after the establishment of the equivalence classes could have been influenced by the repetition of the test. The present experiment used a control condition that measured the effect of test repetition in the absence of an intervention. Thus, any effects of test repetition on score improvements would be factored out by the subtraction of any increase in the control group score from the pretest to the posttest from the gains obtained in the experimental condition. The repetition of the test in the control group produced a 2% increase in test scores. When this estimate is subtracted from the improvements in experimental group scores (37%), the computer-based equivalence intervention accounted for a 35% mean improvement in posttest scores. Thus, test repetition had a minimal effect on the increase in scores on the test after the establishment of equivalence classes. The increase in test scores for participants in the experimental group can be attributed to the induction of the four interaction-indicative equivalence classes.

Social validity and pedagogical implications. The study ended with an evaluation of social validity for the participants in the experimental group. They indicated that the treatment goals were valid, the procedures were acceptable, and their changes in test scores were important. During the debriefing phase of the experiment, many experimental group participants reported that they "finally got" what constituted a statistical interaction. This verbal report is supported by their improved performances from pre- to posttesting. In summary, these postexperimental comments about the procedure support the validity of its usage to teach this subject matter.

Equivalence class formation was an effective method for teaching individuals to identify equivalent representations of the combined effects of two variables on some dependent variable. Given the fact that many students struggle with statistics in particular, the stimulus control technology embodied in the establishment of equivalence classes can provide an important contribution to the longstanding debate about the improvement of the skills deficits these individuals present. Indeed, establishing classes of equivalent stimuli may be an effective technology for remedying both students' inabilities to apply the concepts learned in instruction to real-world problems and the "behavioral weaknesses" identified by Bradstreet (1996) and Seipel and Apigian (2005), respectively.

Factors that influence the likelihood of class formation. Sidman (1987), Carrigan and Sidman (1992), and Johnson and Sidman (1992) have argued that the use of only two comparison stimuli could lead to responding away from a Co− (called a reject relation) rather than responding to an experimenter-defined sample–Co+ relation (called a select relation). If so, responding would give the illusion of control by the relation between a sample and a comparison from the same class and of class formation. In the present experiment, although training and testing were conducted with two comparisons per trial, class-consistent performances were maintained during the four-mix-plus tests, which involved the presentation of trials with two additional negative comparisons. Responding, then, had to be controlled by the relations between the samples and the comparisons that came from the same class as the samples (i.e., by select relations). Thus, four four-member equivalence classes were formed using only two comparisons per trial. This finding is consistent with recent data that showed similar likelihoods of class formation using two, three, and six comparisons per trial (Saunders, Chaney, & Marquis, 2005). Perhaps the establishment of classes using locked pairs is one parameter that increases the likelihood of forming equivalence classes using only two comparisons.

Many studies have shown that equivalence class formation is optimized with classes that have only one nodal stimulus and that have sample-as-node or comparison-as-node training structures instead of linear series training structures (Arntzen & Holth, 1997; Green & Saunders, 1998; Saunders & Green, 1999). In the present experiment, however, all participants formed four four-member equivalence classes with rapidity even though they contained two nodal stimuli instead of one and had linear series training structures. These results raise questions regarding the validity of the general view mentioned above. Perhaps it was the use of a simple-to-complex training and testing protocol and the use of semantically meaningful stimuli that were responsible for the reliable and rapid formation of equivalence classes that contained a few nodes and had a linear series training structure.

Generalization of relations in equivalence classes. In many situations, it is necessary to establish behavioral repertoires that are expected to occur in contexts other than those in which the behavior is trained (Stokes & Baer, 1977). Within the context of education, a student is expected to respond correctly to appropriate and novel examples that differ from the stimuli or relations used during formal instruction.

In the present experiment, the generalization of the relations in the interaction-indicative equivalence classes to novel exemplars was assessed in five ways. One involved determining whether the within-class relations remained intact when tested in the context of novel negative comparisons. This circumstance was evaluated with the results of the four-mix-plus test. Specifically, it is possible that the relations among the stimuli in one equivalence class would remain intact only when tested in the presence of the negative comparisons used in the training trials. The results of the four-mix-plus tests proved that the relations in each class remained intact even when tested in the context of comparisons drawn from classes not used during training. These data then demonstrated the generalization of the emergent relations to new contexts that varied in terms of negative exemplars. Other tests assessed the generalization of relational control when the within-class relations contained stimuli that were variants of the stimuli used to establish the classes, when test trials were presented in a format that differed from that used during computer-based instruction, when test trials contained a different number of choices from which to select the comparison that was from the same class as the sample, and when the order of test questions was controlled by the participant rather than by the experimenter. Generalization to these four modes of testing was assessed concurrently with the performances recorded on the post-class-formation paper-and-pencil tests. Specifically, the tests contained questions that differed in content from the stimuli used when forming the corresponding equivalence classes during computer-based instruction. The format of the questions in the paper-and-pencil test differed in many ways from the trial format used during the formation of equivalence classes. If choices in the test are equated to the comparisons presented in the class-formation procedures, the two differed in terms of using two versus four choices per question or trial. Finally, whereas the participant did not control the order of trial presentations during the computer-based four-mix and four-mix-plus tests, the participant was free to scan the questions in the paper-and-pencil test in any order and to change answers to any question prior to submitting it. In most cases, participants responded with high levels of accuracy on the post-class-formation paper-and-pencil tests.

Generalized equivalence classes. The performances mentioned above demonstrated the generalization of the relations among the stimuli in each of the equivalence classes to novel exemplars presented in novel formats. This sort of generalization is also characteristic of generalized equivalence classes, classes that contain sets of perceptually disparate stimuli and other stimuli that are perceptual variants of the former stimuli (Adams et al., 1993; Belanich & Fields, 2003; Branch, 1994; Fields & Reeve, 2000; Lane, Clow, Innis, & Critchfield, 1998). Thus, the classes that emerged in the present experiment were generalized equivalence classes.

The generalization that occurred to novel stimulus exemplars in the present experiment was also reported by Ninness et al. (2006) but not by Cowley et al. (1992) and Lynch and Cuvo (1995). A number of studies have identified training and testing parameters that broaden the range of variants that come to function as members of generalized equivalence classes. (Belanich & Fields, 2003; Fields et al., 1991, 2002; Fields & Reeve, 2001; Galizio, Stewart, & Pilgrim, 2004). Perhaps the generalization problems reported by Cowley et al. and Lynch and Cuvo could be overcome by the inclusion of the above-mentioned parameters in replications of their experiments.

Limitations of the present study. This experiment had four limitations. First, it formed classes with only four exemplars. An interaction, however, can have other representations such as bar graphs, tables of data, and summary statements of factorial ANOVAs. The expansion of class size to include these exemplars and their variants would extend a student's ability to identify the wide range of representations of interactions that would be encountered in natural settings. Second, to understand interactions, a student should be able to identify different representations of an interaction, which uses a selection-based or receptive repertoire, and also describe an interaction verbally or in written form, which uses a production-based or expressive repertoire. The present study explored the emergence of the former but not the latter repertoire. Third, the present experiment did not determine how different modes of instruction such as equivalence class formation, listening to traditional lectures, and self-study of textbook material affect the acquisition of knowledge of statistical interactions. Fourth, the present study demonstrated the feasibility of using equivalence-based instruction to teach one particular college-level subject matter. A similar approach might be used to establish understanding of the contents of other academic subject matters. Additional research would be needed to address each of these limitations.

Individualized education and behavioral diagnostics. Equivalence class procedures can isolate specific relational deficits among the stimuli that should be functioning as members of a particular interaction-indicative equivalence class. For example, although a student may accurately identify a particular type of interaction when given a graph and a description of the effects of the variables depicted in that graph, the same individual might not identify that type of interaction when given a description of the graph. Once discovered, that information might be used to design a minimal intervention that should induce all of the deficient or missing relations in a class. In short, a system of behavioral diagnostics (Sidman, 1986) could be used to develop tailor-made training programs that would correct the stimulus control deficiencies in an individual's behavioral repertoire with a minimal amount of training and testing—an individualized instruction process that is largely absent in standardized group-oriented teaching curricula. Such a strategy, then, should lead to the development of a technology of teaching (Skinner, 1968) and a personalized system of instruction (Keller, 1968; Pear & Crone-Todd, 1999).

REFERENCES

Adams, B. J., Fields, L., & Verhave, T. (1993). Formation of generalized equivalence classes. *The Psychological Record*, *43*, 553–566.

Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record*, *47*, 309–320.

Belanich, J., & Fields, L. (2003). Generalized equivalence classes as response transfer networks. *The Psychological Record*, *53*, 373–414.

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician*, *50*, 69–78.

Branch, M. (1994). Stimulus generalization, stimulus equivalence, and response hierarchies. In S. C. Hayes, L. J. Hayes, M. Sato, & K. Ono (Eds.), *Behavior analysis of language and cognition* (pp. 51–70). Reno, NV: Context Press.

Carrigan, P. F., & Sidman, M. (1992). Conditional discrimination and equivalence relations: A theoretical analysis of control by negative stimuli. *Journal of the Experimental Analysis of Behavior*, *58*, 183–204.

Christopher, A. N., & Marek, P. (2002). A sweet tasting demonstration of random occurrences. *Teaching of Psychology*, *29*, 122–125.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Connell, J. E., & Witt, J. C. (2004). Applications of computer-based instruction: Using specialized software to aid letter-name and letter-sound recognition. *Journal of Applied Behavior Analysis*, *37*, 67–71.

Cowley, B. J., Green, G., & Braunling-McMorrow, D. (1992). Using stimulus equivalence procedures to teach name-face matching to adults with brain injuries. *Journal of Applied Behavior Analysis*, *25*, 461–475.

Davidson, G. V., & Kroll, D. L. (1991). An overview of research on cooperative learning related to mathematics. *Journal of Research in Mathematics Education*, *22*, 362–365.

de Rose, J. C., de Souza, D. G., & Hanna, E. S. (1996). Teaching reading and spelling: Exclusion and stimulus equivalence. *Journal of Applied Behavior Analysis*, *29*, 451–469.

Fields, L., Matneja, P., Varelas, A., Belanich, J., Fitzer, A., & Shamoun, K. (2002). The formation of linked perceptual classes. *Journal of the Experimental Analysis of Behavior*, *78*, 271–290.

Fields, L., & Reeve, K. F. (2000). Synthesizing equivalence classes and natural categories from perceptual and relational classes. In J. C. Leslie & D. Blackman (Eds.), *Experimental and applied analysis of human behavior* (pp. 59–83). Reno, NV: Context Press.

Fields, L., & Reeve, K. F. (2001). A methodological integration of generalized equivalence classes, natural categories, and cross-modal perception. *The Psychological Record*, *51*, 67–87.

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

Fields, L., Reeve, K. F., Rosen, D., Varelas, A., Adams, B. J., Belanich, J., et al. (1997). Using the simultaneous protocol to study equivalence class formation: The facilitating effects of nodal number and size of previously established equivalence classes. *Journal of the Experimental Analysis of Behavior*, *67*, 367–389.

Fields, L., & Verhave, T. (1987). The structure of equivalence classes. *Journal of the Experimental Analysis of Behavior*, *48*, 317–332.

Fields, L., Verhave, T., & Fath, S. (1984). Stimulus equivalence and transitive associations: A methodological analysis. *Journal of the Experimental Analysis of Behavior*, *42*, 143–157.

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

Green, G., & Saunders, R. R. (1998). Stimulus equivalence. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 229–262). New York: Plenum.

Guercio, J. M., Podolska-Schroeder, H., & Rehfeldt, R. A. (2004). Stimulus equivalence technology to teach emotion recognition skills to adults with acquired brain injury. *Brain Injury*, *18*, 593–601.

Hinde, R. J., & Kovac, J. (2001). Student active learning methods in physical chemistry. *Journal of Chemical Education*, *78*, 93–99.

Imam, A. (2006). Experimental control of nodality via equal presentations of conditional discriminations in different equivalence protocols under speed and no-speed conditions. *Journal of the Experimental Analysis of Behavior*, *85*, 107–124.

Johnson, C., & Sidman, M. (1992). Conditional discriminations and equivalence relations: Control by negative stimuli. *Journal of the Experimental Analysis of Behavior*, *59*, 333–347.

Keller, F. S. (1968). Good-bye teacher. *Journal of Applied Behavior Analysis*, *1*, 79–89.

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

Lane, S. D., Clow, J. K., Innis, A., & Critchfield, T. S. (1998). Generalization of cross-modal stimulus equivalence classes: Operant processes as components in human category formation. *Journal of the Experimental Analysis of Behavior*, *70*, 267–280.

LeBlanc, L. A., Miguel, C. F., Cummings, A. R., Goldsmith, T. R., & Carr, J. E. (2003). The effects of three stimulus-equivalence testing conditions on emergent US geography relations of children diagnosed with autism. *Behavioral Interventions*, *18*, 279–289.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, *58*, 981–1007.

Lynch, D. C., & Cuvo, A. J. (1995). Stimulus equivalence instruction of fraction-decimal relations. *Journal of Applied Behavior Analysis*, *28*, 115–126.

Mulhern, G., & Wylie, J. (2004). Changing levels of numeracy and other core mathematical skills among psychology undergraduates between 1992 and 2002. *British Journal of Psychology*, *95*, 355–370.

Nasser, F. (1999). Prediction of statistics achievement. In *Proceedings of the International Statistical Institute 52nd Conference* (Vol. 3, pp. 7–8). Helsinki, Finland.

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

Pear, J. J., & Crone-Todd, D. E. (1999). Personalized systems of instruction in cyberspace. *Journal of Applied Behavior Analysis*, *32*, 205–209.

Peden, B. F. (2001). Correlational analysis and interpretation: Graphs prevent gaffes. *Teaching of Psychology*, *28*, 129–131.

Rosenthal, B. (1992). No more sadistics, no more sadists, no more victims [editorial]. *UMAP Journal*, *13*, 281–290.

Saunders, R. R., Chaney, L., & Marquis, J. G. (2005). Equivalence class establishment with two, three, and four-choice matching-to-sample by senior citizens. *The Psychological Record*, *55*, 539–559.

Saunders, R. R., & Green, G. (1999). A discrimination analysis of training-structure effects on stimulus equivalence outcomes. *Journal of the Experimental Analysis of Behavior*, *72*, 117–137.

Seipel, S. J., & Apigian, C. H. (2005). Perfectionism in students: Implications in the instruction of statistics. *Journal of Statistics Education*, *13*, Retrieved February 16, 2006, from http://www.amstat.org/publications.jse/v13n2/seipel.html

Sidman, M. (1971). Reading and audio-visual equivalences. *Journal of Speech and Hearing Research*, *14*, 5–13.

Sidman, M. (1986). The measurement of behavioral development. In N. A. Krasnegor, D. B. Gray, & T. Thompson (Eds.), *Advances in behavioral pharmacology: Vol. 5. Developmental behavioral pharmacology* (pp. 43–52). Hillsdale, NJ: Erlbaum.

Sidman, M. (1987). Two choices are not enough. *Behavior Analysis*, *22*, 11–18.

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

Sidman, M., Kirk, B., & Willson-Morris, M. (1985). Six-member stimulus classes generated by conditional-discrimination procedures. *Journal of the Experimental Analysis of Behavior*, *43*, 21–42.

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

Simon, J. L., & Bruce, P. (1991). Resampling: A tool for everyday statistical work. *Chance*, *4*, 23–32.

Skinner, B. F. (1968). *The technology of teaching*. Englewood Cliffs, NJ: Prentice Hall.

Smeets, P. M., & Barnes-Holmes, D. (2005). Establishing equivalence classes in preschool children with one-to-many and many-to-one training protocols. *Behavioural Processes*, *69*, 281–293.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis*, *10*, 349–367.

Ward, L. G., & Kaflowitz, N. G. (1986). Issues in research training … again? *Counseling Psychologist*, *14*, 139–145.

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement*, *4*, 141–155.

Received January 7, 2008
Final acceptance June 18, 2008
Action Editor, Chris Ninness