Effects of an Online Stimulus Equivalence Teaching Procedure on Research Design Open-Ended Questions Performance of International Undergraduate Students

Ana Carolina Sella · Daniela Mendonça Ribeiro · Glen W. White

Abstract The purpose of the present study was to assess if training in matching-to-sample (MTS) tasks would yield not only new MTS performance but also written topography-based responses involving research design names, definitions, notations, and examples in four international undergraduates. Thirty-six experimental stimuli—composed of nine research design names and their corresponding definitions, notations, and examples—were presented in a MTS format during teaching and emergent conditional relations testing sessions. The six topography-based response probes were composed of nine open-ended questions each. Participants learned all conditional relations, showed emergence of symmetric and transitive relations, and of topography-based responses. The present study provides some steps towards the use of stimulus equivalence for international delivery of content.

Keywords Stimulus equivalence · Research methods · Online instruction

Many behavior analysts have been working towards translating highly effective teaching procedures (which were developed and tested mostly in laboratories and laboratory-like settings) into applied and service-delivery contexts, such as schools (Johnson and Street 2004) and college classrooms (Neef et al. 2011), online education (Walker and Rehfeldt 2012), among other settings.

Stimulus equivalence is among the procedures that have empirical evidence on its efficacy and efficiency to teach

A. C. Sella (✉) · G. W. White
University of Kansas, 1000 Sunnyside Avenue, Room 4089, Lawrence, KS 66045, USA
e-mail: carolsella@yahoo.com.br

D. Mendonça Ribeiro
Universidade Federal de São Carlos, São Carlos, Brazil

different skills to different populations in different settings (de Rose et al. 1996; Fienup et al. 2010; Fienup and Critchfield 2010). It is an attempt to explain how the myriad of arbitrary relations among signs and their referents, which characterize human symbolic functions, are formed. Equivalence-based instructions are considered important because they aim at teaching generatively (Fienup et al. 2010; Fienup and Critchfield 2010). This implies programming procedures in a way that involves direct teaching of a few conditional discriminations that will yield untaught performances (Fienup et al. 2010; Green and Saunders 1998; Sidman 1971). The possibility of yielding new, untaught behaviors is very important when one considers the limited instructional time that one might have to teach a given content. As discussed by Lovett et al. (2011), Walker et al. (2010), and Walker and Rehfeldt (2012), this possibility becomes even more noteworthy when one can yield topography-based responses from teaching selection-based responses.

In the last few years, several studies have investigated the feasibility of using of stimulus equivalence in higher education instruction (Critchfield and Fienup 2010; Fields et al. 2009; Fienup et al. 2010; Fienup and Critchfield 2010, 2011; Lovett et al. 2011; Ninness et al. 2005, 2006, 2009; Walker and Rehfeldt 2012; Walker et al. 2010). Target topics of instruction included statistics (Critchfield and Fienup 2010; Fields et al. 2009; Fienup and Critchfield 2010, 2011), brain–behavior relations (Fienup et al. 2010), mathematical formulas and their graphed analogues (Ninness et al. 2005, 2006, 2009), disabilities (Walker et al. 2010), and single-subject designs (Lovett et al. 2011; Walker and Rehfeldt 2012). Overall, the dependent variables of interest in these studies included percentage of correct responses, average number of correct responses and number of correct trials in testing, number of trials (or block of trials) to mastery criterion in teaching sessions, average time to complete the tasks, total time of engagement in tasks, errors during pretest, generalization to

novel relations and responses, and social validity of the procedures. Teaching procedures included MTS tasks, train to mastery, and accuracy feedback. Additional procedures included, but were not limited to, introductory lectures, computer assisted instruction, error correction procedures, and differential reinforcement with gradual fading. The teaching procedures have been presented in a variety of formats, such as computerized; paper and pencil; online; and live, oral instruction.

These studies advanced the application of stimulus equivalence to teach complex verbal behaviors (i.e., college-level topics) to advanced learners. The studies' results were successful in teaching participants to a level of performance accuracy in taught relations, in showing the emergence of many novel relations, in demanding very little student time investment. In addition, these studies provided demonstrations of stimulus equivalence efficacy and efficiency under different conditions (i.e., different instruction formats, material formats, and in different settings). For example, while Ninness et al. (2005) used instructor-generated explanations (i.e., lecture) before the MTS tasks, Fienup and Critchfield (2010) did not provide this type of explanation, thus offering a demonstration of the efficacy of equivalence procedures in a more direct way. Additionally, some studies taught and tested the target relations in a multiple-choice format (Fienup and Critchfield 2010) while other studies targeted teaching selection-based responses and tested topography-based responses (Lovett et al. 2011; Walker and Rehfeldt 2012; Walker et al. 2010).

Specifically, Lovett et al. (2011) evaluated the efficacy of a stimulus equivalence protocol to teach single-subject designs. The authors presented names (A), definitions (B), graphical representations (C), and clinical vignettes (D) of withdrawn, multiple baseline, alternating treatment, and changing criterion single-subject designs. MTS teaching sessions (A-B, A-C, A-D), symmetry (B-A, C-A, D-A), transitivity (B-C, C-B, C-D, D-C), and generalization tests (A-C', A-D') were computer-based (programmed using Microsoft Visual Basic 2008). Tact tests (topography-based responses) were presented using flash cards (C-A and D-A, then C'-A and D'-A). Lovett et al. also administered paper-and-pencil pre- and posttests that allowed the authors to compare the results of the equivalence protocol to the results of a group that was taught in a traditional lecture format. Additionally, these authors administered a social validity protocol to evaluate satisfaction with instruction. Walker and Rehfeldt (2012) used the same stimuli as Lovett et al. (2011) to evaluate the efficacy of an MTS protocol to teach single-subject design methodology to graduate-level professionals. The authors taught A-B, A-C, and B-D relations directly. Neither symmetry, transitivity, nor symmetry of transitivity relations were tested. Testing tasks involved what the authors named tact relations (CA; C'1-A, C'2A) and intraverbal relations (BA, DA, D'1-A, D'2-A, D-B, D'1-B, D'2-B). All teaching and testing was delivered thorough Blackboard. Finally, Walker et al. (2010) examined the effects of an equivalence-based teaching protocol on the emergence of topography-based derived stimulus relations. The authors presented names (A) and definitions (B) of disabilities, a disability's primary cause (C), and an effective treatment or service (D) for the particular disability. The authors taught A-B, A-C, and C-D relations directly and tested B-A, C-A, D-A (Final Test 1) and A-B', A-C', A-D' (in Final Test 2).[1]

Aiming at advancing the research on the effects of selection-based teaching protocols on written topography-based responses of distant learners, the purpose of the present study was to assess if MTS teaching tasks would yield not only emergent MTS performance, but also written topography-based responses. In this study, nine names, nine definitions, nine notations, and nine examples of group research designs were grouped into three types of research designs (pre-, quasi-, and true-experimental designs). Stimuli were defined through a concept analysis (Tiemann and Markle 1990). For each type of research design, MTS teaching involved A-B, A-C, and A-D relations. Then, symmetry, transitivity, and symmetry of transitivity relations were tested. After each type of research design was trained and tested in an MTS format, open-ended questions were presented to assess topography-based written responses. The open-ended questions were comprised of three parts: A, B, and C. As in Walker et al. (2010)'s Final Test 2, Parts A and B of the open-ended questions assessed if, when presented with a *design name*, the participant would write its definition (Part A) and its notation (Part B). Part C assessed if, when presented with a *design name*, the participant would create a new example, different from what was presented in the MTS teaching tasks. In addition to the research design names, Parts A, B, and C contained key words that described the relevant aspects of the stimuli that were to be included in the written responses (Lovett et al. 2011). All training and testing tasks were presented through Adobe Connect®.

Method

Participants

Participants were four Brazilian undergraduates, John (19 years old), Sarah (22 years old), and Mary and Barbara (23 years old), whose primary language was Portuguese. They were enrolled in at least one class in the social or behavioral sciences, and reported that they had not had any advanced

---

[1] Even though the authors used A-B, A-C, and A-D to define the relations tested in Final Test 2, the " ' " was added to highlight the fact that what was requested from the participants was a topography-based response.

classes on research methods. Participants were instructed to read, sign, scan, and return the consent form—as approved by the *Human Subject Committee of Lawrence* (HSCL # 19431)—by e-mail if they wished to participate. In exchange for participating, participants received a book, a PDF package with research related articles, and a participation certificate.[2]

Participants were selected due to their performance in an online reading comprehension test (Loft 2009), which was administered to assure that they had the initial reading comprehension repertoire necessary to participate in the study (i.e., at least six out of nine correct responses in the reading comprehension test). John emitted six, Barbara emitted seven, Sarah emitted eight, and Mary emitted nine correct responses. In addition to scoring at least six out of nine questions correctly in the reading test, participants remained eligible for the study only if their percentage of correct responses in each one of the Probes 1, 2, and 3 was lower than 20 % (probes are described below). All participants had scores lower than this criterion.

John and Mary were exposed to all phases of the procedure. Barbara and Sarah were exposed to Probes 1, 2, 3, glossary, MTS teaching and testing sessions of pre-experimental designs, and Probe 4. After Probe 4, these two participants told the experimenter that they needed to withdraw from the study due to personal issues.

Settings

Two virtual environments were used for data collection: Skype ™ and Adobe Connect®.

Skype ™ is a software application that allows users to make voice and video calls over the Internet, to exchange information, to share screens, and to access other communication functions. Skype ™ was used to host the reading comprehension test.

Adobe Connect® is a computer program that can be used to deliver information through presentations, online training modules, and web conferencing. Additionally, Adobe Connect® can record data on participants' responses to different types of questions. Adobe Connect® was used to host the session content, to present all trials and instructions, and to record the data from both the MTS tasks and the open-ended questions. Since Adobe Connect® is based on Adobe Flash®, all participants were required to have Adobe Flash® installed on their computers to access the links to the sessions. All links were sent by e-mail. The software Adobe Captivate® was used to program all sessions.

Experimental Stimuli

Group research designs were targeted as experimental stimuli, since these designs are widely used and considered the "gold

standard" in the behavioral and social sciences (Cozby 2008; Odom et al. 2005). Given that in the research literature different group design names are used to identify similar designs, Campbell and Stanley[3]'s (1963) book was used to establish which design names would be part of stimulus set A (design names) in this study.

Campbell and Stanley (1963)'s book describes several group research designs: three pre-experimental designs, three true experimental designs, and several quasi-experimental designs. To control for the number of stimuli included in each stimulus class, besides the three pre-experimental designs and the three true experimental designs, three quasi-experimental designs were randomly selected. The nine group research designs presented in this study were (a) pre-experimental designs: the one-shot case study, the one-group pretest-posttest design, the static group comparison; (b) quasi-experimental designs: nonequivalent control-group design, counterbalanced design, the multiple time series design; and (c) true experimental designs: the pretest-posttest control group design, the Solomon four-group design, and the posttest-only control group design.

Once these nine research designs names were defined, ten concept analyses (Tiemann and Markle 1990; Twyman et al. 2005) were conducted: one for group research designs in general, and one for each of the nine research designs that would comprise the stimulus classes. As described by Sota et al. (2011), Tiemann and Markle (1990), among others, the concept analysis resulted in the shared and varying properties of the members of our stimulus classes. The shared properties are hereafter called *critical attributes*; the varying properties, *variable attributes* (Tiemann and Markle 1990). Searches for the critical and variable attributes were conducted using group research design and the nine research design names as key words.

Two types of stimuli derived from the ten concept analyses: (a) MTS stimuli[4] (27 stimuli which were grouped into three sets)—nine research design definitions (Set B), nine research design notations (Set C), and nine research design (brief) examples (Set D), and (b) key words used as prompts in the open-ended questions (i.e., participants, dependent variable, independent variable, and randomization). "Appendix A" displays all the experimental stimuli used during the study.

As suggested by Walker et al. (2010), the stimulus format presentation within and across sets of stimuli was kept as similar as possible to try to avoid establishing discriminations based on stimulus format and/or number of words.[5]

---

[2] Money payment is not allowed in research studies in Brazil.

[3] When this study was designed, Campbell and Stanley's (1963) book was the most cited reference in the area.
[4] The nine research design names had been previously determined as described at the beginning of the "Experimental Stimuli" section.
[5] Notations are preestablished graphic representations of research designs. Therefore, they cannot be manipulated in terms of its format and components.

Dependent Variables

The primary dependent variables of interest were percentage of correct responses in (a) teaching sessions; (b) symmetry, transitivity and symmetry of transitivity sessions; and (c) open-ended questions (Parts A, B, and C). Additionally, responses to Part C of the open-ended questions (which instructed participants to generate examples that were not presented during MTS tasks) were analyzed according to (a) different examples generated across probes and (b) different examples generated across questions.

Experimental Design

A multiple-probe design across the three types of research designs was implemented on an individual basis (Barlow et al. 2009; Horner and Baer 1978). Performance in the three types of research designs (pre-experimental designs, quasi-experimental designs, and true-experimental designs) was measured three times before the introduction of the intervention and again after each part was taught to each participant through MTS tasks. As described in the "Experimental Stimuli", each type of group research design was composed of three different research designs: pre-experimental designs encompassed the one-shot case study, the one-group pretest–posttest design, and the static-group comparison. Quasi-experimental designs encompassed the nonequivalent control-group design, the counterbalanced design, and the multiple time series design. Finally, the true experimental designs encompassed the pretest–posttest control-group design, Solomon four-group design, and posttest-only control-group design.

Figure 1 depicts the general experimental phases, performance criteria, and contingency definition in each experimental phase, including both open-ended and MTS tasks.

Procedures

All sessions were hosted virtually and participants could access them at any time, from any computer (as long as Adobe Flash was installed). A session was defined as a block of teaching or testing trials, and the number of trials and trial format varied according to the experimental phase (sessions are described in detail below). Participants were instructed to complete two MTS sessions in one day (one teaching and one testing session). They took an average of 10 min to go through each MTS session. In probe session days, participants were instructed to complete only one session: these sessions involved nine open-ended questions that took an average of 25 min to be answered. When a participant finished either the two consecutive MTS sessions or a probe session, he/she e-mailed the experimenter. A link to a new session was sent if performance criterion was met; if not, a link to a reteaching

session was sent instead. If a participant did not write the experimenter within three days of the sending of a link, he/she was prompted by e-mail to work on the tasks. The total number of sessions each participant was exposed to depended on performance. Participants that went through all experimental phases completed the procedure within 24 to 28 days of its onset (i.e., reading test).

Glossary

This activity was optional and it was developed to give participants access to basic research vocabulary. The glossary was presented after the three baselines probes, directly before the presentation of the first MTS teaching session. The target vocabulary terms were *reliability*, *independent variable*, *dependent variable*, *correlation*, *cause-and-effect*, *validity*, *notation*, *selection bias*, *random numbers table*, *experimental control*, *sample*, *randomization*, and *control group*.
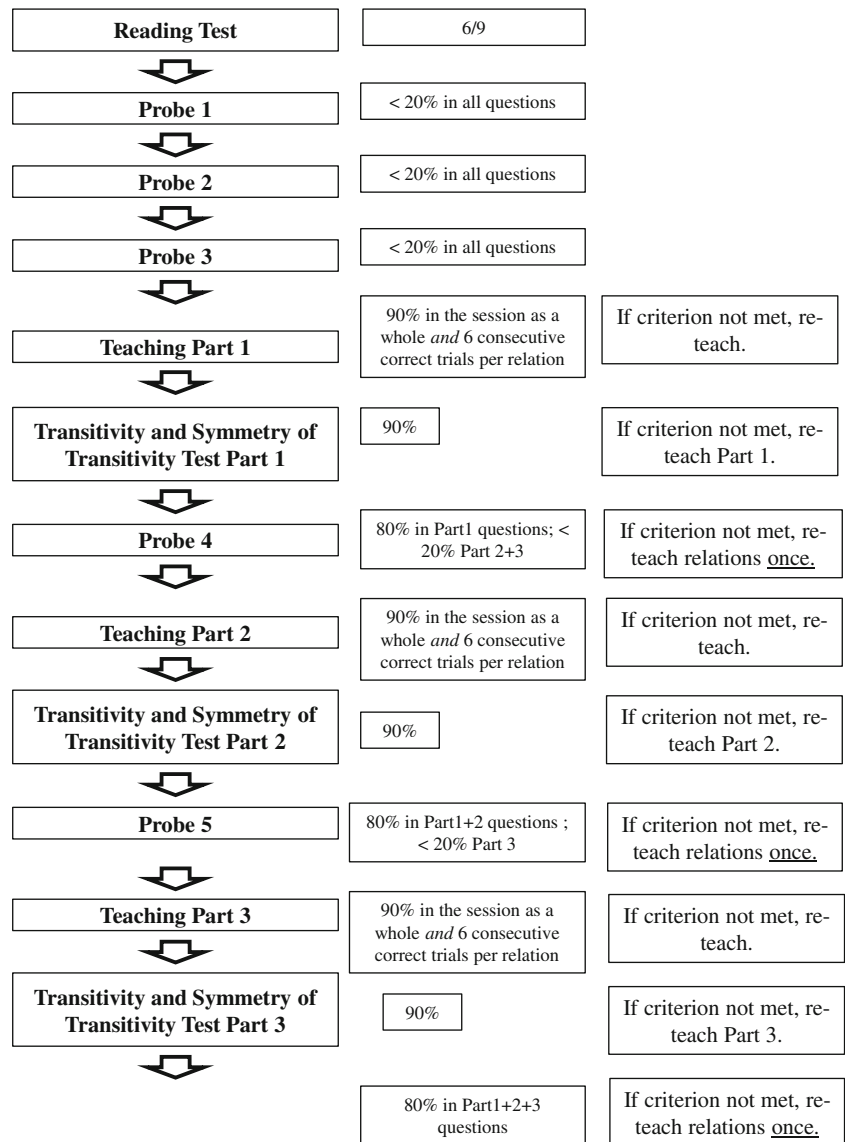
The glossary was composed of three Adobe Captivate slides containing MTS tasks. The definition of each term was presented on the left part of the screen, with the letters A, B, or C in front of the definition. The terms were presented on the right part of the screen, and the participant clicked on a drop-down arrow to the right of the terms to select which letter matched the term's name to its definition. Accuracy feedback was provided on-screen: correct responses resulted in the presentation of a 3 cm×2 cm green rectangle as the background for the word *correct* at the top right corner of the screen. Incorrect responses resulted in a 3 cm×2 cm pink rectangle, with the words *incorrect—try again* at the bottom right part of the screen. Participants received the correct answers by e-mail, so they could refer to them at any moment throughout the study. All participants went through and scored 100 % correct responses in the glossary.

Probe Sessions

A total of six probes were conducted to verify participants' performance in open-ended questions before (Probes 1, 2, and 3) and after each one of the three MTS teaching and testing conditions (Probes 4, 5, and 6). The first three probes were used as baseline measures. The other three probes were used as measures against which changes in performance were evaluated. At the beginning of each probe session, the following instruction was provided:

> Today you will answer open-ended questions. You can: (a) answer the questions in English; answer the questions in your own language; use a mix of languages if you need to borrow some expressions. Be sure to answer ALL items of the questions before you move onto the next question. If you do not know the answer, write "I do not know the answer" before moving onto the next

Fig. 1 Experimental phases, performance criteria, and contingencies of the study

Reading Test	6/9	
Probe 1	< 20% in all questions	
Probe 2	< 20% in all questions	
Probe 3	< 20% in all questions	
Teaching Part 1	90% in the session as a whole *and* 6 consecutive correct trials per relation	If criterion not met, re-teach.
Transitivity and Symmetry of Transitivity Test Part 1	90%	If criterion not met, re-teach Part 1.
Probe 4	80% in Part1 questions; < 20% Part 2+3	If criterion not met, re-teach relations once.
Teaching Part 2	90% in the session as a whole *and* 6 consecutive correct trials per relation	If criterion not met, re-teach.
Transitivity and Symmetry of Transitivity Test Part 2	90%	If criterion not met, re-teach Part 2.
Probe 5	80% in Part1+2 questions ; < 20% Part 3	If criterion not met, re-teach relations once.
Teaching Part 3	90% in the session as a whole *and* 6 consecutive correct trials per relation	If criterion not met, re-teach.
Transitivity and Symmetry of Transitivity Test Part 3	90%	If criterion not met, re-teach Part 3.
	80% in Part1+2+3 questions	If criterion not met, re-teach relations once.

question. You will not be allowed to go back to previous questions.

After this instruction, nine open-ended questions were presented, one for each of the nine research designs. Each open-ended question presented a design name and key words (derived from the concept analysis of group research designs) that prompted what should be contained in the answers. All open-ended questions were presented in the following format:

Part A[6]: "What features define the [*design name*] design? (Be sure to write: i. when the dependent variable is measured; ii. when the independent variable is introduced; iii. how many groups are needed and if there is a control group; iv. if there is randomization)." Part B: "What is the notation that represents this design? (Be sure to illustrate: i. when the dependent variable is measured; ii. when the independent variable is introduced; iii. the group(s); iv. the randomization, if applicable)." Part C: "Please, provide an example of this design, different from the one given in the multiple-choice questions, and be sure to include: i. what is the dependent variable(s) and when it will be measured (observed); ii. what is the independent variable(s) and when it is introduced; iii. who are the participants, how many participants you will have, and if the participants will be divided into groups; iv. will there be randomization?"

After answering one question, the participant clicked on "submit" and the next question was presented. Correct or

---

[6] Part A tested name—written definition relations; Part B, name—written notation relations; and Part C, name—written new example relations.

incorrect responses were followed only by the next question. After answering the nine questions, the participant e-mailed the experimenter to request the link to the next experimental condition. Responses to open-ended questions were scored according to a 73-page rubric. After scoring the questions, the experimenter sent the link to the next experimental condition or the link relative to the relations that needed to be retaught. In open-ended question tasks, mastery criterion varied depending on the experimental phase. During the first three open-ended question tasks (i.e., baseline probes), participants that had more than 20 % correct responses were discontinued from the study. For probes after MTS teaching and testing tasks, mastery criterion of 80 % of correct responses was required in Parts A and B of the questions encompassing research designs that had already been taught. Failure to meet 80 % of correct responses in Parts A and B led to MTS tasks reteaching. There was no performance criterion for Part C of any of the open-ended questions, since these required novel responses that were not targeted during teaching sessions.

MTS Teaching Sessions

These sessions targeted at A-B, A-C, and A-D relations. Each MTS teaching session targeted one of the three types of group research designs (i.e., pre-, quasi-, or true-experimental designs). The first teaching session regarding pre-experimental designs was introduced right after the Glossary. The first teaching session regarding quasi-experimental designs was introduced right after the fourth probe. And the first teaching session regarding true-experimental designs was introduced right after the fifth probe.

At the beginning of each teaching session, the following instruction[7] was provided:

> These are multiple-choice tasks. You will be shown a research design name at the top of the screen. Then, three choices will be presented. The alternatives give you definitions of the research design names. Choose the definition that you think best matches the design name at the top. I will let you know every time you make a correct choice or I will ask you to try again.

After the six-name (A)–definition (B) trials were presented, instructions on the next set of trials (A-C) were presented: "Now, the alternatives will give you notations of the research design names. Choose the notation that you think best matches the design name at the top."

Another block of at least six trials was presented. Finally, instructions on the name (A)–examples (D) were presented:

"Now, the alternatives give you examples of the research designs. Choose the example that you think best matches the design name at the top."

Each teaching session was comprised of at least 18 trials, but the total number of trials was automatically increased if participants emitted an incorrect response. A new block of six trials was presented every time there was an error. The new block of trials included only the relation in which there was an error (A-B, A-C, or A-D). The 18 trials were subdivided into three blocks. The first six trials presented design name–design definition relations (A-B relations); the next six trials contained design name–design notation relations (A-C relations); and the last six trials encompassed design name–design example relations (A-D relations). Trials were presented in a MTS format: a sample stimulus at the top of the screen and three comparison stimuli at the bottom of the screen. All stimuli were simultaneously presented. The sample stimulus, the position of all comparison stimuli, and the position of the correct answer were quasirandomized: The position was never the same in more than two consecutive trials.

During MTS teaching sessions, accuracy feedback was provided in the same way as it was in the Glossary tasks. However, for incorrect responses, not only visual feedback was provided but also the same trial was presented as a correction trial (not scored as correct or incorrect) and the program automatically presented a new set of six trials, as described previously.

Mastery criterion in teaching sessions was two-folded: For the session as a whole, the criterion was 90 % of correct responses. For each one of the three conditional relations taught in a session (i.e., A-B, A-C, and A-D), the participant had to emit six consecutive correct responses (a new set of six trials was presented until this criterion was met). After criterion was met in a teaching session, a testing session was presented.

MTS Testing Sessions

These sessions were designed to test the emergence of symmetry (B-A, C-A, D-A), transitivity (B-C, C-D) and symmetry of transitivity (C-B, D-C) relations.[8] Each testing session was composed of 21 trials in which each conditional relation was presented three times, one trial for each conditional discrimination pair being tested (e.g., one trial of B1-A1, one of B2-A2, one of B3-A3). At the beginning of a session, this instruction was provided: "Now you will go through 21 questions related to what you learned in the last session. Remember, the following questions will not provide feedback on correct or incorrect responses, so be sure to pay attention to your

---

[7] It is important to mention that instructions on note taking were not provided.

[8] BD and DB relations were not tested since pilot data showed that participants responded at 70 % of correct responses or more in these relations, before MTS teaching was presented.

responses." Trials were presented in a MTS format, similar to the one in teaching trials. However, during the tests, correct or incorrect responses were followed only by the next trial. Conditional relations and the position of correct answers presentation were also quasi-random. Mastery criterion in MTS testing tasks was 90 % of correct responses. If criterion was not met, MTS teaching sessions were presented again.

Figure 2 depicts MTS teaching and testing tasks on its upper half; the lower half of the figure shows the tested open-ended tasks.

Social Validity

Participants were encouraged to give feedback on the experimental stimuli and the experimental phases at any moment: in all the e-mails that contained links to teaching or testing sessions, the last sentences included requests to inform on any problems that might have occurred during sessions, and/or suggestions on stimulus presentation. In addition, at the end of the study, participants were sent a social validity questionnaire through e-mail which contained nine affirmative propositions about the procedures. Six out of the nine propositions were Likert-type scaled. The scale ranged from 1 (*strongly disagree*) to 5 (*strongly agree*) and referred to (a) ease of use; (b) online sessions and time and space flexibility; (c) usefulness of information; (d) links sent on time; (e) importance of feedback on teaching sessions; and (f) recommendation to other people. The last three questions were open ended and asked about (g) the most useful and (h) the least useful features of the procedures, and (i) requested additional suggestions for changes.

Interobserver Agreement

Interobserver agreement (IOA) was assessed for all six open-ended probes, for at least one participant per probe. IOA was calculated for both (a) percentage of correct responses in Part A, B, and C and for (b) Part C in regard to examples themes across and within probes. Selection of participant's probes was random.

In regard to percentage of correct responses in probes (Parts A, B, and C), IOA was calculated in two different forms. The first one consisted of assigning each one of the 12 items comprising the open-ended questions either *agreement* (value=1) or *disagreement* (value=0) and dividing the number of agreements by the agreements plus disagreements. Then, the results were multiplied by 100. Using this calculation, IOA was 100 % for Probes 1, 2, 3, and 6; 93.5 % for Probe 4; and 91.7 % for Probe 5. In the second form of calculation—since items in the open-ended questions could be scored as 0, 0.25, 0.5, 0.75, or 1—for items in which there was not exact agreement (exact agreement counted as 1), the smaller score was divided by the higher score to find the partial agreement for a given item. The agreements and partial agreements were added and divided by 72 (total number of items in a probe), and multiplied by 100. Agreement was 100 % for Probes 1, 2, 3, and 6 and 95.1 % for Probes 4 and 5. Regarding different examples generated across probes and different examples generated across questions in Part C, agreement was 100 %.

Results

Baseline results (Probes 1, 2, and 3) for all participants and posttests results (Probe 4 for all participants; Probes 5 and 6

Fig. 2 Summary of the relations taught and tested in pre-experimental designs, quasi-experimental designs, and true-experimental designs. *Black arrows* show taught relations and *gray arrows* show tested relations
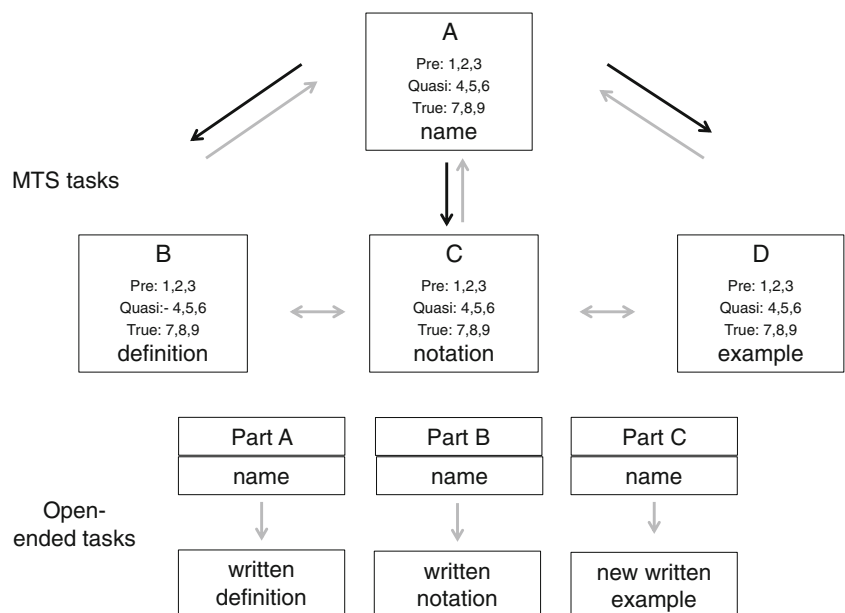
for John and Mary) in pre-, quasi-, and true-experimental designs are shown in Fig. 3. Figure 4 displays the percentage of correct responses in symmetry (B-A, C-A, D-A), transitivity (B-C, C-D), symmetry of transitivity (C-B, D-C), and overall scores in each one of the MTS testing sessions, for all participants.

Probes

As depicted in Fig. 3, prior to teaching sessions, all participants scored below 20.0 % of correct responses in Probes 1, 2, and 3. In Probe 4, right after the pre-experimental designs were taught, they scored between 89.5 % and 94.8 % of correct responses in Parts A and B, and 37.5 % to 87.5 % of correct responses in Part C, in the three questions regarding pre-experimental designs. There was no change in performance in questions regarding quasi- and true-experimental designs. In regard to their new examples (Part C of open-ended questions), all participants used one *theme* to illustrate the three different pre-experimental designs. For example, even though John defined the different features of each one of the three pre-experimental designs in his examples, he used *diuresis* as the dependent variable (DV) and *beer* as the independent variable (IV) in all three examples. The same pattern of responding was observed in Mary's, Barbara's, and Sarah's responses: Mary used *blood sugar* (DV) and *breakfast* (IV); Barbara used *stress levels* (DV) and *go through surgery* (IV); and Sarah used *performance in story reading* (DV) and *reinforcement* (IV) in all three examples they provided in Part C.

Results of Probe 5—introduced after quasi-experimental designs were taught—show that John scored 100 % of correct responses in Parts A, B, and C of questions regarding pre-experimental designs. He scored 99.0 % in Parts A and B and 93.7 % in Part C of quasi-experimental designs. There was no change in performance in true experimental designs questions. Mary scored 92.7 % in Parts A and B and 58.3 % in Part C of pre-experimental designs, and 67.7 % in Parts A and B and 37.5 % in Part C of quasi-experimental designs. Thus, she did not meet the criterion to move to the next experimental phase. So, quasi-experimental designs MTS teaching and testing tasks were re-presented. After criterion was met in the MTS tasks, Probe 5 was reintroduced. Mary scored above 70 % of correct responses in pre-experimental questions and 80.2 % in Parts A and B, and 68.7 % in quasi-experimental questions. Regarding the examples given in Part C of Probe 5, both participants varied the theme of their examples when compared to those in Probe 4. However, they used one theme across the six probe questions for which they provided examples. John used *blood sugar* (DV) and *lasagna* (IV) in all six examples. Mary used *movement amplitude* (DV) and *a new stretching program* (IV) across the six examples she provided in both Probe 5 and Probe 5II.

Finally, in Probe 6, after three true-experimental designs were taught, John scored 99 % of correct responses in Parts A and B and 100 % in Part C of all pre-, quasi-, and true-experimental designs. Mary scored above 91 % of correct responses in Parts A and B, and above 79 % of correct responses in Part C of pre-, quasi-, and true-experimental designs.

It is worthwhile to mention that Mary used the same example theme from Probe 5 (*movement amplitude* [DV] and *a new stretching program* [IV]) in all Probe 6 questions. John maintained his pattern of responding, using a new theme throughout Probe 6 (*blood pressure* [DV] and *coffee* [IV]).

MTS Teaching Sessions

All participants met mastery criterion (90 % of correct responses). The total number of teaching trials was 56 and 90 for John and Mary—who finished the study—and 24 and 18 trials for Barbara and Sarah—who were exposed only to teaching of pre-experimental designs.

It is important to note that when Mary was in Trial 14 of the first teaching session, her internet signal was disconnected, so the four last trials were not presented to her (A-D relations). She went on to the MTS testing session (she had received both links together) and did not meet performance criterion then. Consequently, she was exposed to a new MTS teaching session on pre-experimental designs. Additionally, she did not meet performance criterion in Probe 5, so she was exposed to an additional teaching on quasi-experimental designs. Mary scored 100 % of correct responses during each one of the additional teaching sessions.

MTS Testing

As depicted in Fig. 4, Barbara and John scored 100 % of correct responses and Sarah scored 95.2 %, showing the emergence of symmetry, transitivity, and symmetry of transitivity in pre-experimental designs. Mary scored 66.6 % of correct responses during the pre-experimental design MTS testing session. As Mary did not meet the criterion, pre-experimental designs teaching tasks were presented again, followed by a new testing session. In this testing session, Mary scored 90.4 % of correct responses.

Only John and Mary were exposed to quasi- and true-experimental designs MTS testing sessions, in which both participants scored 100 % of correct responses.
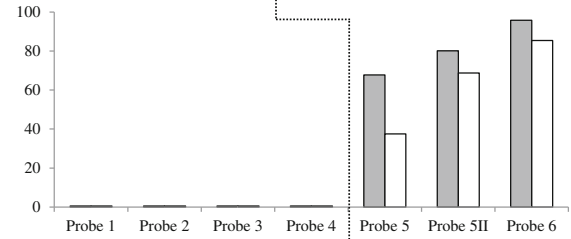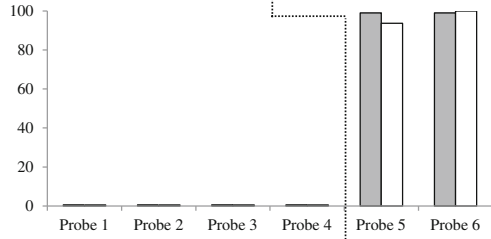
Fig. 3 Participants' percentage correct responses in probes, in each type ▶ of design: Performance in pre-experimental designs is at the top of the figure, quasi-experimental designs are in the center of the figure, and true-experimental designs are at the bottom of the figure
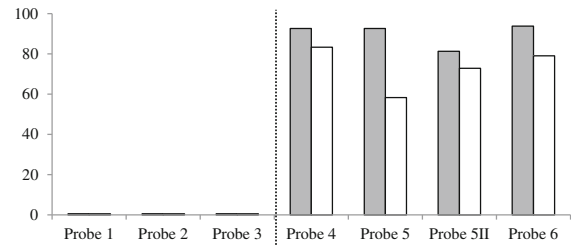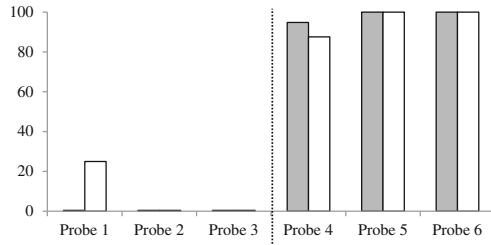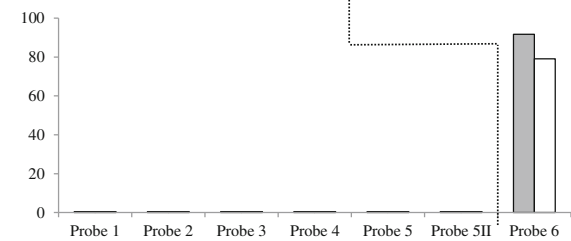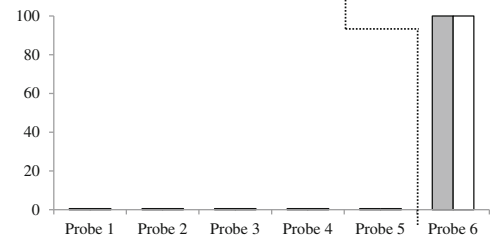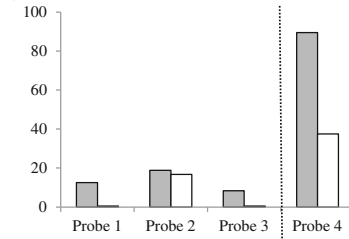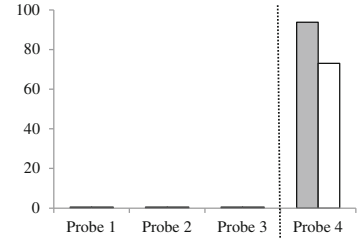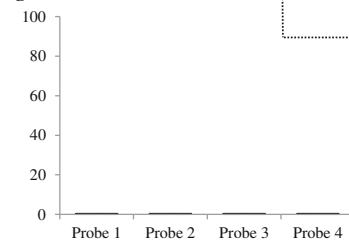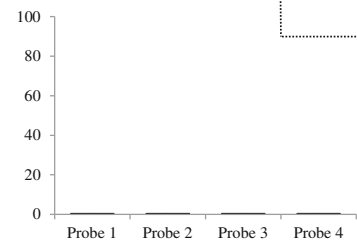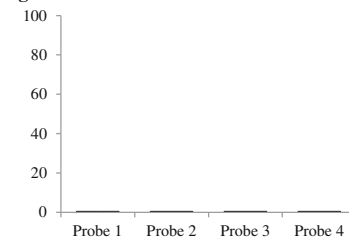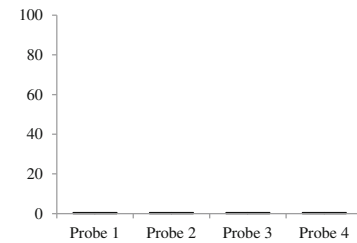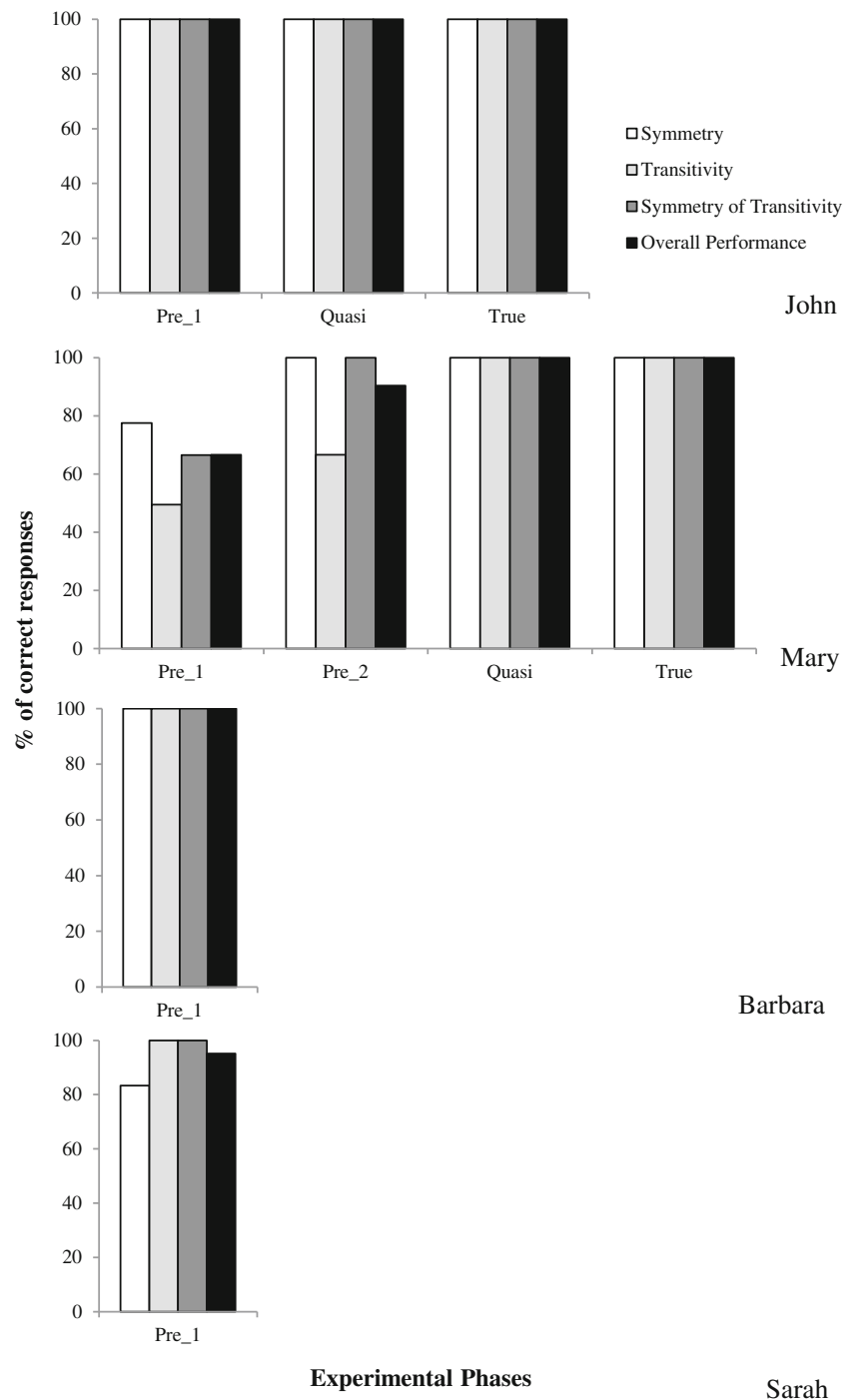
Fig. 4 Percentage of correct responses in symmetry (B-A, C-A, D-A), transitivity (B-C, C-D) and symmetry of transitivity (C-B, D-C) relation tasks for pre-, quasi-, and true-experimental designs, as well as the overall score in each one of the testing sessions, for all participants

Social Validity

The two participants who were exposed to the whole study and answered the social validity questionnaire gave high ratings to the tutorial. John rated all six Likert-type scaled propositions as 5, in a 1 (*strong agree*) to 5 (*strongly agree*) scale. Mary rated four of the six propositions as 5 and two as 4. In the open-ended propositions, both participants provided information on what to keep in the multiple-choice questions

formatting (John) and what to change in the probes—Mary suggested a decrease in the number of probes, because "the open-ended questions become tiring."

Discussion

The purpose of this study was to evaluate the effects of a stimulus equivalence instructional package on international

undergraduates' performance in conditional discrimination and open-ended tasks that involved group research design names, definitions, notations, and examples. The present study adds to the literature on the applications of stimulus equivalence technology in higher education in a few aspects: first, in Part C of the open-ended questions, participants were requested to generate creative examples. Even though Lovett et al. (2011) and Walker and Rehfeldt (2012) tested response generalization to novel stimuli, neither one of these studies had requested open-ended creative responses. Second, the stimuli were defined through a concept analysis. A concept analysis is a systematic way to find shared properties of stimuli that goes beyond using a book or article to define target stimuli. Third, as discussed by Lovett et al. (2011), key words describing the relevant aspects of the stimuli were included in the questions regarding the open-ended questions. Finally, all tasks were programmed in Adobe Captivate and delivered through Adobe Connect, thus showing the possibility of using other technologies (different from the ones used in previous studies) to deliver equivalence-based teaching protocols.

The results indicate the possibility of using stimulus equivalence technology in international online education settings, using programs that are readily available for the public. Moreover, the results from the social validity measures suggest that this format of instruction, using Skype, Adobe Captivate, and Adobe Connect might be an acceptable way of delivering online instruction. As highlighted by Fienup and Critchfield (2011) and Lovett et al. (2011), it is still necessary to compare this form of instruction delivery to other ones, including video conferences, podcasts, and webinars, among others.

This study used a multiple-probe design, which allowed for the monitoring of performance in open-ended questions before and after the introduction of MTS teaching and testing tasks. Performance in open-ended questions only increased after direct teaching in the MTS tasks. Participants met performance criterion in Parts A and B in all probes, except for Mary, who was exposed to new MTS teaching and testing tasks before meeting criterion in Probe 5. One might argue that the topography-based responses for Parts A and B are a result of participants taking notes during MTS teaching tasks. This procedure did not aim at teaching note-taking behaviors or requiring participants to take and submit notes since this would increase response costs. Additionally, since the study was done online and with international students, it was decided not to give any instructions on note taking, since controlling these behaviors would not be feasible. Future studies should address this question and seek for ways to either control note-taking behaviors or examine its possible effects on the emergence of topography-based responses. On another hand, the presentation of the keywords might have functioned as thematic probes (Skinner 1953), providing supplementary strength to the target responses. As suggested by Lovett et al. (2011), the presentation of verbal rules describing the relevant aspects of the stimuli increases the probability that the target response will occur.

Participants also provided new examples in Part C of the open-ended questions (i.e., creative topography-based responses). These findings add to others in the literature showing the emergence of topography-based responses after selection-based training (Lovett et al. 2011; Walker et al. 2010). The fact that participants provided new examples for all research designs suggests that this method of instruction may result in the emergence of behaviors that usually are desired for college students: the participants were able to write about what was directly taught; and they were able to generate examples that could be applied to professional situations they might encounter. In this study, although the generated examples had an autoclitic frame similar to the one that had been taught in the MTS tasks and prompted by the key words, the examples' themes varied from participant to participant (Bandini and de Rose 2006; Matos and Passos 2010; Skinner 1957/2002).

It should be noted that in addition to comparing this form of instruction delivery to other forms, investigating the reliability of readily available programs, and controlling note-taking behavior, future studies could investigate other questions, such as whether participants would maintain both MTS and topography-based performances a month or more after the procedure was concluded, as discussed by Walker et al. (2010) and assessed by Walker and Rehfeldt (2012). Additionally, the present study did not test for potential class merger between all nine research design names, definitions, notations, and examples. Future studies could evaluate whether all nine stimuli of each set became part of the same equivalence class and, at the same time, they should evaluate whether—given the nine stimulus names, definitions, notations, and examples—participants can discriminate among pre-, quasi-, or true-experimental designs.

Many studies have been investigating the use of equivalence-based instruction with advanced learners. Despite their advances, the studies identified several questions that still need to be addressed if equivalence-based instructions are to be used successfully in the context of higher education, both in classroom and online courses. For example, Fienup and Critchfield (2010), Walker et al. (2010), Walker and Rehfeldt (2012), among others, have emphasized that equivalence protocols still need to be systematically compared to other educational protocols. Additionally, studies such as Walker et al. (2010) noted that participants' textual repertoire in these studies was probably very sophisticated, so similar outcomes would probably be demonstrated only in participants with sufficient textual proficiency, thus, pointing to another question on which prerequisite skills might be necessary for equivalence-based instruction success. The

application possibilities for equivalence-based protocols are infinite, especially when instruction time is very limited and learners are recruited from various regions of the world. Connecting new hardware and software technologies with equivalence-based instruction might be a way to assure that more people know and make use of an instructional technology that has been shown to work in many settings, with different populations, to teach different subject matters. In addition, comparing different technologies and educational protocols can help in the choice-making process: data can guide nonexperts on which technologies and educational protocols might be more efficient for different contexts and learners.

In a world in which different hardware and software technologies change quickly, and a myriad of educational protocols are available, research is important to provide scientific evidence to help people determine best practices for different educational settings.

Appendix A

	Design name (Set A)	Design explanation or definition (Set B)	Notation (Set C)	Example (Set D)
1	The One-Shot Case Study	The dependent variable is measured (O) only after the independent variable (X) is introduced. The independent variable is introduced before the measure of the dependent variable. Only one group is needed and there is no control group. There is no randomization.	X O1	The dependent variable is heart rate and it is measured after the independent variable is introduced. The independent variable is jogging and it is introduced before measuring the heart rates. The participants are 25 students who will not be divided into groups. There is no randomization.
2	The One-Group Pretest–Posttest Design	The dependent variable is measured before (O1) and after (O2) the independent variable (X) is introduced. The independent variable is introduced after the first measure of the dependent variable. Only one group is needed and there is no control group. There is no randomization.	O1 X O2	The dependent variable is heart rate and it is measured before and after the independent variable is introduced. The independent variable is jogging and it is introduced after the first measure of heart rates. The participants are 25 students who will not be divided into groups. There is no randomization.
3	The Static Group Comparison	The dependent variable is measured (O) for both groups only after the independent variable (X) is introduced to the experimental group. The independent variable is introduced to the experimental group before the measure of the dependent variable. Two groups are needed; one is the control group. There is no randomization.	X O1 __ __ __ __ __ O1	The dependent variable is heart rate and it is measured for both groups after the independent variable is introduced to the experimental group. The independent variable is jogging and it is introduced to the experimental group before heart rates are measured in both groups. The participants are 50 students who will be divided into two groups: experimental group and control group. There is no randomization.
4	Nonequivalent Control-Group Design	The dependent variable is measured for both groups before (O1) and after (O2) the independent variable (X) is introduced to the experimental group. The independent variable is introduced to the experimental	O1 __ X __ O2 O1 O2	The dependent variable is heart rate and it is measured for both groups before and after the independent variable is introduced. The independent variable is jogging and it is introduced to the experimental group after the first

	Design name (Set A)	Design explanation or definition (Set B)	Notation (Set C)				Example (Set D)
		group after the first measure of the dependent variable. Two groups are needed; one is the control group. There is no randomization.					measure of heart rates. The participants are 50 students who will be divided into two groups: experimental group and control group. There is no randomization.
5	Counterbalanced Design	The dependent variable is measured for all groups (O), after each one of the four independent variables (X1, X2, X3, X4) is introduced for each experimental group. Each independent variable is introduced to all groups but in a different order for each group. Four groups are needed, but there is no "true control group", since the independent variable is introduced for all groups. There is no randomization.	Time 1 Time 2 Time 3 Time 4 Group A X1O X2O X3O X4O Group B X2O X4O X1O X3O Group C X3O X1O X4O X2O Group D X4O X3O X2O X1O				The dependent variable is heart rate and it is measured for all groups, after each independent variable is introduced for the groups. The independent variables can be jogging (X1), swimming (X2), dancing (X3), and walking (X4) and each one of them is introduced to all groups, but in a different order for each group. The participants are one hundred students who will be divided into four groups. There is no randomization.
6	The Multiple Time Series Design	The dependent variable is measured several times (O), for both groups, before and after the independent variable (X) is introduced to the experimental group. The independent variable is introduced to the experimental group after several measures of the dependent variable. Two groups are needed; one is the control group. There is no randomization.	O O O OXO O O O O O O O O O O O				The dependent variable is heart rate and it is measured several times, for both groups, before and after the independent variable is introduced. The independent variable is jogging and it is introduced to the experimental group after several measures of heart rates. The participants are 50 students who will be divided into two groups: experimental group and control group. There is no randomization.
7	The Pretest-Posttest Control Group Design	The dependent variable is measured for both groups before (O1) and after (O2) the independent variable (X) is introduced to the experimental group. The independent variable is introduced to the experimental group after the first measure of the dependent variable. Two groups are needed; one is the control group. There is randomization.	R O1 X O2 R O1 O2				The dependent variable is heart rate and it is measured for both groups before and after the independent variable is introduced. The independent variable is jogging and it is introduced to the experimental group after the first measure of heart rates. The participants are 50 students who will be randomly assigned to either one of two groups: experimental group and control group. There is randomization.
8	The Solomon Four-Group Design	The dependent variable is measured for two groups before (O1 and O3) and after (O2 and O4) the independent variable (X) is introduced to the experimental groups. For the other two groups, the dependent variable is measured only after (O5 and O6) the independent variable is introduced to the experimental groups. The independent variable is introduced to the two experimental groups. For the first experimental group (Group A), the independent variable is introduced after the first measure of the dependent variable;	Group A R O1 X O2 Group B R O1 O2 Group C R X O2 Group D R O2				The dependent variable is heart rate, it is measured before and after the independent variable is introduced for two of the four groups; it is measured only after in the other two groups. The independent variable is jogging and it is introduced to the experimental groups differently: for Group A it is introduced after the dependent variable is measured. For Group C, it is presented before the dependent variable is measured. The participants are 100 students who will be randomly assigned to either one of four groups:

Design name (Set A)	Design explanation or definition (Set B)	Notation (Set C)	Example (Set D)
	for the other experimental group (Group C), it is introduced before. Four groups are needed; two are control groups. There is randomization.		experimental group A or C, control group B or D. There is randomization.
9 The Posttest-Only Control Group Design	The dependent variable is measured for both groups only after (O) the independent variable (X) is introduced to the experimental group. The independent variable is introduced to the experimental group before the measure of the dependent variable. Two groups are needed; one is the control group. There is randomization.	R X O1 R O1	The dependent variable is heart rate and it is measured for both groups after the independent variable is introduced to the experimental group. The independent variable is jogging and it is introduced to the experimental group before heart rates (dependent variable) are measured in both groups. The participants are 50 students who will be randomly assigned to either one of two groups: experimental group and control group. There is randomization.

References

Bandini, C., & de Rose, J. (2006). Os processos [The processes]. In *A abordagem behaviorista do comportamento novo* [*The behavioral approach to new behavior*] (pp. 63–84). Santo Andre, SP: ESETec.

Barlow, D., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior for change* (3rd ed.). Boston, MA: Pearson Education.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Cozby, P. (2008). *Methods in behavioral research*. New York, NY: McGraw-Hill.

Critchfield, T., & Fienup, D. (2010). Using stimulus equivalence technology to teach statistical inference in a group setting. *Journal of Applied Behavior Analysis, 43*(4), 763–768.

de Rose, J., de Souza, D., & Hanna, E. (1996). Teaching reading and spelling: Exclusion and stimulus equivalence. *Journal of Appled Behavior Analysis, 29*(4), 451–469.

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*(3), 575–593. doi:10.1901/jaba.2009.42-575.

Fienup, D., & Critchfield, T. (2010). Efficiently establishing concepts of inferential statistics and hypothesis decision making through contextually controlled equivalence classes. *Journal of Applied Behavior Analysis, 43*(3), 437–462. doi:10.1901/jaba.2010.43-437.

Fienup, D., & Critchfield, T. (2011). Transportability of equivalence-based programmed instruction: Efficacy and efficiency in a college classroom. *Journal of Applied Behavior Analysis, 44*(3), 435–450. doi:10.1901/jaba.2011.44-435.

Fienup, D., Covey, D., & Critchfield, T. (2010). Teaching brain-behavior relations economically with stimulus equivalence technology. *Journal of Applied Behavior Analysis, 43*(1), 19–33. doi:10.1901/jaba.2010.43-19.

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

Horner, R., & Baer, D. (1978). Multiple-probe technique: A variation on the multiple baseline. *Journal of Applied Behavior Analysis, 11*(1), 189–196.

Johnson, K., & Street, E. (2004). *The Morningside model of generative instruction: What it means to leave no child behind*. Beverly, MA: Cambridge Center for Behavioral Studies.

Loft, K. (2009). *The virtues of venom*. Retrieved from http://fcit.usf.edu/fcat10r/home/sample-tests/virtues-of-venom/index.html.

Lovett, S., Rehfeldt, R., Garcia, Y., & Dunning, J. (2011). Comparison of a stimulus equivalence protocol and traditional lecture for teaching single-subject designs. *Journal of Appled Behavior Analysis, 44*(4), 819–833. doi:10.1901/jaba.2011.44-819.

Matos, M., & Passos, M. (2010). Emergent verbal behavior and analogy: Skinnerian and linguistic approaches. *The Behavior Analyst, 33*(1), 65–81.

Neef, N., Perrin, C., Haberlin, A., & Rodrigues, L. (2011). Studying as fun and games: Effects on college students' quiz performance. *Journal of Applied Behavior Analysis, 44*(4), 897–901. doi:10.1901/jaba.2011.44-897.

Ninness, C., Rumph, R., McCuller, G., Vasquez, E., Harrison, C. A. F., & Bradfield, A. (2005). A relational frame and artificial neural network approach to computer-interactive mathematics. *The Psychological Record, 55*(1), 135–153.

Ninness, C., Barnes-Holmes, D., Rumph, R., McCuller, G., Ford, A. M., Payne, R., & Elliott, M. P. (2006). Transformations of mathematical and stimulus functions. *Journal of Applied Behavior Analysis, 39*(3), 299–321.

Ninness, C., Dixon, M., Barnes-Holmes, D., Rehfeldt, R. A., Rumph, R., McCuller, G., & McGinty, J. (2009). Constructing and deriving reciprocal trigonometric relations: A functional analytic approach. *Journal of Applied Behavior Analysis, 42*(2), 191–208. doi:10.1901/jaba.2009.42-191.

Odom, S., Brantlinger, E., Gersten, R., Horner, R., Thompson, B., & Harris, K. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*(2), 137–148.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research, 14*(1), 5–13.

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: The MacMillan Company.

Skinner, B. F. (2002). *Verbal behavior*. Cambridge, MA: B. F. Skinner Foundation. Original work published 1957.

Sota, M., Leon, M., & Layng, T. (2011). Thinking through text compre-
    hension II: Analysis of verbal and investigative repertoires. *The
    Behavior Analyst Today, 12*(1), 12–20.

Tiemann, P., & Markle, S. (1990). *Analyzing instructional content* (4th
    ed.). Champaign, IL: Stipes.

Twyman, J., Layng, T., Stikeleather, G., & Hobbins, K. (2005). A
    non-linear approach to curriculum design: The role of behav-
    ior analysis in building an effective reading program. In W.
    L. Heward et al. (Eds.), *Focus on behavior analysis in

education* (Vol. 3, pp. 55–68). Upper Saddle River, NJ:
    Merrill/Prentice Hall.

Walker, B., & Rehfeldt, R. (2012). An evaluation of the stimulus equiv-
    alence paradigm to teach single-subject design to distance education
    students via Blackboard. *Journal of Applied Behavior Analysis,
    45*(2), 329–344. doi:10.1901/jaba.2012.45-329.

Walker, B., Rehfeldt, R., & Ninness, C. (2010). Using the stimulus equiv-
    alence paradigm to teach course material in an undergraduate rehabil-
    itation course. *Journal of Applied Behavior Analysis, 43*(4), 615.