perspectivas em análise do comportamento
ISSN 2177-3548
instituto par ciências do comportamento
www.revistaperspectivas.org

Teaching Academic Concepts: The Effects of Equivalence-Based Instruction

Hanne Augland[1], Torunn Lian[1], Erik Arntzen[1]

[1] Oslo Metropolitan University – Oslo/Norway | **Título abreviado**: Fusão Cognitiva em Cuidadores: Revisão de Escopo | **Endereço para correspondência:** OsloMet – Oslo Metropolitan University, Department of Behavioural Science, P.O. Box 4 St., Olavs plass, N-0130 Oslo, Norway | **Email**: augland.hanne@gmail.com | **doi**: 10.18761/PAC02016an115025

Abstract: The present study examined the effect of equivalence-based instruction (EBI) on academic concept formation. Twenty-six college students experienced matching-to-sample (MTS) training to establish five measurable dimensions of behavior with four members each. We arranged a within-subject pre- and post-design. In addition, half of the participants experienced pre-training for two concepts, while the other half did not. The participants underwent training in a one-to-many training structure of five 4-member classes. All participants formed equivalence classes in the MTS format. Furthermore, all participants showed improved performance on fill-in and multiple-choice post-tests. The results indicate that EBI can promote generative outcomes in participants. However, pre-training some relations for half of the participants did not enhance the equivalence formation. The results in the study replicate earlier studies in that EBI effectively establishes academic concepts and facilitates generative outcomes. Additionally, it adds to the existing literature by repeatedly testing each potential emergent relation to determine whether participants formed equivalence classes.

Keywords: stimulus equivalence, university students, academic concepts, equivalence-based instruction, generative outcome

Resumo: O presente estudo examinou o efeito da instrução baseada em equivalência **(EBI)** na formaçãode conceitos acadêmicos. Vinte e seis estudantes universitários participaram de um treinamento de matching-to-sample (MTS) para estabelecer cinco dimensões mensuráveis do comportamento, cada uma composta por quatro membros. Foi empregado um delineamento intrassujeitos com pré- e pós-teste. Além disso, metade dos participantes passou por um pré-treinamento para dois conceitos, enquanto a outra metade não recebeu esse pré-treinamento. Os participantes foram submetidos a um treinamento em uma estrutura de ensino do tipo um-para-muitos, envolvendo cinco classes com quatro membros cada. Todos os participantes formaram classes de equivalência no formato MTS. Ademais, todos apresentaram melhora no desempenho em pós-testes do tipo preenchimento de lacunas e de múltipla escolha. Os resultados indicam que a EBI pode promover desempenhos generativos nos participantes. Entretanto, o pré-treinamento de algumas relações para metade dos participantes não aumentou a formação de equivalência. Os resultados do estudo replicam achados anteriores ao demonstrar que a EBI estabelece de forma eficaz conceitos acadêmicos e facilita desempenhos generativos. Adicionalmente, o estudo contribui para a literatura existente ao testar repetidamente cada relação potencialmente emergente para determinar se os participantes formaram classes de equivalência.

Palavras-chave: equivalência de estímulos, estudantes universitários, conceitos acadêmicos, instrução baseada em equivalência, desempenho generativo

The stimulus equivalence paradigm describes how arbitrarily related stimuli can acquire the same behavioral function. The defining features of stimulus equivalence; reflexivity, symmetry, and transitivity should emerge without direct training (Sidman & Tailby, 1982). After establishing at least two overlapping conditional discriminations, probes for emergent relations are presented. Hence, if selecting B1 in the presence of A1, and selecting C1 in the presence of B1 is directly trained, then selecting A1 in the presence of B1 would be an example of symmetry. Selecting C1 in the presence of A1 would be an example of transitivity, and finally, selecting A1 in the presence of C1 would be an example of combined symmetry and transitivity, sometimes referred to as equivalence. The last property, reflexivity, is demonstrated if A1 is selected in the presence of itself. The reflexivity relations are typically not tested for, as they are regarded as a prerequisite (Sidman, 1994).

Even though research on stimulus equivalence originated in an applied setting, focused on establishing socially relevant behaviors (Sidman, 1971), there was a quick a shift towards basic research. Nevertheless, recent years have witnessed an increase in research focused on socially significant behaviors, utilizing technology grounded in the knowledge from stimulus equivalence. Stimulus equivalence technology is now frequently referred to as equivalence-based instruction (EBI). Over the past two decades, the literature on EBI has expanded significantly, addressing a wide range of skills across diverse populations (see Brodsky & Fienup, 2018; Pilgrim, 2020; or Rehfeldt, 2011 for reviews). Much of the interest in this teaching format is linked to the basic premise of EBI: that by teaching a few relations in a class, new and untrained relations will emerge (Sidman, 1994).

Most EBI studies have employed a matching-to-sample (MTS) procedure to establish the relevant baseline relations and subsequently tested for emergent relations in the same format. If a participant demonstrates all emergent relations following MTS training, it is typically concluded that an equivalence class—or, in academic contexts, a concept—has been formed. However, Albright et al. (2015) pointed out that MTS performance is insufficient to infer a functional repertoire. Thus, if the participants could apply the repertoire in new and different situations, it would strengthen the functionality of the repertoire. Therefore, assessing whether MTS training and test performance can enhance outcomes beyond the MTS format, would increase confidence in EBI as an effective instructional format. This seems especially important when teaching academically relevant concepts. Consequently, some researchers have included measurements of what has typically been described as generalization assessments. However, in instances where both stimuli and response modalities vary (e.g., when participants initially engage in MTS training and are then required to type in their responses), the underlying mechanisms involved become less clear. Therefore, we use generative outcome as a general term to make a distinction from primary stimulus generalization, and to emphasize that the assessment involves testing some variation of the experiences encountered during reinforcement.

Examples of generative outcome would be multiple-choice and tact performance. Performance on multiple-choice is often referred to as selection-based responding, whereas performance on tact probes is referred to as topography-based responding (Michael, 1985). Hereon, the terms selection-based and topography-based responding will be used. Some studies have solely assessed selection-based behavior as the generative outcome. For example, Critchfield and Fienup (2010) showed an increase in scores on multiple-choice questions after establishing two statistical terms through MTS training for all participants. Critchfield (2014) extended these results by reusing most of the multiple-choice questions from Critchfield and Fienup (2010) when establishing conditional discriminations involving two potential class members. Instead of the traditional MTS procedure, the participants experienced online note cards containing short descriptions followed by a multiple-choice test. The generative outcome for the 60 students participating in the study improved on the 32 multiple-choice questions. Furthermore, the participants in study by Albright et al. (2015) underwent training and testing involving MTS procedures that focused on concepts related to statistical variability. Following this training, they demonstrated improved performance on a paper-

and-pencil multiple-choice test that included textual definitions similar to, but not identical to, the stimuli used during training. Together, the results in these studies indicate that generative outcomes in the form of multiple-choice generally improve as a function of MTS training.

Even though the MTS format is valuable in demonstrating the efficacy of EBI in producing generative outcomes in a multiple-choice format, it still represents selection-based responding. College students are also expected to talk and write accurately about academic concepts. Thus, some studies have assessed generative outcomes in the form of speaking and writing performance following MTS training and test. Albright et al. (2016) employed EBI to teach behavioral functions to college students. In addition to the MTS format, they assessed performance on multiple-choice tests that included variations of stimuli from training, as well as oral tests that involved naming the correct operant function based on a description, a graph, or a vignette. The results indicated that EBI improved performance on both multiple-choice and oral tests, with participants performing better on selection-based (multiple-choice) than on topography-based (oral tests) questions. Along the same lines, Reyes-Giordano and Fienup (2015) arranged both selection- and topography-based tests in their study. Both tact and intraverbal behavior were assessed as part of topography-based responding. The results aligned with Albright et al. (2016) reporting improved performance on generative tests.

Lovett et al. (2011) also demonstrated generative outcome following selection-based training. The participants first established different single-subject designs as conditional discriminations and were then asked to tact the vignettes and graphs experienced in training. In addition, they assessed tact relations for untrained vignettes and graphs. The results on the tact test (topography-based) varied across the tested relations. Three out of four participants reached high levels of accurate tacting for trained and novel graphs. Results also showed tact of vignettes, though on lower levels. However, in contrast to the abovementioned studies, the participants performed better on topography-based testing than on selection-based testing. It is worth noting that 11 participants were tested only with multiple-choice testing, while four additional participants completed tact probes before the multiple-choice test.

When testing only a selection of the participants, it is hard to conclude on the potential difference between modalities. It is necessary to test both modalities to see how the different operants may vary within the same participants. It is possible that the higher outcome for topography-based responding in Lovett et al. (2011) was influenced by a low number of participants, and by the fact that the different modalities were tested across different participants. An important next step would be to assess all participants across all modalities, enabling the drawing of meaningful conclusions regarding the differences among the modalities. Nevertheless, the abovementioned studies altogether show that EBI improves accuracy outside the MTS format. These findings are essential in applied settings. However, different studies did find various outcomes regarding selection- and topography-based generative outcomes.

Even though a few studies have investigated the effect of EBI on both topography- and selection-based repertoires, more research seems necessary to get a broader understanding of the phenomenon. Fienup et al. (2010) highlighted the necessity to further assess topography-based behavior after selection-based training, as the standard in teaching is to produce topography-based responding. Hence, it is important to evaluate potential topography- and selection-based repertoires as generative outcomes for participants who form equivalence classes.

Moreover, to our knowledge, just a few studies have implemented EBI as part of a university course (Augland et al., 2020; Critchfield, 2014; Wiskow et al., 2024). Wiskow et al. (2024) and Critchfield (2014) embedded the EBI training and testing online, and found an overall increase at post-test. Further, Augland et al. (2020) found a profound effect of EBI conducted in class, compared to another student active learning format, when comparing time spent in learning. Even though these studies are an important contribution to the field, more data on EBI in applied settings seems important.

The final aspect that has influenced the design of the study is that, the course instructors reported that exam results from previous years suggest that distinguishing between *count* and *frequency* is more

challenging than distinguishing between the other measurable dimensions. Based on this information, we incorporated an exploratory component into the study. We aimed to explore whether training the name of the measurement units; frequency and count to its corresponding descriptions, prior to introducing the relations in the remaining classes would impact students' performance.

To sum up, the purpose of the present experiment was to evaluate the effects of EBI integrated in a bachelor's course in behavior analysis on (a) equivalence class formation in MTS format and (b) students' generative outcomes after EBI measured as performance on fill-in and multiple-choice tasks. In addition, we wanted to explore the potential effects of providing pre-training with AB relations for two of the measurable dimensions.

Method

Participants

The experimenter (first author) recruited 27 university students between the ages of 19 and 49 for the study. One participant withdrew consent resulting in 26 participants, 20 women and 6 men. They were all enrolled in a bachelor's program in behavior analysis in their second year and attended their first methods course. They were not exposed to the relevant academic concepts earlier in their bachelor's program but might have been exposed to broader definitions earlier on, in other contexts.

Participants received an informed consent form that included information about the experiment, their right to withdraw without any negative consequences, and that data would be anonymized. At onset of the experiment, the experimenter emphasized that the trained academic concepts would be of interest for the upcoming exam, that participation in the study was voluntary, and that they could withdraw at any time without any consequences. At the onset of the training, participants were instructed not to talk to one another and to turn off their cellular phones. Participants were all debriefed and offered individual feedback on their results once the data analysis was ready.

Setting and Material

The present study was approved by the Norwegian Center of Research Data (reference no. 527551). The course instructor agreed to include MTS training and testing as part of the course, and participants actively gave consent to let the data be included in the study.

The classroom was equipped with rows of stationary computers, and the computer cabinets were placed between the screens to prevent participants from viewing each other. Participants were seated 0.5–1 meter apart. The experimenter was placed in front of the classroom, and two research assistants were in the back to be able to monitor the participants.

We employed HP 70032 stationary computers running Microsoft Windows 10 in the MTS training and testing. The computers had a 24-inch screen and an external mouse connected through a USB port. Custom-made software controlled stimulus presentation throughout the training of conditional discriminations. In addition, the program automatically registered and saved all critical information; trial type, correct/incorrect comparison choices, and programmed consequences.

Design

We arranged a mixed group and within-subject pre- and post-test design to evaluate the effects of EBI on equivalence class formation and generative outcome. One group received pre-training of two AB relations of two classes, while the other received training on all AB relations (No pre-training group). To assess the potential impact of pre-training, participants were randomly assigned to one of two conditions by pulling a number from an envelope. Furthermore, to balance the number of trials, participants in the No-pretraining group experienced 25 trials consisting of all name definition-relations (AB-relations).

Dependent Measures

Yields, the number of participants forming equivalence classes was used as the dependent measure for equivalence class formation. Furthermore, percentage correct on the pre-and post-test scores was used to evaluate the effects of EBI on the generative outcome. The participants were given 1 point for

every correct fill-in and multiple-choice question, and the percentages were calculated by dividing the number correct by the maximum number of points and multiplying by 100. In addition, these scores were split into selection-based (multiple-choice) and topography-based (fill-in questions) to reveal potential differences. Finally, as an additional measure, social validity on a 7-point Likert scale was scored to assess participants' confidence in their knowledge and time spent in the training.

Procedure

Tests for Generative Outcome

Since participants did not have any knowledge about dimensions of behavior, they were presented with the following text in the online program before the multiple-choice questions and fill-ins were presented: "You can measure observable behaviors. Today, you will learn how behaviors can be measured and when to use the different methods. All questions are relevant for the upcoming exam."

Figure 1 shows an overview of experimental phases. The pre-and post-tests for generative outcomes were designed in an online program, and was conducted directly followed by the MTS test, ensuring that participants could not communicate with each other prior to the test. The same questions were included in both the pre- and post-tests, with all fill-in questions presented before the multiple-choice questions. The participants saw one question at a time and had to click "next" on the screen to see the next question. They could move back and forth between questions on the different pages. The pre- and post-tests consisted of 20 questions testing BA, CA, and DA relations. Since the experiment was conducted as part of the ordinary college course, only symmetry relations were assessed to prevent the experiment from being too lengthy. Four questions were targeted toward each measurable dimension in the conditional discriminations. Of these, two questions were fill-in questions, while two others were multiple-choice questions that included all five dimensions as response alternatives. Examples of multiple-choice questions and fill-ins are given in Appendix A. Typically, a vignette would be followed by a question asking which dimension would be suitable for the given case. In addition to this, participants were asked to fill in the correct dimension for different vignettes, or they were presented with a target behavior and required to identify the appropriate measurable dimension.

Interobserver Agreement

The answers by the participants were automatically registered and downloaded after the experiment was completed to calculate inter-observer agreement (IOA). The first and second authors conducted a trial-by-trial IOA to check for agreement on the participants' given answers to the multiple-choice and fill-in questions. This was calculated by dividing the number of agreed trials by the total number. The interobserver agreement was 98.7% for the pre-test and 98.8% for the post-test.

Equivalence-Based Instruction

Stimuli. Table 1 provides an overview of the stimuli. The stimuli were arranged as five potential classes, corresponding to the concepts of count, frequency, latency, inter-response time, and duration. The table is designed to present the concepts in columns 2–6, and the rows show the different members in each potential class. The different rows correspond with the name of the concepts to be formed (A-stimuli), a description of the measurement (B-stimuli), written vignettes (C-stimuli), and a paraphrase of the definition, including conditions determining when each measure is appropriate (D-stimuli). B- and D-stimuli were both written as definitions in different wording to ensure a more flexible stimulus control and thereby promote generative outcomes. The definitions are in accordance with Cooper et al. (2007) in the course curriculum.

All stimuli were presented on the screen in black letters on a white background. All stimuli had a click-sensitive area of approximately 4 cm in width and 4 cm in height.

Instructions. At the onset of the conditional discrimination training, a Norwegian version of the following instructions appeared on the screen:

When the experiment begins, a word or a sentence will appear on the screen. You should click on it. Alternative answers will then appear on the

Table 1. Stimulus Set

	1	2	3	4	5
A	Count	Frequency	Latency	Inter-response time	Duration
B	How often a response occurs	How often a response occurs per unit of time	The time from a stimulus is presented until a response is emitted	The time from a response ends until a new one begins	How long does a response last
C	Lars jumps three times	Lars jumps three times in one minute	From Ida says "jump" to Lars jumps it takes three seconds	Three seconds elapse from Lars jumps to the next time he jumps	Lars jumps a series of jumps that last for three seconds
D	Used when the number of occurrences, regardless of time, is of interest	Used when the number during a certain period of time is of interest	Used when the time from discriminative stimulus to response is of interest	Used when the time between responses is of interest	Used when one is interested in how long a response goes on

Note. The figure displays stimuli used in conditional discrimination training. 1-5 represents the to-be-formed classes, and A-D represent the members.

screen. You are to choose one of them. For each trial the software will inform whether your choice was correct or incorrect. You will no longer receive feedback for every trial when all tasks are learned. However, it is still possible to get everything correct based on what you have already learned. Do your best to get everything correct. Good luck! Press Start to begin the experiment.

Phase 1. Table 2 shows the order of introduction of baseline trials and the relations tested for. Participants experienced either conditional discrimination training on count and frequency (Pre-training group) or between all classes (No pre-training). The baseline relations were presented in a One-to-Many (OTM) training structure where A-stimuli served as a sample. The following presents the different trial types as letter and number combinations. The first letter-number combination refers to the sample stimulus and is separated from the five comparison stimuli by a slash. The correct comparison stimulus is underlined. For the participants who did not experience pre-training, the subsequent trials were presented: A1/B1B2B3B4B5, A2/B1B2B3B4B5, A3/B1B2B3B4B5, A4/B1B2B3B4B5, and A5/B1B2B3B4B5). For the participants who experienced pretraining, A1/B1B2 and A2/B1B2 were presented. The participants were trained in discrimination between count and frequency; each trial type was presented five times in a block, and the mastery criterion was set to 98% correct. To control for the minimum number of times, the participants were exposed to each trial type, participants in the No pre-training group experienced 25 trials with feedback on correct and incorrect trials, in this phase regardless of performance.

Phase 2. Training after Phase 1 was the same for both groups. We arranged a simultaneous training and test protocol in which all baseline relations were trained and mastered according to the preset criterion before testing for emergent relations. Fifteen baseline conditional discriminations were established in an OTM training structure. The baseline trials were introduced in a serialized manner, implying that all AB-relations and AC-relations were introduced separately and mastered to criterion before the relations were presented in a mixed training block. Finally, the AD relations were introduced separately and mastered before participants experienced a mix of AB, AC, and AD relations. Each trial started with presenting a sample stimulus in the middle of the screen. A click on the sample stimulus was immediately followed by five comparison stimuli in a circle around the sample stimulus on the screen. The location of the comparison stimuli alternated across trials. Following a comparison choice, a programmed consequence in blue letters on a white background was presented in the middle of the screen for 500ms. Correct comparison choices were followed by the Norwegian words for "good," "perfect," "correct," and the like in

Table 2. Overview of Experimental Phases

Experimental Phases	Trial Types	Min. Trials	% Criterion
Pre-training/No pre-training	A1B1 and A2B2/A1B1, A2B2, A3B3, A4B4 and A5B5	10/25	98/none
AB relations	A1B1, A2B2, A3B3, A4B4 and A5B5	15	>98
AC relations	A1C1, A2C2, A3C3, A4C4, and A5C5	15	>98
Mixed Phase	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, and A5C5	30	>98
AD relations	A1D1, A2D2, A3D3, A4D4, and A5D5	15	>98
Mixed Phase	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, A5C5, A1D1, A2D2, A3D3, A4D4, and A5D5	45	>98
Thinning of Programmed Consequences			
75% probability	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, A5C5, A1D1, A2D2, A3D3, A4D4, and A5D5	45	>98
50% probability	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, A5C5, A1D1, A2D2, A3D3, A4D4, and A5D5	45	>98
0% probability	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, A5C5, A1D1, A2D2, A3D3, A4D4, and A5D5	45	>98
Test for Emergent Relations			
Directly Trained	A1B1, A2B2, A3B3, A4B4, A5B5, A1C1, A2C2, A3C3, A4C4, A5C5, A1D1, A2D2, A3D3, A4D4, and A5D5	(45)	>90
Symmetry Trials	B1A1, B2A2, B3A3, B4A4, B5A5, C1A1, C2A2, C3A3, C4A4, C5A5, D1A1, D2A2, D3A3, D4A4, and D5A5	(45)	>90
Equivalence Trials	B1C1, B2C2, B3C3, B4C4, B5C5, B1D1, B2D2, B3D3, B4D4, B5D5, C1B1, C2B2, C3B3, C4B4, C5B5, C1D1, C2D2, C3D3, C4D4, C5D5, D1B1, D2B2, D3B3, D4B4, D5B5, D1C1, D2C2, D4C4, and D5C5	(90)	>90

Note. The table displays the different phases of training and testing. Trial types show the trial types presented in each phase, with the first combination of letter and number indicating the sample stimulus and the combination following the slash indicating the correct comparison. Min. Trials indicate the minimum number of trials given of the procedure. Parentheses indicate that the number of trials is set and not a minimum.

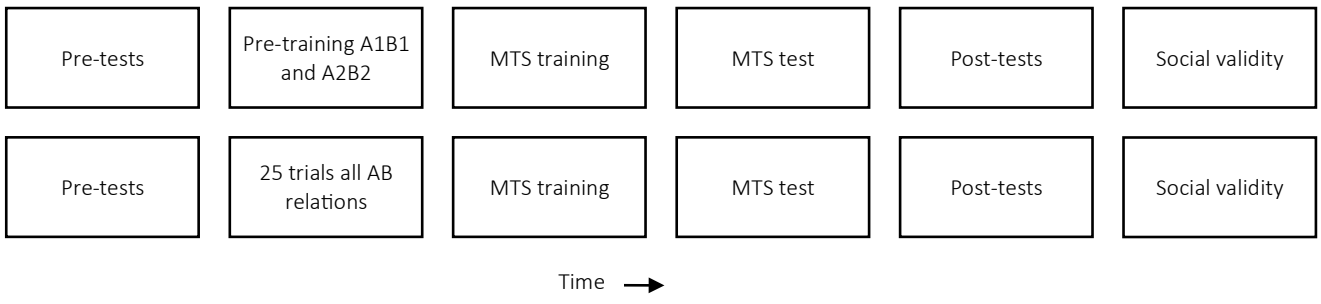

Figure 1. Overview of Experimental Phases

Note. The figure provides an overview of the different experimental phases. The top row displays phases for the Pre-training group, while the bottom row shows phases to the No pre-training group.

the middle of the screen. Incorrect responses were followed by the Norwegian word for "wrong," presented in the middle of the screen. A 500ms inter-trial interval followed each trial. The different trial types were presented in random order four times in each training block. The criterion for proceeding to the next training block was a 98% correct comparison choice. Due to a program error, P17976 (No pre-training) had a criterion of 90% correct throughout training.

Thinning of Programmed Consequences. The participants then experienced training blocks in which programmed consequences were gradually thinned from 100% to 0% before test for emergent relations. The steps were 75%, 50%, and 0% probability of programmed consequences per block. The consequences were controlled by the program.

Test for Emergent Relations. The test for stimulus equivalence presented all possible emergent relations, and probes for maintenance of baseline relations, symmetry, and equivalence trials were tested for in a mixed block. Each trial type was presented in a random order three times, constituting 180 trials. No programmed consequences were presented during the test. The criterion to conclude that equivalence classes were formed was a minimum of 90% correct comparison choices for baseline conditional discriminations, symmetry, and equivalence, respectively.

Social Validity. The social validity questionnaire was presented as a continuation of the post-test and consisted of four questions. Participants were asked to rate each question on a 7-point Likert scale. The questionnaire is shown in Appendix B. Each question is listed along with the scale. Higher scores indicate a more positive evaluation of the EBI. The questionnaire also included questions to assess participants' confidence in their knowledge and time commitment.

Results

Equivalence Class Formation

Table 3 shows the individual results for stimulus equivalence. The upper panel shows participants in the Pre-training group, while the bottom panel shows the results for the participants who did not experience pre-training. The columns in Table 3 show the percentage performance for baseline, symmetry, and equivalence trials. Bold numbers indicate that the participant reached the equivalence criterion of 90% correct. Table 3 shows that all participants reached the equivalence criterion. Furthermore, two participants, one in each group, had one trial type with more than one incorrect response. P17954 in the Pre-training group had two incorrect responses on the D5C5 relation, while P17975 in the No pre-training group failed twice on the B1C1 relation.

Generative Outcome

Figure 2 shows the participants' mean percentage of correct responses for the generative outcomes, before and after EBI. All participants improved from pre- to post-test. The mean pre-test score is 29.6% (range 15%–60%), while the mean post-test score for all participants is 86.7% (range 55%–100%), with a mean increase of 57.1%. The circles indicate individual scores for pre-tests and the squares for post-tests. The participant with the highest improvement was P17966, who had a change in score from 20% in the pre-test, to 100% in the post-test. Conversely, P17961showed the lowest improvement from pre- to post-test at 15%.

Further, Figure 3 shows pre- and post-test scores split into dark grey bars indicating selection-based responding and light grey bars showing the mean topography-based score in percent correct. Before training, the participants' mean score for selection-based behavior was 48.8% (range 20%–70%). After EBI, the mean score increased by 31.2% to 80% (50%–100%). As mentioned, P17961 had the lowest overall change from pre- to post-test. This is reflected when looking at the selection-based behavior separately. This participant had lowered his score from pre- to post-test by 10%. However, the score for topography-based behavior was positive, with an improvement of 40%. Conversely, P17957, P17964, and P17966 improved their score for selec-

Table 3. Results of MTS Training and Test

		Percent Correct Test		
Participant #	**Trials Pre-training**	**BSL**	**SY**	**EQ**
Pre-training				
17966	10	**100**	**100**	**100**
17958	10	**100**	**100**	**100**
17964	10	**100**	**100**	**100**
17953	20	**100**	**100**	**99**
17965	10	**100**	**100**	**99**
17957	19	**100**	**100**	**99**
17963	20	**100**	**100**	**97**
17952	10	**100**	**100**	**96**
17961	10	**100**	**98**	**100**
17977	20	**100**	**98**	**100**
17960	20	**100**	**98**	**99**
17954	10	**100**	**96**	**96**
17956	10	**98**	**99**	**99**
17962	20	**96**	**98**	**96**
No pre-training				
17973	25	**100**	**100**	**100**
17955	25	**100**	**100**	**100**
17951	25	**100**	**100**	**100**
17969	25	**100**	**100**	**100**
17972	25	**100**	**100**	**99**
17971	25	**100**	**100**	**99**
17959	25	**100**	**100**	**99**
17974	25	**100**	**98**	**99**
17967	25	**100**	**98**	**93**
17975	25	**98**	**100**	**92**
17976	25	**96**	**100**	**98**
17970	25	**93**	**100**	**99**

Note. The first column shows participant numbers, and the second column displays the number of pre-training trials for the Pre-training group and the default set of trials for the No pre-training group. BSL denotes the correct percentage of baseline trials under test. SY and EQ denote the same for symmetry and equivalence trials. Bold numbers indicate that responding is in accordance with the stimulus equivalence index.

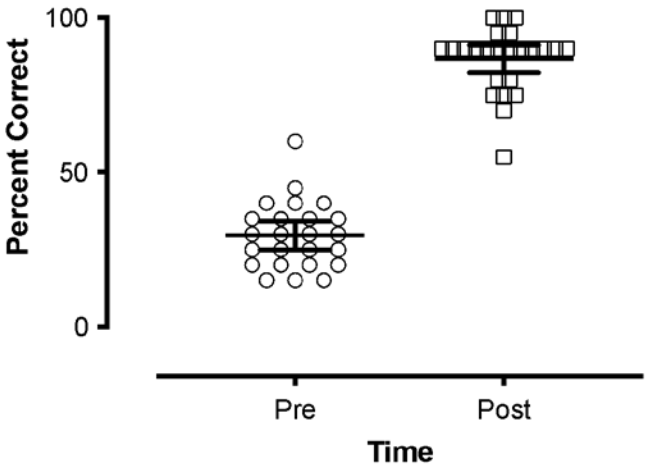

Figure 2. Pre and Post Generative Outcome

Note. The open circles indicate each participant score pre EBI, while the open squares indicate each participant post EBI. The mean pre- and post-scores, as well as the confidence interval (95%), are indicated by the horizontal lines.

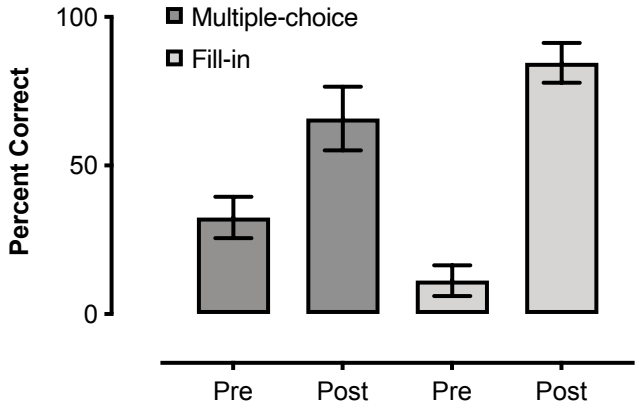

Figure 3. Results on Multiple-Choice and Fill-in.

Note. The figure shows pre- and post-scores split into dark grey bars indicating the mean correct multiple-choice questions and light grey bars showing the mean correct fill-in questions.

tion-based behavior by 60% from pre- to post-test. When analyzing data for all participants forming equivalence classes, the pre-score for topography-based behavior was lower than for selection-based behavior, with a mean of 10.4% (range 0%–60%). The increase of 82.9% led to a mean of 93.3% (range 60%–100%) correct on the post-test. All of the participants improved from pre- to post-test. Eight participants (no. 17956, 17962, 17964, 17965, 17966, 17969, 17970, and 17972) had no correct on the pre-test, and all responses were correct after formation of the equivalence classes.

Statistics

In addition, we performed paired-samples t-tests with bias-corrected and accelerated bootstrapping (BCa) to correct for any bias or skewness in the data (Field, 2024). The selection-based questions increased on average from the pretest (M = 48.57, SE = 2.88) to the posttest (M = 80.00, SE = 3.27). This difference, MD = 31.25, BCa 95% CI [23.22, 39,17] was significant, t(23) = 7.72, p <.001, and represented an effect of Hedges' g = 1.92, 95% CI [1.14, 2.69]. The topography-based questions increased on average from the pretest (M = 10.42, SE = 2.95) to the posttest (M = 93.33, SE = 2.11). This difference, MD = 82.92, BCa 95% CI [77.08, 89,17] was significant, t(23) = 23.46, p <.001, and represented an effect of Hedges' g = 6.13, 95% CI [4.18, 8.07].

Pre-training and Equivalence Class Formation

Six participants met the mastery criteria for A1B1 and A2B2 trials within 20 trials, while the rest mastered after the minimum number of trials. As shown in Table 3, 14 out of 14 participants responded in accordance with stimulus equivalence in the Pre-training group. Among participants in the No pre-training group, 12 out of 12 participants showed stimulus equivalence. Based on our criteria, eight of 14 participants discriminated between the classes after 10 trials, while the rest did so after 20 trials. All the participants in the No pre-training group underwent 25 extra trials, including all classes. Hence, we did not find any difference between the groups.

Social Validity

Social validity tests were completed immediately after the EBI. The participants' mean score for confidence in their knowledge was 4.1 (range: 1–6, mode: 4). The degree to which they would prefer to be taught using this instructional method was scored with a mean of 3.8 (range 1–7, mode: 4). Further, the participants scored how appropriate the time commitment for EBI as an instructional method in relation to the amount they felt they

had learned to a mean of 3.7 (range: 1–7, mode: 4). Lastly, the participants scored the length of the EBI training to a mean of 2,4 (range: 1–6, mode: 1).

Discussion

The present study aimed to investigate the effects of EBI on equivalence class formation as well as generative outcomes in multiple-choice and fill-in assessments. The results show that all participants formed equivalence classes, even subjected to stringent equivalence criterion, indicating class formation. These findings not only enhance our understanding of the effects of EBI in facilitating equivalence class formation but also support its potential implementation in traditional educational settings. Consequently, this study contributes to and expands the existing body of evidence regarding the effectiveness of EBI in higher education.

Additionally, after completing the MTS training and testing, the participants showed new topography-based responses by writing the names of various dependent measures (corresponding to symmetry relations in MTS probes) during a computer-administered test. Participants also showed improved performance on multiple-choice tests involving relations between novel vignettes descriptions and measurement names.

These results add to a growing evidence base showing that following selection-based training with overlapping conditional discriminations, (a) untrained stimulus-stimulus relations within the trained stimulus set readily emerge in MTS, (b) responding established and emerged in the MTS format can occur in the presence of novel variants of stimuli and in untrained modalities, (c) accurate selection-based behavior to novel variants of stimuli can occur without direct instruction. The generative traits of EBI are a valuable outcome and well justified in any educational setting, even with advanced learners. However, we failed in obtaining a differential outcome regarding the pre-training of baseline relations.

Equivalence Class Formation

All participants reached the equivalence criterion of 90% correct for the maintenance of baseline trials, symmetry, and equivalence relations, respectively. P17954 had two incorrect comparison choices on a total of five D5C5 probes, and P17975 erred twice on B1C1 probes. For these stimulus-stimulus relations, the participants showed only intermediate accuracy, and it is reasonable to question whether this is sufficient to conclude that the relations emerged. Furthermore, in an applied setting, the primary interest would be to establish robust repertoires of academically significant behaviors. If a professional provides incorrect answers when asked to name a phenomenon or identify the correct academic concept, it will likely raise questions about their competence and their ability to effectively address the task at hand. Thus, applied researchers should carefully consider the equivalence criterion, and supplementary instruction should be provided for stimulus-stimulus relations that show low or intermediate accuracy despite overall high accuracy. A strength of the current study is that each relation is tested three times in the MTS format to determine whether participants formed classes. This rigorous testing approach challenges the assumption that symmetry relations are intact if presumably more complex relations are demonstrated (e.g., Fienup et al., 2010; Sidman, 1994).

Generative Outcome

Aggregated results on generative outcomes showed an increase in correct responding from pre- to post-test. These results support previous studies (e.g., Albright et al., 2015; Lovett et al., 2011; Reyes-Giordano & Fienup, 2015) showing that visual-visual MTS procedures can be used to teach labeling of core academic concepts and to promote the emergence of topography- and selection-based behaviors. Overall, EBI seems to contribute to the understanding of basic academic concepts in broader contexts than the MTS format.

A more detailed analysis of generative outcome showed that the increase from pre- to post-test was higher for topography-based responding in the form of fill-in tasks than for selection-based responses in the multiple-choice format. Earlier studies have found different outcomes on selection- and topography-based generative outcomes. Some conclude on higher outcomes for topography-based responding (Lovett et al., 2011), while other find higher outcomes on selection-based responding

(Albright et al., 2016; Reyes-Giordano & Fienup, 2015). The results of the studies mentioned differ in terms of the generative outcome of topography- and selection-based behavior. Hence, it is uncertain how MTS training and testing affect generative outcomes in different modalities. Nevertheless, for students in higher education, the generative outcomes, particularly topography-based responding, holds significant importance. Consequently, it is vital for future experiments involving the implementation of EBI to identify the independent variables that influence the various modalities of generative outcomes. Several factors may contribute to the differences observed in selection- and topography-based responding, potentially related to pre-experimental history or the varying training and testing protocols employed. Additionally, it is essential to consider that the complexity of the stimuli used in EBI within higher education may impact the generalizability of the findings. Although the present experiment does not directly address these variables, its results contribute to the growing evidence base regarding EBI in higher education, highlighting the necessity for further research to bridge this knowledge gap.

Effects of Pre-training

The decision to investigate the effects of pre-training A1B1 and A2B2 relations was made based on reports from teachers indicating that students in previous years struggled to differentiate between frequency and count, and adjustments to the experiment were made shortly before the start date. The idea was that establishing name-description relations for count and frequency before introducing other relations could enhance discrimination between the two categories. This approach would reduce the likelihood of numerous errors during initial training, thereby increasing the probability of achieving the desired stimulus control.

Nonetheless, we were unable to differentiate the conditions based on pre-training with regard to the outcomes in class formation. There may be several reasons for this lack of success. First, although previous students have encountered difficulties in discriminating between count and frequency, this may or may not hold for participants in this study. Since participants were randomly assigned to various conditions, and no individual tailoring of stimuli was made, we cannot ascertain whether this was indeed the case for the present participants, or how it might have influenced the results. Furthermore, the balancing of number of trials might have contributed to no difference between the groups. Finally, we arranged an OTM training structure in the present experiment. The OTM structure is associated with a high equivalence yield (Arntzen, 2012) and may have masked the effects of pre-training.

In an academic context, students are often required to discriminate among and generalize within a range of stimulus classes. Although we failed in arranging conditions that could shed light on the effects of pre-training certain relations in the present study, we believe it is a research avenue worthy of further exploration. Individually tailored stimulus sets, and a linear series training structure to reduce the influence of ceiling effects should be prepared as part of such investigations.

Social Validity Questionnaire

The social validity questionnaire models questions asked in earlier studies (e.g., Lovett et al., 2011) and gives a valuable snapshot of how participants assess the procedures immediately upon completion, and results show a relatively low evaluation. However, it is essential to exercise caution when interpreting these findings. The EBI experiment was conducted in the very early phases of the course. The participants did not have any background with the academic concepts, and it is possible that they would have scored differently if the experiment had been conducted as part of, for example, exam preparation, when they had a broader understanding of the importance of the concepts taught. Some earlier studies have only tested each relation once (e.g., Zaring-Hinkle et al., 2016). The continuous testing protocol in the present experiment, along with extensive reading, could have led to fatigue, possibly contributing to the lower scores. Future studies may consider arranging the training and testing into shorter sessions, as well as consider having different people, like the course instructor, evaluate the procedure's acceptability.

Limitations and Further Research

Design

The present experiment was conducted within 5 hours, including breaks. The strength of this arrangement is that it reduces the possible influence of maturation and history as alternative explanations. On the other hand, some of the participants showed behavior that could be interpreted as boredom towards the end of the session, which might influence attending towards stimuli and correct responding. The present experiment was restricted by practicalities such as the implementation of the training as part of the ordinary instruction during a course. To enhance flexibility, an online solution, such as the one implemented by Walker and Rehfeldt (2012), should be further investigated.

EBI in College Courses

The present study contributes to exciting knowledge on EBI in higher education. The results are promising in the sense that EBI can potentially contribute to more effective ways of teaching in larger groups of students, not only in a laboratory setting. However, further research is essential to determine the most effective ways to implement EBI as part of an ordinary college course. Moreover, several researchers have highlighted the insufficient implementation of EBI across various educational settings (Brodsky & Fienup, 2018; Fienup et al., 2024). It seems to be a gap between what is known to be efficient and the implementation of these procedures. One possible explanation for this might be the lack of teaching educators the science of learning, which hinders their ability to recognize the benefits and possibilities of basic research (Sidman, 1994). The development of a mobile application designed to enable teachers to plot stimuli and train the relationships among them could serve to address this barrier. Nevertheless, it remains further efforts to ensure that the stimuli are thoughtfully designed, and the procedures thoughtfully implemented.

Moreover, practical circumstances prevented the collection of follow-up data for MTS, multiple-choice, and fill-in the present study, and limited the conclusions that can be drawn. When establishing academic skills, it would be of interest to see whether the skills are maintained throughout the semester.

Concluding Remarks

The present experiment aimed to investigate the effects of arranging stimulus equivalence technology in teaching undergraduate students basic measurable dimensions of behavior. The results showed that all participants formed equivalence classes. Furthermore, the participants improved their scores on multiple-choice tasks with variants of the training stimuli and fill-in tasks. As such, the present results add to previous studies demonstrating a promising effect of arranging EBI when teaching academic concept formation in higher education. It is also a step towards implementing EBI as part of an ordinary university course. We did not, however, succeed in obtaining conclusive results regarding the effect of arranging pre-training. Further research on the effects of EBI should consider investigating this further, as well as assessing the possible maintenance of the established classes.

References

Albright, L., Reeve, K. F., Reeve, S. A., & Kisamore, A. N. (2015). Teaching statistical variability with equivalence-based instruction. *Journal of Applied Behavior Analysis*, *48*(4), 883–894. https://doi.org/10.1002/jaba.249

Albright, L., Schnell, L., Reeve, K. F., & Sidener, T. M. (2016). Using stimulus equivalence-based instruction to teach graduate students in applied behavior analysis to interpret operant functions of behavior. *Journal of Behavioral Education*, *25*(3), 290–309. https://doi.org/10.1007/s10864-016-9249-0

Arntzen, E. (2012). Training and testing parameters In formation of stimulus equivalence: methodological issues. *European Journal of Behavior Analysis*, *13*(1), 123–135. https://doi.org/10.1080/15021149.2012.11434412

Augland, H., Lian, T., & Arntzen, E. (2020). Comparing a student active learning format to equivalence-based instruction. *European Journal of Behavior Analysis*, *21*(2), 1–20. https://doi.org/10.1080/15021149.2020.1752513

Brodsky, J., & Fienup, D. (2018). Sidman goes to college: a meta-analysis of equivalence-based instruction in higher education. *Perspectives of*

Behavior Science, *41*(1), 95–119. https://doi.org/https://doi.org/10.1007/s40614-018-0150-0

Cooper, J. O., Heron, T., & Heward, W. L. (2007). *Applied Behavior Analysis* (2 ed.). Pearson/Merrill Prentice Hall.

Critchfield, T. S. (2014). Online equivalence-based instruction about statistical inference using written explanation instead of match-to-sample training. *Journal of Applied Behavior Analysis*, *47*(3), 606–611. https://doi.org/https://doi.org/10.1002/jaba.150

Critchfield, T. S., & Fienup, D. M. (2010). Using stimulus equivalence technology to teach statistical inference in a group setting. *Journal of Applied Behavior Analysis*, *43*(4), 763–768. https://doi.org/10.1901/jaba.2010.43-763

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

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

Fienup, D. M., Reeve, K. F., & Colasurdo, C. R. (2024). Developing equivalence-based instruction i higher education: research, practical considerations, and collaborating with content experts. In A. A. De Souza & D. E. Crone-Todd (Eds.), *Behavior analysis in higher education Applications to teaching and supervision* (pp. 121–144). Vernon Press.

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

Michael, J. (1985). Two kinds of verbal behavior plus a possible third. *The Analysis of Verbal Behavior*, *3*, 1–4. https://doi.org/10.1007/bf03392802

Pilgrim, C. (2020). Equivalence-based Instruction. In J. O. Cooper, T. E. Heron, & W. L. Heward (Eds.), *Applied Behavior Analysis* (3 ed., pp. 452–496). Pearson.

Rehfeldt, R. A. (2011). Toward a technology of derived stimulus relations: an analysis of articles published in the Journal of Applied Behavior Analysis, 1992–2009. *Journal of Applied Behavior Analysis*, *44*(1), 109–119. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3050465/pdf/jaba-44-01-109.pdf

Reyes-Giordano, K., & Fienup, D. M. (2015). Emergence of topographical responding following equivalence-based neuroanatomy instruction. *The Psychological Record*, *65*(1), 495–507. https://doi.org/10.1007/s40732-015-0125-4

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research*, *14*(1), 5–13. https://doi.org/10.1044/jshr.1401.05

Sidman, M. (1994). *Equivalence Relations and Behavior: A Research Story*. Authors Cooperative.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

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

Wiskow, K. M., Subramaniam, S., & Montenegro-Montenegro, E. (2024). A comparison of individual and group equivalence-based instruction delivered via Canvas. *Journal of Applied Behavior Analysis, 57*(1), 262–274. https://doi.org/10.1002/jaba.1025

Zaring-Hinkle, B., Carp, C. L., & Lepper, T. L. (2016). An evaluation of two stimulus equivalence training sequences on the emergence of novel intraverbals. *The Analysis of Verbal Behavior*, *32*(2), 171–193. https://doi.org/10.1007/s40616-016-0072-4

Manuscript information

Submitted on: 19/09/2025
Accepted on: 01/02/2026
Associate Editor: Leandro Boldrin

Appendix A

Examples of Fill-in and Multiple-choice Questions

Fill-in questions
You measure the … of the relevant behaviors if it is essential to know how long an epileptic seizure lasts Correct answer: duration
If you measure how long it takes from asking a person how much 4x12 is until you get the answer, this is called... Correct answer: latency
Which measure should be used if the time from the onset of a discriminative stimulus until the response, is of interest? Correct answer: latency
Multiple-Choice questions
You work with a child who often asks for candy. You want to implement an intervention that aims to increase the time between each time the child asks. Which measure would be natural to use? Correct answer: inter response time
You have 12 forks, 12 knives, and 11 spoons. In which measure would you state this? Correct answer: count
You attend a colloquium group, and Silje constantly interrupts you when you have an essential point. You decide to record how many times it happens during a day. This is an example of...? Correct answer: frequency

Note. The table shows examples of fill-in and multiple-choice questions and the correct answers in the pre-and post-tests.

Appendix B

Social Validity

How confident do you feel in your knowledge of different behavior measurements?						
1	2	3	4	5	6	7
Not at all			Somewhat confident			Very confident
Rate the degree to which you would prefer to be taught using this instructional method						
1	2	3	4	5	6	7
Don't prefer			OK			Strongly prefer
How appropriate was the time commitment for this instructional method in relation to the amount you feel you have learned?						
1	2	3	4	5	6	7
Not at all			Somewhat appropriate			Very appropriate
How do you feel about the length of this instructional method?						
1	2	3	4	5	6	7
Too long			OK			Too short