The Psychological Record
https://doi.org/10.1007/s40732-019-00364-3

ORIGINAL ARTICLE

Social Categorization and Stimulus Equivalence: A Systematic Replication

Rebekka C. W. Strand[1] · Erik Arntzen[1]

Abstract

A systematic replication of Watt, Keenan, Barnes, and Cairns (1991) was conducted on two groups of Norwegian soccer supporters. The 24 participants were trained conditional discriminations for the emergence of three 3-member equivalence classes, when members of two of the classes were assumed to be part of the participants' preexperimental history. The stimuli used in these classes were pictures or names of soccer players relevant to their own team or the rivaling team. Participants were trained in a linear training structure before the test. The test was split into three test blocks. Test Block 1, a replication of Watt et al.'s (1991) equivalence test, Test Block 2 an adapted generalization test and Test Block 3 an updated equivalence test. The results in Test Block 1 replicated what was found in 1991, but Test Block 3 did not replicate the same results. In Test Block 2, participants scored as expected and the response patterns were distinctly different between the test groups and the Control Group. Also, the time used to finish the experiment by the soccer team supporters were significantly higher than by participants who had no interest in soccer. This difference was also reflected in the reaction times the participant showed on the emergent relations in test blocks 1 and 3. A correlation was found between the number of expected scores on the questionnaire and the number of passes in Test Block 1. However, no correlation was found in the number of participants who passed in Test Block 3. The study by Watt et al. (1991) was not found to generalize to the context in the current study. However, the extended parts in the study had some promising results on how social categorization can be studied in the derived stimulus paradigm.

Keywords Social categorization · Stimulus equivalence · Preexperimental history · Reaction time · Questionnaire · Prior learning

Stimulus equivalence or the derived relations paradigm is a process which results in the emergence of untrained relations, also known as derived relations or emergent relations. A stimulus equivalence test has been proven as a valid procedure, as a systematic, operational and empirically verifiable way of testing for the emergence of emergent relations in humans (McIlvane, Kledaras, Gerard, Wilde, & Smelson,2018; McIlvane, 2013). It allows researchers to directly study each specific relation and control for other variables that might affect the emergence of the emergent relations. Including a matching-to-sample (MTS) procedure to train conditional discriminations between stimuli, with at least two conditional relations (e.g., AB and BC) with a common member. The test phase in turn evaluates the emergence of relations that are symmetrical (BA and CB), transitive (AC), and equivalence (CA). Reflexivity might also be evaluated (AA, BB and CC), but in general this property is not necessary to evaluate the emergence of other relations (Sidman & Tailby, 1982).

Social categorization can be interpreted into behavior analysis and expanded upon through the process of stimulus equivalence. It makes it possible to look into if preexperimental history interferes with the emergence of emergent relations. By testing for the number of correct responses on emergent relations a unit of measurement of complex human behavior becomes observable. If participants respond differently when the stimuli are assumed to be a part of their preexperimental history an inference could be explained. However, there is a lack of studies on social categorization within the derived relations paradigm.

Many labels have been used as social categorization, preexperimental history, bias, racial attitudes, stereotyping, social contexts, and prior learning (Adcock et al., 2010; de

✉ Rebekka C. W. Strand
rebekka.w.strand@gmail.com

Erik Arntzen
erik.arntzen@equivalence.net

[1] Oslo Metropolitan University, Oslo, Norway

Published online: 02 January 2020

Carvalho & de Rose, 2014; Dixon, Rehfeldt, Zlomke, & Robinson, 2006; Haydu, Camargo, & Bayer, 2015; Peoples, Tierney, Bracken, & McKay, 1998; Watt et al., 1991). These labels are not precise or observable. and a common behavioral definition should be established to solve this problem. In the present article preexperimental history was chosen to describe the suggested interference with emergent relations. Preexperimental history seems to be a more functional term, because it does not refer to anything mentalistic and refers to the general history of learning for each participant before the experiment (Haydu et al., 2015).

There are some studies within the derived relations paradigm that makes it possible to study the effects of preexperimental history (Adcock et al., 2010; Haydu et al., 2015; Kohlenberg, Hayes, & Hayes, 1991; Watt et al., 1991). Watt et al. (1991) conducted the first experiment on how the preexperimental history might interfere with the emergence of derived relations. Watt et al. found that participants did not form equivalence classes when the experimenter defined classes consisted of stimuli related to their preexperimental history. Participants belonged to three different social groups in Northern Ireland. The experimental method consisted of a pretraining phase, a training phase, and two test categories (see Table 1). The training structure they used was linear, A to B and B to C (Arntzen, Grondahl, & Eilifsen, 2010; Arntzen, 2012). Watt et al. found that all of the six Northern Irish Protestants failed to respond in accordance with stimulus equivalence. Five of the Northern Irish Catholics also failed the equivalence test, but the other seven scored in accordance with the experimenter defined classes.

An issue worth mentioning in the Watt et al. (1991) study is that they did not include a symmetry test in accordance with a linear training structure (BA and CB) and a full equivalence test (CA). Sidman (2000) would therefore argue that the claim by Watt et al. is false. Participants might be able to form equivalence classes, but the test would not qualify as a test for equivalence classes. Because of this an extension should be included to test for equivalence classes with the modern criteria.

One of the more recent studies conducted by Haydu et al. (2015) is a convincing example on how to use a full stimulus equivalence procedure to research the effects of preexperimental history on the emergence of emergent relations. They used the MTS procedure on 28 men from three different soccer clubs in Brazil. They tested for all the properties of the equivalence classes, with training (AB and BC), a symmetry test (BA and CB), a transitivity test (AC), and an equivalence test (CA). They were trained to match club emblems (A) to abstract paintings (B), and then to match the same abstract paintings with the words "Good," "Poor," or "Regular" (C). The words "Good," "Poor," or "Regular" were presented systematically so that the main rival would be matched with "Good" and their own club matched with "Poor" for each of the three groups. The way they presented these words in the conditional discrimination procedure were to test for how preexperimental history with the soccer emblems and the words would interfere with the emergence of emergent relations. They found that none of the participants passed the test for equivalence and transitive relations between the club emblems and words. The results in Haydu et al. (2015) corresponded with the results of Watt et al. (1991). However, the studies may not be comparable, because the method used in the original study were questionable at best, a replication of Watt et al. (1991) is therefore necessary for scientific precision.

One of the problems to both of the studies, that were described earlier, is that none of them made attempts to validate the way in which participants identified themselves. By incorporating a prescreening to investigate the social group that participants identified as, one can make some assumptions about how the participants identified themselves as (Slaton, Hanley, & Raftery, 2017). Furthermore, to include a postscreening, a bipolar questionnaire using a Likert scale, one can aim to validate those assumptions, and investigate whether scored on the stimulus equivalence test correlate with scores on the bipolar questionnaire. This might be a way of solving the problem of uncertainty of whether the equivalence test results are due to methodic errors or to the participants' preexperimental history (Critchfield & Perone, 1993; Critchfield, Tucker, & Vuchinich, 1998). The current study will possibly add to the validity of the experimenter defined stimulus classes and add information that might be relevant when interpreting certain response patterns.

Although some of the participants showed the emergence of equivalence classes in the test groups in the study by Watt et al. (1991), a difference might have been found in the amount of time they spent on the study compared to the English participants (Kurzban, Tooby, & Cosmides, 2001). Although, it is discussed if reaction times (RT) is relevant to traditional stimulus equivalence tests, the variable of RT might be considered to be more relevant in a study on preexperimental history due to the focus on previously taught behavior or private events. RT has been shown to be a possible indirect measure of private events, indirect priming and "remembering" in the derived stimulus paradigm, and could therefore be an addition to the methods of studying how preexperimental histories might interfere with emergent relations (Arntzen et al.,. 2010; Eilifsen & Arntzen, 2009; Fields et al., 2002; Leppänen & Hietanen, 2004; Barnes-Holmes et al. (2005); Vaidya, Hudgins, & Ortu, 2015).

Systematic replications are important because they can validate or disprove previous studies (Sidman, 1960). The Watt et al. (1991) study has been cited many times, but research methods have evolved and no replications has been conducted since the original study. The purpose of this study was to systematically replicate Watt et al. (1991) using a different social group and compare the results. Next, adapting the

method to see if an equivalence test would show different results in order to see if RT as a variable might give a better picture of the processes of how preexperimental history interferes with the formation of equivalence classes and see if questionnaires could add to studies on preexperimental history.

Method

Participants

Twenty-six participants were assigned into three groups, one Control Group and two test groups. Twenty-four participants were included in the final results. Participants were recruited through a survey that was published on the group's private Facebook pages. When they had filled out the survey either on the Facebook page, at the university laboratory, or at the soccer stadium, if they responded according to set criteria, they were called and asked to pick a date and time to meet. The minimum age for participation was set to 18. The age range of the Control Group was relatively young (M = 26.75) whereas the Test Group 1 (M = 41.7) and Test Group 2 (M = 35) was generally older. Age ranged from 19 to 62 years old with a mean age of 35. Both test groups consisted of relatively more male (85%) participants and the Control Group was gender balanced (50%). To increase the probability in the recruiting process, participants received a gift card with the value of 100 kr.

Participants were allocated to the different groups based on the survey. If they answered that they had little or no interest in soccer they were assigned to the Control Group. If they responded that they supported one of the teams and disliked the other, and also responded that they either: had a tattoo of their team, attended all home games and some away games or more, or if they had been in a physical altercation regarding their team, then they were included in Test Group 1 or 2. Of these participants, if they supported VIF they were placed in Test Group 1 and if they supported LSK they were placed in Test Group 2. In order to avoid prompting string reactions, a time limit was set on the experiment such that if participants had not reached the Test Block by the end of the 60th min they would be excluded from the data and also asked if they wanted to quit the experiment. The study was assessed and accepted by the Ethical Center for Research in Norway before it was conducted (Norsk Senter for Forskningsdata, 2018).

One of the participants had prior experience with a similar stimulus equivalence study and was therefore excluded from the final data set. The other participant that was excluded from the data set did not reach the Test Block within the 60th min and did not reach the criteria for inclusion. None of the other participants were excluded.

Instructions were provided on how to use the computer. They were informed that they could end the experiment at any time with no negative consequences and that the instructor would come in after an hour to check in on them. Before beginning the test, they were asked to read and sign a document of agreement, which also stated that they would stay anonymous. When the experiment ended, participants were given a second document of agreement in accordance with the Norwegian Center for Research's (NSD) requirements after being fully informed of what the study entailed (Norsk Senter for Forskningsdata, 2018).

Setting

The experimental sessions were conducted in two locations: in a quiet room with a table and a chair in a cubicle at the laboratory at Oslo Metropolitan University, and in a meeting room with a round table and few pictures on the walls at the soccer stadium for LSK in Lillestrøm. Participants were left alone while the experimenter waited outside. Participants were allowed to drink coffee during the experiment, but they were asked to put away their phone and bag. The length of experimental sessions varied from 15 min to 69 min, with a mean length of 33.5 min.

Materials

The experiments were conducted on HP Elitebook laptop computers running Windows 7 operating system. The computers had 17-in. screens and external Dell computer mice were used to control the mouse cursor. All aspects of the training and testing for equivalence relations and to establish discriminative functions were controlled by parameters in a custom-made software program. The software program conducted automatic data recording on the number of trials, the stimulus relations that were trained, the responses to the sample stimuli, RT, the correct/incorrect comparison choice, and the provision of programmed consequences on each trial. Finally, the software counted the number of trials for both training and testing. Verbal reports were recorded through a prepared bipolar questionnaire on the computer. The Likert scale had a minimum score of 1, corresponding with the word "uncomfortable," and maximum score of 5, corresponding with the word "comfortable," in which participants clicked on one of the numbers to every stimulus presented. The questionnaire contained information about what stimulus might be seen as aversive, neutral, or positive, and on how they experienced the experiment.

Stimuli Figure 1 shows the stimuli used. The stimuli were chosen after a pilot was conducted with three participants prior to the current study. The stimuli were defined as three classes with four to five members. The defined classes were pictures

Figure 1. Visual examples of the stimuli used in the experiment. The vertical letters show the number of classes, the horizontal numbers show the number of members. The scarf represents the generalization stimulus in Test Block 1. The three members A3't, B3't and C3't represents the comparison used in Test Block 2.

and names from the two opposing soccer teams, as well as some abstract pictures and one abstract name. Class A were stimuli related to LSK, assumed positive for Test Group 2. The name chosen in this class was Frode Kippe. Class B were neutral symbols and the neutral name Devon Larsen. Class C were stimuli related to VIF, assumed positive for Test Group 1. There were three symbols in this class and two names, the names were John Carew and Freddy Dos Santos. For simplification, the names in the different classes were written as LSK_name_1, Novel_name_1, VIF_name_1, and VIF_name_2 in Figure 1. The scarf, as shown in Figure 1, is written as C3' to specify that it is the first generalization stimulus introduced. In the article the names, LSK_name_1, Novel_name_1, and VIF_name_2 are referred to as A3't, B3't, and C3't to specify that they are only introduced in the generalization test and are also a secondary presentation of generalization stimuli after C3'.

Design

A group design was used with one Control Group and two test groups.

Data Collection

The computer program collected all data in the training and testing condition. Data was also collected through indirect analysis in two questionnaires. The first questionnaire was used as a tool for recruitment. Some of these questions were such as "which soccer team is the best in Norway?" and "do you have a tattoo of your favorite team?" The second questionnaire was based on a Likert scale and was meant to test for discomfort or comfort of the participant's reactions to the different stimuli. This included seven questions, with pictures of each stimulus from class A and C, where they could respond on a scale with 5 points from "uncomfortable" to "comfortable." An added question was also included where they were asked to give an added feedback if there was anything that they wanted to elaborate on.

Procedure

The conditional discriminations were administered by the MTS program, in a linear training structure using simultaneous matching. The programmed consequences were thinned in three steps. The programmed consequences were at 100% in the first step, in the second step the programmed consequences were gradually thinned from 100% to 0%, and in the third step, the test condition, no consequences were presented. The presentation of consequences, depending on the experiment stage, would pop up on the screen after a response. The consequences were present for 1,000 ms with each presentation. Following the consequences an intertrial interval (ITI) lasted for 1,000,ms.

In the three test conditions 10 novel arrangements of stimuli were presented in the two first test blocks and in the third Test Block 9 additional novel arrangements were presented (see Table 1). Participants who formed stimulus classes according to the criteria were scored as PASS, and participants who did not form stimulus classes according to criteria were scored as FAIL. The criteria for PASS was set as 100% correct score on Test Block 1 and Test Block 2. Before starting the experiment, an instruction was presented on the screen in Norwegian.

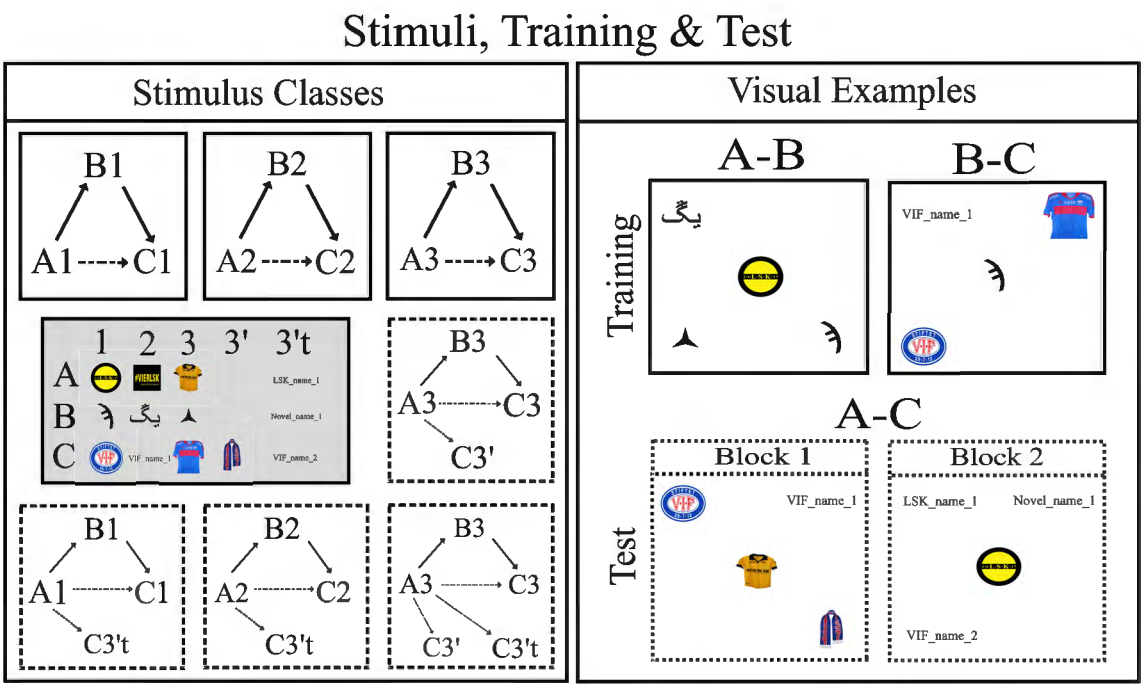

Figure 2. In the right square there are visual examples of the training and Test Block 1 and 2 in accordance with Watt et al. (1991). In the left square there are visual examples of the stimulus class relations. The trained relations are represented with full lined arrows, and the tested relations with dotted lined arrows. The full lined squares within the bigger boxes represent the three equivalence classes. The dotted lined boxes represent the classes with the extra test stimuli.

Stage 1 Training with continuous programmed consequences was conducted in two phases. The first phase was the A to B training, A as sample to B as comparison. The A class was defined as visual stimuli related to the soccer team LSK and the B class were arbitrary stimuli, in the form of unknown figures. In precise terms, three A stimuli and three B stimuli were presented randomly until all stimulus combinations were successfully completed in two successive cycles. The second phase of Stage 1 was then introduced where the training order, B to C, was presented. Where the sample B was an abstract symbol, and the comparison C was the stimuli related to the VIF team. The criterion for mastery was set to 30 trials correct with 15 trials in each block for completion, a mastery criterion of 100% correct before next stage was presented (see Table 2).

Stage 2 Training was then introduced between AB and BC relations with intermittent consequences in random order. All previously trained combinations in Stage 1 were presented at least twice and the programmed consequences were thinned from 100% to 0%, from 100% to 75% to 25% and then to 0%. The criterion for mastery was set to a minimum of 150 trials correct in succession, equaling five blocks with 30 trials per block for completion. If the participant successfully matched all relations twice with a 100% score rate within the 0% programmed consequences condition, they would advance to Stage 3, the test.

Stage 3 The test was then introduced and included three different test blocks that were presented in succession. The first block included baseline relations AB and BC, transitivity relations, AC, and a novel stimulus C3'. The second Test Block was a generalization test. The third Test Block was an equivalence test.

Test Block 1 The first Test Block included presentation of all baseline relations AB and BC, to test for transitivity AC and generalization C3'. At least 10 presentations of

Table 1 A visual presentation and comparison of the training and test blocks in Watt et al. (1991) and the current study

Study	Training (LS)	Test Block 1	Test Block 2	Test Block 3
Watt et al. (1991)	AB BC	AB BC AC + C3'	A -> C3't	
Current Study	AB BC	AB BC AC + C3'	A -> C3't	AB BC AC CB BA CA

C3' and C3't are abbreviations of the stimuli introduced as generalization stimuli.

Table 2 Number of correct responses in all test blocks per participant. In Test Block 1, there were 60 mixed trials, Test Block 2 had 30 mixed trials and Test Block 3 had 90 mixed trials. Percentage of correct responses are also shown. The numbers to the left of the columns represent the participants. The left number per participant across relations, shows correct responding, whereas the number to the right shows the maximum number of possible responses. The abbreviations stand for the relations tested per Test Block.

Participants	Test Block 1				Test Block 2	Test Block 3				
	BL	Tran. (AC)	Gen. 1	(% cor. responses)	Gen. 2	BL	Sym.	Tran. (AC)	Tran. (CA)	(% cor. responses)
Control Group										
18300	30 of 30	20 of 20	10 of 10	100	9 of 30	N/A	N/A	N/A	N/A	N/A
18301	30 of 30	20 of 20	10 of 10	100	9 of 30	N/A	N/A	N/A	N/A	N/A
18302	30 of 30	20 of 120	10 of 10	100	10 of 30	N/A	N/A	N/A	N/A	N/A
18303	30 of 30	12 of 20	10 of 10	86.6	6 of 30	N/A	N/A	N/A	N/A	N/A
18304	30 of 30	20 of 20	10 of 10	100	8 of 30	30 of30	30 of 30	15 of 15	15 of 15	100
18305	30 of 30	10 of 20	10 of 10	83.3	14 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18306	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18307	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18308	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18309	28 of 30	11 of 20	8 of 10	78.3	14 of 30	30 of 30	28 of 30	11 of 15	12 of 15	93.3
Test Group 1										
18310	29 of 30	20 of 20	10 of 10	88.3	0 of 30	N/A	N/A	N/A	N/A	
18311	30 of30	13 of 20	10 of 10	98.3	1 of 30	N/A	N/A	N/A	N/A	
18312	30 of 30	19 of 20	10 of 10	98.3	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18313	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18314	25 of 30	19 of 20	8 of 10	86.6	1 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18315	30 of 30	17 of 20	10 of 10	95	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18316	30 of 30	10 of 20	0 of 10	66.6	0 of 30	30 of 30	30 of 30	12 of 15	12 of 15	933
Test Group 2										
18317	30 of 30	0 of 20	10 of 10	66.6	0 of 30	30 of 30	28 of 30	8 of 15	11 of 15	933
18318	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18319	30 of 30	18 of 20	10 of 10	96.6	9 of 30	30 of 30	30 of 30	14 of 15	11 of 15	100
18320	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	94.4
18321	29 of 30	20 of 20	10 of 10	983	4 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100
18322	30 of 30	10 of 20	0 of 10	66.6	0 of 30	29 of 30	29 of 30	15 of 15	15 of 15	97.7
18323	30 of 30	20 of 20	10 of 10	100	0 of 30	30 of 30	30 of 30	15 of 15	15 of 15	100

Some of the participants were not able to do Test Block 3 and are shown as N/A.

each baseline relation were mixed in with five presentations of transitivity and generalization trials. Test Block 1 was set to 60 possible trials where 20 trials were baseline trials, 20 trials were transitivity trials, and 10 trials were generalization trials, in a mixed order. The criterion for mastery was defined as 100% correct, 60/60 responses.

Test Block 2 This part of the test included a generalization test. The comparison stimulus in this Test Block were three novel names—A3't, B3't, and C3't—that had not been presented in any other part of the training or Test Block 1. C3't was the only comparison stimulus defined as a correct response in this Test Block. Only the stimuli from class A was used as sample stimuli. Each relation was presented 10 times before the next Test Block was introduced. The number of possible trials in this Test Block was 30. The criterion for mastery was set as 33% correct responses with a leniency of 25–50%.

Test Block 3 This Test Block included baseline relations, AB and BC, symmetry relations, BA and CB, transitivity relations AC, and equivalence relations CA. This Test Block did not include the generalization stimuli. The number of possible trials in this Test Block was set to 90, 30 trials were baseline trials, 30 trials were symmetry trials, 15 trials were transitivity trials, and 15 trials were equivalence trials. The criterion for mastery was defined as 100% correct, 90/90 responses (see Figure 3 for visual examples).

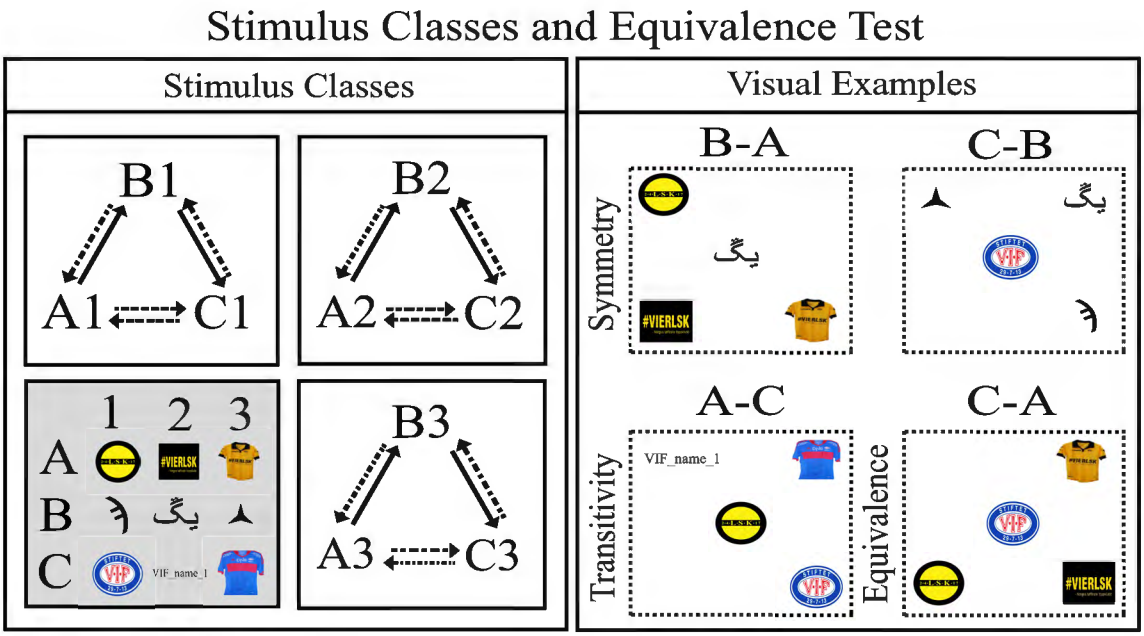

Figure 3. The right square is a visual example of the different emergent relations included in Test Block 3. The left square shows visual examples of all trained and tested relations, including all test blocks. The full lined squares within the left box represent the three full equivalence classes. The full lines represent the trained relations and the dotted lines represent the emergent relations that were tested.

Results

Trials to Criterion

See Figure 5 for average number of trials during training and thinning for all participants the groups. The number of average trials during training was generally less in the Control Group (M = 66.75, range 45–120), than in Test Group 1 (M = 165, range 60–300) and Test Group 2 (M = 122, range 90–375). The significance criterion was set as $p < .05$. There was a significance effect between the Control Group and Test Group 1 t (54) = 3.39 p, < .001. However, there was no significant effect between the Control Group and Test Group 2, t(54) = -1.35, $p < .09$. There was also no significance effect between Test Group 1 and Test Group 2 with a t(36) = 0.81, $p < .21$.

In general, there were fewer average trials during the thinning blocks in the Control Group (M = 189, range 150–270) than in Test Group 1 (M = 270, range 180–540) or in Test Group 2 (M = 282, range 150–480). There was a significant effect between the Control Group and Test Group 1, t(54) = 1.92, $p < .03$, and between the Control Group and Test Group 2 t(54) = 1.92, $p < 0.03$. There was no significant effect between the two test groups t(36) = 0.19, $p < .42$.

Test Blocks 1 and 3

See the bottom half of Figure 4 for a visual presentation of the following results. The percentage of participants who passed with a hundred percent test score in Test Block 1 was 70% in the Control Group, 14% in Test Group 1 and 42% in Test Group 2. The Fisher exact test was conducted on the passes versus fails and showed no significant difference between the groups in Test Block 1. The Control Group and Test Group 1 showed no significant effect $p < .15$, neither did it show any significant results between the Control Group and Test Group 2, $p < .64$, or between Test Group 1 and Test Group 2, $p < 1$.

In Test Block 3, the percentage of participants who passed with a 100% test score was 83.3% in the Control Group, 80% in Test Group 1, and 57% in Test Group 2. The Fisher exact test showed no significant difference between the groups in Test Block 3. There was no effect between the Control Group and Test Group 1, p <1, the effect between the Control Group and Test Group 2 approached significant, $p < .56$, and similar results were observed between the two test groups, $p < .58$.

See the top half of Figure 4 for the average percentage correct scored per group in Test Block 1. The average percentage correct per group was high in the Control Group (M = 92.5%, range 78–100%), but was not as high in Test Group 1 (M = 76.6, range 66–100%) and Test Group 2 (M = 88%, range 66–100%). A t-test was conducted on the percentage correct in Test Block 1 to compare the results between the groups. The test showed no statistical significance between the Control Group and Test Group 1; t(53) = 0.26, $p < .39$, between the Control Group and Test Group 2; t(53) = -1.04, $p < .15$, or between the two test groups; t(36) = 0.69, $p < .24$.

The average percentage correct scored per participant, on the test on emergent relations in Test Block 3, was very high in all groups. The percentage correct scored was similar in the Control Group (M = 98.8%, range 93–100%), Test Group 1

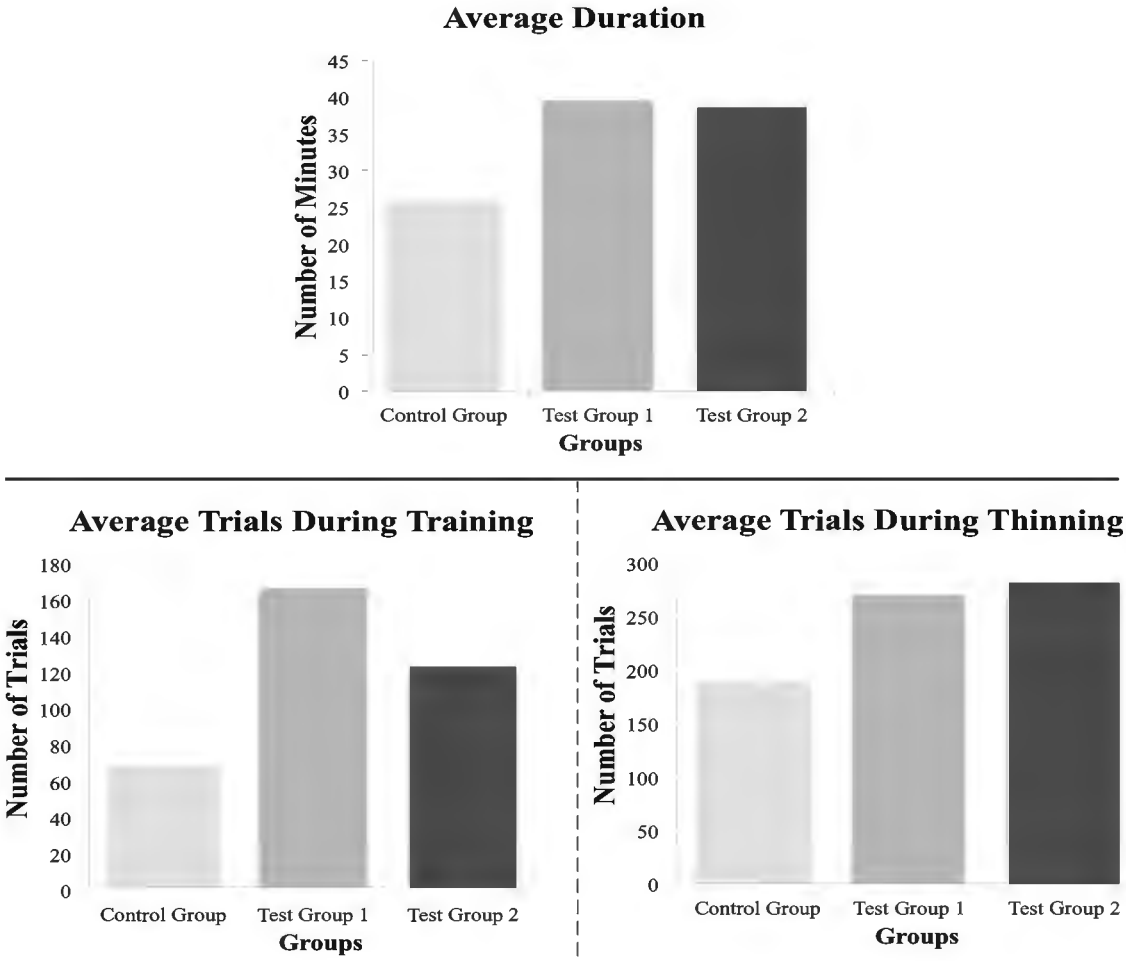

Figure 4. The top figure over the full line shows the average minutes, duration, used for the full experiment in each group. The y-axis shows number of minutes and the x-axis shows the different groups. The bottom left figure shows the average number of trials used in training across all groups. The bottom right figure shows the average number of trials used in thinning across all groups. The y-axis represents the number of trials and x-axis represents the different groups in both of the bottom figures.

($M = 98.6\%$, range 93–100%), and Test Group 2 ($M = 97.5\%$, range 93–100%). A *t*-test was conducted to compare the results between groups, and they showed that there was no significant difference in the results. There was no significant effect between the Control Group and Test Group 1, $t(20) = 0.12$, $p < .45$; the Control Group and Test Group 2, $t(30) = 0.6$, $p < .27$; and the two test groups, $t(24) = 0.42$, $p < .33$.

Test Block 2

The results from Test Block 2 are presented in Figure 6 as a visual presentation of response patterns. A visual analysis of the response matrix in Figure 6 shows a distinct pattern in the two test groups, where they choose A3't more than any other comparison, although the distribution is more random in the Control Group. Though a few of the participants in all groups deviate from the pattern, participant 18301 in the Control Group consistently responded to a different comparison per different sample, A1 to A3't, A2 to B3't, and A3 to C3't. Participant 18307 in the Control Group shows the same pattern as in the test groups with all responses on A3't, this participant also responded to the questionnaire with a preference to the A stimuli (see Figure 8). Participant 18315 in Test Group 1 distributed responses equally between A3't and B3't but showed no responding to C3't. Participant 18320 in Test Group 2 did not respond consistently on A3't as the other participants in the Control Group but responded more distributed on all three comparisons. When compared within the groups the Control Group had a mean percentage of 43% (range 0–50%) correct, Test Group 1 had a mean percentage of 0.9% (range 0–3.3%) correct and Test Group 2 had a mean percentage of 6% (0–30%) correct A statistical analysis of the number of correct responses of 30% (25–50%), defined as responding in accordance with experimenter-defined classes, were set

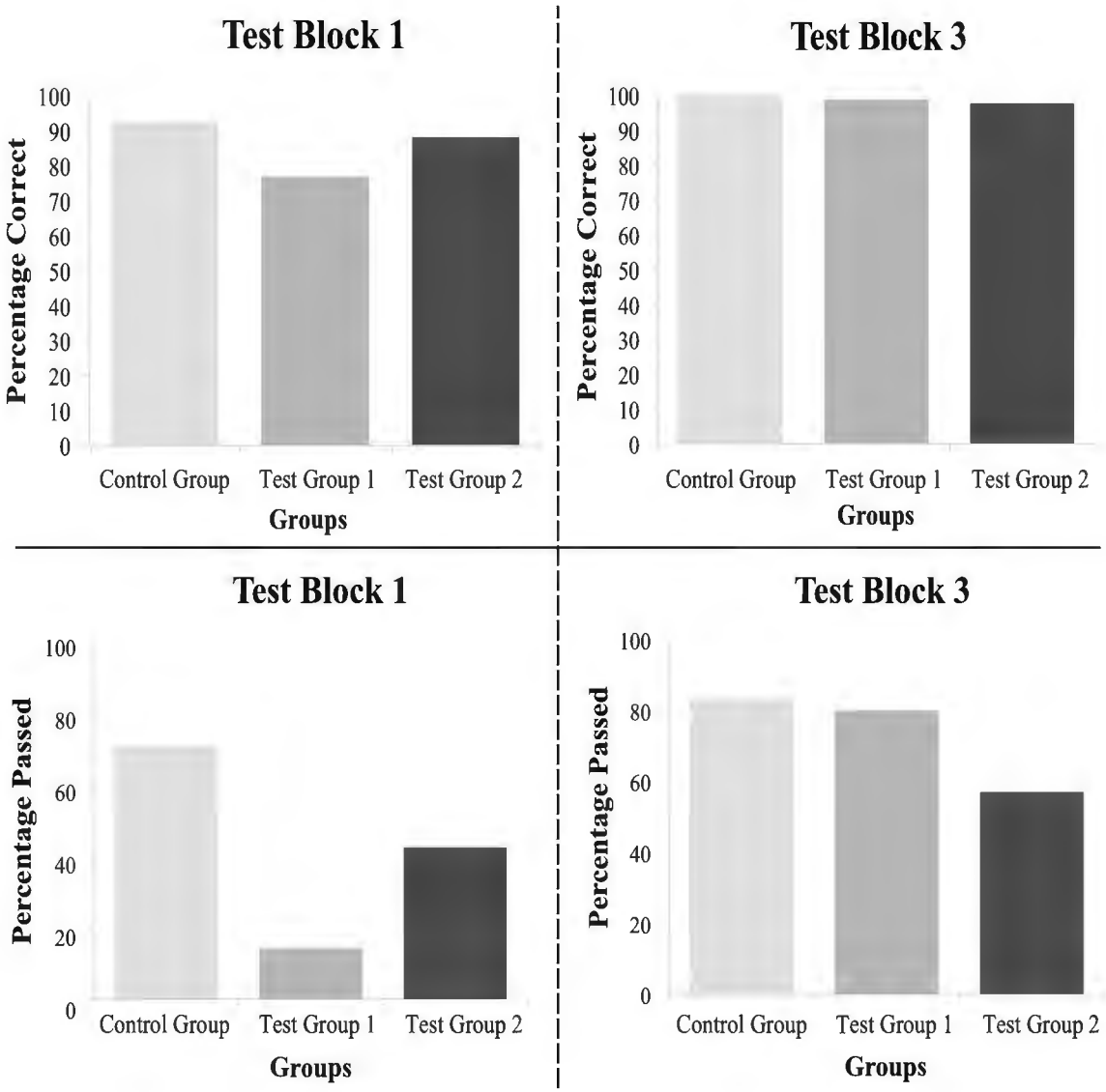

Figure 5. The top two figures show the average percentage correct across all groups. The left figure shows results from Test Block 1 and the right from Test Block 3. The bottom two figures below the full line shows the average percentage of participants per group that responded in accordance with stimulus equivalence. The bottom left figure shows the percentage passed in Test Block 1. The bottom right figure shows the percentage passed in Test Block 3.

against the number of deviating results 0% (0–25%), defined as not responding in accordance with experimenter-defined classes, between groups. The Control Group had eight participants corresponding with the set criteria and two participants who did not, Test Group 1 had zero participants corresponding with the set criteria and Test Group 2 had one participant corresponding with the set criteria and six that did not. The Fisher-exact test was used to compare the different groups. The results showed a statistically significant effect between the Control Group and Test Group 1, $p < 0.0023$. Likewise, the comparison between the Control Group and Test Group 2 showed a significant effect, $p < .0152$. however, the result between the two test groups did not show any significant effects, $p < 1$.

Duration and Reaction Times

See Figure 4 for the average duration spent by participants for the whole experiment. In the Control Group, participants spent fewer minutes on average than the two test groups (M = 25.8, range 15–32). Test Group 1 spent a few more minutes (M = 39.5, range 25–69) than Test Group 2 (M = 38.7, range 26–50). The t-test between the Control Group and Test Group 1 showed a significant effect. The biggest effect was seen between the Control Group and Test Group 2; t(53) = -3.68, $p < .001$, and a similar result was seen between the Control Group and Test Group 1; t(53) = -2.6, $p < .009$. There was no statistical significance between Test Group 1 and 2; t(36) = 0.13, $p > .45$.

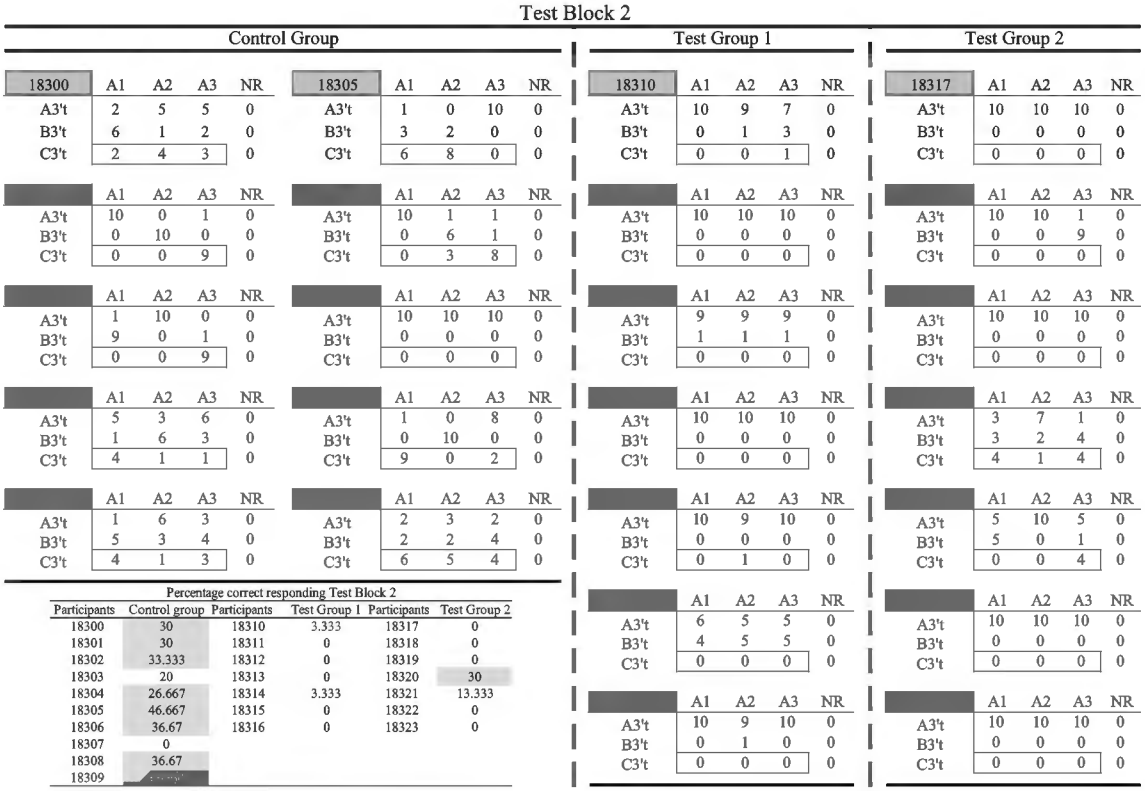

Figure 6. This is a visual presentation of responses made by participants during Test Block 2. The numbers in grey squares represent the number of participants per group. The two bars to the left represent the Control Group, the two longer bars represent the two test groups. NR stands for "no responses" during tests. A1, A2, and A3 are the sample used, A3't, B3't, and C3't was the stimuli used as comparison. The three numbers that are inside squares within the matrices next to C3't represents the responses that were defined as correct. The smaller figure on the bottom left shows the percentage correct responding per participants in all groups. The grey squares in this smaller graph show the participants who reached the defined criteria for mastery.

See Table 4 for the average RT per participant on transitive and equivalence relation in Test Block 1 and Test Block 3, across participants who formed equivalence classes and those who failed to do so. In the Control Group eight participants were quicker in RT in Test Block 3 compared to Test Block 1. Two participants did not increase their RT between test blocks. P18309, who did not pass the test, slowed down on the equivalence relations. Although P18309 slowed down, their RT was quicker than P18308, who passed the test. On the other hand, P18307 had a consistently short RT and passed both test blocks. Three participants in Test Group 1 reduced their RT in Test Block 3 compared to Test Block 1. These participants also went from failing Test Block 1 to passing Test Block 3. However, P18316 showed an increase in RT during Test Block 3 and did not pass in any of the test blocks. In general, six out of the seven participants in Test Group 1 had slow RT. The only participant that consistently had a quick RT was the only participant to also pass both test blocks in this group (P18313). As in Test Group 1, participants in Test Group 2 that increased in RT in Test Block 3 failed Test Block 1 but passed Test Block 3. Likewise, the two participants who had a consistent RT were generally quicker than the other participants in the group and passed both test blocks. Also, the two participants who failed both test blocks likewise showed a slower RT in Test Block 3 than in Test Block 1. P18322 was one of these participants and was also the participant who had the slowest RT of all participants in all groups. Over half of all participants in all groups had a generally slower RT to the generalization relation, A3C4.

See Figure 7 for the average RT of all groups to the transitive and the equivalence relations in Test Block 1 and Test Block 3. A significant difference in RT was found between groups in the transitive relations in Test Block 3, $M = 2.2$ $SD = 0.8$, $F(5.5) = 1.57$, $p < .04$. However, no differences were found in the RT on transitive relations in Test Block 1, $M = 2.6$ $SD = 0.5$, $F(0.01) = 0.003$, $p > .9$, or on RT on the equivalence relations in Test Block 3, $M = 2$ $SD = 0.6$, $F(2.9) = 0.68$, $p > .12$. Test Group 1, $M = 2.1$ $SD = 0.7$, $F(1.7) = 0.88$, $p > .25$, and Test Group 2, $M = 2.7$ $SD = 0.57$, $F(0.98) = 0.32$, $p > .37$, had a more consistent RT across Test Block 1 and Test Block 3, whereas the Control Group, $M = 2.1$ $SD = 0.55$, $F(14.2) = 1.2$, $p < .01$, varied more between the two test blocks.

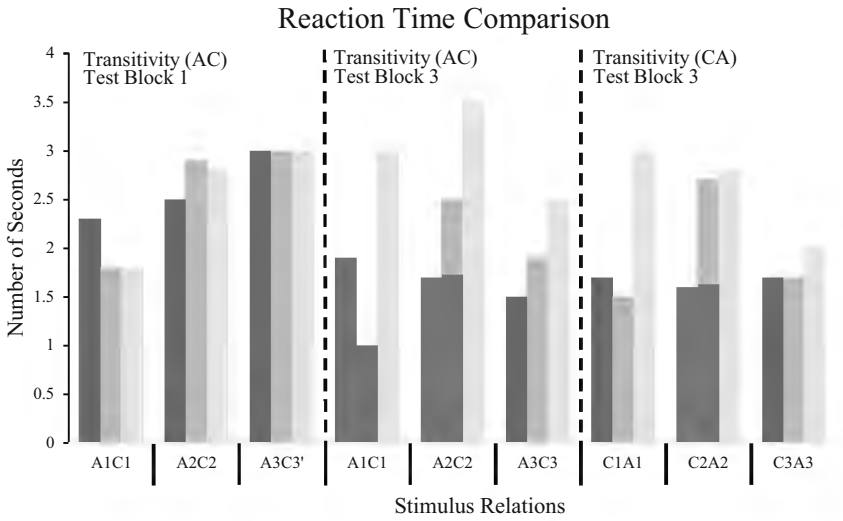

Figure 7. The figure is a presentation of the average RT of all participants responses within each group on transitive relations. The RT on transitive relations were calculated for Test Block 1 and Test Block 3. The darkest bar represents the Control Group, the middle bar represents Test Group 1 and the lightest bar represents Test Group 2. The y-axis represents number of seconds and the x-axis represents each transitive relation tested.

Questionnaire

A visual presentation and a chi square test were conducted on the results of the bipolar questionnaire. See Figure 8 for a visual presentation of the results. The visual presentation shows that the test groups have distinct response patterns that were opposite to each other, whereas the Control Group shows a less distinct pattern. This corresponds to the hypothesis that the test groups would respond negatively to rival team stimuli and positively to their own team's stimuli. However, some outliers can be seen in the written name stimuli, C2, C3't, A3't, and B3't. The chi square test, as seen in Table 3, confirms that the visual pattern was statistically significant. The significance criterion was set as $p < .01$. The preexperimental

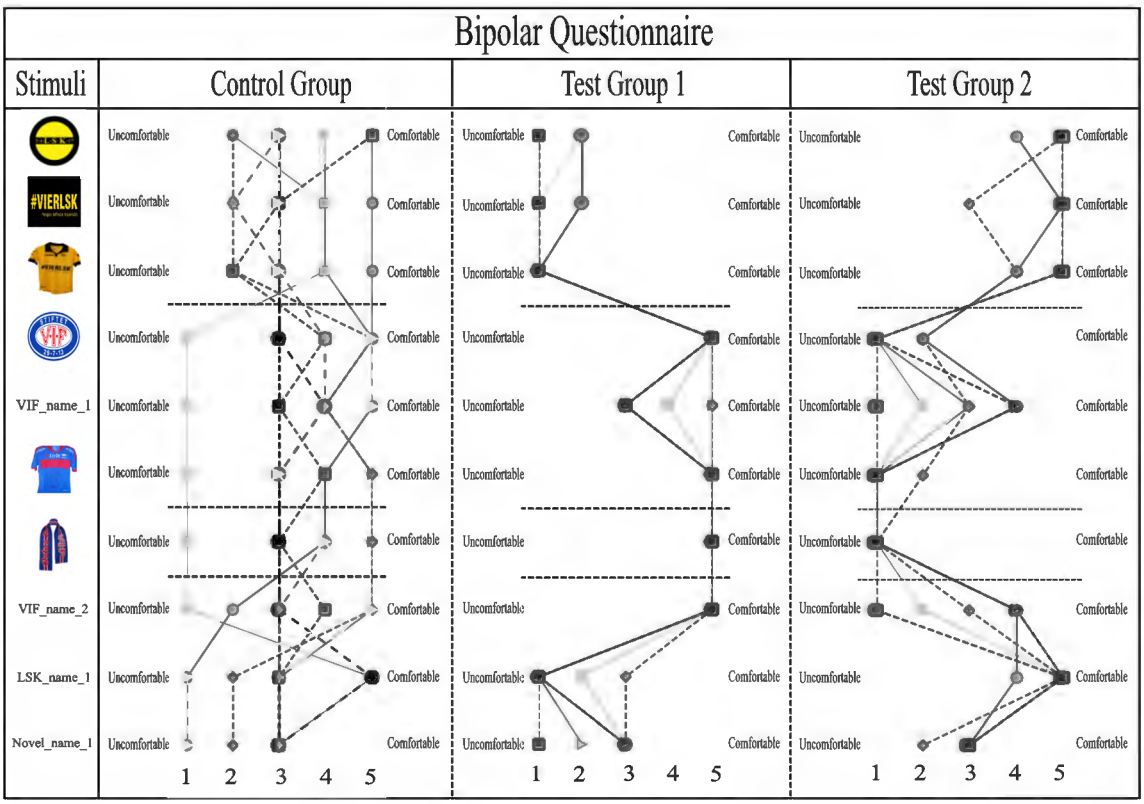

Figure 8. The figure represents the responses participants made on the bipolar questionnaire after the experiment. The left column of pictures represents the stimuli presented per question, and the columns of points to the right of the picture represents the responses made in each group. The left figure shows the Control Group, the middle Test Group 1, and the right Test Group 2. The word "uncomfortable" to the left and "comfortable" to the right of each figure represents the adjectives, where the dimension between them, was scored as a Likert scale of 1 to 5 as shown at the bottom of each figure on the x-axis.

Table 3 Pearson Chi Square test results on the bipolar questionnaire across all stimuli and all groups

Groups		Stimuli									
		Round Yellow	Black Square	Yellow shirt	Round Blue	VIF_name_1	Blue Shirt	Blue Scarf	VIP_name_2	LSK_name_1	Novel_name_1
Control Group	Pearson Chi Square	4.168	4.168	4.168	4.168	4.168	4.168	4.168	4.168	4168	2.088
	df	9	9	9	9	9	9	9	9	9	9
	p	.135	.135	.135	.287	.287	.135	.287	.406	.135	.001
Test Group	Pearson Chi Square	1.653	0.872	0.872	0.872	2.204	0.872	0.872	0872	0.872	0.872
	df	6	6	6	6	6	6	6	6	6	6
	p	.028	.008	.00001	.00001	.272	.00001	.00001	.00001	0.15	.015
Test Group 2	Pearson Chi Square	0.872	0.872	0.872	0.872	2.204	0.872	0.872	2.204	0.872	0.872
	df	6	6	6	6	6	6	6	6	6	6
	p	.0006	.0006	.008	.008	0.683	.0006	.00001	0.446	.0006	.0006

Both test groups did not show expected results when presented stimulus C2.

history seemed to affect participants' responses on the questionnaire in Test Groups 1 and 2, X^2 (5, N = 6) = 0.872, p <.001. The responses were not scored as significant to the stimuli C2 in either group, X^2 (2, N = 6) = 2.204 p > .683 in Test Group 1, and C3't in Test Group 2, X^2 (5, N = 6) = 2.204 p > .446. The preexperimental history did not have any effect on the Control Group responses, X^2 (5, N = 9) = 4.168 p > .135. However, the Control Group responses to B3't did show an outlier, X^2 (5, N = 9) = 2.088 p < .001.

Discussion

The purpose of this study was to systematically replicate and extend Watt et al. (1991) by investigating how preexperimental history effects the formation of equivalence classes. The modifications in the present study were: (1) if an equivalence test would show different results; (2) to see if reaction time is a variable that should be included in future studies on preexperimental history; and (3) to see if a bipolar

Table 4 The average RT, per participant on the Transitive relations during Test Block 1 and Test Block 3. The darker fields represent the part of the test in which the participants did not form equivalence classes.

		Reaction Time								
		Test Block 1			Test Block 3					
	Participants	A1C1	A2C2	A3C3'	A1C1	A2C2	A3C3	C1A1	C2A2	C3A3
Control Group	18300	2.8	1.9	3.6	N/A	N/A	N/A	N/A	N/A	N/A
	18301	1.5	4.2	2.3	N/A	N/A	N/A	N/A	N/A	N/A
	18302	1.3	1.5	2	N/A	N/A	N/A	N/A	N/A	N/A
	18303				N/A	N/A	N/A	N/A	N/A	N/A
	18304	1.5	2.7	2.5	1.2	1	1	1	1.8	1.4
	18305				3.8	2.2	1	2.4	1.4	1.2
	18306	1.8	1.8	6.5	1.4	2.2	1.4	1	1.5	1.4
	18307	1.1	2	1.5	1.4	1.4	1	1.2	2	1.2
	18308	4.5	2.6	4.6	2.4	1.8	3	2.2	1.4	2.8
	18309	1.3	1.8	2.6	1.2	1.8	1.8	2.6	1.4	2.4
Test Group 1	18310	2.9	2.3	3	N/A	N/A	N/A	N/A	N/A	N/A
	18311	2.5	3.4	3.5	N/A	N/A	N/A	N/A	N/A	N/A
	18312	1.4	3.9	2.6	1	3.2	1.5	1	2	1.6
	18313	1.4	2.2	2.4	1.2	1.6	1.6	1.2	2	1.2
	18314	1.7	2.6	4.3	1.6	1.6	2.4	1.4	1.8	2
	18315	1.5	3.5	3.4	1	2	1.8	1.25	4.5	1.6
	18316	1.6	2.7	2.4	1.4	4	2.25	2.7	3.4	2.2
Test Group 2	18317	1.7	2	2	4	4.8	1.6	1.8	4.75	1.4
	18318	1.5	2	2.5	1	2.2	1.4	1.6	1.5	1.8
	18319	2.7	4.8	2.5	1.4	2.2	2.7	2.2	2	1.6
	18320	1.6	2.5	4.9	1.8	2.4	2.8	1.4	3.6	2
	18321	2.2	2.7	4.5	1.4	2.4	1.6	1.4	1.8	1.4
	18322	1.6	3.7	2.1	12.5	9	5.5	11.8	4.2	5
	18323	1.6	1.9	4.1	1.4	1.6	1.8	1.4	2	1.8

Some of the participants were not able to do Test Block 3 and are shown as N/A.

questionnaire would add to studies on preexperimental history.

Twenty-four participants were trained in conditional discriminations and tested for the emergence of three 3-member equivalence classes in a linear structure. Fourteen of those participants were taught to match stimuli that were part of their preexperimental history. Testing for emergent relations was done in three blocks: the first block consisted of a transitivity test and generalization test, the second block consisted of an arbitrary generalization test, and the third block consisted of an equivalence test. The study found that participants in the two test groups responded more correctly and quicker than the Control Group (participants with no interest in soccer), suggesting that preexperimental history interferes with the formation of equivalence classes. However, the results did not correspond with the results in Watt et al. and some interesting aspects of preexperimental history were found in relation to RT and trials to criterion.

Trials to Criterion

In Stage 1 The number of trials used per participant to reach the set criterion of the conditional discrimination training varied among the different groups. Participants in the Control Group and Test Group 2 had a consistently lower number of trials to criterion than participants in Test Group 1, with a mean of 66.7 trials to criterion in the Control Group and a mean of 122 trials to criterion in Test Group 2. This suggests that participants in Test Group 1 had a harder time matching than the other groups, with a mean of 165 trials to criterion. Because all participants were exposed to the same stimuli and training structure, these results suggest that the preexperimental history in Test Group 1 may have interfered with the conditioning of new relations. However, because Test Group 2 did not show similar results, one can argue that the order in which stimuli were first presented in training may be a contributing factor to this difference between the test groups. This also corresponds with the results in Watt et al. (1991) where Test Group 1 also consistently scored lower than Test Group 2. It can be argued that there were no such differences in Haydu et al. (2015), which used the OTM training structure. This needs to be investigated further because the difference might be a result of the linear training structure and not because of preexperimental history (Arntzen et al., 2010).

In Stage 2 Whereas the Control Group used almost the minimum number of trials needed to reach criterion in Stage 2, with a mean of 189 trials to criterion, participants in Test Group 1, with a mean of 270 trials to criterion, and Test Group 2, with a mean of 282 trials to criterion, used a considerably higher number of trials to reach the criterion during Stage 2. This suggests that the participants in the two test groups did not maintain the relations they learned in Stage 1 as well as the Control Group. Furthermore, it suggests that preexperimental history also interfered with how they maintained the conditioned relations. It can be argued that the fact that the two test groups had different scores in trials to criterion in Stage 1 is another confirmation of this. As seen in previous studies, different training structures affects the number of repetitions in training and the likelihood that the participants responded in accordance with stimulus equivalence (Arntzen et al., 2010; Holth & Arntzen, 1998). The results in the two test groups does not correspond with the expected number of trials as seen in the theory of linear training structures. However, the results in the Control Group correspond with this theory, which might also explain why some participants in the Control Group did not form equivalence classes.

Equivalence Formation

Equivalence test as in Watt et al. (1991) In the current experiment, 3 out of 7 participants responded in accordance with stimulus equivalence in Test Group 1 (14%), only 1 out of 7 participants passed in Test Group 2 (42%), and 7 out of 10 passed In the Control Group (70%). This corresponds with Watt et al.'s (1991) results. However, a Fisher exact test showed no significant differences between the groups. There was no significance different in how many responded in accordance with stimulus equivalence in either number or percentage. A visual analysis, as seen in Figure 4, seems to show very different results among the Control Group and the test groups, but statistically this difference is not as clear. Although these results do not correspond with what they claim to have found in Watt et al., the number of participants who responded in accordance with stimulus equivalence in test blocks 1 and 2 compared to the Control group are almost identical. Another point about the results in Test Block 1 is that the test does not test for symmetry or equivalence relations (CA). So, at best one can argue that preexperimental history interferes with transitive relations (AC). If we are to gain a fuller understanding of how preexperimental history affects how humans learn, then a fuller test should be conducted.

Equivalence test The results of Test Block 3 did not correspond with the assumption that preexperimental history interferes with the emergence of stimulus classes. Almost all participants in all groups responded in accordance with stimulus equivalence. Five out of six participants in the Control Group passed (83%), four out of five passed in Test Group 1 (80%), and four out of seven passed in Test Group 2 (57%). These results paint an interesting picture that is contrary to Watt et al.'s (1991) results. The results of the Fisher exact test confirmed that there were no significant differences between the groups, which suggests that preexperimental history did not interfere with the emergence of stimulus classes. However, the

point should be made that studying preexperimental history also entails a difficulty in complete across-study generalization due to differing social contexts. Taking that into consideration, we can only assume that the results are a across-study generalization, and therefore more studies on this matter should be conducted in multiple different social contexts in order to confirm or debunk these results.

Another interesting part of these results is that participants who did not pass in Test Block 1 passed in Test Block 3 with no extra training. Studies on preexperimental histories, with an equivalence test as in Test Block 3, has shown that participants do not form equivalence relations (Haydu et al., 2015; Peoples et al., 1998). This suggests that the delay between the two test blocks may be why more participants show the emergence of stimulus classes in Test Block 3 as a delayed emergence of equivalence classes (Arntzen & Nartey, 2018; Holth & Arntzen, 1998). However, a comparison is not possible because these studies used an OTM training structure. More studies should be conducted to investigate how the delay between tests might affect the number of participants who respond in accordance with stimulus equivalence.

Generalization test The generalization test in Test Block 2 was changed and is an extension of the test in Watt et al. (1991). Only colored pictures were originally chosen as stimuli for the generalization test. However, participants in the pilot study responded to the colors of the stimulus instead of the trained relations. One name was therefore intermixed in the baseline test and three names were chosen for the generalization test.

Distinct response patterns were found among the names presented for comparison in the generalization test in Test Block 2. These patterns were distinctly different among the test groups and the Control Group: the Control Group had a mean percentage of 43% correct, Test Group 1 had a mean percentage of 0.9% correct, and Test Group 2 had a mean percentage of 6% correct. All participants failed to show the experimenter defined classes as expected. The response pattern in Test Block 2, if a participant had no relation to the names presented, was expected to be random. Random responding was defined as 30% correct. The test groups consistently responded incorrectly by clicking on A3't in response to every sample, which suggests that their preexperimental history interfered with their responses. Also, it suggests that although they were trained to match A to B and B to C, they did not retain this training when the comparison stimuli were different in the generalization test even when they scored correct on the generalization stimulus in Test Block 1. As expected, the lack of preexperimental history in the Control Group resulted in random responses to the comparison stimuli. This can be explained as being due to their inability to distinguish among the comparison stimuli, because the participants did not have any prior history with them. In contrast, participants in the test groups should have been able to recognize which stimulus fit into which class, but participants still only responded to the A stimuli despite training. This suggests that preexperimental history interferes with the emergence of new relations. More studies need to be conducted to investigate if such a generalization test could be used to test for prior histories to stimuli.

Duration and reaction time The number of minutes spent on the experiment per participant was substantially different between the Control Group and the test groups. The Control Group used a mean of 25 min to finish the experiment. Test Group 1 used a mean of 39 min and Test Group 2 used a mean of 38 min. This corresponds with the difference in the number of trials used per participants during training. In the test groups, some of the participants used almost twice the number of minutes to complete the experiment as participants in the Control Group. Suggesting that RT could be interpreted as a consequence of preexperimental history (Sidman 1994).

RT was also investigated to see if preexperimental history might interfere with the amount of time participants used between clicking on the sample and clicking on one of the three comparisons in the test blocks. The relations that were observed when taking RT into consideration was the transitive (CA) and equivalent relations (AC) in Test Block 1 and Test Block 3. Interesting patterns were found in all groups, and the RT was dependent on whether they passed the test, failed the test, and had any preexperimental history with the stimuli. In all groups, those participants who either had consistent RTs over trials and test blocks, or reduced RTs from Test Block 1 to Test Block 3, passed Test Block 3. Out of these patterns, four participants in Test Group 1, one participant in the Control Group, and two participants in Test Group 2 passed Test Block 1. Those participants who passed Test Block 1 and Test Block 3 had consistent RTs across all relations, except for the generalization relation (A3/C3'), which had slower RTs across almost all participants who passed both test blocks. Another finding was that all participants who had slower RTs in Test Block 3 compared to Test Block 1 failed both test blocks. Furthermore, almost all participants in the test groups had slower RTs than participants in the Control Group.

Some clear patterns were found in the comparison of RTs versus the emergence of stimulus classes. These patterns may be a distinct measure for preexperimental history within stimulus equivalence studies. It can be argued that private behavior that consists of longer chains of relations, such as preexperimental history, could be what interferes with forming equivalence classes. This might explain why participants in the test groups had slower RTs to the comparison stimuli than participants in the Control Group, even when they

responded in accordance with stimulus equivalence (Catania, 2013, pp. 376–390; Sidman 1994; Leppänen & Hietanen, 2004; Vaidya et al., 2015). Furthermore, these results might build upon results by Barnes-Holmes et al. (2005) on semantic priming within the derived stimulus paradigm and how RT can be a measure of indirect priming or direct priming, which could in turn be related to how the responses in Test Block 2 were seen to follow distinct patterns. RT might be a way to study private events indirectly and could shed light on how preexperimental history interferes with emergent relations.

Self-report and comparison In the current experiment, participants mostly reported as expected on the bipolar questionnaire. The Control Group consistently scored the stimuli on the questionnaire with a mean score of 3.1. Test Group 1 and Test Group 2 consistently scored opposite extremes; for example, on the LSK stimuli, Test Group 1 scored them with an average of 1.2 and Test Group 2 scored them with an average of 4.7. Similarly, the opposite extreme was found when Test Group 2 scored the VIF stimuli with an average of 1.6 and Test Group 2 scored the same stimuli with an average of 4.6. These reports also correlated with the number of participants who passed and failed Test Block 1 in each group. Even though more participants in Test Group 2 passed Test Block 1, it still fit with the results from the questionnaire because the self-reports from Test Group 2 were not as clear as the reports from Test Group 1. However, the questionnaire results did not correspond with the results in Test Block 3. Although participants reported themselves as strong opposers of the opposite team, they still managed to respond in accordance with stimulus equivalence with the same stimuli. This leads to the question of how valid questionnaires are, given that they are based on self-reports and not observable behavior (Domeniconi, de Rose, & Perez, 2014; Lane & Critchfield, 1996). Although self-reports are not seen as valid measurements on their own in behavior analysis, they can be valuable in providing general information about participants' expectations of themselves and how it deviates from their behavior. Also, self-reports can be a valuable tool for collecting information about stimuli, and determining which one should be used in stimulus equivalence experiments. This is confirmed by the results in the current experiment.

Participants in the test groups responded on the questionnaire that the soccer players that were used as names during the test, especially VIF_name_1, were generally liked by both soccer teams. In general, all soccer player names that were used were not scored as 5 or 1 as expected on the scale of the questionnaire. It is interesting that the participants in the test groups had the most difficulty with matching VIF_name_1 to the sample stimuli. This suggests that although questionnaires should not be considered a valid way of studying human behavior, they can be sources of information that can guide experimenters to some understanding of how some stimuli might affect participants' responses.

Limitations

More than one experimental setting was used to conduct the test due to a difficulty in recruiting enough participants for each test group. Although the rooms used were similar, the second setting was at a soccer stadium and this in itself might have affected the results. Likewise, a difference in the age and gender between the test groups and the Control Group might have affected the RTs of the participants, because younger participants are more used to computers (Vaportzis, Clausen, & Gow, 2017). However, this might also be a natural consequence of doing studies on preexperimental history as the number of participants are limited.

Another limitation, as mentioned earlier, is the use of a linear training structure, because older studies have proved them to be less efficient than OTM and many-to-one training structures (Arntzen et al., 2010; Sidman, 2000, p. 144). Also, the use of a group design in behavior analysis might be argued to be less valid than some single-case research designs, but the current design was chosen because it was more relevant when doing a systematic replication (Barlow, Nock, & Hersen, 2009; Cooper, Heron, & Heward, 2014). Furthermore, the current design made more sense because a study on preexperimental history is a comparison of social contexts in groups and should be studied in groups.

A limitation might also be found in the analytic descriptions and statistical methods. The analysis has focused on individual descriptions of behavior to show singular differences in the data, especially when the emphasis is on how preexperimental histories interferes with emergent relations. Furthermore, although a power analysis was not conducted in the current experiment, it should not deter others from comparing the current study to future studies.

Other limitations can be found in the stimuli used in the experimenter-defined classes. The stimuli used might not be stimuli associated with the participants' preexperimental histories, as seen in the results for the name VIF_name_1. However, the questionnaire might be a way to counter this problem. Another problem with the stimuli was that there were similar shapes and colors across classes, which means that participants might have been responding to shapes or color instead of training. If that was the case, however, then most participants would have failed on the generalization stimulus in Test Block 1, which was not the case.

Some aspects of the original study were not included such as low-pitched and high-pitched sounds paired with the programmed consequences. Likewise, the pretraining stage was not included. This might be used as an argument against the validity of this study as a systematic replication. However, there are studies that show that a sound paired with the programmed consequences is not necessary for participants to be able to discriminate between correct and wrong responses (Sidman & Tailby, 1982). The results from this study also

confirm that most participants did establish the relations that were trained, although some did not show the emergence of emergent relations, without the pairing of the sounds. Also, studies have shown that pretraining in simple stimulus equivalence experiments do not show any significant difference compared to participants who did not receive pretraining (LeBlanc, Miguel, Cummings, Goldsmith, & Carr, 2003). Furthermore, when preexperimental history is the focus of the experiment a limitation of preexposure to the experimental variables is always preferable.

Summary

Preexperimental history did interfere with the emergence of equivalence classes although almost all participants responded in accordance with stimulus equivalence. Also, the use of a linear training structure made it unclear if participants who did not respond in accordance with stimulus equivalence did not do so due to their preexperimental history or due to the training structure. Participants in the different groups differed in RT during the test although they had similar results in the formation of equivalence classes. RT should therefore be included in future studies on preexperimental history. Participants in the test groups had distinct response patterns that differed from the Control Group during the generalization test. The adapted generalization test should be replicated as a method of testing for preexperimental histories. The questionnaire did not correlate with the results found in the equivalence test, but it gave information that could guide the assumption of certain response patterns to certain stimuli and the generalization test. The results found in this experiment confirm that more studies are needed in this area to challenge the current experiment, previous research, and methods.

Compliance with Ethical Standards Informed consent was obtained from all individual participants included in the study. The authors declare that they have no conflict of interest. Furthermore, that all procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data and Materials The authors declare that data and materials will be shared if requested. The data and materials can be found through contacting the first author.

References

Adcock, A. C., Merwin, R. M., Wilson, K. G., Drake, C. E., Tucker, C. I., & Elliott, C. (2010). The problem is not learning: Facilitated acquisition of stimulus equivalence classes among low-achieving college students. *The Psychological Record, 60*, 43–55. https://doi.org/10.1007/BF03395693.

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

Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The effects of different training structures in the establishment of conditional discriminations and subsequent performance on tests for stimulus equivalence. *The Psychological Record, 60*, 437–461. https://doi.org/10.1007/BF03395720.

Arntzen, E., & Nartey, R. K. (2018). Equivalence class formation as a function of preliminary training with pictorial stimuli. *Journal of the Experimental Analysis of Behavior, 110*, 275–291. https://doi.org/10.1002/jeab.466.

Barlow, D. H., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior for change*.

Barnes-Holmes, D., Staunton, C., Whelan, R., Barnes-Holmes, Y., Commins, S., Walsh, D., et al. (2005). Derived stimulus relations, semantic priming, and event-related potentials: Testing a behavioral theory of semantic networks. *Journal of the Experimental Analysis of Behavior, 84*, 417–433. https://doi.org/10.1901/jeab.2005.78-04.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2014). *Applied behavior analysis* (2nd ed.). England: Pearson.

Critchfield, T. S., & Perone, M. (1993). Verbal self-reports about matching to sample: Effects of the number of elements in a compound sample stimulus. *Journal of the Experimental Analysis of Behavior, 59*, 193–214. https://doi.org/10.1901/jeab.1993.59-193.

Critchfield, T. S., Tucker, J. A., & Vuchinich, R. E. (1998). Self-report methods. In *Handbook of research methods in human operant behavior* (pp. 435–470). New York, NY: Springer. https://doi.org/10.1007/978-1-4899-1947-2_14.

de Carvalho, M. P., & de Rose, J. C. (2014). Understanding racial attitudes through the stimulus equivalence paradigm. *The Psychological Record, 64*, 527–536. https://doi.org/10.1007/s40732-014-0049-4.

Dixon, M. R., Rehfeldt, R. A., Zlomke, K. R., & Robinson, A. (2006). Exploring the development and dismantling of equivalence classes involving terrorist stimuli. *The Psychological Record, 56*, 83–103. https://doi.org/10.1007/BF03395539.

Domeniconi, C., de Rose, J. C., & Perez, W. F. (2014). Effects of correspondence training on self-reports of errors during a reading task. *The Psychological Record, 64*, 381–391. https://doi.org/10.1007/s40732-014-0009-z.

Eilifsen, C., & Arntzen, E. (2009). On the role of trial types in tests for stimulus equivalence. *European Journal of Behavior Analysis, 10*, 187–202. https://doi.org/10.1080/15021149.2009.11434318.

Fields, L., Reeve, K. F., Matneja, P., Varelas, A., Belanich, J., Fitzer, A., & Shamoun, K. (2002). The formation of a generalized categorization repertoire: Effect of training with multiple domains, samples, and comparisons. *Journal of the Experimental Analysis of Behavior, 78*, 291–313. https://doi.org/10.1901/jeab.2002.78-291.

Haydu, V. B., Camargo, J., & Bayer, H. (2015). Effects of preexperimental history on the formation of stimulus equivalence classes: A study with supporters of Brazilian soccer clubs. *Psychology Neuroscience, 8*, 385. https://doi.org/10.1037/h0101276.

Holth, P., & Arntzen, E. (1998). Stimulus familiarity and the delayed emergence of stimulus equivalence or consistent nonequivalence. *The Psychological Record, 48*, 81–110. https://doi.org/10.1007/BF03395260.

Kohlenberg, B. S., Hayes, S. C., & Hayes, L. J. (1991). The transfer of contextual control over equivalence classes through equivalence classes: A possible model of social stereotyping. *Journal of the Experimental Analysis of Behavior, 56*, 505–518. https://doi.org/10.1901/jeab.1991.56-505.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of*

the National Academy of Sciences, 98, 15387–15392. doi:https://doi.org/10.1073/pnas.251541498

Lane, S. D., & Critchfield, T. S. (1996). Verbal self-reports of emergent relations in a stimulus equivalence procedure. *Journal of the Experimental Analysis of Behavior, 65*, 355–374. https://doi.org/10.1901/jeab.1996.65-355.

LeBlanc, L. A., Miguel, C. F., Cummings, A. R., Goldsmith, T. R., & Carr, J. E. (2003). The effects of three stimulus-equivalence testing conditions on emergent US geography relations of children diagnosed with autism. *Behavioral Interventions: Theory Practice in Residential Community-Based Clinical Programs, 18*, 279–289. https://doi.org/10.1002/bin.144.

Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research, 69*, 22–29. https://doi.org/10.1007/s00426-003-0157-2.

McIlvane, W. J. (2013). Simple and complex discrimination learning. In I. G. Madden, W. V. Dube, T. D. Hackenberg, G. P. Hanley, & K. A. Lattal (Eds.), *APA handbook of behavior analysis* (Vol. 2, pp. 129–163). Washington, DC: American Psychological Association. https://doi.org/10.1037/13938-006.

McIlvane, W. J., Kledaras, J. B., Gerard, C. J., Wilde, L., & Smelson, D. (2018). Algorithmic analysis of relational learning processes in instructional technology: Some implications for basic, translational, and applied research. *Behavioural Processes, 152*, 18–25. https://doi.org/10.1016/j.beproc.2018.03.001.

Norsk Senter for Forskningsdata. (2018). Retrieved from http://www.nsd.uib.no

Peoples, M., Tierney, K. J., Bracken, M., & McKay, C. (1998). Prior learning and equivalence class formation. *The Psychological Record, 48*, 111–120. https://doi.org/10.1007/BF03395261.

Sidman, M. (1960). *Tactics of scientific research* (Vol. 16.2). University of Michigan: Basic Books.

Sidman, M. (1994). Equivalence relations and behavior: A research story. Boston, MA: Authors Cooperative.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior, 74*, 127–146. https://doi.org/10.1901/jeab.2000.74-127.

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*, 5–22. https://doi.org/10.1901/jeab.1982.37-5.

Slaton, J. D., Hanley, G. P., & Raftery, K. J. (2017). Interview-informed functional analyses: A comparison of synthesized and isolated components. *Journal of Applied Behavior Analysis, 50*, 252–277. https://doi.org/10.1002/jaba.384.

Vaidya, M., Hudgins, C. D., & Ortu, D. (2015). Conditional discriminations, symmetry, and semantic priming. *Behavioural Processes, 118*, 90–97. https://doi.org/10.1016/j.beproc.2015.05.012.

Vaportzis, E., Clausen, M. G., & Gow, A. J. (2017). Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study. *Front Psychol, 8*, 1687. https://doi.org/10.3389/fpsyg.2017.01687.

Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41*, 33–50. https://doi.org/10.1007/BF03395092.