





# Effectiveness of Different Training and Testing Parameters on the Formation and Maintenance of Equivalence Classes: Investigating Prejudiced Racial Attitudes

Táhcita M. Mizael<sup>1</sup>  · João H. de Almeida<sup>1</sup>  · Bryan Roche<sup>2</sup>  · Julio C. de Rose<sup>1</sup> 

Accepted: 6 September 2020 / Published online: 19 October 2020

© Association for Behavior Analysis International 2020

## Abstract

This study aimed to verify the role of three parameters on the formation of equivalence classes between Black faces and a positive symbol, in children who demonstrated negative bias toward Black faces in a pretest. Maintenance was also verified 6 weeks after equivalence tests. Forty-six children (11 Black; 27 girls) who demonstrated racial bias in a pretest were divided into four groups. All groups first learned AB relations (A1 and A2 were, respectively, a positive and a negative symbol, and B were abstract stimuli) and then BC relations (C1 was a Black face and C2 was an abstract stimulus). The Control Group then advanced immediately to equivalence tests (AC, and CA, without differential consequences). For the Mixed Training Group, a block of trials mixing AB and BC relations, with differential consequences, preceded equivalence tests. For the Feedback Reduction Group, equivalence tests were preceded by a trial block mixing AB and BC relations, but with feedback in 50% of trials. The Symmetry Group received symmetry tests after training of each baseline relation. Thirty-three children showed class formation relating Black faces and the positive symbol, and 27 maintained at least one of the equivalence relations after 6 weeks. Average biases toward Black faces were positive in a posttest, for participants who formed equivalence classes, and remained negative for those that did not form classes. The Control Group showed less pronounced bias reduction and maintenance of relations after 6 weeks, suggesting that these outcomes may be affected by training parameters.

**Keywords** Racial prejudice · Stimulus equivalence · Maintenance of equivalence classes · Training and testing parameters · Racial bias

Táhcita Mizael was supported by a doctoral fellowship from the São Paulo Research Foundation (Grant # 2015/10225-5), and João de Almeida was supported by a postdoctoral fellowship from the São Paulo Research Foundation (Grant #2014/01874-7). Julio de Rose was supported by a Research Productivity Grant from the National Research Council (CNPq). This manuscript is based on one of the studies of the doctoral dissertation presented by the first author to the Graduate Program in Psychology at Universidade Federal de São Carlos (Brazil). This research was part of the scientific program of Instituto Nacional de Ciência e Tecnologia sobre Comportamento, Cognição e Ensino (INCT-ECCE), supported by CNPq (Grants 573972/2008-7 and 465686/2014-1) and FAPESP (Grants 2008/57705-8 and 2014/50909-8). We thank Professor Deisy de Souza, leader of INCT-ECCE, for her encouragement and support for this research, and her careful review of a previous version of this article. We are also indebted to two anonymous reviewers for thoughtful remarks about a previous version of the manuscript.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s40732-020-00435-w>) contains supplementary material, which is available to authorized users.

✉ Julio C. de Rose  
julioderose@yahoo.com.br

<sup>1</sup> Universidade Federal de São Carlos (UFSCar), São Carlos, Brazil

<sup>2</sup> National University of Ireland, Maynooth, Ireland

The stimulus equivalence paradigm (e.g., Sidman, 1994, Sidman & Tailby, 1982) is a well-known method for investigating symbolic relations in the laboratory. The most used procedure to investigate equivalence class formation is called matching-to-sample (MTS). A common version of MTS to investigate equivalence is: in the presence of a sample stimulus (e.g., A1), two or more comparison stimuli are displayed (e.g., B1, B2, Bn), and choices of a specific stimulus (e.g., B1) are reinforced; in the presence of another sample stimulus (e.g., A2, . . . An), choices of B2, . . . Bn, respectively, are reinforced. Next, the B stimuli are now used as samples, so that given B1, B2, . . . Bn, choices of C1, C2, . . . Cn are reinforced. Tests in the absence of differential consequences then verify emergent relations, i.e., relations not directly trained, but derived from training. Demonstration of the formation of  $n$  equivalence classes usually requires three types of tests, typically conducted in the absence of differential consequences. Symmetry tests verify functional reversibility between sample and comparison (given B as a sample, the participant chooses the correspondent A stimulus; given C as a sample, the participant chooses the correspondent B stimulus).

Transitivity tests verify relations between stimuli that were indirectly related (choosing the corresponding C comparison given A samples). Tests for *symmetrical transitivity* verify selections of the A samples given the corresponding C comparisons.<sup>1</sup>

When stimuli are equivalent, they are interchangeable in some contexts, such as when the word “ice cream” can be used to denote an actual ice cream in conversation. Also, when a stimulus has a given function (either acquired in the preexperimental history or experimental training), and this function is extended to equivalent stimuli, this outcome is called transfer of functions. In the literature, there is considerable evidence of transfer of functions such as discriminative, eliciting, or consequential (e.g., de Rose, McIlvane, Dube, Galpin, & Stoddard, 1988; Dougher, Augustson, Markham, Greenway, & Wulfert, 1994; Hayes, Kohlenberg, & Hayes, 1991).

One of the many topics that can be studied with the stimulus equivalence paradigm is prejudice. In this paradigm, prejudice can be conceived as the establishment of equivalence relations between a given stigmatized group or individual and negative attributes. The majority of studies within the behavior-analytic literature about prejudices, biases, and/or stereotypes have used what Mizael, de Almeida, Silveira, and de Rose (2016) called the Conflicting Relations Paradigm (CRP). The CRP is a design in which AB and BC relations are trained and then, AC and CA relations are tested, where both A and C are familiar stimuli and presumably opposite to each other in accordance to sociocultural patterns (e.g., Catholics and Protestants; a positive attribute and Black individuals). In the CRP, therefore, responding in accordance with the experimental equivalence classes conflicts with the participants’ presumed learning history.

Most studies that used the CRP found a low yield of equivalence classes, presumably because the preexperimental history interfered with responding in accordance with the experimental classes (e.g., Barnes, Lawlor, & Smeets, 1996; de Carvalho & de Rose, 2014; Haydu, Camargo, & Bayer, 2015; Moxon, Keenan, & Hine, 1993; Watt, Keenan, Barnes, & Cairns, 1991). Some researchers, however, acknowledge the learning history as an important variable in explaining these results, but also point to the training and testing parameters used as another potential variable that could account at least partially for the results (e.g., de

Carvalho & de Rose, 2014; Mizael et al., 2016; Strand & Arntzen, 2020).

In basic, translational and applied research using the stimulus equivalence paradigm, training parameters can vary widely. Several studies have demonstrated that certain parameters can enhance the yield of equivalence class formation or reorganization (e.g., Adams, Fields, & Verhave, 1993; Arntzen & Holth, 1997; Bortoloti & de Rose, 2009; de Almeida & de Rose, 2015; Fields, Arntzen, Nartey, & Eilifsen, 2012). Applying the findings from the literature abovementioned, de Carvalho and de Rose (2014) recruited four children who demonstrated racial biases in a screening test. These children then learned to relate a positive symbol (A1) to an abstract one (B1), and then that one to faces of Black<sup>2</sup> individuals (C1). In the other class, a negative symbol (A2) was related to an abstract stimulus (B2), which was related to another abstract stimulus (C2). Because this was an exploratory study, different parameters were used for each child. Although just one participant formed the classes, the researchers suggested some parameters that could enhance equivalence class formation.

Building upon this, Mizael et al. (2016) recruited 13 children who showed racial biases in a screening test and added three parameters to the training procedure of de Carvalho and de Rose (2014): a mixed training of the baseline relations after training AB and BC relations, a feedback reduction phase before testing for equivalence class formation, and symmetry tests after training each baseline relation, i.e., the simple-to-complex protocol (Adams et al., 1993). Unlike most studies with the CRP, Mizael et al. found that all 13 children formed equivalence classes. In addition, nine participants continued to relate the positive sample to the Black face as comparison in an additional test (named AC3), in which White faces were also available as choices. In this test, the positive (A1) and negative symbols (A2) were samples, and comparisons were a Black face (C1), an abstract symbol (C2), and a White face (C3). There was also evidence of transfer of functions from the positive symbol to the Black faces: on the pretest, the White faces were considered positive and the Black, negative. However, on the posttest, both faces were considered positive, and the difference between the evaluations of the Black and White faces was no longer statistically significant.

Therefore, in an attempt to identify the necessary and sufficient conditions for equivalence class formation when prospective class-members have conflicting preexperimental functions, the present study aimed to verify the role of three training parameters used in the study of Mizael et al. (2016; mixed training, feedback reduction, and symmetry tests) in the formation of

<sup>1</sup> Choosing the A stimulus given the corresponding C stimulus attests both the properties of symmetry and transitivity and is sometimes called “equivalence” (Fields, Verhave, & Fath, 1984). The mathematical definition of equivalence also requires a demonstration of reflexivity: giving each stimulus as a sample in order to select an identical comparison (cf. Sidman, 1994; Sidman & Tailby, 1982). However, Saunders and Green (1992) argued persuasively that the results of such tests are confused with generalized identity matching, and tests of reflexivity have rarely been conducted in recent research on equivalence.

<sup>2</sup> In Brazil, the terms *branco* (White) and *negro* (Black) are used by the official agencies and the individuals themselves to refer to the race or color of an individual. Because the research was carried out in Brazil, we used the closest translation of the terms.

equivalence classes between Black faces and a positive symbol in children who demonstrated racial bias in a screening pretest. Participants were distributed into four groups. Each group had AB training followed by BC training. Training for three groups included one of the aforementioned parameters. None of these parameters were included for the Control Group (Group C), in which participants were only trained to relate AB and then BC. All participants were then tested for equivalence class formation (AC and CA tests) and then given a modified test (AC3) to check if, in the presence of the positive symbol, participants would continue to choose one of the Black faces even when a White face was available as a third comparison stimulus (along with a Black face and an abstract stimulus). A comparison of pre- and posttest biases was also carried out. To check for maintenance of emergent relations, participants from all groups were tested again for symmetry and equivalence 6 weeks later.

## Method

### Participants

Seventy-eight children were initially screened. Children were only recruited if they demonstrated racial bias in a prescreening test (See Procedure). Hence, 46 children (27 girls), aged 8–10, were recruited from a public school in a medium-sized city in the State of São Paulo (Brazil). Participants were divided into four groups (see Procedure). Eleven children were Black, and the remaining 35 were White.<sup>3</sup> The project was approved by the ethics committee of the University, and parents signed a consent form before each child could engage in the tasks. Table 1 shows the age, sex and participants' skin color in each group.

### Setting, Equipment, and Stimuli

Participants' data were collected in the school's toy library. A Dell Inspiron 14.550 computer, with Intel Core i3 processor, with the MTS III software (Wallace, 2003) presented stimuli and recorded responses. Sessions lasted approximately 10–15 min. All sessions occurred daily and individually, and the experimenter was present and sat next to each participant throughout the study.

Stimuli were the same used in the Mizael et al. (2016) study: A1 was a hand making a thumbs-up sign; in A2, the same hand making a thumbs-down sign; B1, B2, and C2 were abstract stimuli; C1 comprised four pictures of faces of Black people (two men [C1.1 and C1.2] and two women [C1.3 and C1.4]). C3 comprised four faces of White people (also two men and two women: C3.1, C3.2, C3.3, and C3.4). The faces

were shown one at a time; C3 was only used in the screening and in the AC3 posttest. The pictures were obtained from <http://faceresearch.org>, and displayed a face of a White or Black individual, with an apparent age of 20–30 years, with no apparent emotional expression, on a standard light-grey background.

### Procedure

**Matching-to-Sample trials** Samples appeared (one per trial) on the upper center section of the screen. A click on the sample produced the presentation of two or three comparison stimuli (depending on the phase). They appeared in a row on the bottom of the screen, and the sample remained on the screen until a comparison was selected (simultaneous matching). A display of moving stars on the computer screen and a high pitch sound followed correct responses whereas a black screen followed incorrect responses (see Fig. 1). After every correct response during training trials with feedback, the experimenter dropped a marble into a cup. Each marble corresponded to a point. Those were accumulated along the study and exchanged for small prizes at the end. No differential consequences were programmed in test trials. The last phase of training for the Feedback Reduction Group (Group R) had differential consequences in 50% of the trials (see Procedure). The intertrial interval was 1.5 s.

Table 2 shows the training and testing phases, the number of trials, the relations trained/tested, and the learning criterion on each phase.

**Screening: criterion to identify racial bias** The AC3 test is an MTS test format in which comparison stimuli were a Black face (C1), an abstract symbol (C2), and a White face (C3), and samples were the positive (A1) or negative symbols (A2). This test was a 16-trial block with no programmed differential consequences. Participants were instructed to select, in each trial, the comparison stimulus that matched the most with the sample. Thus, they had to select among the Black face, White face, and the abstract symbol. The male and female faces were presented an equal number of times, in a randomized sequence.

The same equation used by Mizael et al. (2016)'s was employed to identify bias ( $b$ ):

$$[b = [(W+) - (B+)] + [(B-) - (W-)]x(-1)] * -1$$

where  $b$  is the bias index,  $W+$  and  $W-$  stand for frequency of selections of White faces for, respectively, the positive and negative samples, and  $B+$  and  $B-$  are frequencies of selections of Black faces for the positive and negative samples, respectively. For ease of interpretation, a multiplication by  $-1$  was performed, so that positive values indicated positive biases towards Black faces and negative values indicated negative

<sup>3</sup> The participants' skin color classification was given by the experimenter (heteroattribution), who is a Black woman.

**Table 1** Participants' Description (Age, Sex Assigned at Birth and Skin Color), Separated by Group

Participant	Sex	Age	Skin Color	Participant	Sex	Age	Skin Color
M1	F	10	W	S1	F	9	W
M2	M	10	W	S2	M	10	W
M3	F	10	W	S3	M	9	B
M4	M	10	B	S4	F	8	W
M5	M	8	B	S5	F	8	W
M6	F	9	W	S6	M	10	W
M7	F	10	W	S7	M	9	W
M8	M	8	W	S8	F	9	B
M9	F	8	W	S9	F	9	W
M10	F	10	B	S10	F	10	W
M11	F	9	W	S11	M	10	W
M12	F	10	W	S12	F	10	W
R1	M	9	W	C1	F	9	W
R2	F	9	B	C2	M	10	W
R3	M	10	W	C3	M	8	W
R4	F	8	W	C4	M	9	W
R5	F	8	W	C5	M	9	W
R6	F	10	W	C6	F	9	W
R7	F	10	W	C7	F	8	W
R8	M	10	B	C8	F	10	W
R9	F	8	B	C9	F	8	W
R10	M	9	B	C10	M	10	B
R11	M	10	B	C11	F	10	W

Note. M = mixed training group; R = feedback reduction group; S = symmetry test group; C = control group; F = female; M = male; W = White B = Black

biases towards the same faces. An index of zero designated absence of bias. Participants were only recruited if a bias index of -4 or lower was obtained.

**Pretraining** A pretraining familiarized participants with the MTS procedure. Familiar stimuli were used to train three arbitrary relations (X1Y1, X2Y2, and X3Y3). These stimuli were only used in this phase. Trial blocks had 15 trials each, 5 with each sample (X1, X2, X3), in a randomized sequence. Each trial presented the three comparison stimuli (Y1, Y2, and Y3). The learning criterion to advance to the training was at least 14 correct trials in a block. If a participant did not reach criterion in five trial blocs, his or her participation ended.

**Training and Symmetry Tests** During all training phases, blocks had 16 trials and the criterion was at least 15 correct choices in two consecutive blocks. The learning criterion on symmetry tests was a maximum of one error in a block (the training/test sequence would be repeated if the criterion was not attained, but this never happened in this study). The

sequence of trials in each training block was randomized with the restriction that all samples occurred equally often.

Training for Group C began with the AB relation, until participants reached the criterion. The BC relation was then trained until criterion. Tests were conducted immediately after. The other groups had different types of additional training phases (see Table 2). The Symmetry Test Group (Group S) had tests for symmetry after achieving criterion on each of the baseline relations. The Mixed Training Group (Group M) had a mixed training of AB and BC relations after achieving criterion for the BC relation; the R group also had this mixed training of AB and BC relations, with the difference that differential feedback occurred in 50% of trials in a randomized sequence.

There was no maximum number of blocks to terminate a child's participation. To guarantee that participants voluntarily wanted to continue doing the tasks, the experimenter frequently asked if they wanted to keep going.

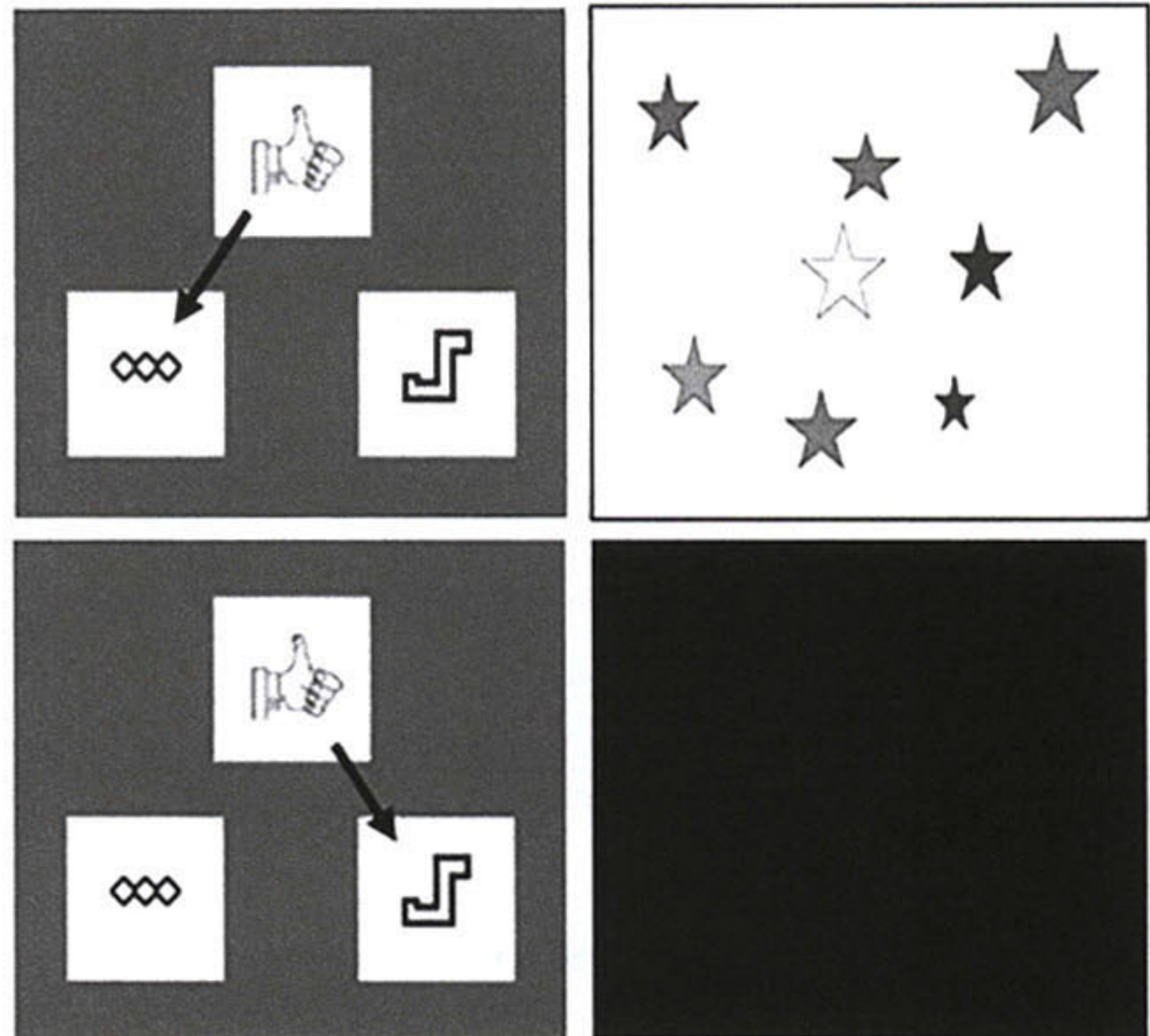
**Equivalence Tests** These tests verified the emergence of AC and CA relations. Two 16-trial test blocks (16 AC and 16 CA) were conducted twice with each participant, comprising a total of four testing blocks. The criterion for immediate equivalence class formation was at least 15 correct trials in at least three test blocks. Equivalence tests were always conducted one day after attaining criterion on the last training phase (mixed training for the M group, feedback reduction training for the R group, CB symmetry for the S group, and BC training for the C group).

**Posttests** The AC3 posttest followed the equivalence tests. This test was identical to the AC3 pretest. It aimed to check if participants that formed the intended equivalence classes would change their responses when a White face was available as a third comparison-stimulus in AC test trials. Participants were tested in two 16-trial blocks.<sup>4</sup>

**Maintenance Tests** All participants were tested for symmetry, transitivity, and equivalence 6 weeks after the initial test. The maintenance test was comprised of a 64-trial block, with 32 symmetry trials (16 BA and 16 CB) and 32 equivalence trials (16 AC and 16 CA). Baseline relations were not included in the tests to avoid retraining the relations (Rehfeldt & Root, 2004; Saunders, Wachter, & Spradlin, 1988b). This test had no programmed differential consequences. Criterion for maintenance was at least 14 correct responses in each relation. Note that BA and CB relations had been tested only for Group S. For the other

<sup>4</sup> As in the study of Mizacl et al. (2016), the AC3 test was followed by the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006) to assess racial bias. The IRAP results revealed no racial bias, but they are difficult to interpret because no similar measure was taken in the pretest. For this reason, the procedure and results with the IRAP are presented in the Online Resource.

**Fig. 1** Display of training trials. The top panel displays the programmed consequence for correct responses (e.g., A1B1), and the bottom panel displays the programmed consequence for incorrect responses (e.g., A1B2). The black arrows were not shown during trials



groups, the term *maintenance* is not strictly applicable to these tests, although we may surmise that participants who passed AC and CA tests had derived the prerequisite symmetrical relations.

## Results

### Dependent Variables

Main dependent variables were changes in the bias index (*b*) between pretest and posttest, yield of equivalence classes for each group (number of participants that did or did not show equivalence classes), and percentage of participants that maintained the classes after 6 weeks. Performance in the AC3 posttest was also assessed.

### Equivalence Yield and Maintenance

Tables 3, 4, 5, and 6 show participants' performances during equivalence tests, AC3 posttests, and maintenance tests. Two of 46 participants did not finish the training phase. R11 decided voluntarily to withdraw, and M12 said she wanted to quit after an inquiry by the experimenter. Therefore, a total of 44 participants learned the baseline relations and went through equivalence tests and AC3 posttests. This report will focus only on the results for equivalence and maintenance. However, results for acquisition are also presented in the Online Resource (Tables A, B, C, and D).

Thirty-three participants formed the equivalence classes in accordance with the proposed criterion (71.7% of participants overall). Test results are described for each group.

**Mixed Training Group** Table 3 shows that eight participants (M1–M8) attained criterion for equivalence class formation. On the AC3 posttest, five participants (M1–M5) maintained their responses by selecting the Black faces given the positive symbol, even with a White face available as comparison. The remaining participants related the positive symbol with the White faces in 32% of the trials (see Table Y in the Online Resource). In the maintenance tests, four of the eight participants that had formed the classes attained criterion in the four relations tested: M1, M2, M6, M7.

**Feedback Reduction Group** As Table 4 shows, eight participants (R1–R8) demonstrated equivalence class formation. On the AC3 posttest, five participants (R1–R5) maintained their performances; R4 achieved the criterion only on the second AC3 test, falling short of criterion on the first test by one point. Participants who did not achieve criterion on this test related the positive symbol with the abstract stimulus in 33.3% of trials and with the White faces in 29.4% of trials (see Online Resource, Table Y). Six participants attained criterion for maintenance: R1, R2, R3, R4, R6, and R7.

**Symmetry Group** Table 5 shows that 10 participants (S1–S10) demonstrated equivalence class formation in all four tests. Five participants (S1, S2, S6, S7, and S8) maintained their responses on the AC3 posttest; most responses inconsistent

**Table 2** Training and Testing Sequence, Containing the Number of Trials, Trained/Tested Relations and Learning Criterion in Each Phase

Phase	Trials	Criterion	Relations
Pretest	15	---	AC
Pretraining	15	15 of 16 trials	X1Y1/X2Y2/X3Y3
AB Training	16	15/16 trials (2x)	A1B1/A2B2
BA Symmetry Test (Group S only)	16	15 of 16 trials	B1A1/B2A2
BC Training	16	15/16 trials (2x)	B1C1/B2C2
CB Symmetry Test (Group S only)	16	15 of 16 trials	C1B1/C2B2
ABBC Mixed Training (Group M only)	16	15/16 trials (2x)	A1B1/A2B2/ B1C1/B2C2
Baseline Review 50% feedback (Group R only)	16	15/16 trials (2x)	A1B1/A2B2/ B1C1/B2C2
AC Test	16	15 of 16 trials	A1C1/A2C2
CA Test	16	15 of 16 trials	C1A1/C2A2
Posttest	16	---	A1C1/A2C2
Maintenance Tests	16	15 of 16 trials	B1A1/B2A2 C1B1/C2B2 A1C1/A2C2 C1A1/C2A2

Note. 2x means each participant had to achieve this criterion in two consecutive blocks

with training were choices of the White face given the positive symbol (60.6% of trials; see Table Y, Online Resource). Five participants (S1–S5) showed maintenance of all relations.

**Control Group** Table 6 shows that seven participants (C1–C7) formed the intended equivalence classes on all four tests. The main difference between this and the other groups' performances seems to be in the responses on the AC3 posttest: of the seven participants who formed the equivalence classes, only two (C1 and C4) maintained their previous responses, relating the Black faces with the positive symbol. The other participants related, in most trials, the positive symbol with the abstract stimulus (34.8% of trials) or with the White faces (30.2% of trials; see Table Y in the Online Resource). Only

two participants (C1 and C2) showed maintenance of all relations.

Table 7 summarizes the outcomes of the four groups in AC/CA, AC3, and maintenance tests, as compared with the reference study by Mizael et al. (2016). For each of these tests, the table presents the percentage of participants in each group that attained criterion. Maintenance data are presented as two percentages: "% tot" is the percentage of participants who attained criterion for maintenance, considering the total number of participants in the group; "% pass" is the percentage of participants who formed classes that maintained them 6 weeks later.

Groups R, S, and M had approximately similar outcomes for equivalence, AC3, and maintenance (with Group R showing slightly higher scores for maintenance). Performance of Group C was noticeably poorer in all tests, in particular in the AC3 and maintenance tests.

**Table 3** Group M's performances during testing

P	AC/CA Tests				P/F	AC3 Tests		Maintenance Tests			
	BA	CB	AC	CA		BA	CB	AC	CA		
M1	100	100	100	100	P	100	100	100	100	100	100
M2	100	100	100	100	P	100	100	100	94	100	100
M3	94	100	100	100	P	100	100	94	69	100	100
M4	100	100	100	100	P	100	100	100	50	0	0
M5	94	100	100	100	P	100	100	13	31	100	100
M6	100	100	100	100	P	69	81	100	94	100	100
M7	75	100	100	100	P	44	31	100	100	100	100
M8	100	100	100	100	P	13	19	100	13	6	6
M9	19	25	56	44	F	6	6	100	13	38	63
M10	56	0	0	6	F	13	0	6	0	100	94
M11	6	0	0	0	F	0	0	6	100	13	0

Note. P = participant; P/F = passed or failed the initial equivalence tests

### Comparison of AC3 Pre- and Posttests

To make this comparison, the difference in the bias value ( $b$ ) in the pre- and post-AC3 tests was calculated: to obtain the change in bias, the  $b$  value in the pretest was subtracted from the value in the posttest. Positive values of variation indicate that negative bias toward Black faces reduced in the posttest, whereas negative values indicate an increase in negative bias towards Black faces.<sup>5</sup>

Figure 2 shows the change in bias for all participants. Each bar corresponds to a participant. For each group, participants

<sup>5</sup> It is important to note that a decrease in negative bias towards Black faces is indicated by an increase in the value of the bias index  $b$ , because  $b$  is an index of negative bias (more matchings of the Black faces to the negative symbol as compared to the matchings of the White faces with those symbols).

are ordered from the most positive changes to the less positive (or most negative). Black bars correspond to participants that passed equivalence tests, whereas gray bars correspond to those that did not pass. The hash identifies participants that showed a bias value of  $-4$  or lower in the posttest, and that therefore would still meet the criterion for inclusion in the study.

All participants began the study with a bias value of  $-4$  or lower. Figure 2 shows a positive variation for 37 of the 44 participants. Their negative biases toward Black faces measured by the AC3 test decreased by one to a maximum of 19 points. Only 11 participants would meet criterion for inclusion in the study in the AC3 posttest ( $b \leq -4$ ).

Of the 33 participants that attained criterion for equivalence class formation, only 2 showed an increase in negative bias towards Black faces. Two participants did not change their bias value, whereas 29 showed a decrease in negative bias. For the 33 participants that formed the classes, their negative bias was reduced (i.e.,  $b$  values *increased*) by an average of 8 points (median of 12). For the 11 children that did not form classes, there was also a decrease in negative bias (*increase* in the  $b$  value). This decrease in negative bias for participants who did not form classes was much smaller, with an average of 2.9 points (median of 3).

Table 8 clarifies the choice patterns underlying the decrease in negative bias towards Black faces in the AC3 posttest. The table presents, for each group, choices in the AC3 pre- and posttest among the three comparison stimuli (White faces, Black faces, abstract stimulus) when the positive and negative symbols were samples. In the pretest, when the sample was a positive symbol, participants from all groups chose the White faces in more than 50% of the trials. In the posttest, selections of the White face for the positive sample varied roughly between 20 and 40%. Therefore, the equivalence relation between the positive symbol and the Black faces that emerged after training was, to a certain extent, disrupted when the White face was available as a comparison for trials with the positive sample (see Tables 3, 4, 5, and 6). Nevertheless, the Black face was the most frequent choice for the positive sample in the AC3 posttest for participants in all groups (from a minimum of 36% for Group C to a maximum of 59% for Group M). For the negative symbol as a sample, the Black face was the preferred choice during the pretest, with group preferences ranging roughly around 50–70%. In the posttest, the preference for all groups when the sample was the negative symbol changed toward the abstract symbol, in accordance with the equivalence relation that had been established.

A general bias measure ( $b$ ) has been used up to this point in the description of the results. Henceforth we will focus on separate bias measures for the White and the Black faces, respectively. To calculate a bias index toward the White face, the number of times each participant chose a White face given the negative symbol was subtracted from the number of times

**Table 4** Group R's Performances during Testing

P	AC/CA Tests				P/F	AC3 Tests		Maintenance Tests			
	BA	CB	AC	CA		BA	CB	AC	CA		
R1	100	100	100	100	P	100	100	100	100	100	100
R2	100	100	100	88	P	94	100	100	100	100	100
R3	100	100	94	100	P	100	94	88	88	100	100
R4	100	100	100	94	P	88	100	94	88	100	94
R5	94	100	94	81	P	94	94	100	50	100	100
R6	100	100	100	94	P	63	81	100	100	100	100
R7	100	94	100	100	P	56	44	100	94	100	100
R8	38	100	100	94	P	38	13	75	94	81	94
R9	44	88	100	88	F	19	31	100	100	13	6
R10	19	19	25	13	F	0	0	94	75	0	0

Note. P = participant; P/F = passed or failed the initial equivalence tests

this face was chosen for the positive symbol, so that positive values indicated positive biases toward the White faces. A similar calculation was performed for the Black faces.

Figure 3 compares the mean bias values on pre- and posttest for participants who formed the intended equivalence classes and participants who did not form them. A one-way ANOVA comparing pretest and posttests evaluations of Black and White faces revealed main effects [ $F(7,172) = 25.9$ ;  $p < 0.0001$ ]. An Uncorrected Fisher's LSD allowed for multiple comparisons among the results in the pre- and posttest.

As it can be seen in Fig. 3, for participants who formed the classes, there was a statistically significant difference between the evaluations of Black faces on pre- and posttest. On pretest,

**Table 5** Group S's Performances during Testing

P	AC/CA Tests				P/F	AC3 Tests		Maintenance Tests			
	BA	CB	AC	CA		BA	CB	AC	CA		
S1	100	100	100	100	P	100	100	100	100	100	100
S2	100	94	100	100	P	100	94	100	100	100	100
S3	100	100	100	100	P	56	56	94	100	100	100
S4	100	100	100	100	P	63	56	100	94	100	100
S5	100	100	100	100	P	31	75	100	100	100	100
S6	100	100	100	100	P	94	100	100	31	75	94
S7	100	100	100	100	P	100	94	0	25	94	100
S8	100	100	100	100	P	94	100	94	0	0	6
S9	100	94	100	100	P	69	38	100	31	0	0
S10	100	100	100	100	P	0	0	38	25	44	50
S11	19	19	44	31	F	25	31	56	63	9	50
S12	38	6	0	0	F	6	6	100	6	0	0

Note. P = participant; P/F = passed or failed the initial equivalence tests

**Table 6** Group C's Performances during Testing

P	AC/CA Tests				P/F	AC3 Tests		Maintenance Tests			
	BA	CB	AC	CA		BA	CB	AC	CA		
C1	100	100	100	100	P	100	100	100	100	100	94
C2	94	100	100	100	P	88	81	94	100	94	94
C3	100	100	100	100	P	44	63	100	50	94	94
C4	100	100	100	100	P	100	100	0	25	94	100
C5	94	94	100	100	P	56	44	75	100	100	100
C6	100	100	100	100	P	81	69	6	6	100	100
C7	100	94	100	100	P	44	50	100	19	44	13
C8	50	25	13	25	F	6	13	100	50	19	31
C9	44	0	0	0	F	38	50	94	6	0	0
C10	0	0	0	0	F	1	1	100	0	6	0
C11	0	0	0	0	F	0	0	81	0	0	0

Note. P = participant; P/F = passed or failed the previous test

those faces were more related with the negative symbol than the positive one, and in posttest, they were more related with the positive symbol than with the negative one ( $p < 0.0001$ ). Regarding the White faces, there was a small difference between the evaluations on pre- and posttest (means in the posttest were less positive), but this difference was not statistically significant ( $p = 0.2003$ ). For participants who did not form the classes, Black faces were considered negative both in pre- and posttest, although means were less negative in the posttest. However, this difference was not statistically significant ( $p = 0.1364$ ). There was a decrease in the White faces' evaluations on the posttest, compared to pretest, and this difference was statistically significant ( $p = 0.0409$ ). The Uncorrected Fisher's LSD test showed significant effects for the evaluations of Black faces for participants who formed the classes ( $p < 0.0001$ ) and also for the evaluations of the White faces for participants who did not form the classes ( $p < 0.05$ ). All the remaining comparisons were not significant.

Data in Fig. 4 are averages for each group. As evidenced in the statistical analyses, for all groups, the differences in the faces' evaluations were statistically significant only on the pretest. Also, no gender bias was identified on pre- or posttest. That is, participants did not significantly choose one gender more than another during those trials.

Comparing the bias levels for White and Black faces in the pre- and posttest, the Mann-Whitney test showed significant differences in the pretest for all groups ( $U = 8$ ;  $p < 0.001$  for the M Group;  $U = 1$ ;  $p < 0.01$  for the R Group;  $U = 0$ ;  $p < 0.001$  for the S Group; and  $U = 3.5$ ;  $p < 0.001$  for the C Group). In the posttest, differences for all groups were no longer statistically significant ( $U = 54$ ;  $p > 0.05$  for the M Group;  $U = 34.5$ ;  $p > 0.05$  for the R Group;  $U = 71.5$ ;  $p > 0.05$  for the S Group; and  $U = 58.5$ ;  $p > 0.05$  for the C Group).

## Discussion

The aim of this study was to evaluate the effectiveness of three training and testing parameters (mixed training of baseline relations, feedback reduction training and symmetry tests) on the formation of equivalence classes between Black faces and a positive symbol in children who demonstrated racial bias in a screening pretest. Data showed that the use of any of the three parameters increased the number of participants that formed the classes and that reduced their racial biases from pre- to posttest.

### Equivalence Class Formation

Regarding equivalence class formation, data showed that the group with most participants forming the intended classes was the S Group: 10 of 12 participants demonstrated equivalence class formation in all four tests (83% of participants). On the other groups, the percentage of participants who formed the classes varied between 63% (Group C) and 72% (Group M). Indeed, some work published has already pointed out the symmetry tests as an important variable to the formation of equivalence classes. Sidman, for instance, reported some cases in which participants formed the classes only after demonstrating symmetry of the trained relations (e.g., Sidman, Wilson-Morris, & Kirk, 1986; Sidman, 1994). Other researchers obtained similar results, pointing to the emergence of symmetry relations as a condition for the formation of equivalence classes (e.g., Fields, Adams, Newman, & Verhave, 1992; Saunders et al., 1988a, b; Stromer & Osborne, 1982). Overall, the yield of equivalence classes in this study was higher than in most studies using the CRP. However, the yield was not as high as the one obtained in the study by Mizael et al. (2016), who combined all three parameters of the present study and found that all 13 participants showed equivalence class formation.

**Table 7** Percentage of Participants from Each Group Who Attained Criterion on the Three Different Tests (Equivalence, AC3, and Maintenance), as Compared with the Reference Study by Mizael et al. (2016)

	Reference Study (%)	M (%)	R (%)	S (%)	C (%)
AC/CA	100	73	80	83	64
AC3	69	45	50	42	18
Maintenance (% tot)		36	60	42	18
Maintenance (% pass)		50	63	50	29

Note. In the reference study (Mizael et al., 2016) the experimenters did not measure maintenance of the relations. M = mixed training group; R = feedback reduction group; S = symmetry group; C = control group; For maintenance, % tot is percentage of the total number of participants in the group who showed maintenance, and % pass is percentage of participants that passed AC/CA tests who showed maintenance



It has been assumed that typical studies with the CRP showed a low yield of equivalence classes because the preexperimental relations interfered with the experimentally designed classes. However, as noted, for instance, by Mizael et al. (2016), most studies with the CRP lacked symmetry tests, baseline revisions and feedback reduction. The present results suggest that adding at least one of these parameters may increase the equivalence yield. It is premature, however, to conclude that this is the only variable responsible for the difference in yield. Most studies with the CRP were carried out with adults, who presumably have a longer history with preexperimental relations that conflict with the experimentally designed classes. Hence, future studies should investigate the effect of training parameters in studies with adults. Also, there is a possibility, in the present study, that the higher yield for Groups M and R is due to the added number of training trials in the mixed baseline (AB/BC) phase.

### AC3 Test

If results in the AC3 test are considered in terms of participants who achieved criterion or not, then the results show that the AC transitive relations that emerged in the training context were, to a certain extent, disrupted when a White face became also available as a comparison. However, the results indicate that the AC3 outcomes cannot be considered in such dichotomous fashion. Transitive relations that emerged from training were those between the negative sample and the abstract comparison, and between the positive sample and the Black face. Table 7 shows that these were still the predominant choices in the AC3 posttest, even if they were not exclusive. White faces were frequently matched to the positive sample in the AC3 posttest, albeit less frequently than the Black faces. Also, Black faces were matched to the negative sample, although at a much lower frequency than during the pretest. It seems that the results of the AC3 posttest are interpreted more accurately when, rather than taking results in an all-or-none fashion, the frequencies of selections of Black and White faces are used to construct indices that vary along a continuum, such as the  $b$  value used as a criterion to select participants and to generate data in Fig. 2, as well as the specific bias indices used to generate data for Figs. 3 and 4. Individual data in Fig. 2 show some cases in which individual performance deviate from the group averages, such as the data from participant S10, who formed the classes but showed a large increase in negative bias on posttest. Some participants that did not pass the tests for equivalence class formation, on the other hand, showed relatively large values of bias reduction, such as S12 and R9. These discrepancies are probably due, in part, to false negative results in equivalence tests, such as for R9, who nearly attained criterion.

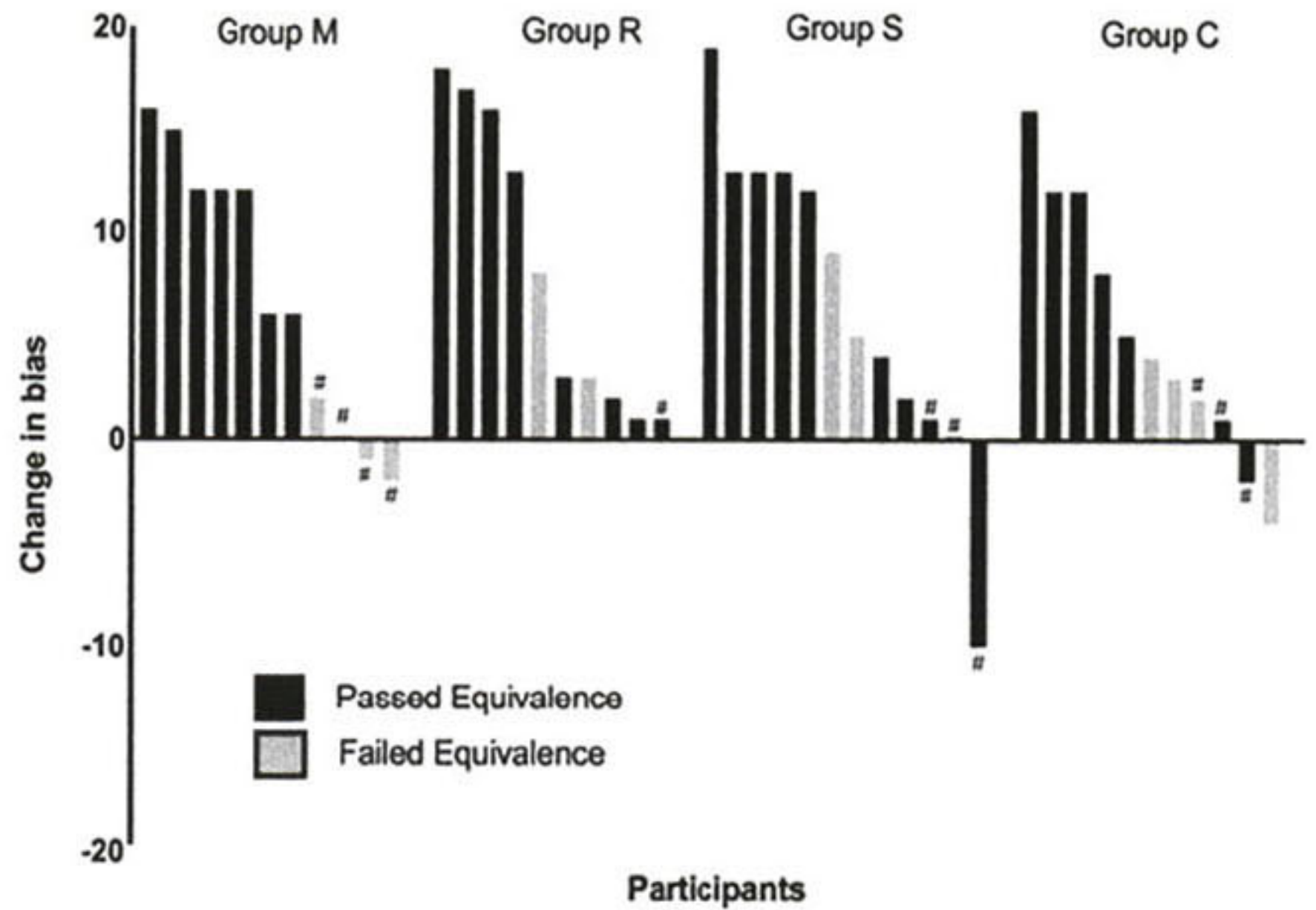
### Comparison of Bias on Pre- and Posttest

The comparison of bias levels on pre- and posttest showed that there was a bias reduction after training for all groups. Statistical analyses that compared the evaluations of White and Black faces before and after training showed a statistically significant difference in the evaluations on the pretest but not on the posttest. The evaluations of the Black faces from all participants who formed the classes in all groups were highly negative in the pretest and became positive and not significantly different from the White faces' evaluations in the posttest. For participants who did not form the classes, on the other hand, biases toward Black faces remained negative in the posttest. However, even for these participants there was still a visually apparent reduction of the negative bias towards Black faces, although it was not statistically significant. For these participants, there was also a statistically significant reduction in the evaluation of the White faces. The reason for this is not clear. This may be related to the forced choice requirement of the test in which participants had to choose only one comparison in each trial. In this sense, it could be important to delineate, in future studies, a test that allows a participant to choose more than one comparison in the same trial. An example could be the use of a sorting procedure (e.g., Amd, de Oliveira, Passarelli, Balog, & de Rose, 2018; Fields, Arnzten, & Moskness, 2014; Varelas & Fields, 2017), where participants could sort together all stimuli they believed to go together.

Group C was the one with the lowest correlation between equivalence class formation and reduction of negative biases. A hypothesis that could help to understand the low performance of participants of this group is that they received the lowest number of training trials/exposures (the other groups were exposed to at least 32 trials more than this group). In the literature, this variable (number of training trials) is deemed an important variable to increase both the yield of equivalence class formation and transfer of functions (e.g., Bortoloti, Rodrigues, Cortez, Pimentel, & de Rose, 2013). Therefore, future investigations should examine parametric variations with equal number of training trials to check this possibility.

Because group C had a smaller number of training trials, participants of this group also had less exposure to Black faces than the other groups during training. This could be a factor responsible for the relatively poorer outcomes of this group in terms of bias reduction. However, results of Figs. 2 and 3 indicate that equivalence formation was the major determinant of bias reduction. For participants that formed the classes, an equivalence relation between Black faces and a positive symbol emerged, whereas this relation did not emerge for participants that did not show class formation. When participants are grouped by equivalence outcome, regardless of training condition, the group that formed equivalence shows a dramatic change in bias toward Black faces, as shown in Fig. 3. The

**Fig. 2** Change in the bias value (*b*) between pre- and posttest. Each bar represents the subtraction of the value of *b* in the pretest from *b* in the posttest, for an individual participant. Participants within each group are ordered from the highest to the least positive change in bias. The hash indicates participants whose *b* value in the posttest remained equal to or below -4 and, therefore, would still meet criterion for inclusion in the study



group that did not form equivalence, regardless of training condition, do not show a significant bias reduction. These results indicate that class formation, rather than amount of exposure to Black faces was the main variable accounting for change in bias. The individual changes in bias, depicted in Fig. 2 confirm that, despite some individual variations and a few discrepant results, equivalence formation rather than amount of exposure to Black faces accounts for change in bias. This is clear because the participants of Group C that did show class formation display a pattern of change in bias similar to the other groups, despite having less exposure to Black faces.

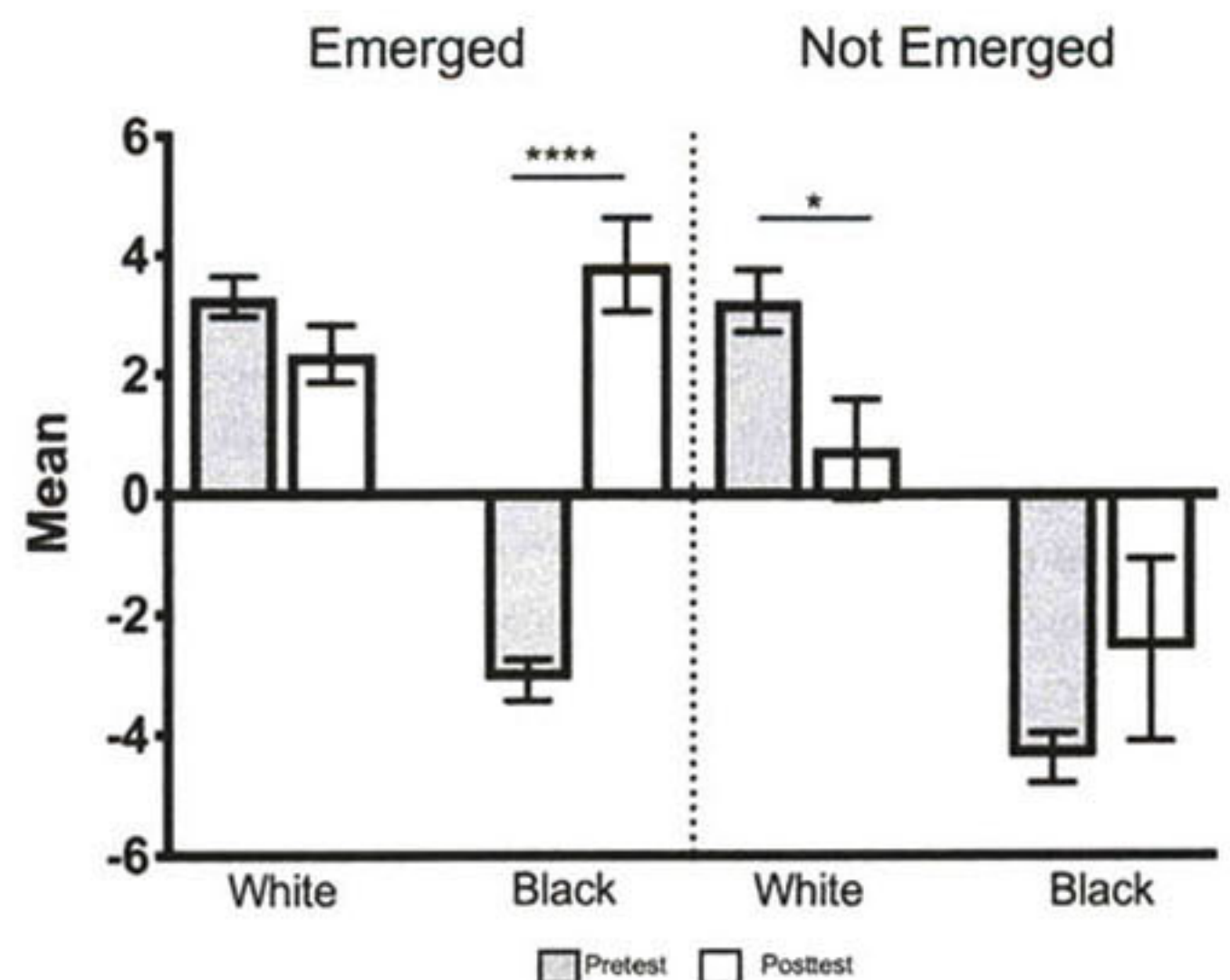
**Maintenance**

From the 33 participants that formed the classes (in all groups), 27 (81.8%) maintained at least one of the equivalence relations. In total, 17 participants maintained all four relations (51.5%). It is interesting that some participants that had not shown equivalence class formation initially, performed in accordance with symmetry, transitivity, or equivalence relations in the maintenance test: from the 13 participants who did not attain criterion, 8 (61.5%) demonstrated symmetry (M9, M11, R10, S12, C8, C9, and C10) or equivalence (M10).

The emergence of BA relations on the maintenance test for participants that did not form the classes can be explained taking into consideration the type of stimuli used on training.

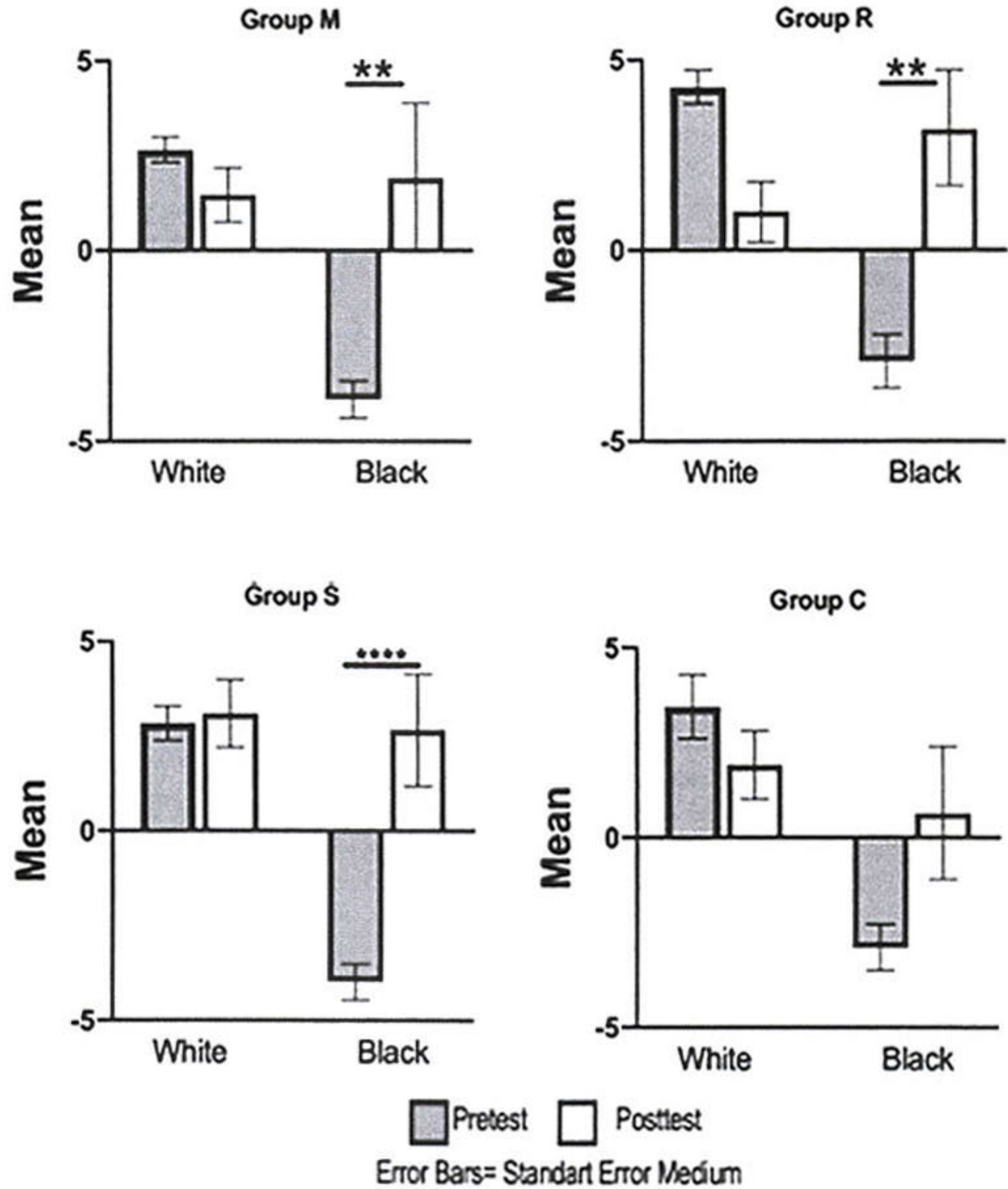
**Table 8** Number and Percentage of Trials in which Participants Choose Each Comparison Stimulus for the Positive and Negative Symbols as Samples in the AC3 Pretest and Posttest

Group		Positive Symbol		Negative Symbol	
		Pre	Post	Pre	Post
M	White	54 (56.2%)	24 (27.2%)	21 (21.8%)	8 (9%)
	Black	21 (21.8%)	52 (59%)	67 (69.7%)	31 (35.2%)
	Abstract	21 (21.8%)	12 (13.6%)	8 (8.3%)	49 (55.6%)
R	White	65 (73.8%)	17 (21.2%)	18 (20.4%)	7 (8.7%)
	Black	11 (12.5%)	46 (57.5%)	43 (48.8%)	14 (17.5%)
	Abstract	12 (13.6%)	17 (21.2%)	27 (30.6%)	59 (73.7%)
S	White	56 (58.3%)	41 (42.7%)	22 (22.9%)	4 (4.1%)
	Black	16 (16.6%)	50 (52%)	64 (66.6%)	18 (18.7%)
	Abstract	24 (25%)	5 (5.2%)	10 (10.4%)	74 (77%)
C	White	51 (57.9%)	26 (29.5%)	13 (14.7%)	5 (5.6%)
	Black	14 (15.9%)	32 (36.3%)	46 (52.2%)	25 (28.4%)
	Abstract	23 (26.1%)	30 (34%)	29 (32.9%)	58 (65.9%)



**Fig. 3** Evaluations of Black and White faces, for all participants who formed vs. did not form the classes, on pre- and posttest. \*= $p < 0.05$ . \*\*\*\*= $p < 0.0001$

**Fig. 4** Evaluations of Black and White faces, for all groups, on pre- and posttest. Each bar is the average of the number of times each participant matched the respective face to the positive stimulus minus matches to the negative stimulus



Although A was a familiar stimulus, its relation with B, an abstract stimulus, is not a conflicting relation in terms of the preexperimental history. The emergence of CB relations for M11, as well as transitivity and symmetrical transitivity for M10, is probably an artifact of a response pattern in which participants *arbitrarily assigned* each comparison to one sample (cf. Saunders, Saunders, Kirby, & Spradlin, 1988a). In a test with two samples and two comparisons, such a pattern will produce scores of either 100% or 0% for different relations, which is consistent with test results of several participants both in equivalence and in maintenance tests.

In summary, the high percentage of participants that maintained their performances on the maintenance test is a positive outcome, especially when we consider the social relevance of this work. Thus, because Group C showed poorer maintenance outcomes, as shown in Table 7, we may conclude that the use of the training parameters increased the number of participants that maintained their performances 6 weeks after the initial tests.

### Implications for the Investigation and Reduction of Racial Prejudice

The preexperimental biases shown by all children recruited for this experiment are an indication of racial prejudice or racism. Guerin (2005) cautioned against the assumption that racism is a cause (possibly internal) of an enormous diversity of overt behaviors. Racism and other prejudices refer to numerous behaviors that do not have a unique cause and may be related to many different variables.

The CRP may be a useful experimental procedure to tackle one of the aspects underlying many forms of prejudiced behavior: equivalence relations between human groups and negative attributes. Thus, the CRP is an experimental model of derivation of relations that conflict with prejudiced *beliefs* or *attitudes*. The present study shows that, at least in the laboratory context, the prejudiced relations between Black faces and a negative symbol may be overcome by an

emergent relation with those faces and a positive symbol. The study also suggests that training parameters influence the probability that the emergent relation will overcome the prejudiced one.

This demonstration opens many lines for future research. It seems particularly important to investigate to what extent the reduction of bias in the experimental context may generalize to other experimental measures and to behavior in naturalistic settings. Thus, Mizael et al. (2016) found a significant increase in the evaluation of Black faces with the Self-Assessment Manikin (Bradley & Lang, 1994), to the point that differences between average evaluations of Black and White faces became no longer significantly different after equivalence class formation. Comparison of pre- and postexperimental measures with an instrument to measure implicit attitudes would also provide an important index. The extent to which these measures may be correlated with, or predict, behavior beyond the laboratory (e.g., in play settings, school) should also be explored. For instance, researchers could verify if bias reduction in the experimental context will increase the time children spend interacting with Black children in the playground or snack time. Other possibilities might be the use of a "doll test" to see if children are more inclined to play with Black dolls or consider them more pretty and nice after reduction of bias in the lab setting. Researchers could also present drawings or pictures of Black and White individuals and ask with which of them children would like to play with/be friends with.

The finding that biases could be overcome in the experimental context opens the way to interventions in naturalistic settings. We may suppose that naturalistic interventions should go beyond the schematic relation between faces and a single symbol. Researchers may design interventions designed to generate derived relations between several aspects of Black individuals (e.g., physical traits, cultural traits, written descriptions) with a range of positive aspects. Also, directly trained (rather than emergent) relations should be explored, as in evaluative conditioning, i.e., changes of the valence of a stimulus (CS) by pairing it to another stimulus (US) with a positive valence (cf. de Houwer, Thomas, & Bayens, 2001; Hughes et al., 2018).

## Limitations

Although the results of the present research showed that participants' choices were consistent throughout training and testing, the use of only two comparison stimuli was a limitation of the present research. The use of only two comparison stimuli may produce false-positive equivalence results, due to choices controlled by relations between sample and S- (cf. Sidman, 1987; Dube & McIlvane, 1996). Also, after participants have been trained in some MTS relations, they may respond to new relations in the absence of feedback by *arbitrarily assigning*

each comparison to a specific sample. With only two choices, this arbitrary assignment will produce either 100% or 0% scores, with 50% probability for each. Scores of 0% indicate sharp stimulus control, although the controlling relations are the opposite to those intended by the experimenter. Whenever several scores of 0% (or close to 0%) are found, one may suspect that they are produced by arbitrary assignment. If this is the case, one may suspect that some scores of 100% may have also been produced by arbitrary assignment.

An important limitation of the present research, already mentioned, is that the measures are limited to the experimental context. Some of the measures were binary yes/no categorizations, such as forming or not equivalence and maintaining or not the classes. These binary categorizations may not accurately reflect changes that occurred along the experiment. This may be the reason why some participants that *did not form* equivalence classes showed high values of bias reduction.

## Compliance with Ethical Standards

**Conflict of Interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Ethical Approval** All procedures performed in studies 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.

**Informed Consent** Informed consent was obtained from parents of all individual participants included in the study.

**Availability of Data and Materials** Online supplementary material is provided with this manuscript. This material contains descriptions of one of the instruments used in this research and individual data from all participants.

## References

- Adams, B. J., Fields, L., & Verhave, T. (1993). Effects of test order on intersubject variability during equivalence class formation. *The Psychological Record*, *43*, 133–152.
- Amd, M., Oliveira, M. A., Passarelli, D. A., Balog, L. C., & de Rose, J. C. (2018). Effects of orientation and differential reinforcement II: Transitivity and transfer across five-member sets. *Behavioural Processes*, *150*, 8–16. <https://doi.org/10.1016/j.beproc.2018.02.012>.
- Amtzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record*, *47*, 309–320. <https://doi.org/10.1007/BF03395227>.
- Barnes, D., Lawlor, H., & Smeets, P. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record*, *46*, 87–107. <https://doi.org/10.1007/BF03395165>.
- Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*, 169–177.

- Bortoloti, R., & de Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record, 59*, 563–590. <https://doi.org/10.1007/BF03395682>.
- Bortoloti, R., Rodrigues, N. C., Cortez, M. D., Pimentel, N., & de Rose, J. C. (2013). Overtraining increases the strength of equivalence relations. *Psychology & Neuroscience, 6*(3), 357–364. <https://doi.org/10.3922/j.psns.2013.3.13>.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy & Experimental Psychiatry, 25*(1), 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9).
- de Almeida, J. H., & de Rose, J. C. (2015). Changing the meaningfulness of abstract stimuli by the reorganization of equivalence classes: Effects of delayed matching. *The Psychological Record, 65*, 451–461. <https://doi.org/10.1007/s40732-015-0120-9>.
- 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>.
- De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853–869. <https://doi.org/10.1037/0033-2909.127.6.853>.
- de Rose, J. C., McIlvane, W. J., Dube, W. V., Galpin, V. C., & Stoddard, L. T. (1988). Emergent simple discrimination established by indirect relation to differential consequences. *Journal of the Experimental Analysis of Behavior, 50*, 1–20. <https://doi.org/10.1901/jeab.1988.50-1>.
- Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior, 62*(3), 331–351. <https://doi.org/10.1901/jeab.1994.62-331>.
- Dube, W. V., & McIlvane, W. J. (1996). Some implications of a stimulus control topography analysis for emergent behavior and stimulus class. In T. R. Zentall & P. M. Smeets (Eds.), *Stimulus class formation in humans and animals* (pp. 197–218). Amsterdam, The Netherlands: North Holland/Elsevier.
- Fields, L., Adams, B. J., Newman, S., & Verhave, T. (1992). Interactions among emergent relations during equivalence class formation. *Quarterly Journal of Experimental Psychology Section B, 45*(2), 125–138. <https://doi.org/10.1080/14640749208401013>.
- Fields, L., Amtzen, E., & Moksness, M. (2014). Stimulus sorting: A quick and sensitive index of equivalence class formation. *The Psychological Record, 64*, 487–498. <https://doi.org/10.1007/s40732-014-0034-y>.
- Fields, L., Amtzen, E., Nartey, R. K., & Eilifsen, C. (2012). Effects of a meaningful, a discriminative, and a meaningless stimulus on equivalence class formation. *Journal of the Experimental Analysis of Behavior, 97*, 163–181. <https://doi.org/10.1901/jeab.2012.97-163>.
- Fields, L., Verhave, T., & Fath, S. (1984). Stimulus equivalence and transitive associations: A methodological analysis. *Journal of the Experimental Analysis of Behavior, 42*(1), 143–157. <https://doi.org/10.1901/jeab.1984.42-143>.
- Guerin, B. (2005). Combating everyday racial discrimination without assuming racists or racism: new intervention ideas from a contextual analysis. *Behavior & Social Issues, 14*, 46–70. <https://doi.org/10.5210/bsi.v14i1.120>.
- 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*, 385–396. <https://doi.org/10.1037/h0101276>.
- Hayes, S. C., Kohlenberg, B. S., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence relations. *Journal of the Experimental Analysis of Behavior, 56*, 119–137. <https://doi.org/10.1901/jeab.1991.56-119>.
- Hughes, S., Barnes-Holmes, D., Van Dessel, P., de Almeida, J. H., Stewart, I., & De Houwer, J. (2018). On the symbolic generalization of likes and dislikes. *Journal of Experimental Social Psychology, 79*, 365–377. <https://doi.org/10.1016/j.jesp.2018.09.002>.
- Mizael, T. M., de Almeida, J. H., Silveira, C. C., & de Rose, J. C. (2016). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record, 66*, 451–462. <https://doi.org/10.1007/s40732-016-0185-0>.
- Moxon, P., Keenan, M., & Hine, L. (1993). Gender-role stereotyping and stimulus equivalence. *The Psychological Record, 43*, 381–394.
- Rehfeldt, R. A., & Root, S. (2004). The generalization and retention of equivalence relations in adults with mental retardation. *The Psychological Record, 54*, 173–186. <https://doi.org/10.1007/BF03395468>.
- Saunders, R. R., & Green, G. (1992). The nonequivalence of behavioral and mathematical equivalence. *Journal of the Experimental Analysis of Behavior, 57*, 227–241. <https://doi.org/10.1901/jeab.1992.57-227>.
- Saunders, R. R., Saunders, K. J., Kirby, K. C., & Spradlin, J. E. (1988a). The merger and development of equivalence classes by unreinforced conditional selection of comparison stimuli. *Journal of the Experimental Analysis of Behavior, 50*, 145–162. <https://doi.org/10.1901/jeab.1988.50-145>.
- Saunders, R. R., Wachter, J., & Spradlin, J. E. (1988b). Establishing auditory stimulus control over an eight-member equivalence class via conditional discrimination procedures. *Journal of the Experimental Analysis of Behavior, 49*(1), 95–115. <https://doi.org/10.1901/jeab.1988.49-95>.
- Sidman, M. (1987). Two choices are not enough. *Behavior Analysis, 22*, 11–18.
- Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.
- Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the test paradigm. *Journal of the Experimental Analysis of Behavior, 37*(1), 5–22. <https://doi.org/10.1901/jeab.1982.37-5>.
- Sidman, M., Wilson-Morris, M., & Kirk, B. (1986). Matching-to-sample procedures and the development of equivalence relations: The role of naming. *Analysis & Intervention in Developmental Disabilities, 6*, 1–19. [https://doi.org/10.1016/0270-4684\(86\)90003-0](https://doi.org/10.1016/0270-4684(86)90003-0).
- Strand, R. C. W., & Amtzen, E. (2020). Social categorization and stimulus equivalence: A systematic replication. *The Psychological Record, 70*. Advanced online publication. <https://doi.org/10.1007/s40732-019-00364-3>.
- Stromer, R., & Osborne, G. (1982). Control of adolescents' arbitrary matching-to-sample by positive and negative stimulus relations. *Journal of the Experimental Analysis of Behavior, 37*(3), 329–348. <https://doi.org/10.1901/jeab.1982.37-329>.
- Varelas, A., & Fields, L. (2017). Equivalence based instruction by group based clicker training and sorting tests. *The Psychological Record, 67*, 71–80. <https://doi.org/10.1007/s40732-016-0208-x>.
- Wallace, B. W. (2003). *Match to Sample Program III [Computer software]*. Worcester, MA: UMass/Eunice Kennedy Shriver Center's Behavioral Sciences Department.
- 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>.