The Psychological Record
https://doi.org/10.1007/s40732-018-0304-1

ORIGINAL ARTICLE

Implicit and Explicit Measures of Transformation of Function from Facial Expressions of Fear and of Happiness via Equivalence Relations

William F. Perez[1,2] • João Henrique de Almeida[2,3] • Julio C. de Rose[2,3] • Andrea H. Dorigon[1] • Eduardo L. de Vasconcellos[1] • Marco A. da Silva[1] • Najra D. P. Lima[1] • Roberta B. M. de Almeida[1] • Rodrigo N. M. Montan[1] • Dermot Barnes-Holmes[4]

Abstract

Studies on equivalence relations have suggested that abstract symbols might acquire emotional functions when related to facial expressions. The present study aimed to investigate the transformation of emotional functions from facial expressions of fear and of happiness to abstract stimuli via equivalence relations. A delayed matching-to-sample task established two equivalence classes between facial expressions of emotions and nonsense abstract stimuli: A1(Fear)-B1-C1-D1; A2(Happiness)-B2-C2-D2. After relational training (AB, AC, CD) and equivalence tests (BD, DB), the participants evaluated the meaning of one nonsense stimulus from each class (D1 and D2) by means of a semantic differential and an Implicit Relational Assessment Procedure (IRAP). Results from both the semantic differential and the IRAP supported the conclusion that the emotional functions of the faces, in terms of fear and happiness, had transformed via the equivalence classes to the D stimuli. Results are discussed in terms of the dynamics of arbitrarily applicable relational responding.

Keywords Equivalence relations · Transformation of function · Semantic differential · IRAP · Facial expression

Behavior analysts have often studied equivalence relations as a way of analyzing symbolic behavior (Sidman, 1994; Sidman & Tailby, 1982). The basic assumption is that such equivalence relations are functionally similar to symbol–referent relations that are so pervasive in natural language (e.g., Sidman, 1994; Hayes, Barnes-Holmes, & Roche, 2001; de Rose & Bortoloti, 2007). In terms of procedure, equivalence relations are typically studied by establishing arbitrary conditional relations between stimuli (e.g., AB, BC) and by testing for novel relations that derive from those previously trained. To test the equivalence between the stimuli that were arbitrarily related to each other during training such stimuli are presented in novel combinations or sequences (e.g., AC, CA, BA, CB), to determine if they function as the "same," for instance, by demonstrating that they are directly related via symmetry (e.g., if A is equivalent to B, then B must be equivalent to A) or indirectly related via a "mediating" node (e.g., if A is equivalent to B and B is equivalent to C, then A must be equivalent to C and C must be equivalent to A).

Studies also suggest that specific behavioral functions may also transfer via equivalence relations. Thus, if a stimulus from an equivalence class is directly trained to function as a discriminative stimulus, for example, the other stimuli in that class may acquire that function without direct training (e.g., Bortoloti & de Rose, 2009; de Rose, McIlvane, Dube, Galpin, & Stoddard, 1988; Dougher, Augustson, Markham, Greenway, & Wulfert, 1994; Hayes, Kohlenberg, & Hayes, 1991; Perez, Fidalgo, Kovac, & Nico, 2015; Perez, Tomanari, & Vaidya, 2015; Perez et al., 2017). This phenomenon, called "derived transfer of function" or "transformation of function" (see Dymond & Rehfeldt, 2000, for a review),

> . . . is compatible to the idea that, in many contexts, we react to symbols as if we were facing the events they refer to. Thus, a stimulus that has (or acquires) a given

✉ William F. Perez
will.f.perez@gmail.com

[1] Paradigma – Centro de Ciências e Tecnologia do Comportamento, São Paulo - SP, Brazil

[2] Instituto Nacional de Ciência e Tecnologia sobre Comportamento, Cognição e Ensino (INCT-ECCE), Rodovia Washington Luís, São Carlos - SP, Brazil

[3] Universidade Federal de São Carlos, São Carlos, Brazil

[4] Ghent University, Ghent, Belgium

Published online: 07 August 2018

> function could be compared to a "referent" and the stimuli equivalent to it could be compared to "symbols" that substitute the referent in certain contexts. (de Rose & Bortoloti, 2007, p. 87; see also Bortoloti & de Rose, 2011)

One of the important implications of the derived transformation of function effect is that it may provide a behavior-analytic way of studying and explaining how previously neutral stimuli acquire emotional functions (or more informally, emotional meaning) without direct experience with those stimuli. Given the importance of facial stimuli in early human development, and throughout the life-span of humans in their social interactions with each other (Parr, Winslow, Hopkins, & De Waal, 2000), the derived transformation of "facial" functions has been seen as an important area of study (e.g., Bortoloti & de Rose, 2008, 2009, 2011, 2012). In the study reported by Bortoloti and de Rose (2009) adult participants were exposed to a matching-to-sample task that aimed to establish three 4-member equivalence classes: A1B1C1D1, A2B2C2D2, and A3B3C3D3. The A stimuli were comprised of pictures of faces that were designated as angry (A1), neutral (A2) or happy (A3); the B, C, and D stimuli were all abstract figures. The participants learned AB, AC, and CD conditional relations followed by equivalence tests (BD and DB tests). After that, they evaluated the "meaning" of each of the three D stimuli to determine if the A functions had transferred to D1, D2, and D3, equivalent to A1 (angry), A2 (neutral), and A3 (happy), respectively. For this purpose, the D stimuli were presented in a semantic differential (Osgood & Suci, 1952; Osgood, Suci, & Tannenbaum, 1957), which is an instrument comprised of multiple scales anchored by opposite adjectives (sad/happy, bad/good, ugly/beautiful etc.). To avoid pretest/posttest reactivity for the experimental group, a control group evaluated the facial expressions (set A) and the same nonsense stimuli (set D) without any prior equivalence training or testing. Results suggested that the participants who successfully completed the equivalence training and testing evaluated the D stimuli as predicted (i.e., D1-Angry; D2-Neutral; D3-Happy). The control group, who had evaluated the actual faces appropriately failed to show any derived transformation effects to the D stimuli. Further studies using a similar methodology found that the transformation of facial or emotional "meaning" measured by the semantic differential: (a) occurs even when the facial expressions are presented in a fraction of a second (Bortoloti & de Rose, 2008), and are modulated by parameters of the MTS task such as (b) the delay to present the comparison stimuli (Bortoloti & de Rose, 2009, 2012; de Almeida & de Rose, 2015); (c) the number of nodes belonging to each equivalence class (Bortoloti & de Rose, 2009); (d) the training structure (Bortoloti & de Rose, 2011); and (e) the overtraining of baseline conditional relations (Bortoloti, Rodrigues, Cortez, Pimentel, & de Rose, 2013).

Most studies using the methodology developed by Bortoloti and de Rose used angry, neutral, and happy faces (de Almeida & de Rose, 2015; Bortoloti & de Rose, 2007, 2008, 2009, 2011, 2012; Bortoloti et al., 2013; Silveira, Mackay, & de Rose, 2018). Perez, de Almeida, and de Rose (2015) investigated the transformation of emotional functions using sad faces. At the time of writing, however, there were no published studies that reported a derived transformation of facial functions indicative of "fear." On balance, numerous studies of derived transformation effects have reported derived fear responses using a variety of measures. The basic method involves establishing an equivalence class and then pairing one or more of the stimuli in that class with the delivery of mild electric shock or other aversive stimuli, and finally testing to determine if other member of the equivalence class not paired with the aversive stimuli produce fear responses (e.g., Augustson & Dougher, 1997; Dougher et al., 1994; Dymond, Roche, & Bennett, 2013; Dymond, Schlund, Roche, & Whelan, 2014; Luciano et al., 2014; Vervoort, Vervliet, Bennett, & Baeyens, 2014). In general, these studies have produced the predicted derived transformation of function effects, thus supporting the basic argument that humans may learn to fear stimuli that have never predicted the presentation of an aversive stimulus.

Another way in which humans may learn to fear stimuli without direct experience of a physically aversive stimulus, such as shock, is through the facial expressions of other individuals (e.g., Olsson & Phelps, 2004, 2007). One potentially important line of research would be to combine the use of facial stimuli showing fear with the derived transformation of function paradigm. The primary purpose of the current study was to initiate this program of research.

Consistent with previous studies of derived transformation of facial (emotional) functions the current study employed semantic differentials to assess transformation effects. A number of recent studies employed a measure or method that emerged directly from the study of derived relations to assess fear and avoidance responses: the implicit relational assessment procedure (IRAP; Barnes-Holmes et al., 2006; Hughes, Barnes-Holmes, & Vahey, 2012). The IRAP has been used primarily to assess the emotional functions of stimuli that were established in the preexperimental environment and a recent meta-analysis of the IRAP in the clinical domain indicates that it has a relatively high level of predictive validity (Vahey, Nicholson, & Barnes-Holmes, 2015). In particular, in the context of fear and avoidance, a number of recent studies have also indicated that the IRAP provides a valid measure of spider-related behavior (e.g., Nicholson & Barnes-Holmes, 2012; Leech, Barnes-Holmes, & Madden, 2016).

In the study reported by Leech et al. (2016), participants were exposed to an IRAP designed to measure avoidance and approach bias towards spiders. Each trial onset presented an image of either a spider or a pet (puppy or kitten), a phrase related to either avoidance (e.g., “I need to escape,” “I need to avoid it”) or approach (“I can touch it,” “I can approach it”), and two relational options, to confirm (“Yes”) or deny (“No”) the relation between the picture and the phrase. The procedure presented blocks of trials in which the participants had to respond consistently with spider-avoidance bias (e.g., *Spider—I need to escape/Yes; Spider—I can approach/No)* and pet-approach bias (e.g., *Pet—I need to escape/No; Pet—I can approach/Yes*). These blocks alternated with inconsistent blocks, in which the participants had to respond with spider-approach bias (e.g., *Spider—I need to escape/No; Spider—I can approach/Yes)* and pet-avoidance bias (e.g., *Pet—I need to escape/Yes; Pet—I can approach/No*). Results suggested that the participants responded faster (shorter mean latency) in blocks consistent with spider-avoidance bias compared to spider-approach bias. Correlational analyses indicated that the bias score on the spider-approach trial type of the IRAP predicted self-reported levels of fear and actual approach behavior towards a live tarantula spider.

As noted above, the IRAP has been used in the context of symbolic relations established in the laboratory (i.e., equivalence classes) with facial expressions (Bortoloti & de Rose, 2012). In Bortoloti and de Rose’s study (Bortoloti & de Rose, 2012), first, two 4-member equivalence classes were established: A1(happy faces)-B1-C1-D1 and A2 (angry faces)-B2-C2-D2. During a subsequent IRAP task, on each trial, a facial expression (A1 or A2) was presented along with a nonsense word (D1 or D2) and two relational response options, “True” or “False.” The participants were required to respond across alternating blocks that were consistent with the equivalence training (A1-D1/True, A1-D2/False, A2-D1/False, A2-D2/True) or inconsistent (A1-D1/False, A1-D2/True, A2-D1/True, A2-D2/False). Results showed that mean response latencies on the consistent blocks were shorter compared to the inconsistent blocks. The results thus indicated that the IRAP performance was sensitive to the symbolic or equivalence relations that were established in the laboratory with the facial stimuli.

One limitation to the study reported by Bortoloti and de Rose (2012) was noted by Perez, de Almeida et al. (2015). Specifically, the stimuli presented during the IRAP task were the stimuli from the equivalence training and testing phase, and therefore the results may indicate that the IRAP was sensitive to equivalence-class formation but not to any transformation of functions arising from the use of the facial stimuli. One way in which to assess the transformation of such emotive functions from the faces to equivalent stimuli would be to present the stimuli from the equivalence classes with words with positive and negative emotional functions in the IRAP. In effect, it would involve presenting in the IRAP the emotionally valenced words displayed in the semantic differential scales used by previous studies with other facial expressions (e.g., Bortoloti & de Rose, 2009, 2011; Bortoloti et al., 2013; de Almeida & de Rose, 2015; Perez, de Almeida, & de Rose, 2015). This strategy was adopted in the current study. In particular, we sought to determine if the IRAP performances would show a response bias in which participants confirmed more quickly than they denied that negative words were equivalent to the stimuli from the fear-face class and positive words were equivalent to stimuli from the happy-face class. Such a result would suggest that the IRAP was sensitive not only to the formation of experimentally induced equivalence classes but also to the derived transformation of emotional (facial) functions.[1] In short, we trained and tested participants for the transformation of positively and negatively valenced functions using faces and then asked them to complete both a semantic differential and an IRAP that sought to determine whether these measures yielded effects consistent with the predicted transformation of facial functions.[2]

Method

Participants

Seventy-seven verbally competent adults (M = 40, F = 37) ranging in age from 18 to 65 years (M = 26) took part in the experiment. Participants were recruited through personal contacts (sample of convenience) and numbered in the order in which they were recruited. Recruitment was conducted in a relatively unsystematic manner and therefore there were no systematic differences between participants in terms of allocated number. The first 42 participants were allocated to the experimental group; the remaining individuals were allocated to the control group. None of them had previously participated in any research involving equivalence relations or the IRAP. Before the experiment began, participants read a term of consent (approved by the Brazilian platform for ethical committees, Plataforma Brasil). At the end of the experimental sessions, they were fully debriefed concerning the goals of the experiment and procedural issues under consideration. They

[1] Recent research has indicated that the IRAP as a context for assessing biases in patterns of arbitrarily applicable relational responding is more complex than originally proposed (e.g., Barnes-Holmes, Finn, Barnes-Holmes, & McEnteggart, 2018; Finn, Barnes-Holmes, McEnteggart, 2018). One of the reviewers of the current article suggested that such findings highlighted a potential methodological flaw in the IRAP. We would argue that methodological flaws in any procedure or measure need to be defined in terms of conceptual or analytic assumptions, which were not specified in the review. We shall, however, consider a specific effect in the Discussion section that emerged in the current study that appears directly relevant to the material presented by Barnes-Holmes, Finn, et al. (2018) and Finn et al. (2018).

[2] The reader should note that the purpose of the current study was to examine the transformation of positively and negatively valenced functions based on fearful and happy faces rather than “fear” and “happy” functions specifically. That is, the words employed in the semantic differential scales and the IRAP were generally positive and negative (“good” versus “bad”) rather than related only to fear and happiness.

received no payment or compensation for participating in the research.

Experimental Setting and Equipment

The experimental sessions took place in a quiet room equipped with a table, a chair, and a laptop computer. Custom-written software in Visual Basic (6.0) presented stimuli, delivered consequences, and recorded participants' responses during the equivalence task. The IRAP ran the latency-based task and calculated D_{IRAP} scores during the last experimental phase.

Among stimuli, there were six facial expressions of fear (A1a, A1b, A1c, A1d, A1e, A1f) and of happiness (A2a, A2b, A2c, A2d, A2e, A2f), portrayed by three different male (a, b, c) and three different female (d, e, f) characters, extracted from the Pictures of Facial Affect© CD-ROM purchased from Paul Ekman's website (Ekman & Friesen, 1976). There were also 18 nonsense black forms on a white background (stimuli B1, B2, B3, C1, C2, C3, D1, D2 and D3).

Procedure

The first 42 participants on the list comprised the Experimental Group. They were given three different tasks. First, they went through a conditional discrimination training that aimed to establish two equivalence classes (see Fig. 1), one involving the faces of fear (A1) and three nonsense stimuli (B1C1D1) and another involving the happy faces (A2) and another three nonsense stimuli (B2C2D2). Second, the participants evaluated the meaning of two nonsense stimuli, one equivalent to the facial expressions of fear (D1) and the other equivalent to the facial expressions of happiness (D2), by means of a semantic differential. Finally, they were exposed to an IRAP task involving stimuli D1 and D2 and some of the positive and negative words from the semantic differential.

The next 35 participants on the list comprised the Control Group. These participants were neither submitted to the equivalence training and testing nor the IRAP. Their goal was to use the semantic differential to evaluate the meaning of the 12 facial expressions and of the same two nonsense stimuli (D1 and D2) evaluated by the Experimental Group. The comparison between the evaluations from the Experimental Group and from the Control Group allowed inferring the transformation of functions from the facial expressions to the nonsense stimuli. Whereas in the Control Group the nonsense stimuli would be expected to have a neutral "meaning," in the Experimental Group these stimuli would have their function affected by the equivalence with the facial expression from the class they belonged to. Thus, the evaluations of the nonsense stimuli by the Experimental Group should be similar to the evaluations of the facial expressions by the Control Group.

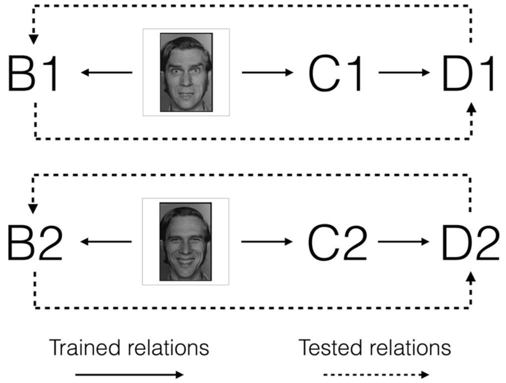

Fig. 1 Resume of equivalence training and testing phase

Equivalence Training and Testing This phase involved a delayed matching-to-sample task (DMTS) that aimed to establish two equivalence classes: A1(fear)B1C1D1 and A2(happy)B2C2D2 (see Fig. 1). Before any trial onset, the participants read a brief instruction for performing the task that described the presentation of the stimuli, the use of the mouse, and the feedback for correct and incorrect responding. Each trial began with the presentation of a sample stimulus (e.g., the face of fear, A1) in the center of the screen. A mouse-click response on the sample stimulus was followed by the withdrawal of that stimulus and the onset of three comparison stimuli after 2 s (e.g., B1, B2, B3) at the bottom of the screen, side-by-side. The selection (mouse-click) of the comparison stimulus programmed to belong to the same class of the sample stimulus (e.g., click on B1, given A1 as sample; click on B2, given A2, etc.) was followed by the withdrawal of all comparisons, by the immediate presentation of the word "CORRECT" in the center of the screen for 1s, and by a sequence of ascending notes; a mouse-click response to any of the other stimuli (B2 or B3) was followed by the withdrawal of all the comparisons, by the immediate presentation of the word "INCORRECT" in the center of the screen for 1s, and by a dissonant sound. A 1-s intertrial interval (ITI) separated the delivery of consequences from the next trial onset. The third comparison stimulus presented across trials (B3, C3 and D3) was never correct; these stimuli were presented to reduce the likelihood of correct responses by rejecting the incorrect stimulus (sample/S-relations; Sidman, 1987; Perez, Tomanari et al., 2015). The presentation of stimuli was randomized; the sample stimuli could not be presented more the three trials consecutively and the same comparison stimulus could not be presented in the same location for more than three trials consecutively.

Conditional relations were presented in the following order: AB, AC, and CD. The training phase began presenting AB trials (A1B1, A2B2) until participants met the mastery criterion of 12 consecutive correct responses; AB trial types (Sample/Comparison1Comparison2Comparison3—the correct comparison is underlined) were A1/B1B2B3 and A2/B1B2B3. A1 and A2 stimuli could be one of six different characters presenting a fearful face (A1a, A1b, A1c, A1d, A1e, A1f) or a happy face (A2a, A2b, A2c, A2d, A2e, A2f), respectively; A1 and A2 stimuli were randomly selected among the six different characters (a–f). After mastering AB trials, AC (A1C1, A2C2) and CD trials (C1D1, C2D2) were taught obeying the same criterion; trial types were A1/C1C2C3, A2/C1C2C3 and C1/D1D2D3, C2/D1D2D3. Once the conditional relations were taught separately, AB, AC, and CD trials were mixed on the same block until participants reached 18 consecutive correct responses.

Mastering baseline relations initiated the equivalence tests. First, the participants read the following instruction: "From now on, the computer will no longer present feedback, but will keep recording your hits and errors." The test block comprised 36 trials, 8 for each tested relation: B1D1, B2D2, D1B1, and D2B2; trial types were: B1/D1D2D3, B2/D1D2D3 and D1/B1B2B3, D2/D1D2D3. The participant's responses during tests were not followed by programmed consequences but the ITI. If scores on equivalence test were below 34 correct responses in 36 trials, the participants were thanked, debriefed, and would not advance to the following experimental phases. Thus, there was neither retraining nor retesting.

Semantic Differential Participants from the Experimental Group who had positive results on the equivalence tests were instructed to evaluate stimuli D1 and D2 by means of the semantic differential (Bortoloti & de Rose, 2009; see also Osgood & Suci, 1952; Osgood et al., 1957). Each scale comprised seven intervals and was anchored by bipolar terms (a pair of opposite adjectives). There were 13 scales formed by 13 different pairs of opposite adjectives, the Portuguese equivalents of sad/happy, tense/relaxed, rough/smooth, slow/fast, ugly/beautiful, heavy/light, negative/positive, passive/active, hard/soft, bad/good, unpleasant/pleasant, poor/rich, and submissive/dominant. The scales represented a series of continua, each going from an adjective to its opposite. The set of scales was printed on an A4 sheet that also depicted one of the D stimuli above them. Participants received three sheets, the first one containing instructions to fill in the scales (Bortoloti & de Rose, 2009) and the others displaying D1 and D2 for evaluation in a randomized order. The position of the positive and negative adjectives was balanced across scales and the left–right position of the polar terms on the sheet was randomized. Participants in the Control Group not only had to evaluate D1 and D2, but also had 12 extra sheets presenting, in each of them, the faces of fear (A1a–f) and the happy faces (A2a–f).

IRAP As illustrated in Fig. 2, the IRAP trials comprised the simultaneous presentation of one abstract stimulus on the top of the screen, one word in the center, and two relational response options in the lower corners. The nonsense stimuli were D1 (equivalent to the faces of fear) or D2 (equivalent to the happy faces); the words, selected from the semantic differential (Factor 1; de Almeida, Bortoloti, Ferreira, Schelini, & de Rose, 2014), were either positive (the Portuguese equivalent of happy, relaxed, beautiful, positive, good, pleasant) or negative adjectives (the Portuguese equivalent of sad, tense, ugly, negative, bad, unpleasant); the response options were two words with fixed positions, the Portuguese equivalent of "True" on the left corner and "False" on the right. Participants were required to choose one of the two response options, pressing the letter "d" on the keyboard to choose the response displayed on the left corner or "k" to choose the one on the right. Correct responses were followed by the withdrawal of all stimuli presented on that trial and a brief 400ms ITI. Incorrect responses were followed by the presentation of a red X on the center of the screen and stimuli were not withdrawn. The trial would end and the ITI began only after the participant had emitted the correct response.

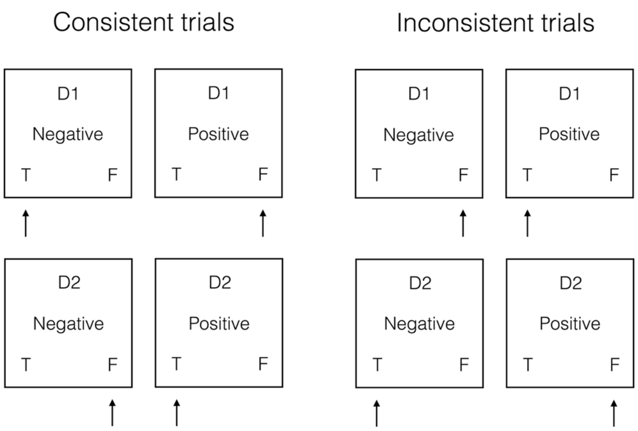

Fig. 2 IRAP trial types presented on consistent and inconsistent blocks. The arrow indicates the correct response programed for each type of trial

Participants were exposed to blocks of 24 trials each, which could be consistent or inconsistent with equivalence relations that were trained. During consistent blocks (see left portion of Fig. 2), each of the following trial types were presented (Label-Target/Correct response option): D1-Negative/True, D1-Positive/False, D2-Negative/False, D2-Positive/True. During inconsistent blocks (see right portion of Fig. 2), that reversed the contingencies of reinforcement for response options, trial types were: D1-Negative/False, D1-Positive/True, D2-Negative/True, D2-Positive/False. For the purpose of communication, we will label the D1 stimulus (equivalent to the facial expression of fear) as "Fear" and the D2 stimulus (equivalent to the facial expression of happiness) as "Happy." Thus, the four IRAP trial types will be referred hereafter as: Fear-Negative, Fear-Positive, Happy-Negative, and Happy-Positive. The trial types were presented an equal number of times in each block and were randomized across trials. Consistent and inconsistent blocks always alternated. Half of the participants began with a consistent block and the other half with an inconsistent block.

The IRAP comprised practice (warm up) and testing blocks. The practice phase began presenting a pair of consistent/inconsistent blocks with accuracy mastery criterion of 80% of correct responses in both blocks. After reaching accuracy criterion, the participants were exposed to another pair of consistent/inconsistent blocks to which a median latency criterion of 2000ms was added. If any participant failed to reach accuracy and latency criteria after three pairs of practice blocks, they were thanked, debriefed, and their data discarded. The participants who achieved accuracy and latency criteria during the training phase went directly to the IRAP testing phase. IRAP testing consisted of a fixed set of three pairs of consistent/inconsistent blocks, presented exactly as described for the end of the training phase. Test blocks were presented with no accuracy or latency criteria required for participants to progress from one block to the next; instead participant's data would be excluded if their accuracy fell below 75% in more than one block, or if their median latency exceeded 2000ms in any test block. At the end of the last test block, a brief message appeared ending the IRAP. Only test blocks were considered in the data analysis and to calculate the D_{IRAP} score.[3]

Results

Table 1 presents results from equivalence training and testing. Participants took from 12 to 160 trials to meet mastery criterion on AB, 12–46 trials on AC, 12–45 trials on CD, and 18–61 trials during mixed (AB+AC+CD) training trials. Only one participant (13) did not complete training and quit the experiment after 176 trials without reaching criterion on the AB training step. During equivalence tests, 34 from 41 participants reached criterion (34/36 or approximately 94% of correct responses). The seven participants who did not pass the equivalence testing (5, 11, 14, 18, 19, 21, and 27) were not assigned to the following experimental phases.[4]

Figure 3 presents results from the semantic differential task. Data analysis was based on the 34 participants who successfully completed the relational training. To proceed with data analysis, each of the thirteen 7-point interval scales comprising the semantic differential of each stimulus received a value from -3 (assigned to the position closest to the negative adjective) to +3 (assigned to the position closest to the positive adjective). Results from the Experimental Group show that D1 (equivalent to fear) had a negative valence whereas D2 (equivalent to happy) had a positive valence in Factor 1 bipolar scales. These evaluations differed from the Control Group in which the nonsense stimuli were evaluated as neutral in most scales. The evaluations from the Experimental Group (D1 and D2) were also close to the evaluations from the Control Group regarding the facial expressions (A1 and A2), which suggest the transformation of functions from the faces (set A) to the originally nonsense stimuli (set D). Statistical analysis employing a Kruskal-Wallis test, suggests that no difference was found between the evaluations of D1 by the Experimental Group and by the evaluations of the faces of fear by the Control Group (Factor 1, $p > .05$); however, the Experimental Group evaluated D2 more positively than the Control Group evaluated the happy faces (Factor 1, $p < .0001$). The effects observed in the comparison between D2 (Experimental Group) and the happy faces (Control Group) were individually considered in each dimension of Factor 1 of the Semantic Differential, by a series of t tests; such analysis revealed a significant difference in only three dimensions: Ugly x Beautiful, $p < .001$; Heavy x Light, $p < .01$; Pleasant x Unpleasant, $p < .01$. In addition, the evaluations from the Experimental Group regarding the nonsense stimuli differed from the evaluations from the Control Group (D1 exp X D1 con $p <. 0001$; D2 exp X D2 con, $p <. 0001$), suggesting a negative and positive valence for D1 and D2 in the evaluations of the Experimental Group, respectively, and neutrality for both stimuli in the evaluations of the Control Group.

IRAP data analysis comprised 24 participants. One participant quit after the semantic differential; four participants did not meet criteria to finish the IRAP training; four participants did not meet criteria to finish the IRAP training; two participants did not meet the accuracy criterion during IRAP test blocks; and two others did not meet both the accuracy and the latency criteria. One participant was excluded from data analysis after having been identified as a significant outlier in Grubbs's test ($p < .05$, $z = 2.28$). The IRAP primary data is the time (in milliseconds) elapsing from the onset of each trial to the first correct response (latency). The latency obtained from the six test blocks (three consistent and three inconsistent) from each participant were then transformed into D_{IRAP} scores (Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). Figure 4 displays the mean D_{IRAP} scores obtained for the Experimental Group in each of the four IRAP trial types: Fear-Negative, Fear-Positive, Happy-Negative and Happy-Positive (see horizontal axis). A positive value on the D_{IRAP} score indicates faster responding on consistent trials when compared to inconsistent trials; a negative value on the D_{IRAP} score indicates faster responding on inconsistent trials when compared to consistent trials. Thus, higher positive

[3] The D_{IRAP} score is the most widely used effect size measure employed with the IRAP and for this reason it was used in the present study, because it allows for relatively direct comparisons between the current findings and previously published IRAP studies. It should be recognized, however, that alternative effect-size measures may be used with the IRAP (see De Schryver, Hussey, De Neve, Cartwright, & Barnes-Holmes, 2018, for a recent example and detailed discussion of effect-size measures).

[4] The reader should note that this level of attrition in participants failing to pass an equivalence test is consistent with previous research, particularly when only a single exposure to the equivalence test is provided, as was the case here.

Table 1 Participants results on equivalence training and testing

	Training (# trials to reach mastery criterion)				Equivalence tests
Participant	AB	AC	CD	Mixed	(% correct responses)
1	34	12	13	18	97
2	38	15	15	25	100
3	26	15	15	29	100
4	22	15	15	18	100
5	17	20	12	39	72
6	20	13	15	18	97
7	22	12	13	18	100
8	160	15	15	18	100
9	30	15	15	18	100
10	48	15	18	34	94
11	14	15	29	18	89
12	20	45	14	18	97
13	176				
14	27	13	28	18	89
15	45	22	17	33	97
16	28	15	21	21	97
17	49	15	13	18	100
18	38	42	21	24	83
19	42	15	12	26	92
20	27	15	15	18	97
21	74	32	34	61	56
22	30	15	15	18	100
23	20	10	9	51	100
24	23	13	15	18	100
25	22	12	12	26	100
26	12	12	21	18	97
27	34	14	20	18	33
28	77	28	21	43	97
29	29	22	19	26	97
30	29	12	14	18	97
31	24	12	45	38	94
32	26	15	15	18	97
33	29	42	15	18	94
34	20	12	12	18	100
35	16	12	16	20	100
36	31	15	15	18	100
37	23	15	12	18	100
38	65	12	15	18	100
39	85	15	15	18	100
40	31	46	25	18	100
41	50	15	15	19	100
42	12	16	24	36	100

IRAP scores correspond to faster responses in those trials that required responding to relations coherent with the equivalence training (as presented in Fig. 1). Results suggest that participants responded faster on most trials consistent with the equivalence training. Significantly shorter latencies on consistent trials were obtained for Fear-Negative/True, Happy-Negative/False and Happy-Positive/True. Repeated one-sample *t* test statistical analysis revealed that three of the four

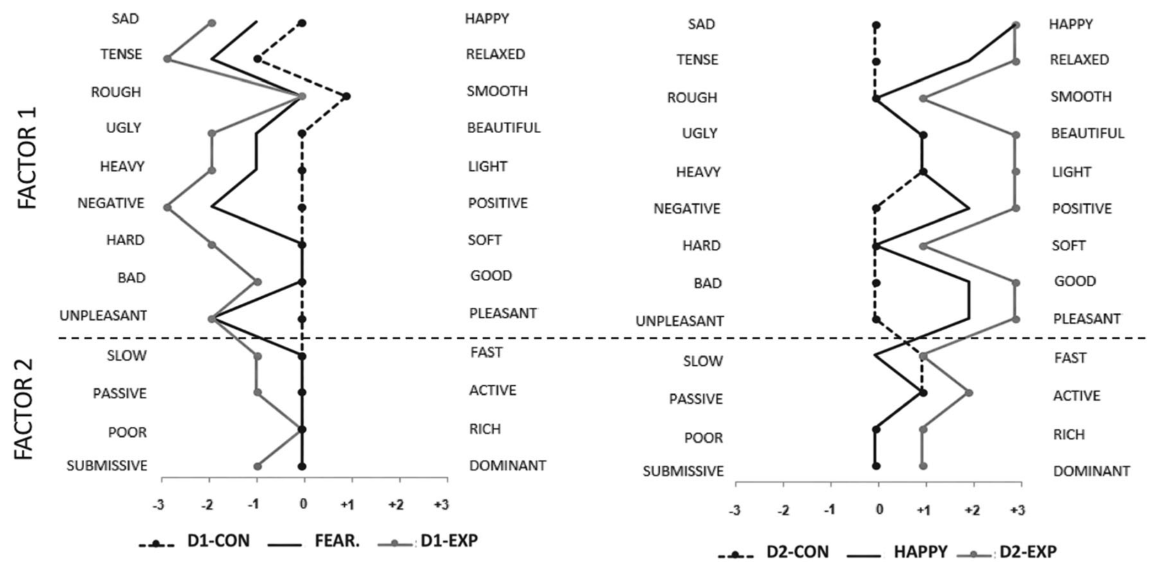

Fig. 3 Median scores on the semantic differential. The grey solid lines represent the median of the evaluations of D1 (left) and D2 (right) obtained for the Experimental Group; the black dashed lines represent the median of the evaluations of D1 (left) and D2 (right) obtained for the Control Group; the black solid lines represent the median of the evaluations of the facial expressions of fear (left) and happiness (right) by the Control Group

trial types presented significant differences compared to zero (Fear-Negative $p = 0.0008$; Fear-Positive $p = 0.2825$; Happy-Negative $p < 0.0001$; Happy-Positive $p < 0.0001$). An analysis of variance, one-way ANOVA, comparing the mean D_{IRAP} scores on each trial type produced a main effect for trial type F (3,23) = 8,358 $p < .0001$ $\eta^2_{partial} = .20$. Bonferroni's multiple comparisons test was employed as a posttest to investigate multiple comparisons in the ANOVA and revealed that results on Happy-Positive trial type significantly differed from Fear-Negative ($p < .0001$) and Fear-Positive ($p < .001$). The remaining comparisons were not statistically significant (all $ps > .05$). In addition, comparing consistent and inconsistent blocks, some variance was observed in the Fear-Positive trial type. An ANOVA 2x4 (block order x trial type) revealed no main effect for the order of blocks (consistent vs. inconsistent) F (3,23) = 2,135 $p = .1036$. Once this evaluation was nonsignificant and the participant sample were split in the half, an additional series of t-tests with adjusted value of p ($\alpha = .01$) confirmed the results of the 2x4 ANOVA showing no significant effect in the individual comparisons between consistent and inconsistent trial types (all $ps > .01$).

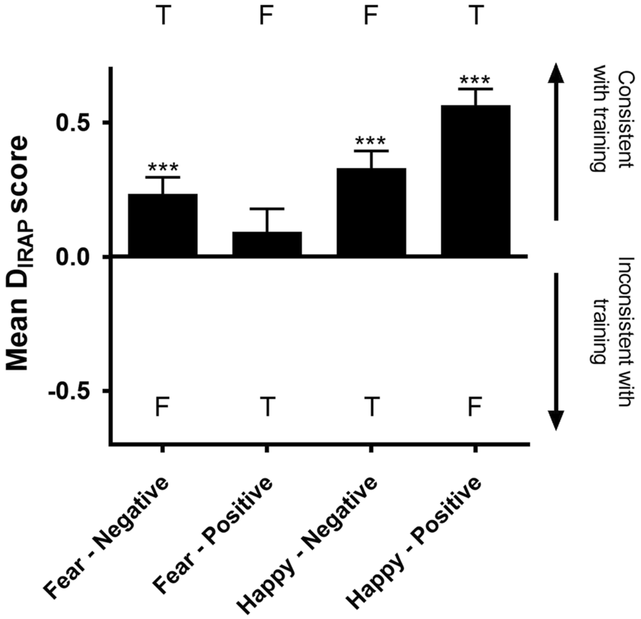

Fig. 4 Mean D_{IRAP} scores obtained from the Experimental Group for each IRAP trial type (horizontal axis). A positive value on the D_{IRAP} score indicates faster responding on trials consistent with equivalence training; a negative value on the D_{IRAP} score indicates faster responding on inconsistent trials

Discussion

The present study aimed to investigate the transformation of emotional functions from facial expressions of fear and of happiness to abstract stimuli via equivalence relations. Two equivalence classes (A1-B1-C1-D1; A2-B2-C2-D2) were first established in which the A stimuli were pictures of fearful (A1) and happy (A2) faces, and the remaining stimuli were all nonsense shapes. The participants then evaluated the two D stimuli from the equivalence classes by means of a semantic differential and an IRAP. Results from both the semantic differential and the IRAP supported the conclusion that the emotional functions of the faces, in terms of fear and happiness, had transformed via the equivalence classes to the D stimuli.

The present study appears to be the first to assess a derived transformation of functions using faces expressing fear (and happiness) with both a semantic differential measure and the IRAP. However, as noted previously, the scales targeted negative (and positive) valence rather than fear (and happiness) specifically. Previous studies have reported broadly similar transformation effects using faces depicting anger, sadness, or disgust (rather than fear; e.g., Bortoloti et al., 2013; Bortoloti & de Rose, 2009; de Almeida & de Rose, 2015; Perez, de Almeida et al., 2015; Silveira et al., 2018). In this sense, the current findings are entirely consistent with previous research.

A possibly interesting difference emerged between the experimental and control groups, in the current study, in terms of their responses to the semantic differential scales. Specifically, the control group produced weaker evaluations of the facial stimuli relative to the experimental group's ratings of the D (nonsense) stimuli. One might expect the opposite outcome, in that the ratings of the faces would be stronger than ratings of derived nonsense shapes. On balance, in rating actual faces the control group may have responded to multiple dimensions of each face, and thus a particular face may have expressed "fear," for example, but also have been considered "beautiful" or "attractive," which would undermine its negative valence on the semantic differential. In contrast, the D stimuli, which were abstract shapes rather than pictures of faces, might have been less likely to possess positive or negative functions based on their physical properties. As a result, the D stimuli may have produced more consistent responses when rating these stimuli semantically. In making this argument, however, it is worth noting that the effect may be somewhat more complex or subtle, given that the use of simultaneous versus delayed MTS procedures have also been found to produce broadly similar differences on semantic differential measures of derived transformation of function effects (Bortoloti & de Rose, 2009, 2011, 2012; de Almeida & de Rose, 2015)

The pattern of trial-type effects observed with the IRAP was generally consistent with the derived transformation of functions. That is, each of the four trial-types were positive and therefore consistent with the predicted bias scores one would expect if the function of D1 and D2 stimuli had been transformed, acquiring the negative valence of the fearful faces and the positive valence of the happy faces, respectively. An interesting effect also emerged, which has been reported previously in the literature. Specifically, a single-trial-type-dominance effect emerged in the current data (see Finn, Barnes-Holmes, & McEnteggart, 2018; Kavanagh, Barnes-Holmes, Barnes-Holmes, McEnteggart, & Finn, 2018). It is critical to note that the size of the bias effect for D2 (equivalent to happy faces) with positive words was significantly stronger than the bias effect for D1 (equivalent to fearful faces) with negative words. Given that the two trial-types share the same response option during consistent blocks (i.e., "True") and during inconsistent blocks (i.e., "False"), a simple explanation that appeals to a positivity bias for "True" over "False" is not possible (see Barnes-Holmes, Finn, McEnteggart, & Barnes-Holmes, 2018). One explanation might be that participants simply found the negative evaluative stimuli less relevant to fear than the positive stimuli were to happiness, and thus the size of the latter trial-type was the larger of the two. On balance, the single trial-type dominance effect has also been reported in IRAP studies in which the valence of the stimuli does not differ from each other and therefore additional variables are likely at play here. As a result, researchers attempted recently to explain the single trial-type dominance effect in terms of the different levels of coherence that the two trial-types involve (see Barnes-Holmes, Finn et al., 2018, for a detailed treatment). In the case of the current study, the response required on both Happy-Positive and Fear-Negative trial-types during consistent blocks is "True" (and "False" during inconsistent blocks). Insofar as the response option "True" is more positively valenced than "False" in natural language, the Happy-Positive trial-type is maximally coherent during consistent blocks of trials (i.e., the label, target, the relation between them, and the required response are all positively valenced). This is not the case for the Fear-Negative trial-type because the label and target stimuli are negatively valenced, although the relation between them is positive, in the sense that negative stimuli cohere. The response option ("True") is again positively valenced and thus the overall coherence of this trial-type is much reduced relative to the Happy-Positive trial-type (see Barnes-Holmes, Finn et al., 2018, and Finn et al., 2018, for detailed discussions of the RFT-based model that has been proposed for the single trial-type dominance effect).

Of course, the foregoing explanation for the single-trial-type-dominance effect remains open to debate, but the effect itself is clearly apparent in the current data. This is important because no other published study, at the time of writing, had reported the effect based on a derived transformation of functions. It appears, therefore, that the derived transformation effect observed in the current study produced IRAP performances that closely resemble those that have thus far only been observed with stimuli that acquired their functions in the natural environment, which of course means that the histories giving rise to those functions remains unknown. In the current study we know exactly how the functions of the D stimuli were established (i.e., via a derived transformation of functions) and because they yielded effects on the IRAP similar to those observed with "natural" stimuli, this strengthens the conclusion that derived transformation is a behavioral process with considerable ecological validity.

The fact that the IRAP effects observed in the current study yielded two patterns indicative of the derived transformation of functions (positive bias scores and a single-trial-dominance effect) seems to be important in the context of using it to compensate for a potential weakness in the use of the semantic differential alone. It could be argued that evidence of derived transformation using a semantic differential could be based, at least to some extent, on demand characteristics because participants could simply provide ratings that are deemed to be in accordance with what they think the researcher is expecting to find (see Chawla & Ostafin, 2007; Furnham, 1986). In contrast, when participants are asked to complete the IRAP they are encouraged to respond as quickly and accurately as possible across all trials and the differential in response latencies is taken as the dependent measure, rather than a subjective rating as is the case with the semantic differential. On balance, recent research has reported that IRAP bias effects may be "faked" or manipulated (e.g., Hughes et al., 2016), but such manipulation appears to require specific instructions about the IRAP itself and how to "fake" a particular performance. Such instructions were not employed in the current study. Furthermore, attempts to engineer the single-trial-type-dominance-effect using instructions failed in a recent study (Finn et al., 2018), and ongoing efforts by our research group to generate the effect "artificially" have proven less than successful. Therefore, it seems highly unlikely that the derived transformation effects observed using the IRAP in the current study can be explained as due to the demand characteristics of the experiment.

On balance, future studies could employ additional measures that have been used to assess fear responses. For example, the IRAP has been used successfully with both electroencephalogram (e.g., Power, Harte, Barnes-Holmes, & Barnes-Holmes, 2017) and electromyography (Roddy, Stewart, & Barnes-Holmes, 2011) measures. Perhaps, future research could attempt to replicate the current finding while also recording neural and/or facial reactions to the stimuli to determine if these also yield evidence of a derived transformation of functions. Other lines of inquiry might also be built out of the current research. For example, a framework for analyzing the dynamics of derived relational responding in terms of multiple levels of relational development and multiple dimensions (see Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017; Barnes-Holmes, Barnes-Holmes, Hussey, & Luciano, 2016; Barnes-Holmes, Boorman et al., 2018; Barnes-Holmes, Finn, et al. 2018) has recently been offered. It is interesting that recent findings using the semantic differential as a measure of the derived transformation of functions suggest that one or more of the dimensions identified in this new framework may be used to influence the ratings obtained on the semantic differential. For example, two of the dimensions, levels of derivation and levels of complexity, have been shown to be relevant. In particular, increases in the complexity in an equivalence class appear to reduce the strength of the transformation of functions on the semantic differential (Bortoloti & de Rose, 2009), whereas decreases in derivation (i.e., overtraining) appear to increase the strength of derived transformation on the differential measure (Bortoloti et al., 2013). Perhaps future research could seek to determine if similar increases and decreases in the "strength" (Hussey, Barnes-Holmes, & Barnes-Holmes, 2015) of the derived transformation of functions would also be observed using the IRAP (i.e., in terms of the size of the D_{IRAP} scores) and/or other measures.

Acknowledgements The present research was supported by means of a research funding from the São Paulo Research Foundation on behalf of William F. Perez (FAPESP 2016/05935-6). The second author was supported by a post-doc fellowship from the São Paulo Research Foundation (FAPESP 2014/01874-7 and 2017/10037-0). Preparation of the manuscript was supported by the Brazilian National Research Council (CNPq, Grants 573972/2008-7 and 465686/2014-1) and the São Paulo Research Foundation (FAPESP, Grant 2008/57705-8), both for Instituto Nacional de Ciência e Tecnologia sobre Comportamento, Cognição e Ensino (INCT-ECCE), chaired by Deisy G. de Souza (UFSCar).

Funding The present research was supported by the São Paulo Research Foundation (FAPESP 2014/01874-7, 2016/05935-6, 2017/10037-0, 2008/57705-8) and by the Brazilian National Research Council (CNPq, Grants 573972/2008-7 and 465686/2014-1); the present study is also part of the research program of the Instituto Nacional de Ciência e Tecnologia sobre Comportamento, Cognição e Ensino (INCT-ECCE).

Compliance with Ethical Standards

Conflict of Interest The authors declare they have 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. The research is approved by the Brazilian platform for ethical committees (Plataforma Brasil, CAAE # 54489116.4.0000.5504).

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

References

Augustson, E. M., & Dougher, M. J. (1997). The transfer of avoidance evoking functions through stimulus equivalence classes. *Journal of Behavior Therapy and Experimental Psychiatry, 3,* 181–191. https://doi.org/10.1016/S0005-7916(97)00008-6.

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational frame theory: Finding its historical and philosophical roots and reflecting upon its future development: An introduction to part II. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 117–128). West Sussex, Wiley-Blackwell.

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From the IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioral Science, 6,* 434–445. https://doi.org/10.1016/j.jcbs.2017.08.001.

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.

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2018). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *Behavior Analyst, 40,* 1–19. https://doi.org/10.1007/s40614-017-0124-7.

Barnes-Holmes, Y., Boorman, J., Oliver, J. E., Thompson, M., McEnteggart, C., & Coulter, C. (2018). Using conceptual developments in RFT to direct case formulation and clinical intervention: Two case summaries. *Journal of Contextual Behavioral Science, 7,* 89–96. https://doi.org/10.1016/j.jcbs.2017.11.005.

Bortoloti, R., & de Rose, J. C. (2007). Medida do grau de relacionamento entre estímulos equivalentes. *Psicologia: Reflexão e Crítica, 20,* 252–258. https://doi.org/10.1590/S0102-79722007000200011.

Bortoloti, R., & de Rose, J. C. (2008). Transferência de "significado" de expressões faciais apresentadas brevemente para estímulos abstratos equivalentes a elas. *Acta Comportamentalia, 16,* 223–241.

Bortoloti, R., & de Rose, J. C. (2009). Assessing the relatedness of equivalent stimuli through the semantic differential. *The Psychological Record, 59*, 563–590. https://doi.org/10.1007/BF03395682.

Bortoloti, R., & de Rose, J. C. (2011). An "Orwellian" account of stimulus equivalence. Are some stimuli "more equivalent" than others? *European Journal of Behavior Analysis, 12,* 121–134. https://doi.org/10.1080/15021149.2011.11434359.

Bortoloti, R., & de Rose, J. C. (2012). Equivalent stimuli are more strongly related after training with delayed matching than after simultaneous matching: A study using the implicit relational assessment procedure (IRAP). *The Psychological Record, 62,* 41–54. https://doi.org/10.1007/BF03395785.

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,* 357–364. https://doi.org/10.3922/j.psns.2013.3.13.

Chawla, N., & Ostafin, B. (2007). Experiential avoidance as a functional dimensional approach to psychopathology: An empirical review. *Journal of Clinical Psychology, 63,* 871–890. https://doi.org/10.1002/jclp.20400.

de Almeida, J. H., Bortoloti, R., Ferreira, P. R. S., Schelini, P. W., & de Rose, J. C. (2014). Análise das propriedades psicométricas de instrumento de diferencial semântico [Psychometric analysis of a semantic differential instrument]. *Psicologia: Reflexão e Crítica, 27,* 272–281. https://doi.org/10.1590/1678-7153.20142720.

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 Rose, J. C., & Bortoloti, R. (2007). A equivalência de estímulos como modelo de significado. *Acta Comportamentalia, 15,* 83–102.

de Rose, J. C., McIlvane, W. J., Dube, W. V., Galpin, V. C., & Stoddard, L. T. (1988). Emergent simple discriminations established by indirect relations to differential consequences. *Journal of the Experimental Analysis of Behavior, 50,* 1–20. https://doi.org/10.1901/jeab.1988.50-1.

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, D. (2018). The PI IRAP: An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure. *Journal of Contextual Behavioral Science, 7*(1), 97–103. https://doi.org/10.1016/j.jcbs.2018.01.001.

Dougher, M., Augustson, E., Markham, M., Greenway, D., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior, 62,* 331–351. https://doi.org/10.1901/jeab.1994.62-331.

Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *Behavior Analyst, 23,* 239–254.

Dymond, S., Roche, B., & Bennett, M. (2013). Relational frame theory and experimental psychopathology. In S. Dymond & B. Roche (Eds.), *Advances in relational frame theory: Research & application* (pp. 199–218). Oakland, CA: New Harbinger.

Dymond, S., Schlund, M. W., Roche, B., & Whelan, R. (2014). The spread of fear: Symbolic generalization mediates graded threat-avoidance in specific phobia. *Quarterly Journal of Experimental Psychology, 67,* 247–259. https://doi.org/10.1080/17470218.2013.800124.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Paul Ekman Group. Retrieved from www.paulekman.com.

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect on the IRAP: Developing a differential arbitrarily applicable relational responding effects (DAARRE) model. *The Psychological Record, 68,* 11–25. https://doi.org/10.1007/s40732-017-0262-z.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7,* 385–400. https://doi.org/10.1016/0191-8869(86)90014-0.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York, NY: Plenum.

Hayes, S. C., Kohlenberg, B. K., & Hayes, L. J. (1991). The transfer of specific and general consequential functions through simple and conditional equivalence classes. *Journal of the Experimental Analysis of Behavior, 56,* 119–137. https://doi.org/10.1901/jeab.1991.56-119.

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science, 1,* 17–38. https://doi.org/10.1016/j.jcbs.2012.09.003.

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the implicit relational assessment procedure. *European Journal of Social Psychology, 46,* 632–648. https://doi.org/10.1002/ejsp.2207.

Hussey, I., Barnes-Holmes, D., & Barnes-Holmes, I. (2015). From relational frame theory to implicit attitudes and back again: Clarifying the link between RFT and IRAP research. *Current Opinion in Psychology, 2,* 11–15. https://doi.org/10.1016/j.copsyc.2014.12.009.

Kavanagh, D., Barnes-Holmes, Y., Barnes-Holmes, Y., McEnteggart, C., & Finn, M. (2018). Exploring differential trial-type effects and the impact of a read-aloud procedure on deictic relational responding on the IRAP. *The Psychological Record, 68,* 163–176. https://doi.org/10.1007/s40732-018-0276-1.

Leech, A., Barnes-Holmes, D., & Madden, L. (2016). The implicit relational assessment procedure (IRAP) as a measure of spider fear, avoidance, and approach. *The Psychological Record, 66,* 337–349. https://doi.org/10.1007/s40732-016-0176-1.

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodrıguez-Valverde, M., Barnes-Holmes, D., Dougher, M. J., & Gutierrez-Martinez, O. (2014). Effects of an acceptance/defusion intervention on experimentally induced generalized avoidance: A laboratory demonstration. *Journal of the Experimental Analysis of Behavior, 101,* 94–111. https://doi.org/10.1002/jeab.68.

Nicholson, E., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62,* 263–278. https://doi.org/10.1007/BF03395801.

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science, 15,* 822–828. https://doi.org/10.1111/j.0956-7976.2004.00762.x.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10,* 1092–1102. https://doi.org/10.1038/nn1968.

Osgood, C. E., & Suci, G. I. (1952). A measure of relation determined by both mean difference and profile information.

Psychological Bulletin, 49, 251–262. https://doi.org/10.1037/h0062981.

Osgood, C. E., Suci, G. I., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Parr, L. A., Winslow, J. T., Hopkins, W. D., & De Waal, F. B. M. (2000). Recognizing facial cues: Individual discrimination by chimpanzees (*Pan troglodytes*) and rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology, 114,* 47–60. https://doi.org/10.1037/0735-7036.114.1.47.

Perez, W. F., de Almeida, J., & de Rose, J. C. (2015). Transformation of meaning through relations of sameness and opposition. *The Psychological Record, 65,* 679–689. https://doi.org/10.1007/s40732-015-0138-z.

Perez, W. F., Fidalgo, A. P., Kovac, R., & Nico, Y. C. (2015). The transfer of Cfunc contextual control through equivalence relations. *Journal of the Experimental Analysis of Behavior, 103,* 511–523. https://doi.org/10.1002/jeab.150.

Perez, W. F., Tomanari, G. Y., & Vaidya, M. (2015). Effects of select and reject control on equivalence class formation and transfer of function. *Journal of the Experimental Analysis of Behavior, 104,* 146–166. https://doi.org/10.1002/jeab.284.

Perez, W. F., Kovac, R., Nico, Y., Caro, D. M., Fidalgo, A. P., Linares, I. ... de Rose, J. C. (2017). The transfer of Crel contextual control (same opposite, less then, more than) through equivalence relations. *Journal of the Experimental Analysis of Behavior, 108,* 318–334. https://doi.org/10.1002/jeab.164.

Power, P. M., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2017). Combining the implicit relational assessment procedure and the recording of event related potentials in the analysis of racial bias: A preliminary study. *The Psychological Record, 67,* 499–506. https://doi.org/10.1007/s40732-017-0252-1.

Roddy, S., Stewart, I., & Barnes-Holmes, B. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology, 41,* 688–694. https://doi.org/10.1002/ejsp.839.

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

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

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

Silveira, M. V., Mackay, H. A., & de Rose, J. C. (2018). Measuring the transfer of meaning through members of equivalence classes merged via a class-specific reinforcement procedure. *Learning & Behavior, 46,* 157–170. https://doi.org/10.3758/s13420-017-0298-6.

Vahey, N. A., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). A first test of the implicit relational assessment procedure (IRAP) as a measure of self-esteem: Irish prisoner groups and university students. *The Psychological Record, 59,* 371–388. https://doi.org/10.1007/BF03395670.

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the implicit relational assessment procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48,* 59–65. https://doi.org/10.1016/j.jbtep.2015.01.004.

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PLoS One, 9*(5), e96569. https://doi.org/10.1371/journal.pone.0096569.