Training and Testing for a Transformation of Fear and Avoidance Functions using the Implicit Relational Assessment Procedure: The First Study

Aileen Leech[a], Jaber Bouyrden[b], Nathalie Bruijsten[c], Dermot Barnes-Holmes[a], & Ciara McEnteggart[a]

[a]Department of Experimental, Clinical and Health Psychology, Ghent University, Ghent 9000, Belgium

[b]Department of Psychology, Thomas More Hogeschool, Antwerp, Belgium.

[c]Radboud University, Nijmegen, The Netherlands.

Corresponding Author: Aileen Leech

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan 2

B-9000 Ghent

Belgium

Abstract

Experiment 1 aimed to establish “fearful” and “pleasant” functions for arbitrary stimuli (geometric shapes) by relating those stimuli to pictures of spiders and pets using a training version of the Implicit Relational Assessment Procedure (IRAP). The transformation of these functions for the arbitrary stimuli was assessed by exposing participants to a ‘traditional’ version of the IRAP, the Fear-IRAP employed by Leech et al. (2016, 2017). A broadly similar pattern of response biases was recorded for the Fear-IRAP as had been observed in the previously published studies. Experiment 1 thus supported the assumed but untested assumption that the relational context provided by the IRAP may both serve to establish and reveal fear-related response biases in arbitrary stimuli. A second experiment attempted to replicate the effects observed in Experiment 1 but using pictures of ‘unfamiliar’ Australian marsupials as arbitrary stimuli. The pattern of results obtained in Experiment 2 failed to replicate the pattern observed in Experiment 1, or that reported in the previously published studies by Leech et al. Overall, the findings suggest a possibly important boundary condition for the IRAP as a training and/or testing context for establishing fear-related response biases for arbitrary stimuli.

Key words: Avoidance; Derived; Fear; IRAP; Testing; Training.

1. Introduction

Over the past 40 years, a growing number of behavior-analytic researchers have sought to study human language and cognition through the medium of derived stimulus relations (Barnes-Holmes, Finn, McEnteggart, & Barnes-Holmes 2018). The seminal work focused on equivalence classes (Sidman, 1971) as a functional-analytic model of symbolic relations (see Sidman, 1994). Other researchers subsequently began to explore a range of derived relations, using this work to develop an account of human language that extended beyond that provided by equivalence relations alone (e.g., Hayes, Barnes-Holmes, & Roche, 2001). One of the key concepts in this latter work was the *derived transformation of functions* (Dymond & Rehfeldt, 2000). As a simple example, imagine a participant who had been trained (e.g., A-B and B-C) and tested (e.g., C-A) for a three-member equivalence class. If a specific stimulus function was then established for the A stimulus, given appropriate contextual cues, that function would then emerge for the C stimulus. As a concrete example, if the A stimulus was paired with a fear-inducing stimulus, such as the delivery of mild electric shock, when the C stimulus was then presented fear responses may be observed although C had never predicted shock. In effect, the function of the C stimulus had been *transformed* into a fear-inducing stimulus based on its equivalence relation to the directly paired (with shock) A stimulus (Dougher, Augustson, Markham, Greenway, & Wulfert, 1994). Numerous studies have reported this basic effect using a variety of functions, procedures, and different types of relations (e.g., Dougher, Hamilton, Fink & Harrington, 2007; Dymond & Barnes, 1995; Whelan & Barnes-Holmes, 2004).

At a conceptual level, the transformation of functions effect has helped to provide a behavior-analytic explanation for the emergence of fear and avoidance responses for stimuli or events that have not been directly paired with an aversive experience (Dymond & Roche, 2009). As such, derived transformation effects have been used to address a weakness in direct conditioning accounts of how irrational fears and phobias may emerge (at least in verbally-

sophisticated humans). Some of the most recent work in this area has begun to explore the relationship between derived fear and derived avoidance and the "treatment" of these effects as functionally independent behaviors. Specifically, two studies reported by Luciano and colleagues (2013, 2014) focused on the derived transformation of fear and avoidance responses via equivalence relations.

The first study (Luciano et al. 2013) established a fear response for stimuli in an equivalence relation using a respondent conditioning paradigm and electric shock as a UCS, and then demonstrated the derived transformation of that function to other members of the equivalence class. In effect, the researchers provided matching-to-sample training designed to establish an equivalence relation among six stimuli (A-B-C-D-E-F) and when the A and B stimuli were paired with shock the E and F stimuli also elicited fear, although they were never directly paired with shock. Subsequently, the fear responses were extinguished for both the respondently conditioned stimuli, and the other members of the equivalence class, by presenting the directly conditioned stimuli in the absence of shock. Critically, however, participants continued to engage in avoidance responding even though the fear had been extinguished (i.e., as measured by skin conductance). In effect, avoidance continued in the apparent absence of fear. In a broadly similar study, Luciano et al. (2014) demonstrated again that it was possible to establish a derived transformation of fear and avoidance functions via equivalence relations, but this time they did not employ an extinction procedure. Rather, they presented an analogue intervention based on acceptance and commitment therapy (ACT; Hayes, Strosahl & Wilson, 1999), which they labelled a defusion protocol. Participants exposed to this protocol continued to show fear responses (as measured using skin conductance) but avoidance responses dropped to near zero. In this second study, therefore, fear continued in the absence of avoidance. Taken together, therefore, the two studies demonstrated the functional independence of fear and avoidance using the derived relations (and transformation of functions) paradigm.

Another line of behavior-analytic research that is very much focused on human language and cognition has also begun to explore the functional independence of fear and avoidance, but using a different paradigm to that employed in typical derived transformation of function studies (see Nicholson & Barnes-Holmes, 2012). Specifically, the research employs a method, known as the Implicit Relational Assessment Procedure (IRAP), which is a computer based task that requires participants to respond quickly and accurately (under time pressure) to sets of stimuli employing a response pattern that may be considered consistent or inconsistent with previous learning histories (see Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). Participants are presented with trials where one of two label stimuli are presented at the top of a computer screen (e.g., either a picture of a spider or a positively valenced picture, such as a puppy), and one of two target stimuli that are presented in the middle of the screen (e.g., "scares me" versus "I like it"). The task for the participant is to choose between two response options, such as "Yes" versus "No", presented at the bottom right and left of the screen. The typical IRAP thus contains four trial-types, which, based on the foregoing examples may be summarized as follows: Puppy-Positive; Puppy-Negative; Spider-Positive; Spider-Negative. All things being equal, the general hypothesis is that relational responding should be quicker and more accurate on history-consistent rather than history-inconsistent blocks of trials. Thus, response latencies should be shorter when participants are required to respond "True" on Puppy-Positive and Spider-Negative, and "False" on Puppy-Negative and Spider-Positive trial-types, relative to the opposite response pattern (e.g., "False" on Puppy-Positive).

A recently published study attempted to measure fear, avoidance, and approach as independent response biases using two IRAPs (Leech, Barnes-Holmes, & Madden, 2016). One IRAP (the Fear-IRAP) contained trials with statements pertaining to fearful versus pleasant reactions to spiders, whereas the other IRAP (the Avoidance-IRAP) contained statements pertaining to avoiding versus approaching spiders. Both IRAPs also presented

pictures of spiders on some trials and pictures of pets (puppies or kittens) on other trials. In the second experiment reported by Leech et al. (2016), performance on the Fear-IRAP predicted self-reported fear but not actual approach behavior on a behavioral approach task (BAT) towards a live spider, whereas performance on the Avoidance-IRAP successfully predicted approach behavior. Overall, therefore, the findings were consistent, in a broad sense, with the data reported by Luciano and colleagues (2013, 2014), in that fear and avoidance/approach responses towards spiders may be conceptualized as two functionally independent classes of behavior.

Although the research reported by Luciano et al. (2013, 2014) and Leech et al. (2016; see also Leech et al., 2017) may be seen as broadly consistent, it is important to note some fundamental differences. The research by Luciano and colleagues involved training and testing for the derived transformation of functions in the experimental context, whereas the research by Leech and colleagues employed the IRAP as a measure of fear and avoidance that had (presumably) been established in the natural environment (i.e., either through direct experience and/or or human language; i.e., derived relational responding). Indeed, the vast majority of published IRAP research has adopted the latter strategy (i.e., using the procedure to assess pre-experimentally established response biases; see Hughes & Barnes-Holmes 2011, for one notable exception). Although a recent meta-analysis has found the IRAP, in the clinical domain, to have a relatively high level of predictive validity (Vahey, Nicholson, & Barnes-Holmes, 2015), it seems important to test the basic assumption that the IRAP may be used to assess fear and avoidance responses for stimuli that have been related to fear-inducing stimuli within the experimental context itself. The primary purpose of the current study was to initiate this line of research.

As an aside, the IRAP has been used to assess laboratory-induced equivalence class formation, when the classes contained facial expressions (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 D1 or D2 (from the equivalence class) and two 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 equivalence relations that were established in the laboratory with the facial stimuli. One limitation to the study reported by Bortoloti and de Rose was noted by Perez, de Almeida, and de Rose (2015). Specifically, the stimuli presented during the IRAP task were the stimuli from the equivalence training and testing phase, and thus 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.

Experiment 1 in the current study aimed to establish "fearful" and "pleasant" functions for arbitrary stimuli by relating those stimuli to pictures of spiders and pets, respectively, using a training version of the IRAP, rather than a traditional stimulus pairing procedure (e.g., respondent conditioning, matching-to-sample, etc.). The Training-IRAP is similar to the traditional IRAP, except the same pattern of relational responding is required across every block of trials. In the current case, participants were presented with images of pets (i.e., kittens or puppies depending on preference) and images of spiders with two arbitrary geometric shapes (i.e., a circle and a square), thus creating four trial-types (e.g., Pet-Circle; Pet-Square; Spider-Circle; Spider-Square). Given that it was a Training-IRAP, participants were required to emit the same response, either "Similar" or "Different" to each trial-type across every block of trials until they reached pre-determined performance criteria. In the current example, therefore, this may have required the following responding: Pet-Circle/Similar; Pet-Square/Different; Spider-Circle/Different; Spider-Square/Similar. The aim

of this training was to establish a relational network in which the circle should acquire a positively valenced pet function and the square should acquire a negatively valenced spider function. The transformation of these functions for the arbitrary shape stimuli was assessed by exposing participants to the Fear-IRAP employed by Leech et al., which required responding in opposing response patterns across blocks of trials (e.g., Circle-"I like it"/"Yes" on one block and Circle-"I like it"/"No" on the next block). As is standard practice in IRAP research, participants were instructed to respond quickly and accurately across all blocks of trials. Participants also completed a behavioral approach task (BAT) using a live spider and questionnaires similar to those employed by Leech et al.

In using a Training-IRAP it is worth noting a potentially important issue. Unlike many studies that aim to establish fear responses for previously neutral stimuli, the Training-IRAP does not involve pairing a neutral stimulus with an unconditioned stimulus (UCS; e.g., pictures of spiders). Pavlovian conditioning paradigms, for example, typically involve presenting the neutral stimulus followed shortly thereafter with a UCS across numerous trials. In contrast, the Training-IRAP involves pairing stimuli, but all of the stimuli are paired with each other equally across blocks of trials. Thus, for example, the pictures of spiders will appear equally often with the circle and the square stimuli. The pairings are differentiated only in terms of the response option that participants are required to choose (i.e., "Similar" or "Different"). If the Training-IRAP establishes a fear function for the stimulus that is responded to as similar rather than different, the effect cannot be explained simply on the basis of stimulus pairings (because spiders will have appeared equally often with square and circle). An explanation for the emergence of a fear function for only one of the shapes must be based, at least to some extent, on the differential relational properties of the two response options. Such an explanation would thus be seen as more consistent with a relational or propositional account of fear acquisition than a traditional associative account (see Hughes, Barnes-Holmes, & De Houwer, 2011). On the grounds of intellectual honesty, it should be

emphasized that the first Experiment reported here was largely exploratory, and thus it was not designed as a formal test of an associative versus relational/propositional account.

The results from the first experiment supported the conclusion that the IRAP (as a training and testing context) may be used to establish and reveal fear-related response biases for arbitrary stimuli. In Experiment 2 we sought to replicate and extend the effects obtained in Experiment 1, but using arbitrary stimuli that differed considerably from two simple geometric shapes. Specifically, two relatively unknown Australian marsupials were employed as the arbitrary stimuli (i.e., a quoll and a quokka). In addition, two standard IRAPs were used to test for the transformation of functions; the Fear- and Avoidance-IRAPs employed by Leech et al. (2016, 2017). The results of Experiment 2 failed to replicate the basic effects obtained in the first experiment, but considered together the two experiments raise some interesting questions for future research, which will be considered in the General Discussion.

1. Experiment 1

2.1 Method

2.1.1 Ethical Considerations

The study reported here was conducted in accordance with the ethical guidelines of Ghent University. Prior to the experiment, participants read and signed a consent form informing them that they could withdraw from the study at any time. Upon completion, participants were fully debriefed.

2.1.2 Participants

Fifty seven undergraduate students attending Ghent University, Belgium, volunteered to participate in the study (N = 57; 42 Females, 15 Males). Participants were paid €10 for their participation. Twenty six participants were eliminated due to their failure to achieve the necessary performance criteria on the IRAPs (see "procedure" section), leaving 21 females and 10 males (N = 31), the results of whom were subject to analysis. The mean age was 20.9

years (*SE* = .374), with a range of 18 – 26 years. The participants completed the study individually in the Department of Experimental, Clinical and Health Psychology at Ghent University. Given the exploratory nature of the research a formal power analysis was not conducted. However, based on the recent meta-analysis of criterion effects for the IRAP in the clinical domain (Vahey, et al., 2015), a sample size greater than 29 is required for first order correlations to achieve statistical power of approximately 0.8.

2.1.3 Materials

The study employed four questionnaires; a Spider Fear Rating Question, the Depression, Anxiety and Stress Scale-21 (DASS); the Acceptance and Action Questionnaire II (AAQ-II); and the Fear of Spiders Questionnaire (FSQ). A Training-IRAP, a Test Fear-IRAP, and a Behavioral Approach Task (BAT) were also employed. All verbal material used in the current study were presented to participants in Dutch.

2.1.3.1 Spider Fear Rating Question.

Prior to completing the questionnaires, participants were asked to rate their fear of spiders on a Likert scale from 1-5, where 1 was "No fear" and 5 was "Very fearful".

2.1.3.2 Depression, Anxiety and Stress Scale (DASS-21; de Beurs, Van Dyck, Marquenie, Lange, & Blonk, 2001).

The Depression Anxiety and Stress Scale is a 21 item self-report questionnaire which covers a range of core symptoms of anxiety, depression and stress. The English version of this scale (Lovibond & Lovibond, 1995), for a non-clinical sample, has demonstrated excellent internal consistencies among its three subscales (Cronbach's Alphas = .82 - .90), good convergent and discriminant validity (*rs* = .70 - .72) and adequate reliability (Cronbach's alpha = .90 - .95) (Henry & Crawford, 2005). The Dutch translation has been reported to yield similar excellent internal consistency.

2.1.3.3 Acceptance and Action Questionnaire – II (AAQ- II 7-item version; Bernaerts, De Groot, & Kleen, 2012).

The Acceptance and Action Questionnaire-II is a 7-item self-report scale which measures acceptance, experiential avoidance and psychological inflexibility. The AAQ yields an overall score with a maximum of 49 indicating *low psychological flexibility* and a minimum of 7 indicating *high psychological flexibility*. The English version of this scale has been shown to have good psychometric properties and good convergent, discriminant, and incremental validity (Bond et al., 2011). The Dutch translation has yielded similar reliability values.

2.1.3.4 Fear of Spiders Questionnaire (FSQ).

The Fear of Spiders' Questionnaire (FSQ; Szymanski & O'Donohue, 1995) is an 18-item self-report scale for assessing spider phobia. The FSQ is capable of assessing both low and high levels of reported spider phobia with high retest reliability (.97) and high internal consistency (Cronbach's Alpha = .92; Symanski & O'Dononhue, 1995).

2.1.3.5 The Implicit Relational Assessment Procedure (IRAP).

The IRAP is a computer based programme that requires participants to respond quickly and accurately to specific stimuli that are deemed either consistent or inconsistent with their prior learning histories and/or response biases. The stimuli are presented in the forms of trials within a series of blocks. The general assumption of the IRAP is that, all things being equal, relational responding should be quicker and more accurate across blocks of trials that require responding that is consistent with the participant's learning history and/or response biases than on blocks that require responding in a manner that is inconsistent with that history and/or response biases. The primary datum from the IRAP is response latency, which is measured in milliseconds, and defined as the time that elapses from the onset of stimuli in each trial to the emission of a correct response. Participants were required to complete two separate IRAPs, one Training-IRAP and one Test-IRAP.

2.1.3.5.1 Training-IRAP.

The Training-IRAP differs from the traditional IRAP in that it presents consistent blocks of trials only. The objective of the Training-IRAP is to train participants to produce a consistent pattern of relational responses to specific mastery criteria based on accuracy and latency (i.e., unlike a traditional IRAP it does not present alternating blocks of trials that require opposite patterns of responding). The label stimuli for the Training-IRAP consisted of one of sixteen images, eight images of pets (i.e., kittens or puppies) and eight images of spiders presented on each trial. Participants were presented with pictures of either kittens or puppies based on which animal they had a preference for. This was determined by simply asking participants which animal they preferred at the beginning of the IRAP setup. The target stimuli consisted of one of two images, one image of a square and the other of a circle presented on each trial. Two Training-IRAPs were employed in the current study in order to counterbalance the type of shape stimulus that was trained to spiders and pets across participants. Specifically, one Training-IRAP was designed to train a relational network in which spiders were established as similar to square and different to circle; and pets were established as similar to circle and different to square. The other Training-IRAP was designed to train a network in which spiders were established as similar to circle and different to square; and pets were established as similar to square and different to circle. The IRAP software presented all stimuli and recorded participant responses. Each trial presented one of two labels (i.e., pictures of spiders or pictures of pets) at the top of the screen, one of two target stimuli (i.e., pictures of a circle or a square), which were presented in the middle of the screen. Two response options (i.e. “similar” and “different”) were also presented on each trial, at the bottom left and right of the screen. After a correct response on each trial, the screen cleared and the next trial appeared. After an inaccurate response, a red X appeared until a correct response was emitted. All spider and pet images were identical to those employed in Leech et al (2016). The Training-IRAP thus comprised of 4 trial-types*:* Pet->Circle, Pet->Square, Spider->Circle, Spider->Square (see Figure 1).

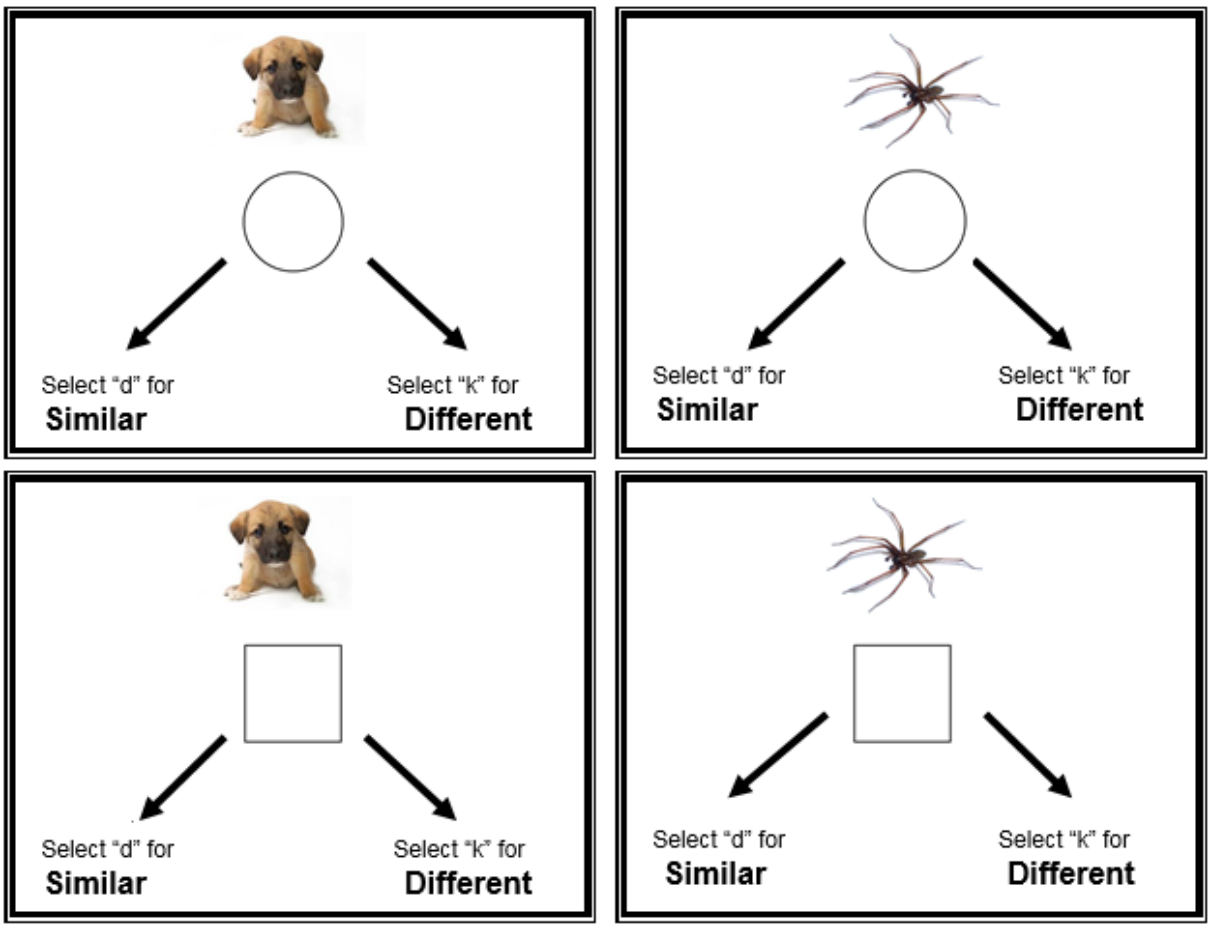

Figure 1. Diagrammatic representation of the four IRAP trial-types. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Pet->Circle, Pet->Square, Spider->Circle, Spider->Square.*

2.1.3.5.2 Test Fear-IRAP.

This IRAP was a "traditional" version, in that it was designed to assess four specific response biases by presenting blocks of trials that required participants to respond in diametrically opposite patterns across successive blocks of trials. The label stimuli for the Fear-IRAP consisted of two images, one square and one circle. Each label stimulus was presented with one of sixteen target statements, eight of which were positive and eight were negative (see Table 1). Examples of the four trial-types for the Fear-IRAP are presented in Figure 2. The response options for the Fear-IRAP were "Yes" and "No".

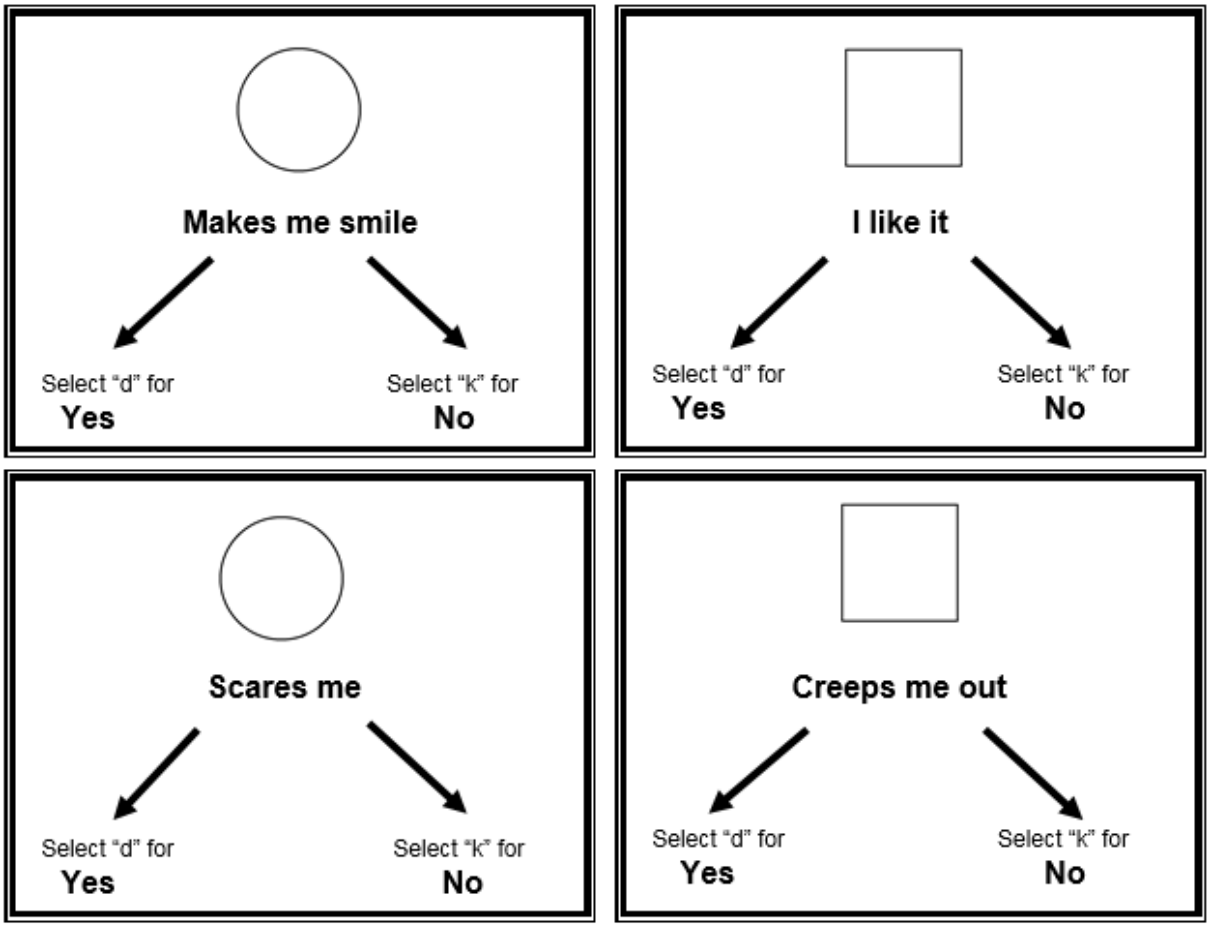

Figure 2. Diagrammatic representation of the four Test Fear-IRAP trial-types. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Pet-Pleasant, Pet-Fear, Spider-Pleasant,* and *Spider-Fear*.

Table 1

Target Stimuli for Test Fear-IRAP

Target Stimuli - Pleasant	Target Stimuli - Fear
Calms me	Scares me
Gives me warm feelings	Makes me uncomfortable
Let's me feel happy	Frightens me
Makes me happy	Makes me anxious
I like it	I hate it
Relaxes me	Gives me stress
Makes me smile	Gives me shivers
Reassures me	I find it annoying

2.1.3.6 Behavioral Approach Task (BAT).

A Chilean Rose tarantula approximately 8cm in diameter (including legs) was employed in the current study. The tarantula was housed in a transparent terrarium with a secure lid. The spider was cared for throughout the study and was used with all participants.

2.1.4 Procedure

There were 4 stages involved in the study: 1. Questionnaires; 2. Training-IRAP (either Spider-Circle or Spider-Square); 3. Test Fear-IRAP; 4. BAT. Participants were randomly assigned to one of the two conditions (Spider->Circle and Spider->Square) before being exposed to the 4 stages of the study.

2.1.4.1 Questionnaires.

Participants completed the four questionnaires in a fixed sequence (Spider Fear Rating Question, the DASS, the AAQ-II, and the FSQ).

2.1.4.2 Training-IRAP.

All participants completed a minimum of one training block. They were advised that each trial would present an image of a spider or a pet at the top of the screen, with an image of a square and a circle in the center, and that their task was to respond to these two things together using one of the two responses options. Participants were informed that their job was to figure out the correct pattern as accurately as possible across each block of trials. The IRAP consisted of blocks of 32 trials, with each of the 4 trial-types presented 8 times within each block. On each trial, a label (e.g., image of spider) appeared on top, a target (e.g., image of square) in the middle, and both response options (i.e., “similar” and “different”) on the bottom left- and right-hand corners of the screen. Participants selected a response by pressing D (for the left option) or K (for the right). After a correct response, the screen cleared and the next trial appeared. After an inaccurate response, a red X appeared until a correct response was emitted. For half of the participants, correct responding on each of the four trial-types was as follows: Pet-Circle/Similar; Pet-Square/Different; Spider-Circle/Different; and Spider-

Square/Similar. For the remaining participants, correct responding required: Pet-Circle/Different; Pet-Square/Similar; Spider-Circle/Similar; Spider-Square/Different. If participants failed to achieve either the accuracy (≥ 87.5%) or the latency criterion (median ≤ 2000 *ms)* on each of the trial-types during a single block, they received automated feedback, and another block was presented. These mastery criteria were applied to ensure that participants were responding with roughly equal speed and accuracy to each of the trial-types. Once participants reached the accuracy and latency criteria, the program ended. Participants were allowed to attempt 5 blocks of trials. If a participant failed to reach both criteria after the 5 blocks the program closed and a message appeared that read "Please contact the researcher". Participants were then debriefed and thanked for their time.

2.1.4.3 Test Fear-IRAP.

Participants who completed the Training-IRAP successfully were exposed immediately to a traditional IRAP (i.e., the Fear-IRAP), similar to that employed by Leech, et al. (2016, 2017). However, rather than presenting pictures of spiders and pets in the IRAP (with positively and negatively valenced statements; e.g., "I like it" versus "It creeps me out), the label stimuli were pictures of the square and circle shapes. In effect, participants were asked to confirm and disconfirm the positivity and negativity of the two types of shape, rather than spiders and pets. In order to define the blocks of trials as consistent versus inconsistent it is necessary to consider both the pre-experimental and Training-IRAP histories of the participants. For present purposes, a history-consistent block of trials required that participants respond in a manner that (i) coordinated a shape, which had been coordinated with spider pictures, with fear-related statements, and (ii) coordinated a shape, which had been coordinated with pet pictures, with positive statements (hereafter referred to as anti-spider blocks, *although no pictures of spiders were actually presented during these blocks*). A history-inconsistent block of trials required that participants respond in the opposite manner; (i) coordinated a shape, which had been coordinated with spider pictures, with positive

statements, and (ii) coordinated a shape, which had been coordinated with pet pictures, with fear-related statements (hereafter referred to as pro-spider blocks, *again no pictures of spiders were actually presented during these blocks*).

Participants were exposed to a maximum number of four pairs of practice blocks, on which they could reach the performance criteria of accuracy ≥ 87.5% and median latency ≤ 2000 *ms* at the trial-type level (i.e., these criteria were applied to each trial-type, not just the overall block). Once participants achieved these criteria, they automatically advanced to the test blocks. No performance criteria were applied for progression through the test blocks, but performance feedback, detailed below, was presented at the end of each block to encourage participants to maintain the practice-block criteria (≥ 87.5% correct and ≤ 2000ms latency at the trial-type level). It is important to emphasize that participants were strongly encouraged to respond as quickly and as accurately as possible across each trial-type and across each block. Also, no reference was made, in any formal or informal instructions, to the shapes as being spider- or pet-related. The only context in which the shapes were related to spiders and pets was during the Training-IRAP described above.

Each practice block and each test block consisted of 32 trials composed of four trial-types, each presented eight times within a block. The four trial-types were defined in terms of a 2x2 combination of the two label stimuli with the two types of target stimuli: Pet trained Shape-Pleasant (hereafter referred to as *Pet-Pleasant*)*;* Pet trained Shape-Fear (hereafter referred to as *Pet-Fear);* Spider trained Shape-Pleasant (hereafter referred to as *Spider-Pleasant*); Spider trained Shape-Fear *(*hereafter referred to as *Spider-Fear).* Examples of these four trial-types are as follows; (i) Pet Trained Shape/"Makes me smile"; (ii) Pet Trained Shape/"Terrifies me"; (iii) Spider Trained Shape/"I like it"; (iv) Spider Trained Shape/"Scares me". The four trial-types were presented in a quasi-random order, such that each trial-type was presented once every four trials (the same trial-type was never presented twice in succession).

On each trial, an image of either the spider-trained shape or the pet-trained shape (i.e., Circle or Square) appeared in the upper center of the screen. Below this, in the center of the screen, a target stimulus appeared (i.e., a fear- or pleasant-related statement). In the bottom third of the screen, the response options were presented *(*i.e., "Yes" and "No"). One response was presented on the bottom right corner; the other was presented on the bottom left corner. These response options alternated randomly across trials with the software ensuring that they did not appear in the same positions for more than three successive trials.

On anti-spider blocks of trials, participants were required to respond "Yes" if the spider-trained shape was presented with a negative statement, or if the pet-trained shape was presented with a positive statement; and to respond "No" if the spider-trained shape was presented with a positive statement, or if the pet-trained shape was presented with a negative statement. On pro-spider blocks of trials, the opposite pattern of responding was required (e.g., to respond "Yes" if the spider-trained shape was presented with a positive statement).

Responses deemed correct for a given block of trials cleared the label, target and response option stimuli; the next set of stimuli appeared 400ms later. Incorrect responses produced a red "*X*" below the target stimulus, which remained on screen (with the label and response option stimuli) until the correct response was emitted. If a participant did not emit a response before 2000ms on any trial, a red exclamation mark ("!") appeared directly below where the red X was presented for incorrect responses, and it remained on screen until a response (correct or incorrect) was emitted.

2.1.4.4 Behavioral Approach Task (BAT).

There were seven steps involved in the BAT, which progressively asked participants to move physically closer to a live tarantula. Each of the BAT steps were read aloud by the experimenter. This was scored from 0 to 7 as participants progressed through each step. The instructions were as follows:

> *The following test is to measure how willing you are to approach a live tarantula. I will ask you if you are willing to do a number of items, one at a time, and if you are willing to do the items, I'll ask you to do so. If at any time you do not want to continue, please feel free to stop.*

The first step involved participants opening the door to the room where the tarantula was kept. If participants failed to complete the first step of opening the door they scored 0, if they opened the door they scored 1 (this score was increased as participants completed the different steps). For the second step, participants were asked if they were willing to enter the room. The third step brought participants closer again and required them to look closely at the tarantula with their face level with the terrarium. The fourth step required participants to place their hands on either side of the terrarium and hold them in place for 10 seconds. The fifth step required participants to sit down and place the terrarium on their lap for 20 seconds whilst also keeping their hands on either side of the terrarium. The sixth step required participants to open the feeding latch in the lid of the terrarium and keep it open. The seventh and final step required participants to place their hand inside the terrarium but ensuring that at no point they physically touched the tarantula.

2.2 Results

2.2.1 Validating the BAT

The correlation between the FSQ and the BAT proved to be relatively strong and significant (r = -.618, $p \leq .0001$), indicating that higher reported levels of fear on the FSQ predicted fewer approach steps on the BAT.

2.2.2 Scoring the IRAP

The primary datum from the IRAP was response latency, which was defined as the time in milliseconds that elapsed from the onset of a trial to the emission of a correct response. Consistent with previously published studies employing the traditional IRAP, the data were screened before being subject to statistical analyses. If a participant's accuracy fell below 87.5%, or if their the median latency exceeded 2000 *ms* across any of the four trial-

types across the six test blocks, the data for this participant were excluded from further analyses. Specifically, participants were allowed to make a maximum of six errors per trial-type across the six test blocks (the median latency for each trial-type within each block had to remain below 2000*ms*). The data from 26 participants were removed on this basis. The latency data from the Fear-IRAP were transformed into *D*-IRAP scores (see Barnes-Holmes, Barnes-Holmes, Stewart & Boles, 2010) (see Table 2).

Table 2

Method for Converting the Response Latencies from Each Participant into D-IRAP Scores

1	The response latencies from only the six test blocks were utilized for the data analysis.
2	Any latency that exceeded 10,000 *ms* was removed from the data set.
3	If the data of a participant contained response latencies of less than 300 *ms* in more than 10% of test block trials, the participant was removed from the data set.
4	Standard deviations were calculated for each of the four trial types per pair of test blocks: four from the response latencies from the first and second test blocks, four from the response latencies from the third and fourth test blocks, and four from the response latencies from the fifth and sixth test blocks.
5	A mean latency score was calculated for each of the four trial types in each test block. This resulted in 24 (consistent and inconsistent) mean latencies for the four trial types over the six test blocks.
6	A difference score was calculated for each of the four trial types in each test block. This was done by subtracting the latency of the anti-spider test block from the corresponding pro-spider test block.
7	Each difference score was divided by its corresponding standard deviation (Calculated in step 4) which yielded 12 *D*-IRAP scores, one for each trial type for each pair of test blocks.
8	Four overall *D*-IRAP scores were calculated for each trial type. This was done by averaging the scores for each of the four trial types across each of the three pairs of test blocks.

Given the forgoing transformation, a larger *D*-IRAP score indicated a greater difference in mean response latencies between the two types of blocks (pro- versus anti-spider

blocks) for each trial-type. In order to facilitate direct comparisons across the spider and pet trial-types, the signs for the *Spider-Fear* and *Spider-Pleasant* trial-types were reversed (i.e., + scores became negative, and – scores became positive). Positive *D*-IRAP scores now indicated a positive bias for both spiders and pets and negative scores indicated a negative bias for both types of stimuli.

2.2.3 Mean Score Analyses

A preliminary mixed repeated measures 2x4 analysis of variance (ANOVA) determined if the counterbalancing variable for type of arbitrary stimulus (circle versus square) yielded a main or interaction effect with the four IRAP trial-types but neither effect was significant (*ps* > .2), and thus this variable was removed from all subsequent analyses. The mean *D*-IRAP scores for the four trial-types are presented in Figure 3. The mean IRAP effect was negative for the *Spider-Fear* trial-type, but positive for the remaining three trial-types. In concrete terms, participants tended to respond "Yes" more quickly than "No" when presented with the derived spider stimulus and a fear statement or the derived pet or spider stimulus and a pleasant statement. When presented with the derived pet stimulus and a fear statement, participants showed a tendency to respond "No" more quickly than "Yes".

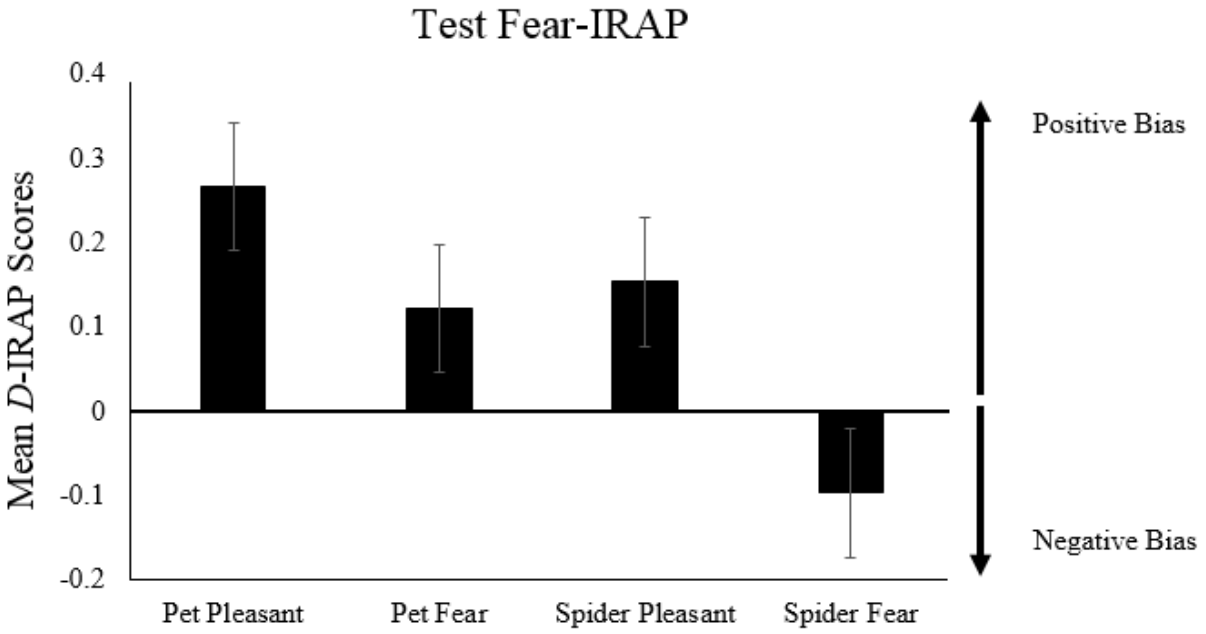

Figure 3. Four mean *D*-IRAP scores for the Test Fear-IRAP with error bars.

A one-way repeated measures analysis of variance (ANOVA) was conducted with trial-type as the repeated measure and this proved to be significant F (3,30) = 6.35, p = .0006, η_p^2 = .17. Post hoc comparisons using Fisher's PLSD tests indicated that the mean scores for the *Pet-Pleasant* (M = .266, SE = .053), *Pet-Fear* (M = .124, SE = .59) and *Spider-Pleasant* (M = .153, SE = .061) trial types were significantly different from the score for the *Spider-Fear* (M = -.097, SE = .059) trial type (remaining *ps* > .2). Four one sample *t*-tests indicated *Pet-Pleasant, Pet-Fear, Spider-Pleasant* trial-type effects were significantly different from zero (*ps* < .043) but the effect for *Spider-Fear* was not (p > .11).

2.2.4 IRAP-Explicit/BAT Correlational Analyses

A correlation matrix was calculated to determine if any of the four trial-types from the IRAP predicted self-reported fear of spiders (on the FSQ) and approach responses on the BAT; none of the correlations were significant (*ps* > .12).

2.2.5 Discussion

The first experiment used a Training-IRAP to establish a relational network containing pictures of pets and spiders and arbitrary stimuli (circle and square). Subsequently, a traditional Fear-IRAP presented the arbitrary stimuli with pleasant and fear-related statements. A broadly similar pattern of response biases was recorded for the Fear-IRAP as had been observed in previously published studies using pictures of pets and spiders (Leech et al., 2016, 2017). The experiment thus supports the widely assumed but largely untested assumption that the relational context provided by the IRAP may both serve to establish and reveal fear-related response biases in arbitrary stimuli.

The current results are consistent with recent research on testing equivalence relations with an IRAP (Bortoloti & de Rose, 2012), but this is the first experiment to demonstrate any sort of derived transformation effect involving fear-related stimuli, thus rendering the current findings potentially important in understanding the basic behavioral processes involved in fear acquisition. Indeed, as noted in the introduction the current derived transformation effects emerged from a training context (i.e., the IRAP) that did not involve differential amounts of stimulus pairings, and thus a simple explanation in terms of associative or Pavlovian learning processes seems untenable.

On balance, when an experiment or study produces or confirms a predicted effect it does seem to be important to determine if there is a contextual variable (or variables) that may reduce or undermine the effect. Such analyses could be important in developing a progressive science that involves reporting both successes and failures within the same study so that future attempts to replicate the effect, and succeed or fail in that regard, may be better understood in terms of the original study.

In Experiment 2, therefore, we attempted to replicate the basic effect observed in Experiment 1 but using arbitrary stimuli that differed considerably from two simple geometric shapes. Specifically, we employed pictures of two relatively unknown marsupials, a quoll and a quokka. We chose these stimuli because they would be almost completely unfamiliar to

European participants but had been used successfully in a study to establish response biases using the implicit association test (IAT; Field & Lawson, 2003). Experiment 2, therefore, was similar to Experiment 1, except that the arbitrary stimuli consisted of a picture of a quoll and a picture of a quokka rather than pictures of a square and a circle. In addition, participants were exposed to two IRAPs, one that focused on fear responses and the other on avoidance responses (similar to Leech et al., 2016, 2017). The purpose of this second experiment, therefore, was to determine if we could replicate the basic effect observed in Experiment 1 for both Fear- and Avoidance-IRAPs but using different arbitrary stimuli. As will become clear, the results of Experiment 2 suggested a possibly important boundary condition for the IRAP as a training and/or testing context for establishing fear-related response biases for arbitrary stimuli.

2. Experiment 2

3.1 Method

3.1.1 Ethical Considerations

The study was conducted in accordance with the ethical guidelines of Ghent University. Prior to the experiment, participants read and signed a consent form informing them that they could withdraw from the study at any time. Upon completion, participants were fully debriefed.

3.1.2 Participants

Seventy one undergraduate students attending Ghent University, Belgium, volunteered to participate in the study (N = 71; 51 Females, 20 Males). Participants were paid €10 for their participation. Forty one participants failed to achieve the necessary performance criteria on the IRAPs (see "procedure" section of Experiment 1), leaving data from 20 females and 10 males (N = 30) for analysis. The mean age was 23.5 years (SE = .879), with a range of 18-43 years. The participants completed the study individually in the Department of Experimental, Clinical and Health Psychology at Ghent University.

3.1.3 Materials

The study employed four questionnaires (i.e. a Spider Fear Rating Question, the Depression, Anxiety and Stress Scale-21 (DASS); the Acceptance and Action Questionnaire II (AAQ-II); and the Fear of Spiders Questionnaire (FSQ); a Training-IRAP, a Test Fear-IRAP and a Test Avoidance-IRAP, and a Behavioral Approach Task (BAT). All verbal material used in the current study were presented to participants in Dutch.

3.1.3.1 Questionnaires.

All questionnaires administered to participants were identical to those employed in Experiment 1.

3.1.3.2 The Implicit Relational Assessment Procedure.

A Training-IRAP and two Test-IRAPs were employed in the current study.

3.1.3.2.1 Training-IRAP.

The Training-IRAP employed in the current study was similar to that used in Experiment 1 with one exception. The target stimuli consisted of one of two images of Australian marsupials (a Quoll and a Quokka) rather than the geometric shapes (a circle and a square) used in Experiment 1 (See Figure 4). The label stimuli (pictures of pets and spiders) and all other aspects of the Training-IRAP were identical to those employed in Experiment 1.

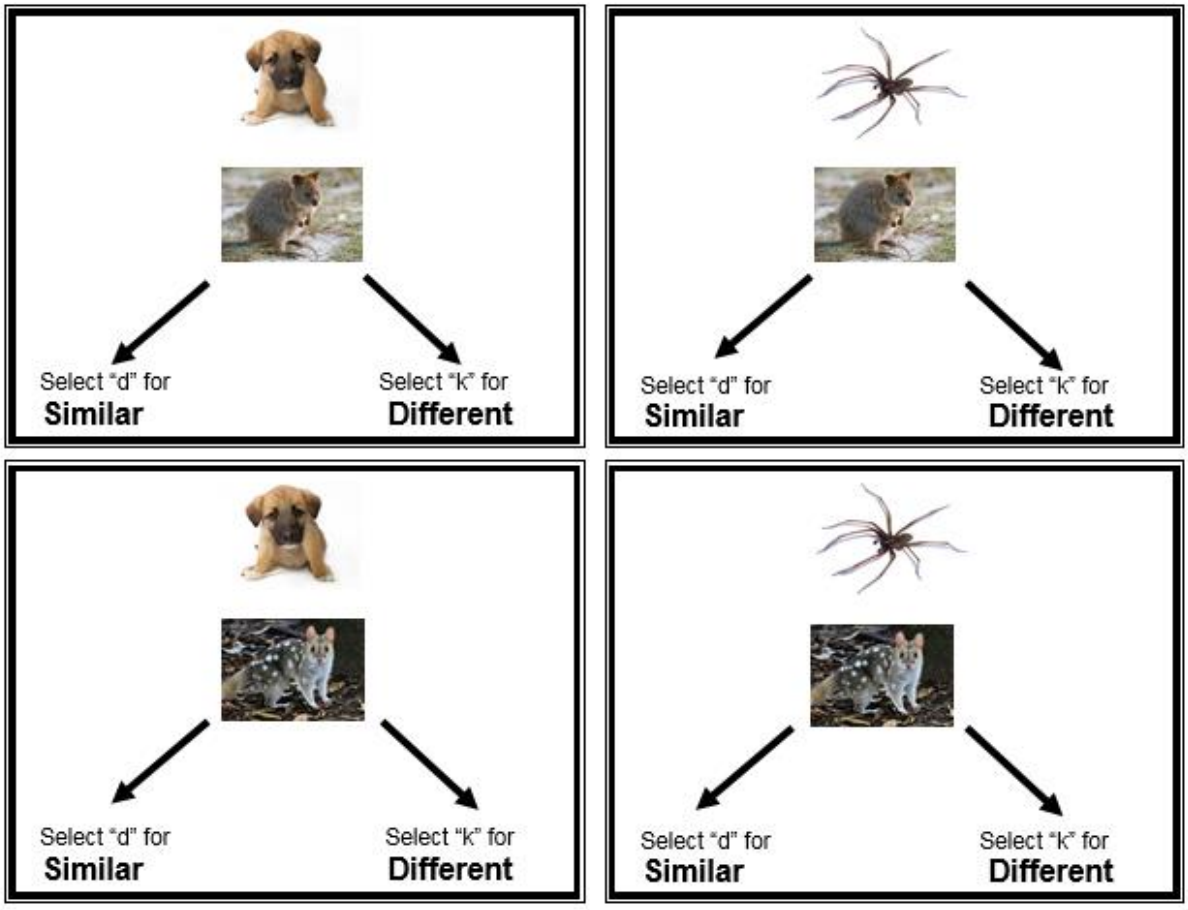

Figure 4. Diagrammatic representation of the four IRAP trial-types. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Pet->Quokka, Pet->Quoll, Spider->Quokka, Spider->Quoll.* The quokka appears in the upper panels and the quoll in the lower panels.

3.1.3.2.2 Test Fear-IRAP.

The Fear-IRAP (hereafter referred to as F-IRAP) employed was similar to that used in Experiment 1, except the target stimuli displayed the images of the Quoll and the Quokka rather than images of circles and squares. All other aspects were identical to Experiment 1. Examples of the four trial-types for the F-IRAP are presented in Figure 5. The response options for the F-IRAP were "Yes" and "No".

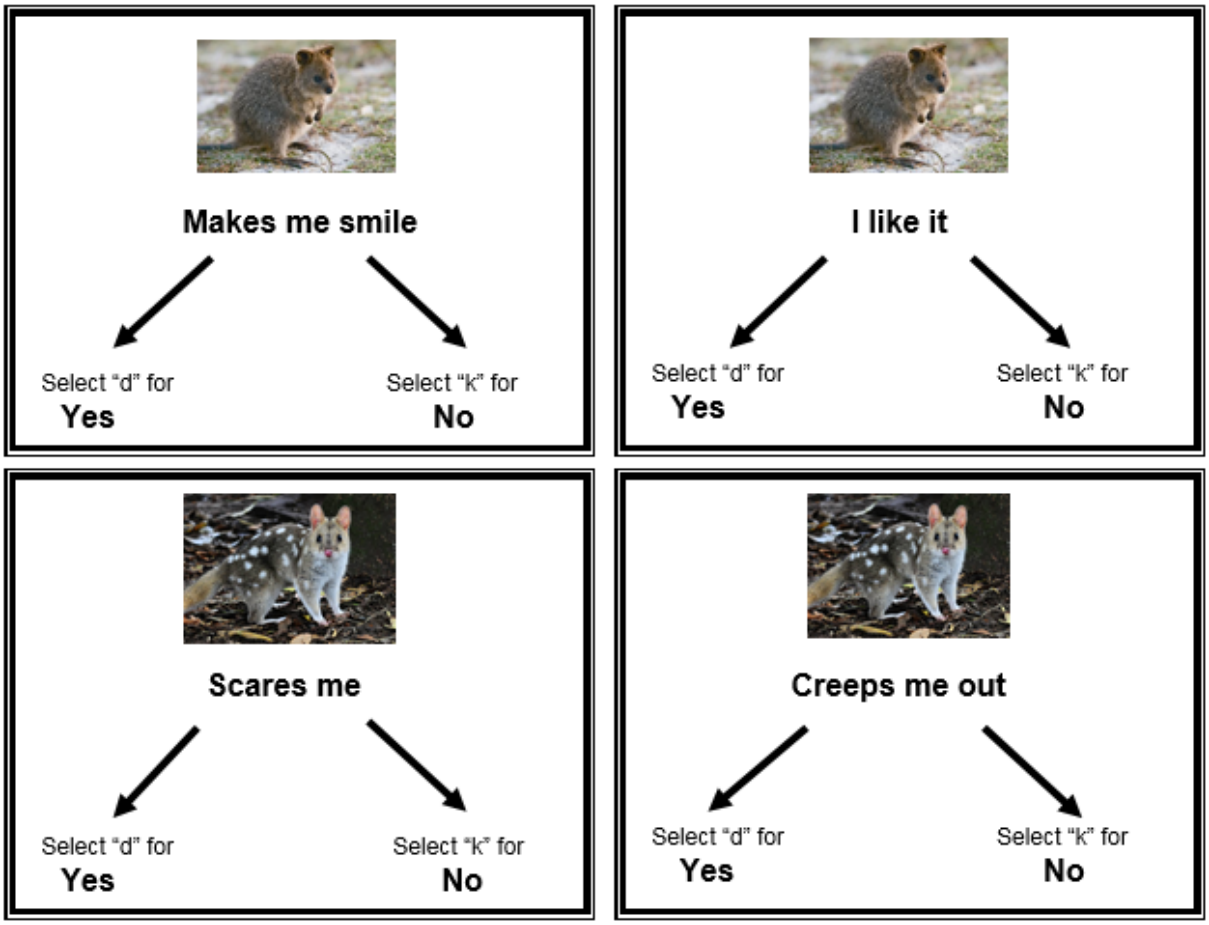

Figure 5: Diagrammatic representation of the four Test Fear-IRAP trial-types. Arrows did not appear on screen. The four IRAP trial types were denoted as: *Pet-Pleasant*, *Pet-Fear*, *Spider-Pleasant*, and *Spider-Fear*.

3.1. 3.2.3. *Test Avoidance-IRAP*.

The Test Avoidance-IRAP (hereafter referred to as the A-IRAP) was similar to the F-IRAP except the target stimuli presented were modified. Specifically, the target stimuli referred to approach and avoidance responses (see Table 3), which were identical to those used by Leech et al. (2016, 2017). The label stimuli used in the A-IRAP were identical to those employed in the F-IRAP (i.e., the quoll and the quokka).

Table 3

Target Stimuli for Test Avoidance-IRAP

Target Stimuli – Approach	Target Stimuli - Avoid
I approach it	I get away
I hold it	I leave
I touch it	I flee
I pick it up	I run away
I carry it	I stand back
I look at it	I avoid it
I play with it	I jump away
I stay	I go away

3.1.4 Procedure

The procedures for the F-IRAP and A-IRAP were similar to those employed in Experiment 1. In relation to the F-IRAP, the four trial-types were defined in terms of a 2x2 combination of the two label stimuli with the two types of target stimuli: Marsupial/Pet-Pleasant (hereafter referred to as *Pet-Pleasant);* Marsupial/Pet-Fear (hereafter referred to as *Pet-Fear*); Marsupial/Spider-Pleasant (hereafter referred to as *Spider-Pleasant*); Marsupial/Spider-Fear (hereafter referred to as *Spider-Fear*). The reader should note that although the trial-types are referred to using "Pet" and "Spider", the actual pictorial stimuli presented during the IRAP were the two marsupials that had previously been related to pictures of pets and spiders (i.e., no pictures of pets or spiders were presented during the test IRAP).

In relation to the A-IRAP, the four trial-types were defined as: Marsupial/Pet-Approach (hereafter referred to as *Pet-Approach);* Marsupial/Pet-Avoid (hereafter referred to as *Pet-Avoid*); Marsupial/Spider-Approach (hereafter referred to as *Spider-Approach*); Marsupial/Spider-Avoid (hereafter referred to as *Spider-Avoid*). Note again, that no pictures of pets or spiders were presented during the test A-IRAP).

3.1.5 Behavioral Approach Task (BAT)

The BAT employed in Experiment 2 was identical to that employed in Experiment 1.

3.2 Results

All aspects of scoring the IRAP and subsequent analyses were similar to those employed in Experiment 1.

3.2.1 Validating the BAT

The correlation between the FSQ and the BAT proved to be relatively strong and significant (r = -.597, p = .0003), indicating that higher reported levels of fear on the FSQ predicted fewer approach steps on the BAT.

3.2.2 Mean Scores Analyses

Two preliminary mixed repeated measures 2x4 ANOVAs were conducted to determine if the counterbalancing variable for type of arbitrary stimulus (quoll versus quokka) yielded significant main effects or interacted with the four IRAP trial-types. In both cases the effects were non-significant (ps > 0.3), and thus this variable was removed from subsequent analyses.

The mean *D*-IRAP scores for the eight trial-types from the F-IRAP and A-IRAP are presented in Figure 6. For the F-IRAP, the mean *D*-IRAP effects showed positive bias scores for the two pleasant trial-types (i.e. *Pet-Pleasant* & *Spider-Pleasant*) and negative scores for the two fear trial-types (i.e. *Pet-Fear* & *Spider-Fear*). For the A-IRAP the *D*-IRAP bias scores were all positive. Two one way repeated measures ANOVAs, one for each IRAP, indicated that the effect for trial-type was significant for the F-IRAP, $F(3,29) = 5.717$, p = .001, ηp^2 = .165, with marginal significance recorded for the A-IRAP, $F(3,29) = 2.585$, p = .058, ηp^2 =.082. The results of a series of post hoc comparisons using Fisher's PLSD tests for both IRAPs are presented in Table 4. The two pleasant trial-types differed significantly from the two fear trial-types for the F-IRAP, and the two approach trial-types differed significantly from the two avoid trial-types for the A-IRAP. In general, therefore, the differential trial-type effects for both IRAPs appeared to be driven largely by the pleasant/fear variable for the F-IRAP and the approach/avoid variable for the A-IRAP, rather than any derived pet or spider functions for the two marsupials. In short, there was limited evidence for the derived

transformation of functions observed in Experiment 2 when Australian marsupials were used as arbitrary stimuli.

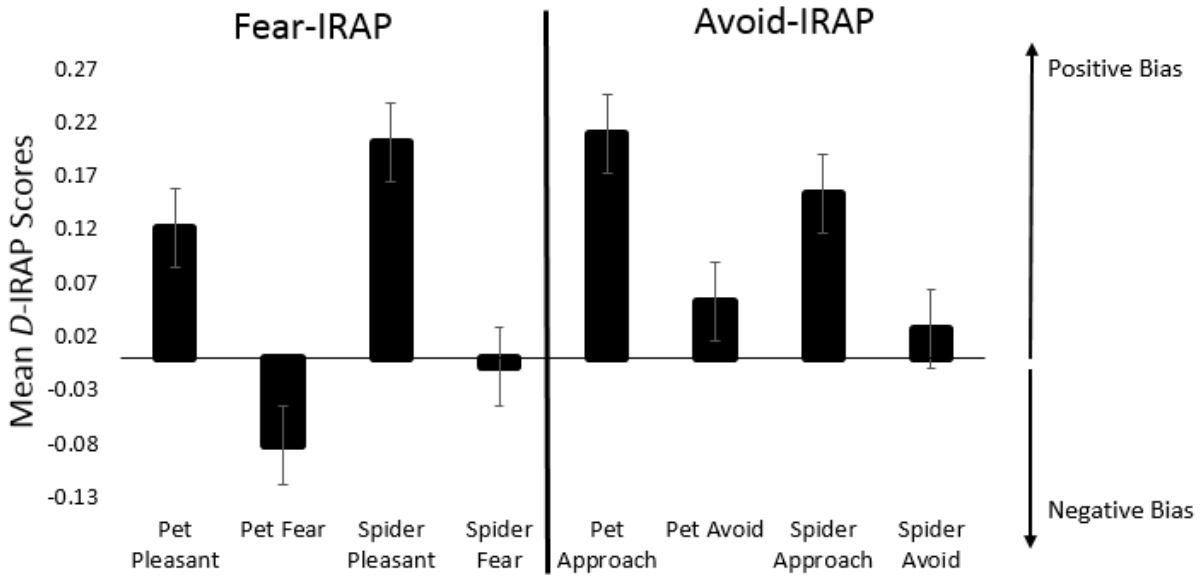

Figure 6. Eight D-IRAP trial-types for the Fear- and Avoidance-IRAPs with error bars.

Table 4

Post hoc comparisons for the Fear- and Avoidance-IRAPs

F-IRAP	Mean Diff	*p* Value	A-IRAP	Mean Diff	*p* Value
Pet Pleasant/Pet Fear	.203	.008	Pet Approach/Pet Avoid	.157	.040
Pet Pleasant /Spider Pleasant	-.08	.292	Pet Approach/Spider Approach	.056	.456
Pet Pleasant/Spider Fear	.13	.088	Pet Approach/Spider Avoid	.183	.017
Pet Fear/Spider Pleasant	-.283	.000	Pet Avoid/Spider Approach	-.101	.186
Pet Fear/Spider Fear	-.073	.332	Pet Avoid/Spider Avoid	.026	.729
Spider Pleasant/Spider Fear	.209	.006	Spider Approach/Spider Avoid	.127	.096

3.2.3 IRAP-Explicit/BAT Correlational Analyses

A correlation matrix was calculated to determine if any of the eight trial-types from the F- and A-IRAPs predicted self-reported fear of spiders (on the FSQ) and approach responses on the BAT; none of the correlations were significant (*ps* > .05).

3. General Discussion

One of the main aims of Experiment 1 was to establish a relational network containing pictures of pets and spiders and arbitrary stimuli (circle and square) within a Training-IRAP context, and subsequently to test for the transformation of functions for the shape stimuli using a Test Fear-IRAP that contained pleasant and fear-related statements. The results of Experiment 1 yielded a broadly similar pattern to that reported by Leech et al. (2016, 2017), suggesting the successful transformation of pleasant and fear functions for the circles and squares in accordance with the relational network containing pictures of pets and spiders. Experiment 2 attempted to replicate and extend the effects observed in Experiment 1 but using pictures of relatively unknown Australian marsupials (a Quoll and a Quokka) as arbitrary stimuli (rather than a circle and square). In addition, two test IRAPs were employed in the second Experiment (a Fear- and an Avoidance-IRAP). Interestingly, the pattern of results recorded for both test IRAPs in Experiment 2 failed to replicate the pattern observed in Experiment 1 or that reported in the previously published studies by Leech et al. Specifically, there appeared to be little evidence for the transformation of functions in the performances obtained from either of the two test IRAPs.

Before continuing it is important to note that although the rate of attrition was relatively high in Experiment 2, relative to Experiment 1 and the studies reported by Leech et al. (2016, 2017), only the data from those participants who achieved the necessary performance criteria in both the training and test IRAPs were included in the final data analyses. Thus, the apparent absence of the transformation of functions effect in Experiment 2 cannot be attributed to a simple failure to train to criterion or to maintain the appropriate stimulus control during the test IRAPs. Or to put it more informally, the participants in

Experiment 2 responded in accordance with the trained relational network (i.e., they knew which marsupials were similar and different to the pets and spiders), but failed to show the differential response latencies indicative of fear and avoidance that could be derived through the network.

The fact that Experiment 1 yielded evidence for a derived transformation of functions, but Experiment 2 did not, appears to identify a potentially important boundary condition when attempting to establish a derived transformation of fear and avoidance functions, at least when using the IRAP as a measure of such transformation effects. That is, establishing baseline training and testing performances using an IRAP may, but does not guarantee, a derived transformation of fear and avoidance functions. Such a conclusion, of course, raises the question of why we observed derived transformation in Experiment 1 (and the fear functions for pictures of spiders in the previously published studies) but not in Experiment 2. One possible explanation, that also may help to explain the relatively high attrition rate, was the topographical overlap between the marsupials and pet stimuli (versus spiders). That is, both marsupials were quite similar, in physical form, to each other and of course to the "cuddly and furry" pictures of the pet stimuli. As such, the discriminative responses required on the training and test IRAPs were relatively difficult to make in Experiment 2. Indeed, it may be that at least some participants focused on an idiosyncratic property of the marsupials such as whether it was spotted or not, or if it had a long or short tail, to learn and to maintain the correct relational responses on the IRAPs. Or more informally, participants learned to categorize the stimuli by focusing on physical properties that did not overlap between the marsupials and the pet stimuli, which increased the difficulty of the task relative to when simple geometric shapes were employed.[1] In addition, the use of shapes may have facilitated

[1] Following a suggestion made by one of the reviewers we calculated the number of blocks that were required to complete the training IRAPs across Experiments 1 and 2. The number of blocks was quite similar across experiments thus suggesting that participants did not find the discriminations with differed dramatically in difficulty between geometric shapes and between the two marsupials.

the derived transformation of fear functions for the spider stimuli. That is, it may have been relatively easy for participants to attribute spider qualities to geometric shapes where a circle could be seen as similar to a spider's bulbous body or a square, with its four sharp corners, being similar to the sharp and "pointy" features of a spider's legs. In other words, the visual properties of the shape images used in Experiment 1 may have evoked, at least to some degree, the visual functions of actual spiders during the test IRAP, which facilitated the observed transformation of fear functions. We shall return to the wider conceptual implications of this issue subsequently.

In addition to the potential impact of the physical features of the stimuli on training and test IRAP performances, it is also worth noting a procedural variable that could be important. Specifically, the performance criteria for both experiments in the current report were employed at the trial-type level rather than the block level, which was used by Leech et al., (2016, 2017). Specifically, participants in both of the current experiments were required to achieve and maintain $\geq$ 87.5% accuracy and $\leq$ 2000 *ms* across each trial-type in any given block, rather than across the total block of trials. Trial-type level criteria are relatively strict in the sense that participants are only allowed to make one error per trial-type per block of trials. The previously published studies by Leech et al., employed more "relaxed" criteria at the block level where participants could make more than one error per trial-type per block. Thus, in principle, participants could respond at chance level (e.g., 50% correct) on one trial-type provided they responded at 100% correct on the remaining three trial-types (i.e., 75% correct at the block level). Although this pattern of trial-type effects would be quite rare in the data we have collected over the years with the IRAP, the impact of using relatively relaxed (block level) versus strict (trial-type level) performance criteria remains to be determined. It is also worth noting that the level of instruction and/or verbal prompting typically employed in previous IRAP studies, including Leech et al., was much reduced in the current experiments (because they were focused on training and not just testing IRAPs). Again, this procedural

variable may help to explain the higher attrition rates. On balance, it is important to note that the attrition rates were higher in Experiment 2 (58%), relative to Experiment 1 (46%), and thus the impact of the marsupials as stimuli was likely a contributing factor.

The arbitrary stimuli employed in Experiment 2 were adapted from a study by Field and Lawson (2003), which presented the relatively unknown Australian marsupials in an IAT. As noted in the introduction, Field and Lawson. reported clear and significant IAT effects using these stimuli. That is, participants produced bias scores indicating positive valence for the animal that was presented with positive information, and/or negative valence for the animal presented with negative information. Given that we sought, in a broad sense, to replicate the findings reported by Field and Lawson, but clearly failed to so, it seems important to consider the numerous differences between the original study and Experiment 2 reported here. First, and most obviously, the study by Field and Lawson employed IATs whereas the current study used IRAPs, and these procedures differ in many respects. Second, the original study employed a population of children aged between 6-9 years, whereas Experiment 2 employed a population of undergraduate university students. Third, in the original study the children were first shown images of the quoll and the quokka and subsequently given a story about each of the animals. One of the stories contained positive information about one animal and the other story contained negative information about the other animal. In contrast, participants in Experiment 2 of the current study were given no information regarding the animals prior to exposure to the training and test IRAPs, and were required simply to engage with the tasks and respond according to the feedback provided. At the present time, therefore, it is difficult, if not impossible, to determine exactly what variable or variables were critical in producing the effects reported by Field and Lawson that were not present in the current study; that is, the measure, the sample (children versus adults), or the procedure used to establish the valence functions of the stimuli, or some combination thereof.

At this point, however, it seems worth noting that the contrast between the two studies may seem counter-intuitive. Specifically, the study by Field and Lawson (2003) successfully produced differential valence functions for the two types of marsupial using a context (a story about the two animals) which was very different to the testing context (the IAT). In contrast, in Experiment 2 of the current study the training and testing contexts were very similar (i.e., they were both IRAPs). At face value, therefore, one might have predicted that the current study would have at least yielded similar if not stronger effects than the original research. On balance, one might argue that the "story-telling" context of the original study embedded the marsupials in a far more elaborate or complex relational network than the relatively simple network established by the Training-IRAP. Perhaps appropriate transformation effects would have been obtained if the "story-telling" approach had been adopted in the current study. At the very least, therefore, the current findings are salutatory in that what would appear to be the most obvious research outcome is not necessarily what emerges from the data. Perhaps future research could pursue this line of inquiry.

One criticism of the current research might be that although one experiment demonstrated a clear transformation of function effect, the second one did not, and there was no third study that attempted to "resolve" the failure to replicate. The reason we chose not to pursue the issue with a third experiment becomes clear when one considers all of the potential variables that may have been involved in the failure to replicate. Specifically, the three previous paragraphs work through many of these potentially important variables and it is clear that the issue is far from simple. Thus any attempt to resolve the failure to replicate could well require many additional experiments. Indeed, a thorough analysis of the variables involved would likely require a relatively extended program of research. At this point, therefore, it seems important to share our research findings and allow other researchers the opportunity to explore the many variables that may be involved when a derived transformation of functions fails to emerge using the IRAP as a training and testing context.

As noted in the *introduction,* if a Training-IRAP successfully established a fear function for the shape that was deemed relationally similar to spiders (but not for the shape that was deemed similar to pets) it would be difficult to explain the effect in purely associative terms. The results of Experiment 1 were relatively clear cut in this regard. Assuming that the effect could be replicated (at least using shapes rather than unfamiliar and highly similar marsupials), the Training-IRAP may prove to be a potentially useful procedure for exploring the distinction between relational/propositional versus associative accounts of fear acquisition. For example, it has been argued that relations or propositions have truth values, whereas associations do not (see Hughes et al., 2011, for a detailed discussion). If the Training-IRAP is seen as a context that requires learning, at least in part, via relations/propositions, rather than simple associations, the version employed here could be developed to explore a far wider range of relational/propositional learning effects. For example, multiple training IRAPs could be used to establish a more complex relational network than was generated here (see reference below to combinatorial entailment).

The current findings may also be directly relevant to an important conceptual distinction within RFT. Specifically, the theory has always distinguished between entailing and the transformation of functions, each involving separate types of contextual control (Hayes et al., 2001). Any instance of relational framing thus involves relating stimuli in the context of a particular cue (a Crel) such as “same as,” “different from”, “opposite to”, “bigger/smaller than” and a contextual cue (a Cfunc) that selects a function that is transformed in accordance with the relation. For illustrative purposes, imagine that an English-speaking child learns that “perro” is the Spanish word for “dog” and the Flemish word for “perro” is “hond”. As a result, “hond” may be entailed with “dog” (via the frame of coordination), if the child correctly answers the question “What is the Flemish word for dog?” Different transformations of function may occur, however, if the child is then asked questions, such as “What does your hond look like?” versus “What does your hond smell like?” In the

first case, the Cfunc “look like” may evoke some of the visual properties of the family dog, whereas the Cfunc “smell like” may evoke some of the olfactory properties of the same dog. In effect, the two types of contextual control that define relational framing may interact independently. In the foregoing example, the Crel “this word means that word” establishes a frame of coordination between “dog” and “hond” (via “perro”), but the Cfuncs (“look like” and “smell like”) evoke different functions for “hond” in accordance with the same entailed relation.

The current findings suggest, if only tentatively, that the present study involved identifying a condition under which both Crel and Cfunc contextual control was observed (Experiment 1) and a condition under which Crel but not Cfunc control was recorded (Experiment 2). Specifically, all participants who successfully completed the experiment responded in accordance with the entailed relational network (successful Crel control), but only when evidence of spider fear (for the arbitrary stimuli) emerged was appropriate Cfunc contextual control involved. Of course, one might object that the current study did not require the complete entailing process, as defined by RFT, because all of the relations were “directly” trained and thus at best we were working with mutual rather than combinatorial entailment. A future study may thus attempt to replicate the current findings but using training and testing IRAPs that require both mutual and combinatorial entailing and the transformation of functions. In any case, the current study highlights the potential benefits of exploring the IRAP as a context for *training* and testing derived relational responding rather than simply as a measure of so-called implicit cognition.

Compliance with Ethical Standards

Conflict of Interest: Aileen Leech declares that she has no conflict of interest. Jaber Bouyrden declares that he has no conflict of interest. Nathalie Bruijsten declares that she has no conflict of interest. Dermot Barnes-Holmes declares that he has no conflict of interest. Ciara McEnteggart declares that she has no conflict of interest.

Ethical Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

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

Funding

This article was prepared with the support of an Odysseus Group 1 grant (2015 – 2020) awarded to the forth author by the Flanders Science Foundation (FWO) and a doctoral research scholarship awarded to the first author.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the implicit relational assessment procedure (IRAP) and the relational elaboration and coherence (REC) model. *The Psychological Record, 60 (3)*, 527–542. https://doi.org/10.1007/BF03395726

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2017). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *The Behavior Analyst*, 1-19. https://doi.org/10.1007/s40614-017-0124-7

Bernaerts, I., De Groot, F., & Kleen, M. (2012). De AAQ-II (Acceptance and Action Questionnaire-II), een maat voor experiëntiële vermijding: normering bij jongeren. *Gedragstherapie, 45*, 389-400.

Bond, F. W., Hayes, S. C., Baer, R., Carpenter, K., Guenole, N., . . . Zettle, R. (2011). Preliminary psychometric properties of the acceptance and action questionnaire-II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy*, *42*(4), 676–88. http://doi.org/10.1016/j.beth.2011.03.007

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(1), 41-54. https://doi.org/10.1007/BF03395785

deBeurs, E., vanDyck, R., Marquenie, L., Lange, A, & Blonk, R. W. B. (2001). De DASS; een vragenlijst voor het meten van depressie, angst en stress. *Gedragstherapie, 34*(1), 35–53.

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. http://doi.org/10.1901/jeab.1994.62-331

Dougher, M. J., Hamilton, D. A., Fink, B. C., & Harrington, J. (2007). Transformation of the discriminative and eliciting functions of generalized relational stimuli. *Journal of the Experimental Analysis of Behavior, 88*(2), 179-197. https://doi.org/10.1901/jeab.2007.45-05

Dymond, S., & Barnes, D. (1995). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more than, and less than. *Journal of the Experimental Analysis of Behavior, 64*(2), 163-184. https://doi.org/10.1901/jeab.1995.64-163

Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst, 23*(2), 239-254. https://doi.org/10.1007/BF03392013

Dymond, S., & Roche, B. (2009). A contemporary behavior analysis of anxiety and avoidance. *The Behavior Analyst, 32*(1), 7-27. https://doi.org/10.1007/BF03392173

Field, A. P., & Lawson, J. (2003). Fear information and the development of fears during childhood: Effects on implicit fear responses and behavioural avoidance. *Behaviour Research and Therapy, 41*(11), 1277-1293. https://doi.org/10.1016/S0005-7967(03)00034-2

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 Press.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy: An experiential approach to behavior change.* New York: NY. Guilford Press.

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology, 44*(2), 227–39. https://doi.org/10.1348/014466505X29657

Hughes, S., & Barnes-Holmes, D. (2011). On the formation and persistence of implicit attitudes: New evidence from the implicit relational assessment procedure (IRAP). *The Psychological Record, 61*(3), 391-410. https://doi.org/10.1007/BF03395768

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorising in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record, 61*(3), 465-496. https://doi.org/10.1007/BF03395772

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*(3), 337-349. http://doi.org/10.1007/s40732-016-0176-1

Leech, A., Barnes-Holmes, D., & McEnteggart, C. (2017). Spider fear and avoidance: A preliminary study of the impact of two verbal rehearsal tasks on a behavior–behavior relation and its implications for an experimental analysis of defusion. *The Psychological Record, 67*(3), 387-398. https://doi.org/10.1007/s40732-017-0230-7

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. http://doi.org/10.1016/0005-7967(94)00075-U

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodríguez-Valverde, M., Barnes-Holmes, D., . . . Gutierrez, G. (2013). Extinction of aversive eliciting functions as an analog of exposure to conditioned fear: Does it alter avoidant responding? *Journal of Contextual Behavioural Science, 2 (3-4),* 120-134. http://dx.doi.org/10.1016/j.jcbs.2013.05.001

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodríguez-Valverde, M., Barnes-Holmes, D., . . . Gutierrez-Martinez, G. (2014). Effects of an acceptance/diffusion intervention on experimentally induced generalised avoidance: A laboratory demonstration. *Journal of the Experimental Analysis of Behaviour, 101 (1),* 94-111. http://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 (2),* 263–278. http://doi.org/10.1007/s40732-016-0176-1

Perez, W. F., de Almeida, J. H., & de Rose, J. C. (2015). Transformation of meaning through relations of sameness and opposition. *The Psychological Record, 65*(4), 679-689.

Szymanski, J. & O'Donohue, W. (1995). Fear of spiders questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry*, *26*(1), 31-34. http://dx.doi.org/10.1016/0005-7916(94)00072-T

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

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

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. http://doi.org/10.1016/j.jbtep.2015.01.004

Whelan, R., & Barnes-Holmes, D. (2004). The transformation of consequential functions in accordance with the relational frames of same and opposite. *Journal of the Experimental analysis of Behavior*, *82*(2), 177-195. https://doi.org/10.1901/jeab.2006.113-04