Four self-related IRAPs: Analyzing and interpreting effects in light of the DAARRE model

Audrey Pidgeon[1], Ciara McEnteggart[2], Colin Harte[2], Dermot Barnes-Holmes[2, 3] and Yvonne Barnes-Holmes[2]

[1] Department of Psychology, National University of Ireland - Maynooth, Maynooth, Co. Kildare, Ireland

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

[3]School of Psychology, Ulster University, Coleraine, Northern Ireland, UK

Corresponding Author:

Colin Harte

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan, 2

9000 Ghent

Belgium

Email: Colin.Harte@UGent.be

Authors' Note This article was prepared with the support of an Odysseus Group 1 grant awarded to the fourth author by the Flanders Science Foundation (FWO). At the time of data collection, the first author was affiliated with the National University of Ireland, Maynooth. Currently, however, she is employed at Ability West, Blackrock House, Salthill, Galway, Ireland.

Abstract

Two studies are presented that involved exploring four different versions of the implicit relational assessment procedure (IRAP) to target self-relevant stimulus relations. Experiment 1 employed stimuli from previous research that used the IRAP to target stimulus relations pertaining to self, and self-esteem in particular. Experiment 2 aimed to explore the use of different types of stimuli (i.e., pictures as well as words), that again focused on self-related stimulus relations, and their potential correlations with measures of self-esteem and psychological distress. Experiment 1 yielded broadly similar findings to those reported previously. Experiment 2 showed that only one trial-type from the IRAP using pictures depicting success versus failure correlated with the measures of self-esteem and psychological distress; none of the remaining 11 trial-types across the 3 IRAPs yielded any significant correlations. The current findings may be seen as relatively progressive when presented in the context of a theoretical model that may be used, albeit in a post-hoc manner, to interpret the specific IRAP response patterns obtained in the current and previously published research. Specifically, an in-depth RFT conceptual analysis of the findings using a recently proposed model of IRAP effects is presented.

KEYWORDS: DAARRE model; IRAP; self-relevant stimuli; RFT

The current research presents two studies that involved exploring four different versions of the implicit relational assessment procedure (IRAP), each of which focused on self-relevant stimulus relations. The IRAP is derived directly from a behavior-analytic account of human language and cognition, known as relational frame theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). Although, by no means a requirement, it seems important to interpret results of IRAP research in terms of the theory that generated the method. In the case of the current article, therefore, the two studies presented here illustrate the value in exploring and developing the IRAP as a tool for assessing relational responses with regard to self, and, critically, also presenting an RFT-based conceptual analysis of the key findings.

On the grounds of intellectual honesty, it is important to note that the empirical research reported herein was conducted approximately eight years ago, but the conceptual analyses are based on work that has emerged only in recent years (e.g., Barnes-Holmes, Finn, Barnes-Holmes, & McEnteggart, 2018). Relatedly, when the empirical research was conducted, the main focus was on attempting to develop the IRAP as a measure of implicit self-esteem, but in light of more recent conceptual developments the focus on self-esteem *per se* has been replaced with a broader focus on "self-relevant" relational responding. In other words, the presentation of the data is not concerned with developing a measure of self-esteem, but rather seeking to generate an increasingly sophisticated understanding and treatment of the IRAP as a method for assessing the various properties of derived relational responding in general. It is important to distinguish, therefore, between the original purpose of the study for which the current data were collected and the presentation of the data in the current article (note, however, that for ease of communication we may sometimes refer simply to the "current study" rather than using the more accurate but cumbersome term "the current presentation of the data from the original study").

For RFT, the main conceptual unit of analysis is the derived stimulus relation. The IRAP was designed to provide a measure of the strength or probability of such relations, particularly those that had been established pre-experimentally in the natural verbal environment (see Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008). The core principle behind the IRAP is that, all things being equal, participants will tend to respond more quickly with regard to stimulus relations that are consistent with their individual learning histories, than those that are not. This difference in response accuracies and latencies is frequently referred to as the IRAP effect (a description of an IRAP is provided below). Critically, the term IRAP effect, or response bias, as employed throughout the current article, should not be seen as a proxy for an implicit attitude as defined in cognitive or social psychology. Rather, the terms *IRAP effect* and *response bias* simply indicate a tendency to respond in one specific direction over another (for a detailed theoretical account see Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010).

Since its development, many studies have demonstrated the utility of the IRAP in measuring response biases in a wide range of areas such as, race (e.g., Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010; Drake, et al., 2010), gender (e.g., Cartwright, Hussey, Roche, Dunne, & Murphy, 2017), religion (e.g., Hughes, Barnes-Holmes, & Smyth, 2017; Scheel, Roscoe, Schaewe, & Yarbrough, 2013), age (e.g., Cullen, Barnes-Holmes, Barnes-Holmes, & Stewart, 2010), and forensic investigation (e.g., Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009). The measure has also shown utility in predicting racial group status (Power, Harte, Barnes-Holmes & Barnes-Holmes, 2017) and parental smoking status (Cagney, Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2017) over and above that of standard self-report measures. Finally, a meta-analysis of clinically-related IRAP studies reported a relatively high level of predictive validity (Vahey, Nicholson, & Barnes-Holmes, 2015).

The basic structure of the IRAP typically involves presenting four different trial-types in blocks of trials. The four trial-types are used to create a 2x2 crossover design in which two separate label stimuli are presented with two separate target stimuli. On each trial, participants are required to choose one of two response options, indicating the stimulus relation between the label and target stimulus. Thus, for example, an IRAP might present a positively valenced label stimulus (at the top of the screen) with a positively valenced target stimulus (in the center of the screen) on one trial, with the response options "True" and "False" (at the bottom left-and right-hand sides of the screen). This trial-type may be denoted as *Positive-Positive;* the remaining three trial-types would be denoted *Positive-Negative, Negative-Positive,* and *Negative-Negative.* During some blocks of trials participants are required to respond in a history-consistent manner; choosing "True" on *Positive-Positive* and *Negative-Negative* trial-types and "False" on *Positive-Negative* and *Negative-Positive* trial types. On other blocks of trials the opposite response pattern is required (e.g., responding "False" on a *Positive-Positive* trial-type). The difference in a combined metric of accuracy and latency between history-consistent versus history-inconsistent blocks of trials yields an IRAP score or effect; typically four such scores are presented and analyzed.

The current study explored the impact of different types of stimuli employed within the IRAP in terms of the extent to which they produce, or fail to produce, specific response biases with regard to self-relevant stimulus relations. The research also sought to explore the extent to which such IRAP performances correlated with measures of psychological distress and self-esteem. Experiment 1 employed the stimuli from the first IRAP study that targeted stimulus relations pertaining to the self (Vahey et al., 2009). Anecdotal verbal reports by some participants indicated that they found specific features of the IRAP particularly challenging. In an effort to address this concern, Experiment 2 aimed to explore the use of

different types of stimuli than those employed in the original Vahey et al. study, but involved relations that were also deemed to be self-relevant. Selection of these stimuli was thus largely exploratory. Specifically, three IRAPs were completed by participants, one of which presented label stimuli in text format and two that presented label stimuli in picture format. That is, a "Self-regard" IRAP presented positive and negative attributes (in text format); an "Emotions" IRAP presented pictures of positive or negative emotions; and an "Achievement" IRAP presented pictures indicating success or failure. Within all three IRAPs, the target stimuli consisted of the phrases, "Like me" and "Not like me" and the response options "True" and "False".

As noted above, one of the primary aims of the current article is to consider the response patterns we obtained across the four separate IRAPs in the context of recent RFT-based conceptual analyses. These recent analyses may be traced back to one of the earliest IRAP studies, which noted a differential trial-type effect that could be explained, in part, by the pre-experimental functions of the two response options. In this early study, an IRAP was used to assess the response biases of white participants toward white and black individuals (Barnes-Holmes, Murphy, et al., 2010). Specifically, participants were presented with the words "Safe" or "Dangerous" as label stimuli at the top of the screen, with a picture of either a white or a black man holding a gun as target stimuli in the center of the screen, and the response options "True" and "False". The trial-types were denoted as *White-Safe, White-Dangerous, Black-Safe,* and *Black-Dangerous.* Although results showed pro-white and anti-black response biases, the anti-black bias was restricted to the *Black-Dangerous* trial-type. That is, when "Dangerous" and pictures of black men holding guns were presented, participants responded "True" more quickly than "False"; but this response bias (responding "True" more quickly than "False") was also observed for the *Black-Safe*

trial-type. In discussing these findings, the authors noted that the IRAP effects may have been influenced by variables beyond those involved in racial bias;

> . . . It is possible, therefore, that a bias toward responding "True" over "False," per se, interacted with the socially loaded stimulus relations presented in the IRAP. If such a response bias does play a role, however, the source of that bias needs to be explained (p. 62).

More recently, the potential interactions among the functions of the stimuli presented in an IRAP have attracted increasing attention, and have led to the development of the Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) model (see Finn, Barnes-Holmes, & McEnteggart, 2018). The DAARRE model is based on the assumption that specific differential trial-type effects may be explained by the interaction between the functional properties of the individual stimuli (e.g., the extent to which they evoke appetitive or aversive reactions) and the relation between the label and target stimuli (i.e., the extent to which they are coordinate or opposite/different). Critically, the functional properties of the response options are also factored into the model. The DAARRE model was generated directly from RFT, and as such, the individual functional properties of the stimuli are labeled as a *Cfunc* property, with the relationship between the label and target is labeled as a *Crel* property. As noted previously, the current studies were conducted some time before the DAARRE model was formulated and thus it would be inappropriate to present it here as a model for predicting the data we obtained. Thus we will present our findings and then use the DAARRE model, in the General Discussion, to interpret our findings in a post-hoc fashion. In doing so, we hope that the current work will be of value to other researchers who intend to employ the IRAP in their own research. We recognize that post-hoc theorizing has its limitations, but these typically apply to situations in which researchers fail to report that they have engaged in, or are engaging in, such theorizing.

Experiment 1

Method

Participants and Setting

Thirty people participated in Experiment 1, 15 females and 15 males. Participants ranged in age from 21 to 40 years, and were recruited via a snowball sampling method. No compensation was offered for participation. Participants completed the experiment on an individual basis in a quiet room (e.g., in the participant's own home or in their place of employment). The experimenter remained present at all times.

Apparatus and Materials

Experiment 1 comprised three self-report measures (The Rosenberg Self-Esteem Scale, the Depression Anxiety and Stress Scales, and a Feeling Thermometer) and an IRAP.

The Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) is a 10-item measure of self-esteem. All items (e.g., "I wish I could have more respect for myself") are rated on a 4-point Likert scale from 1 (Strongly agree) to 4 (Strongly disagree), with a minimum score of 10 and maximum of 40. Higher scores indicate higher self-esteem, while lower scores indicate lower self-esteem. The RSES has demonstrated good internal validity and reliability (e.g., Blascovich & Tomaka, 1991; DeHart & Pelham, 2007).

The *Depression Anxiety and Stress Scales* (DASS; Lovibond & Lovibond, 1965) is a 42-item measure assessing levels of depression, anxiety, and stress (each containing 14 items) in the past week. All items (e.g., "I felt I had nothing to look forward to") are rated on a 4-point scale from 0 (Did not apply to me at all) to 3 (Applied to me very much or most of the time), with a minimum score of 0 and maximum of 42 per scale. Scores are tallied for each

subscale and the sum of all subscale scores yields a total DASS score, yielding a minimum score of 0 and maximum of 126. Higher scores indicate higher levels of psychological distress. The DASS has demonstrated excellent internal consistency (.92 to .97; Antony, Bieling, Cox, Enns, & Swinson, 1998).

The *Feeling Thermometer* is presented pictorially as a visual analog scale and assesses self-warmth from 0°C ("I feel very cold about myself") to 100°C ("I feel very warm about myself"). Participants are asked to rate how warm they feel toward themselves by marking the position on the thermometer that most accurately reflects this feeling.

The current *IRAP* employed the same stimuli that were used by Vahey et al. (2009). The response options used in this IRAP consisted of the participant's name (e.g., "Peter") versus a response option that negated that name (i.e., "Not Peter"). This IRAP is thus referred to as the participant-name IRAP. Although the same stimuli were employed in the IRAP as those used by Vahey, et al (2009), it should not be seen as a replication of that earlier work because so many variables differed between the two studies (e.g., the sample of participants, the instructions provided to participants, the IRAP accuracy and latency criteria, the type of data analyses conducted on the IRAP data).

The participant-name IRAP presented two label stimuli (the words *Similar* or *Opposite*) at the top of the screen with 6 positive adjectives (*good, success, honest, capable, pleasant,* and *confident*) and 6 negative adjectives (*bad, failure, dishonest, worthless, nasty,* and *ashamed*) as target stimuli in the middle of the screen. As noted above, participants were also presented with two response options that comprised each participant's name or the word *Not* followed by the participant's name. The various label-target combinations on the IRAP yield four trial-types; *Similar-Positive, Similar-Negative, Opposite-Positive,* and *Opposite-*

Negative (see Figure 1). The IRAP (2008 version programmed in Visual Basic 6) recorded all response data, including accuracy and latency.

INSERT FIGURE 1 HERE

Fig. 1 Diagrammatic representation of the four IRAP trial-types. The four trial-types were denoted as: *Similar-Positive, Similar-Negative, Opposite-Positive, and Opposite-Negative.* Arrows did not appear on-screen.

Procedure

Self-Report measures. All participants first completed the three self-report measures (RSES, DASS, and Feeling Thermometer). The sequence in which these were presented was randomized across participants.

The IRAP. The IRAP comprised a maximum of four pairs of practice blocks, followed by a set number of three pairs of test blocks. On each trial of the IRAP, a label at the top of the screen (*Similar* or *Opposite*); a target at the center of the screen (e.g., *good* or *bad*), and two response options (*Participant's Name* and *Not Participant's Name)* at the bottom left and right of the screen were presented. Participants responded using either the "d" key for the response option on the left-hand side or the "k" key for the right-hand response option. The response option locations alternated from trial to trial in a quasi-random order, ensuring that they did not remain in the same left-right locations for more than three successive trials.

If a participant chose the response option that was defined as correct within that block of trials, an inter-trial interval of 400 ms was presented, after which the next trial was presented. If a participant chose the response option that was defined as incorrect for that

block of trials the stimuli remained on-screen and a red "X" appeared beneath the target stimulus. Only when the correct response option was selected did the program proceed directly to the 400 ms inter-trial interval, after which the next trial appeared. The pattern of trial presentations, with corrective feedback, continued until the block of 24 trials was completed. Trials were presented in a quasi-random order within each block with the constraint that each label stimulus was presented twice with each target stimulus, across a total of 24 trials. Blocks of history-consistent trials required responding that was in accordance with what would be deemed generally as positive self-regard: *Similar-Positive/Participant's Name; Similar-Negative/Not Participant's Name; Opposite-Positive/Not Participant's Name; Opposite-Negative/Participant's Name.* Inconsistent blocks required the opposite: *Similar-Positive/Not Participant's Name; Similar-Negative/Participant's Name; Opposite-Positive/Participant's Name; Opposite-Negative/Not Participant's Name.* Half of the participants were first presented with a consistent block of trials, while the other half were first presented with an inconsistent block of trials.

After completing a block of trials, the IRAP program provided participants with performance feedback for that block; the feedback was comprised of a message describing how accurately and quickly the participant had responded. The latter metric was determined from stimulus onset to the first correct response, calculated across all 24 trials within the block. Each participant was required to achieve a minimum accuracy of 80 percent correct and a maximum median latency of no more than 2000 ms on each block within a pair. If participants achieved both accuracy and latency criteria on a pair of practice blocks, they immediately proceeded to the first pair of test blocks. If participants failed a pair of practice blocks, practice blocks continued to a maximum of 4 block pairs. Failing to meet the criteria after 4 pairs of practice blocks terminated participation and these data were discarded.

A fixed set of 6 test blocks was presented, with no accuracy or latency criteria required to progress across blocks. Participants were encouraged to maintain the accuracy and latency criteria that they had reached during the practice blocks, by presenting them with obtained percentage correct and median latency at the end of each block.

Results

IRAP Data

In order to pass the practice blocks and move on to test blocks, participants were required to maintain an accuracy level of ≥ 80% correct and a median latency of ≤ 2,000. Exclusion criteria also applied to the test blocks, such that participants were required to maintain an accuracy level of ≥ 79% correct and a median latency ≤ 2,000 ms. on all three pairs of the test blocks.

Consistent with many published IRAP studies, D_{IRAP}-scores were calculated for each of the four trial-types (see Barnes-Holmes, Barnes-Holmes, et al., 2010). Positive D_{IRAP}-scores indicated responding in a manner consistent with positive self-regard (i.e., choosing *Participant's Name* more quickly than *Not Participant's Name* on *Similar-Positive* and *Opposite-Negative* trial-types, and choosing *Not Participant's Name* more quickly than *Participant's Name* on *Similar-Negative* and *Opposite-Positive* trial-types during consistent blocks of trials). Negative D_{IRAP}-scores indicated negative self-regard. Overall, participants demonstrated strong positive D_{IRAP}-scores on the *Similar-Positive*, *Opposite-Positive*, and *Opposite-Negative* trial-types (see Figure 2). That is, participants selected the *Participant's Name* response option more readily than the *Not Participant's Name* response option when presented with *Similar* and positive adjectives and *Opposite* with negative adjectives, and the *Not Participant's Name* response option more readily than the *Participant's Name* response

option when presented with *Opposite* and negative adjectives. The effect for the *Similar-Negative* trial-type approached zero.

INSERT FIGURE 2 HERE

Fig. 2 Mean D_{IRAP} scores for the four trial-types in Experiment 1.

In order to first investigate whether the order in which participants received IRAP blocks (i.e., consistent versus inconsistent first) impacted the D_{IRAP}-scores, a 2 x 4 mixed repeated measures analysis of variance (ANOVA) was conducted, but yielded no significant main nor interaction effect (p's < .67), and thus block order was excluded from subsequent analyses. A one-way repeated measures ANOVA was then conducted and revealed a significant main effect for trial-type, $F(3) = 14.362$, $p < .0001$, $\eta_{p^2} = 0.33$. Six Scheffe post hoc tests indicated that the effects for the *Similar-Positive* trial-type differed significantly from the *Similar-Negative* trial-type ($p < .0001$) and from the *Opposite-Negative* trial-type ($p < .002$). The *Similar-Negative* trial-type also differed significantly from the *Opposite-Positive* trial-type ($p < .001$), while the remaining trial-types did not differ significantly from each other (ps > .1). Four one-sample t-tests confirmed that the *Similar-Positive* trial-type ($t = .464$, $df = 29$, $p < .0001$) and the *Opposite-Positive* trial-type ($t = .284$, $df = 29$, $p < .0003$) were significantly different from zero, however the effects for *Opposite-Negative* ($p = .09$) and *Similar-Negative* ($p = .1$) did not reach significance. Overall, participants showed relatively strong effects when coordinating themselves with positive targets, but not when coordinating themselves with negative targets.

Correlational Analyses

A Pearson correlation coefficient matrix was calculated to assess the relationships among the four IRAP trial-types and the overall *D*-IRAP score (the mean of the four trial-types) with the self-report measures. Out of the 32 correlations, only two were significant (all other *p*'s > .2). Specifically, there was a significant positive correlation between self-warmth on the Feeling Thermometer and the *Opposite-Negative* trial-type ($r = .407$, $p = .025$) and the total *D*-IRAP score ($r = .405$, $p = .026$). That is, the greater the self-warmth, the greater participants' positive self-regard on the IRAP. Given the large number of analyses and small number of significant results, these correlations should be interpreted with extreme caution.

Summary and Discussion

Overall, the findings from Experiment 1 showed broadly similar findings as Vahey et al. (2009) in which IRAP performances correlated significantly with the Feeling Thermometer measure. There were large significant positive effects observed on the *Similar-Positive* and *Opposite-Positive* trial-types, which required choosing different response options within blocks of trials (i.e., own name during *Similar-Positive* and not own name on *Opposite-Positive*). The current experiment failed to find any correlation between IRAP performance and the explicit measure of self-esteem (RSES) and distress (DASS), but this result should be interpreted with caution due to the limited sample size and the likely lack of variance due to all participants being selected from a normative sample. During the debriefing of Experiment 1, some participants commented that they found the IRAP tasks confusing, frequently referring to the fact that they had to choose between selecting their own name and the negation of their own name. In all subsequent IRAPs, therefore, we employed the more standard format of the measure in which the response options consisted of simple confirmatory versus dis-confirmatory terms (i.e., *True* and *False*).

Experiment 2 explored three different IRAPs that sought to target relational responding focused on positive versus negative self-regard, positive versus negative emotional states, and success versus failure. The first IRAP employed only words, whereas the latter two IRAPs employed words and pictures. The stimuli used for the positive versus negative self-regard IRAP (hereafter referred to as the Self-regard IRAP) were based on a study by Scanlon (2014), which employed positive and negative self-descriptions as target stimuli (the Scanlon study aimed to assess self-relevant relational responding in children). The other two IRAPs presented pictures as labels (referred to as the Emotions IRAP and the Achievement IRAP) instead of text. In the Emotions IRAP, the pictures employed were of individuals expressing happiness and sadness, whilst in the Achievement IRAP the pictures depicted individuals in successful or unsuccessful situations. The choice of the stimuli for these two IRAPs were rather arbitrary and exploratory. However independent raters did assess that the pictures we employed sufficiently represented happiness, sadness, success and failure (see below). The target stimuli (the phrases *Like Me* and *Not Like Me*) and response options (*True* and *False*) remained the same across each IRAP. As noted above, Experiment 2 was exploratory in nature and thus we refrained from making any specific predictions.

Experiment 2

Method

Participants and Setting

Thirty individuals participated in Experiment 2, 16 females and 14 males. Participants ranged in age from 21 to 40 years, and were again recruited through snowball sampling, with no compensation offered. All aspects of the setting were identical to Experiment 1.

Apparatus and Materials

The same three self-report measures (RSES, DASS, and Feeling Thermometer) employed in Experiment 1 were used in Experiment 2. Experiment 2 involved three separate IRAPs (the Self-regard IRAP, the Emotions IRAP, and the Achievement IRAP) using an up-dated version of the 2008 IRAP software that now allowed for the use of multiple label stimuli rather than one label (e.g., Dunne, McEnteggart, Harte, Barnes-Holmes, & Barnes-Holmes, 2018).

The *Self-regard IRAP* presented three positive attributes (*Accepted, Popular,* or *Perfect*) and three negative attributes (*Faulty, Broken,* or *Useless*) as label stimuli, along with two phrases (*Like Me* or *Not Like Me*) as target stimuli, and two response options (*True* and *False*). Based on the various label-target combinations, this IRAP yielded the following four trial-types: *Positive-Like Me*; *Positive-Not Like Me*; *Negative-Like Me*; and *Negative-Not Like Me*.

The *Emotions IRAP* presented three pictures of happy faces and three pictures of sad faces as label stimuli, while the target stimuli and response options remained identical to the Self-regard IRAP. Pictures of male faces were used for male participants and pictures of female faces were used for female participants. Based on the various label-target combinations, this IRAP yielded the following four trial-types: *Happy-Like Me*; *Happy-Not Like Me*; *Sad-Like Me*; and *Sad-Not Like Me*.

The *Achievement IRAP* presented three pictures of an individual succeeding in some way (i.e., receiving a degree, getting a job, winning a race, receiving an award, making money, and getting a car) and three pictures of an individual failing in some way (i.e., being fired from a job, being homeless, being in prison, losing a game, losing a sporting match) as label stimuli, with the target stimuli and response options remaining the same as the other two IRAPs. Once again, pictures of males were used for male participants and pictures of females

were used for female participants. Based on the various label-target combinations, this IRAP yielded the following four trial-types: *Success-Like Me*; *Success-Not Like Me*; *Failure-Like Me*; and *Failure-Not Like Me*.

All pictures were first rated by five independent raters (stimuli available upon request from the corresponding author). For the happy and sad faces, raters were asked to indicate on an 8-point Likert scale from -6 (very sad) to +6 (very happy) how happy or sad they found each face to be. For the pictures of people succeeding and failing, raters were asked to again indicate on a similar 8-point Likert scale from -6 (significant failure) to +6 (very successful) the extent to which they thought each picture represented success or failure. All of the pictures used as stimuli were rated </=-4 for those that portrayed sadness or failure and >/=+4 for those that portrayed happiness or success by each rater.

Procedure

Self-Report Measures. This stage of the experiment was identical to Experiment 1.

IRAP. The IRAP instructions, performance criteria, and feedback were identical to Experiment 1. The order in which participants completed the three separate IRAPs was counterbalanced. In light of the findings from Experiment 1, all IRAPs in Experiment 2 commenced with a consistent block. Each IRAP block again presented 24 trials as four trial-types in a quasi-random order with the following constraints per IRAP. For the Self-regard IRAP, each of the six label stimuli appeared twice with each of the two target stimuli. For the Emotions and Achievement IRAPs, each of the 12 label stimuli appeared once with each of the two target stimuli. This ensured that each of the four trial-types was presented six times each within each block.

Results

IRAP Data

The data preparation was identical to Experiment 1 with mean D_{IRAP} scores calculated for the four trial-types for each of the three IRAPs.

Self-regard IRAP. Positive D_{IRAP}-scores indicated responding during consistent blocks, *True* more quickly than *False* on *Positive-Like Me* and *Negative-Not Like Me* trial-types, and responding *False* more quickly than *True* on *Positive-Not Like Me* and *Negative-Like Me* trial-types. Negative D_{IRAP}-scores indicated the opposite pattern (see Figure 3). Overall, participants demonstrated strong positive D_{IRAP}-scores on the *Positive-Like Me* trial-type. That is, participants selected *True* more readily than *False* when presented with positive attributes and *Like Me.* There were weaker positive effects observed on the *Positive-Not Like Me* and *Negative-Not Like Me* trial-types, and a negative weak effect on the *Negative-Like Me* trial-type.

INSERT FIGURE 3 HERE

Fig. 3 Mean D_{IRAP} scores for the four trial-types on the Self-regard IRAP (top), Emotion IRAP (middle), and the Achievement IRAP (bottom) in Experiment 2.

A one-way repeated measures ANOVA was conducted and revealed a significant main effect for trial-type $F(3) = 9.726$, $p < .0001$, $\eta_p^2 = .25$. Six Scheffe post-hoc tests indicated that the effect for the *Positive-Like Me* trial-type differed significantly from the three other trial-types (all p's < .04), while the remaining trial-types did not differ significantly from each other (all ps > .1). Four one-sample t-tests confirmed that the *Positive-Like Me* ($t = .438$, $p < .0001$) trial-type differed significantly from zero, but the three remaining trial-types did not.

Overall, participants showed relatively self-positive response biases when relating positive attributes to themselves (similar to Experiment 1).

Emotions IRAP. Positive D_{IRAP}-scores indicated responding during consistent blocks, *True* more quickly than *False* on *Happy-Like Me* and *Sad-Not Like Me* trial-types, and responding *False* more quickly than *True* on *Happy-Not Like Me* and *Sad-Like Me* trial-types. Negative D_{IRAP}-scores indicated the opposite pattern (see Figure 3). A strong effect was observed on the *Happy-Like Me* trial-type, but a considerably weaker effect was found on the *Happy-Not Like Me*. The effects for the remaining two trial-types approached zero, but the *Sad-Like Me* trial-type was in a negative direction.

A one-way repeated measures ANOVA revealed a significant main effect for trial-type, $F(3) = 13.076$, $p < .0001$, $\eta_{p^2} = .31$. Six Scheffe post-hoc tests indicated that the effect for the *Happy-Like Me* trial-type differed significantly from the three other trial-types (all *p*'s < .02), while the remaining trial-types did not differ significantly from each other (all *p*s > .1). Four one-sample *t*-tests confirmed that the *Happy-Like Me* ($t = .475$, $p < .0001$) and the *Happy-Not Like Me* ($t = .171, 29$, $p < .01$) trial-types were significantly different from zero while the two remaining trial-types were not (*ps* > .4). Overall, participants showed relatively strong positive effects when relating themselves to happy faces.

Achievement IRAP. Positive D_{IRAP}-scores indicated responding during consistent blocks, *True* more quickly than *False* on *Success-Like Me* and *Failure-Not Like Me* trial-types, and responding *False* more quickly than *True* on *Success-Not Like Me* and *Failure-Like Me* trial-types. Negative D_{IRAP}-scores indicated the opposite pattern (see Figure 3). Participants demonstrated a strong positive effect on the *Success-Like Me* trial-type, but a considerably weaker effect on the *Success-Not Like Me* trial-type. The effects for the remaining two trial-types approached zero, with both in a slightly negative direction.

A one-way repeated measures ANOVA revealed a significant main effect for trial-type, F (3) = 6.607, p = .0004, η_{p^2} = .18. Six Scheffe post-hoc tests indicated that the effect for the *Success-Like Me* trial-type differed significantly from the three other trial-types (all p's < .04), while the remaining trial-types did not differ significantly from each other (all ps > .74). Four one-sample t-tests confirmed that the *Success-Like Me* (t = .348, p < .0001) trial-type differed significantly from zero, while the three remaining trial-types did not (all p's > .38). Overall, participants showed a relatively strong positive effect when relating themselves to success.

Correlational Analyses

A Pearson correlation coefficient matrix was calculated to assess the relationships among the IRAP trial-types, the overall *D*-IRAP score, and the self-report measures for each IRAP. There were no significant correlations for the Self-regard IRAP or the Emotions IRAP and the self-report measures (all p's > .06). On the Achievement IRAP, a number of significant correlations emerged, but only for one trial-type. Specifically, the *Failure-Not Like Me* trial-type correlated negatively with DASS Depression (r = -.459, p = .01), DASS Stress (r = -.436, p = .01), and the overall DASS score (r = -.429, p = .02). That is, the more participants showed a response bias toward denying failure, the lower depression, stress and overall psychological distress. Interestingly, this trial-type was positively correlated with the RSES (r = .374, p = .04) and the Feeling Thermometer (r = .384, p = .03). That is, the more participants showed a response bias toward denying failure, the greater their self-esteem and self-warmth.

Summary and Discussion

Across all three IRAPs, significant self-positive effects emerged on the first trial-type (i.e., the trial on which participants were asked to confirm versus deny that positive words or

pictures were like them). It was only on the Emotions IRAP that a second significant effect was observed for another trial-type (i.e., *Happy-Not Like Me*). Only effects on the Achievement IRAP correlated significantly with any of the self-report measures.

General Discussion

Consistent with Vahey et al. (2009), Experiment 1 found that the IRAP effects generally indicated positive self-regard. One of the trial-types indicated negative self-regard but was the weakest of the four and not significantly different from zero. The correlation between IRAP performances and self-warmth on the Feeling Thermometer was also consistent with Vahey et al. However, none of the trial-types (nor the overall *D*-score) correlated significantly with any of the standardized measures of self-esteem or psychological distress. During the course of Experiment 1, ad-hoc self-reports provided by some participants during debriefing indicated that asking participants to select their own name versus the negation of their own name as response options was especially challenging. Indeed, this appeared to be particularly the case for trial-types that involved responding to a double negative (e.g., opposite-worthless-own name/not own name). Thus, subsequent IRAPs in the current study adopted the more traditional format in which the response options were *True* and *False*.

In Experiment 2, we aimed to identify an IRAP that would generate performances that might correlate with standardized measures of self-esteem and/or distress, by employing three different versions that targeted: (1) positive versus negative self-regard; (2) positive versus negative emotional state; and (3) success versus failure. The Self-regard IRAP employed words, whereas the Emotions IRAP and the Achievement IRAP employed both words and pictures. In general, all three IRAPs produced relatively large self-positive effects for the trial-type that involved self-positive relations, with weaker or absent effects for the remaining three

trial-types. Indeed, statistical analyses indicated that the former trial-type differed significantly from the others for all three IRAPs. Performances on the Self-regard and Emotions IRAPs failed to correlate significantly with any of the self-report measures. In contrast, performances on a specific trial-type (*Failure-Not Like Me*) on the Achievement-IRAP correlated with all but one (anxiety) of the self-report measures. That is, a response bias toward denying self-failure predicted lower levels of psychological distress and higher levels of self-esteem and self-warmth.

Overall, the current findings highlight that arguing that the IRAP *per se* is adequate or inadequate as a research tool could be seen as relatively naïve. The IRAP is in one sense an empty frame and its utility will be determined in large part by the stimuli that are inserted into it. In the current study, for example, four different IRAPs were developed and only one of these correlated relatively strongly with measures of self-esteem and psychological distress (i.e., the Achievement IRAP). If nothing else, therefore, the current findings highlight that when choosing to use the IRAP as a tool for research, it will always be critically important to consider whether or not the stimuli that are employed within the procedure overlap sufficiently with the stimulus properties of the domain that is being targeted. In the current case, for instance, it appears that pictures depicting success versus failure were more closely related, functionally, to self-report measures of self-esteem and psychological distress than pictures depicting happy and sad emotions.

At this point, it is important to emphasize that, in Experiment 1, and with all three IRAPs from Experiment 2, a specific trial-type effect emerged that has been reported in previous IRAP research. The effect is known as the *single trial-type dominance effect* (STTDE) because the size of the IRAP effect for one trial-type appears to dominate over the other three trial-types. The DAARRE model was developed primarily in an attempt to explain this effect. As mentioned in the Introduction, the DAARRE model is based on the assumption

that such differential trial-type effects may be explained by the interaction between the functional properties of the individual stimuli (e.g., the extent to which they evoke appetitive or aversive reactions) and the relation between the label and target stimuli (i.e., the extent to which they are coordinate or opposite/different). Critically, the functional properties of the response options are also factored into the model. For illustrative purposes, a simplified graphical representation of the model using the stimuli employed in the Achievement IRAP is presented in Figure 4.

INSERT FIGURE 4 HERE

Fig. 4 The DAARRE model as it applies to the Achievement IRAP. The positive and negative labels refer to the relative positivity of the Cfuncs, for each label and target, the relative positivity of the Crels, and the relative positivity of the RCIs in the context of the other Cfuncs, Crels, and RCIs in that stimulus set.

The images of success likely possess positive Cfunc properties (indicated by + signs next to the label stimuli) relative to the images of failure (indicated by – signs next to the label stimuli)[1]. Similarly, the target stimulus *Like Me* may possess positive Cfunc properties relative to *Not Like Me* (again indicated by + and – signs, respectively, next to the relevant stimuli). The coherence of the relationship between the label and target stimuli (Crel), in terms of the participant's individual history, is indicated with a plus or minus sign. In a normative sample, one would expect pictures of success to cohere more strongly with *Like Me* than *Not Like Me*, but pictures of failure to cohere more strongly with *Not Like Me* than *Like*

[1] In recent publications pertaining to the DAARRE model, a distinction has been made between orienting and evoking functions. However, making such a distinction in the current context seems unnecessary and perhaps unwise, given that the current data were collected many years before the DAARRE model was formulated. As such, the current interpretation remains entirely post-hoc and thus it would seem unwise to "overwork" the theoretical analysis.

Me. Thus, for example, a *Success-Like Me* relation is indicated with a plus sign (i.e., coherence), whereas a *Success-Not Like Me* relation is indicated with a minus sign (i.e., incoherence). Finally, each response option is indicated with a plus or minus sign to denote its likely Cfunc property. Specifically, *True* (+) would typically be used in natural language to indicate coherence and *False* (-) to indicate incoherence. In the current example, therefore, the *Success-Like Me* trial-type (far left of the figure) is the only trial-type in which all of the functions (both Cfunc and Crel) are positive, with the three remaining trial-types all involving some mix of positive and negative functions. Critically, the contrast in coherence versus incoherence between consistent and inconsistent blocks of trials would be largest for the *Success-Like Me* trial-type. More informally, a participant with relatively positive self-esteem would find it very easy to select *True* in the context of the three other plus signs, but quite difficult to select *False*. For the other three trial-types, however, this contrast would be reduced. The same general logic of the DAARRE model could be used to explain the STTDE effect observed for the other three IRAPs (e.g., a picture of a happy face would be indicated with a + sign and a picture of a sad face would be indicated with a – sign). An important caveat, of course, is that all of the functions labeled in Figure 4 are behaviorally determined by past and current contextual histories, and thus are not absolute or inherent in the stimuli themselves.

Another interesting effect that emerged in Experiment 2 for the Achievement IRAP is that only the *Failure-Not Like Me* trial-type correlated with any of the self-report measures. Specifically, responding *False* more quickly than *True* predicted lower levels of self-esteem and higher levels of overall psychological distress. Given that the sample of participants was normative, it may be that the Cfunc properties of the two *Success* trial-types failed to evoke relatively strong differential reactions for participants with high versus low self-esteem/distress because the pictures produced generally equal levels of positivity with regard

to self. In contrast, the Cfunc properties of the two *Failure* trial-types may have evoked relatively strong differential reactions because the label pictures of failure were particularly *threatening* for the low self-esteem participants. Indeed, data from a number of surveys and experimental studies have suggested that low self-esteem individuals are in fact hypervigilant to signs of inadequacy and rejection (Rosenberg & Owens, 2001). Thus, participants with lower self-esteem may have reacted to the IRAP failure pictures with an "oh no that could be me" response that was more or less absent for the higher self-esteem participants.

If this interpretation is correct it would explain why the two positively-labeled trial-types failed to predict the self-report measures. However, it would not explain why only one of the negatively-labeled trial-types correlated with the questionnaires; perhaps the DAARRE model may be useful here. Specifically, the model emphasizes that the Cfunc property of the target differs between the two negatively-labeled trial-types (+ for *Like Me* and – for *Not Like Me*), and thus the Cfunc properties of the target and response option *False* cohere for the *Failure-Not Like Me* trial-type (both – signs), but not for the *Failure-Like Me* trial-type (a + sign and a – sign). As argued above, the pictures of failure may have evoked slightly stronger levels of negativity with regard to self for participants with slightly lower self-esteem/high distress. This is illustrated in Figure 5, where two minus signs indicate the 'threat' reaction of participants with low self-esteem/high distress, whereas only one minus sign is used for participants with high self-esteem/low distress. If the assumption illustrated in Figure 5 is correct, then the more coherent response for participants with low self-esteem/high distress on the *Failure-Not Like Me* trial-type would be *False* (i.e., four minus signs); but this would not be the case for participants with high self-esteem/low distress (i.e., three minus signs). Critically, the differential impact of coherence on the two types may be undermined in the *Failure-Like Me* trial-type because the Cfunc property of the target is positive rather than negative. We are assuming here that the coherence between the properties of target and

response option may dominate as a participant emits a response because they are spatially and temporally contiguous (see Kavanagh, et al., 2019, for a similar argument). Of course, we recognize that the foregoing interpretation is entirely post-hoc and highly speculative, but we offer it here to highlight that the types of effects regularly obtained with the IRAP require a systematic functional analysis of the numerous contextual variables at play during exposure to the task.

INSERT FIGURE 5 HERE

Fig. 5 The DAARRE model as it applies to individuals with high and low self-esteem on the Failure-Not Like Me trial-type on the Achievement IRAP in Experiment 2. Note that the two minus signs for the low self-esteem participants indicate that they may have reacted to the failure pictures as personally more threatening than the high self-esteem participants.

In a related vein, another interesting pattern that emerged for all three IRAPs in Experiment 2 was the fact that the third trial-type (e.g., *Failure-Like Me*) in each case was in a slightly negative direction, whereas the second trial-type (e.g., *Success-Not Like Me*) was in a positive direction. In each case, therefore, it appears the participants tended to confirm that a generic negative label combined with "Like Me" was "True" more quickly than "False." Or in other words, participants produced a response bias confirming that negative labels were like them (trial-type 3); but they also produced a response bias *dis*confirming that positive labels were *not* like them (trial-type 2). This type of pattern was also recently noted by Kavanagh et al. (2019), and is referred to as the *dissonant target trial-type-effect* (DTTTE). To determine if the difference observed across all three IRAPs from Experiment 2 was significant, the data were entered into a 3x4 repeated measures ANOVA (with IRAP type as one factor and trial-type as the second factor). A Scheffe post-hoc test between the second and third trial-types

proved to be marginally significant (p = .05), thus supporting the DTTTE observed across all three IRAPs.

A DAARRE model interpretation of the DTTTE is that participants find it easier to choose the response option that coheres with the target stimulus rather than the response option that does not. Specifically, the coherence in the Cfunc properties between the target stimulus ('Like Me") and the response option ("True") for trial-type 3 tends to produce a response bias towards responding "True" over "False"; while for trial-type 2 the target ("Not Like Me") coheres more with "False" than "True". The DTTTE may be readily understood simply by comparing the plus and minus signs assigned to the target and response-option stimuli in Figure 4, and assuming that participants found it easier to choose the response option with the same sign as the target stimulus. Interestingly, we did not observe a DTTTE in Experiment 1, but in this case the response options were not "True" and "False" but involved the participant's name. It is possible, therefore, that these response options did not possess the same highly differentiated positive and negative Cfunc properties associated with "True" and "False". Once again, we recognize that this post-hoc interpretation is highly speculative, but it has been observed and explained in a similar way in a previously published study (Kavanagh, et al., 2019), and thus it may be useful for other researchers to consider a potential DTTTE in their own IRAP analyses.

At a more general level, the current study may be seen as an attempt to progress research that employs the IRAP towards a more theoretically informed and precise use of the methodology. Specifically, the study involved exploring the extent to which particular types of stimuli (e.g., related to Emotions versus Achievement) correlated with self-reported measures of self-esteem and distress, and also which particular trial-type(s) were involved in the correlations that emerged. Given the current findings, future researchers may be in a better

position to make functional-analytically driven predictions for specific stimulus sets and trial-types than hitherto. We expand upon this claim below.

Given the focus of the research presented in the current article, it may be tempting to question the extent to which the correlations for the Achievement IRAP "genuinely" captured the construct of self-esteem (or psychological distress), but that would miss the point. When adopting the "behavior-as-proxy" approach to psychological research, as is so common in the mainstream, it is always possible to argue that a behavioral measure may be tapping into some other spurious construct. In effect, because no one has direct access to psychological constructs in general, treating behavior as a proxy for a construct always leaves the door open to the criticism that the proxy is not a "pure" measure of that construct. The alternative strategy, the one adopted in presenting the current findings, is that we simply treat the IRAP effects as response biases and do not make any strong claims concerning the extent to which they represent specific constructs (i.e., we do not conclude that the Achievement IRAP is a pure measure of self-esteem). Going forward, however, other researchers may use the specific effects obtained, and the functional-analytic interpretation we have offered, in terms of the DAARRE model, and factor the information into their own empirical and theoretical research. Critically, in presenting the findings in a functional-analytic theoretical framework, as we have done here, they may also be used to expand the relevance of the research to modified versions of the IRAP (e.g., ones with three or more response options, rather than just two) and indeed other methodologies that seek to measure the interactions among Crel and Cfunc properties when participants are required to respond at relatively high speeds.[2]

[2] In arguing that we have presented the current data in a functional-analytic framework, we should emphasize that it remains relatively limited in that regard. For example, the research did not involve single case experimental designs, demonstrating the prediction-and-influence of behavior (with precision, scope, and depth) with individual participants. Furthermore, the DAARRE model itself has recently been integrated into a multi-dimensional multi-level (MDML) framework for analyzing the dynamics of derived relational responding itself (see Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017), thus yielding a hyper-dimensional multi-level (HDML) framework; and this latter framework has led to the proposal of a new conceptual unit of analysis

In closing, we again recognize that the DAARRE model interpretations presented above are entirely post-hoc and speculative, but it seems important to present them here because some of the patterns observed in the current study have been observed previously and interpreted using the DAARRE model in recent articles. Perhaps other researchers, therefore, who are using the IRAP may find the interpretations offered here of some use in attempting to explain and explore similar effects. In any case, it seems important to continue to develop increasingly sophisticated functional analyses of the IRAP in terms of the cluster of variables that produce the patterns we observe with the measure. Indeed, this seems particularly important given that the IRAP has been used widely as a measure in clinical psychology and other domains (e.g., see Vahey et al. 2015).

for an up-dated version of RFT, comprised of relating, orienting, and evoking, known as the ROE (e.g., Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2020). Thus, the presentation of the current data should be seen as *part of a transition* from a relatively mainstream approach to the use of the IRAP as a methodology to a functional-analytic-abstractive oriented approach.

Declarations and Compliance with Ethical Standards

Conflict of Interest: The authors declare no other conflicts of interest. The authors declare they have no other conflicts of interest.

Funding: This article was prepared with the support of an Odysseus Group 1 grant awarded to the fourth author by the Flanders Science Foundation (FWO).

Ethical Approval: All procedures involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards

Informed Consent: Informed consent was obtained from all participants

Availability of Data and Materials: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

References

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, *10*(2), 176. doi: 10.1037/1040-3590.10.2.176

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From 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*(4), 473-483. doi: 10.1016/j.jcbs.2017.08.001

Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2020). Updating RFT (more field than frame) and its implications for process-based therapy. *The Psychological Record.* doi: 10.1007/s40732-019-00372-3

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. doi: 10.1007/BF03395726

Barnes-Holmes, D., Finn, M., Barnes-Holmes, Y., & McEnteggart, C. (2018). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *Perspectives on Behavioral Science (Special issue on Derived Relations), 41*(1), 155-173. doi: 10.1007/s40614-017-0124-7

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The implicit relational assessment procedure (IRAP) as a response-time and event-related potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record, 58*(4), 497-516. doi: 10.1007/BF03395634

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The implicit relational assessment procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record, 60,* 57-66. doi: 10.1007/BF03395694

Blascovich, J. & Tomaka, J. (1991). Measures of self-esteem. In J.P. Robinson, P.R., Shaver, & L.S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1, Measures of personality and social psychological attitudes* (pp. 115-160). San Diego: Academic Press.

Cagney, S., Harte, C., Barnes-Holmes, Barnes-Holmes, & McEnteggart, C. (2017). Response biases on the IRAP for adults and adolescents with respect to smokers and nonsmokers: The impact of parental smoking status. *The Psychological Record, 67*(4), 473-483. doi: 10.1007/s4073-017-0249-9

Cartwright, A., Hussey, I., Roche, B., Dunne, J., & Murphy, C. (2017). An investigation into the relationship between gender binary and occupational discrimination using the Implicit Relational Assessment Procedure. *The Psychological Record, 67*(1), 121-130. doi: 10.1007/s4073-016-0212-1

Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The implicit relational assessment procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record, 59*(4)*,* 591-620. doi: 10.1007/BF03395683

Dawson, D.L., Barnes-Holmes, D., Gresswell, D.M., Hart, A.J., & Gore, N.J. (2009). Assessing the implicit beliefs of sexual offenders using the implicit relational assessment procedure: A first study. *Sex Abuse, 21*(1), 57-75. doi: 10.1177/1079063208326928.

DeHart, T. & Pelham, B.W. (2007). Fluctuations in state implicit self-esteem in response to daily negative events. *Journal of Experimental and Social Psychology, 43*(1), 157-165. doi: 10.1016/j.jesp.2006.01.002

Drake, C.E., Kellum, K.K., Wilson, K.G., Luoma, J.B., Weinstein, J.H., & Adams, C.H. (2010). Examining the Implicit Relational Assessment Procedure: Four preliminary studies. *The Psychological Record, 60*(1), 81-86. doi: 10.1007/BF03395695

Dunne, C., McEnteggart, C., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2018). Faking a race IRAP effect in the context of single versus multiple label stimuli. *The International Journal of Psychology and Psychological Therapy, 18*(3), 289-300.

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect in the IRAP: Developing a differential arbitrarily applicable responding effects (DAARE) model. *The Psychological Record, 68*(1), 11-25. doi: 10.1007/s40732-017-0262-z

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

Hughes, S., Barnes-Holmes, D., & Smyth, S. (2017). Implicit cross-community biases revisited: Evidence for ingroup favoritism in the absence of outgroup derogation in Northern Ireland. *The Psychological Record, 67*(1)*,* 97-107. doi: 10.1007/s40732-016-0210-3

Kavanagh, D., Matthyssen, N., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Vastano, R. (2019). Exploring the use of pictures of self and other in the IRAP: Reflecting upon the emergence of differential trial-type effects. *International Journal of Psychology and Psychological Therapy, 19*(3), 323-336.

Lovibond, S.H. & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Sydney: The Psychology Foundation of Australia.

Power, P.M., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2017). Exploring racial bias in a country with a recent history of immigration of black Africans. *The Psychological Record, 67*(3), 365-375. doi: 10.1007/s40732-017-0223-6

Rosenberg, M. (1965). *Society and the adolescent self-image*. New Jersey: Princeton University Press.

Rosenberg, M. & Owens, T.J. (2001). Low self-esteem people: A collective portrait. In T.J. Owens, S. Stryker, and N. Goodman (Eds.), *Extending self-esteem theory and Research: Sociological and psychological currents* (p. 400-436). New York: Cambridge University Press.

Scanlon, G., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2014). Using the implicit relational assessment procedure (IRAP) to assess implicit gender bias and self-esteem in typically-developing children and children with ADHD and with dyslexia. *Behavioral Development Bulletin, 19*(2), 48-59. doi: 10.1037/h0100577

Scheel, M.H., Roscoe, B.H., Scaewe, V.G., & Yarbrough, C.S. (2014). Attitudes towards Muslims are more favorable on a survey than on an Implicit Relational Assessment Procedure (IRAP). *Current Research in Social Psychology, 22*(3), 22-32.

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*(3)*,* 371-388. doi: 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. doi: 10.1016/j.jbtep.2015.01.004

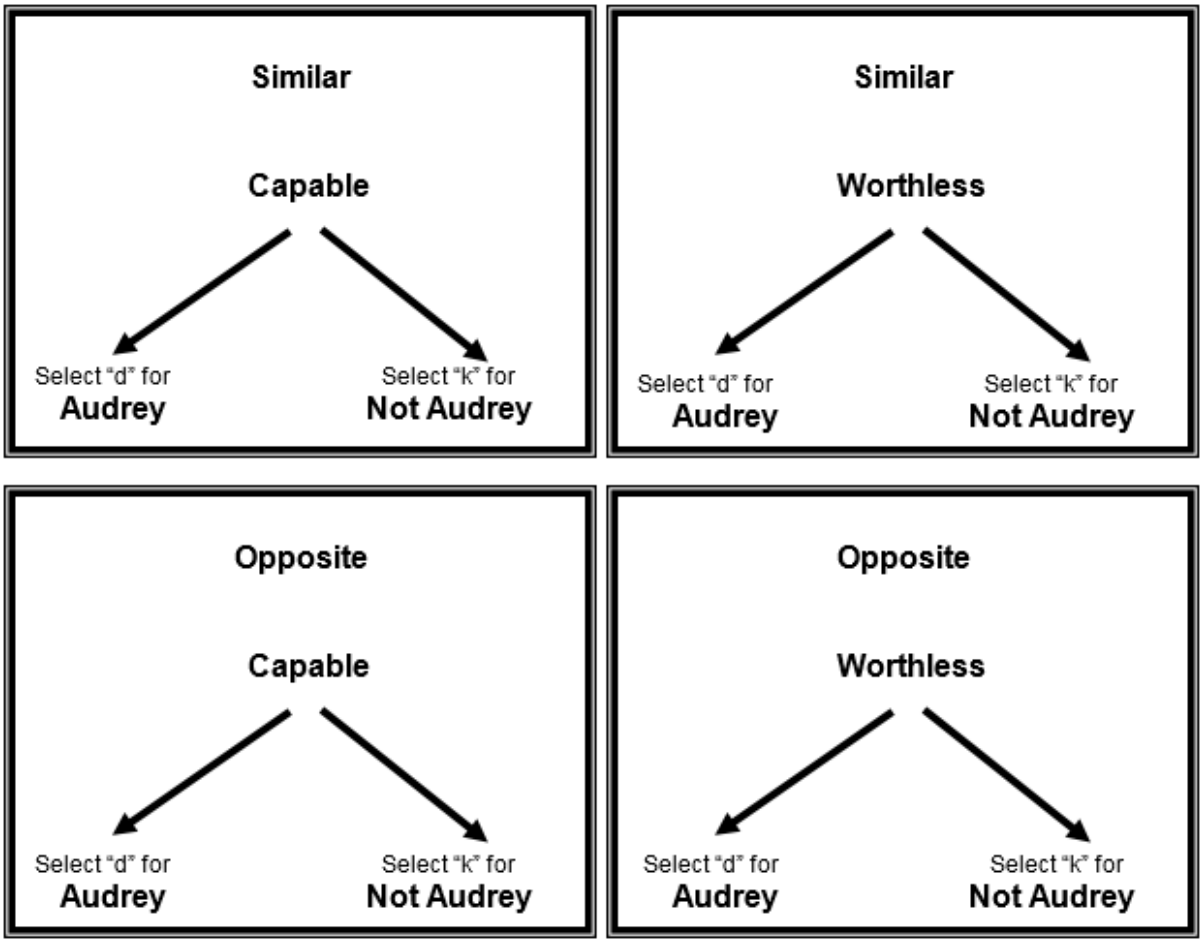
Similar
Capable
Select "d" for
Audrey
Select "k" for
Not Audrey
Similar
Worthless
Select "d" for
Audrey
Select "k" for
Not Audrey
Opposite
Capable
Select "d" for
Audrey
Select "k" for
Not Audrey
Opposite
Worthless
Select "d" for
Audrey
Select "k" for
Not Audrey

Fig. 1

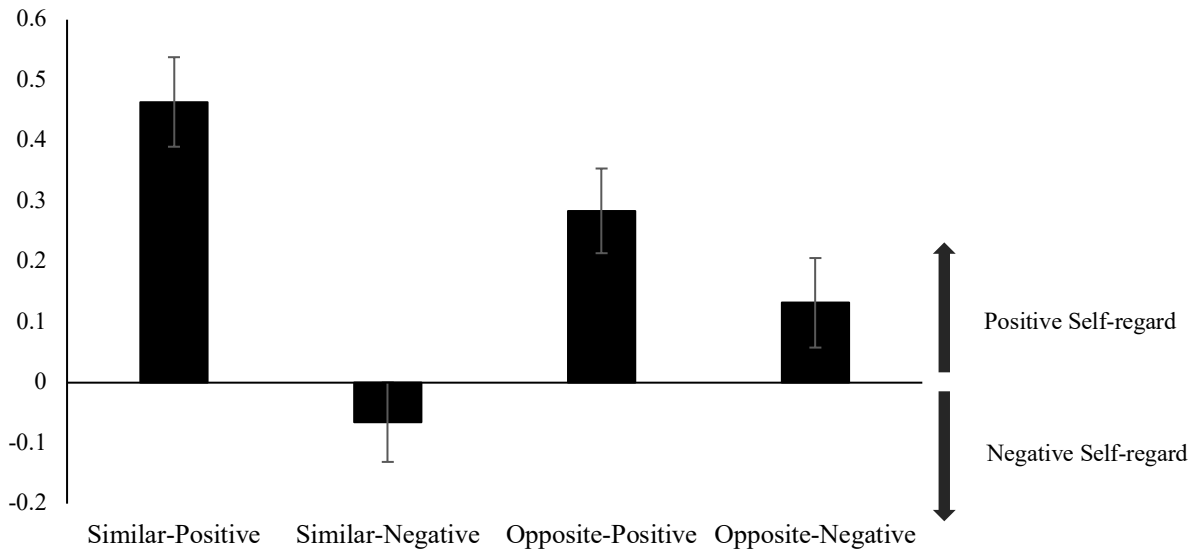
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Similar-Positive
Similar-Negative
Opposite-Positive
Opposite-Negative
Positive Self-regard
Negative Self-regard

Fig. 2

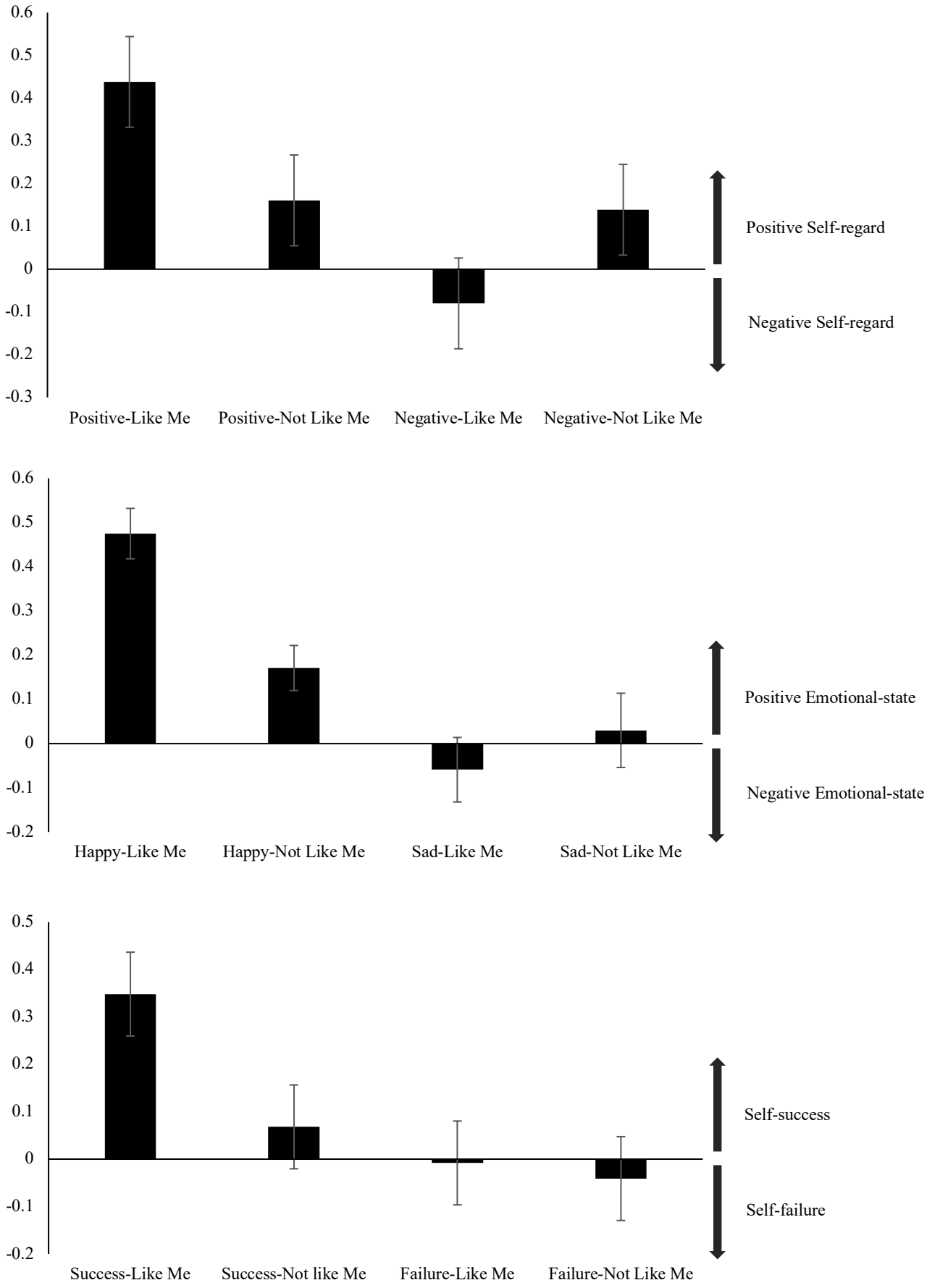
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Positive-Like Me
Positive-Not Like Me
Negative-Like Me
Negative-Not Like Me
Positive Self-regard
Negative Self-regard
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Happy-Like Me
Happy-Not Like Me
Sad-Like Me
Sad-Not Like Me
Positive Emotional-state
Negative Emotional-state
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
Success-Like Me
Success-Not like Me
Failure-Like Me
Failure-Not Like Me
Self-success
Self-failure

Fig. 3

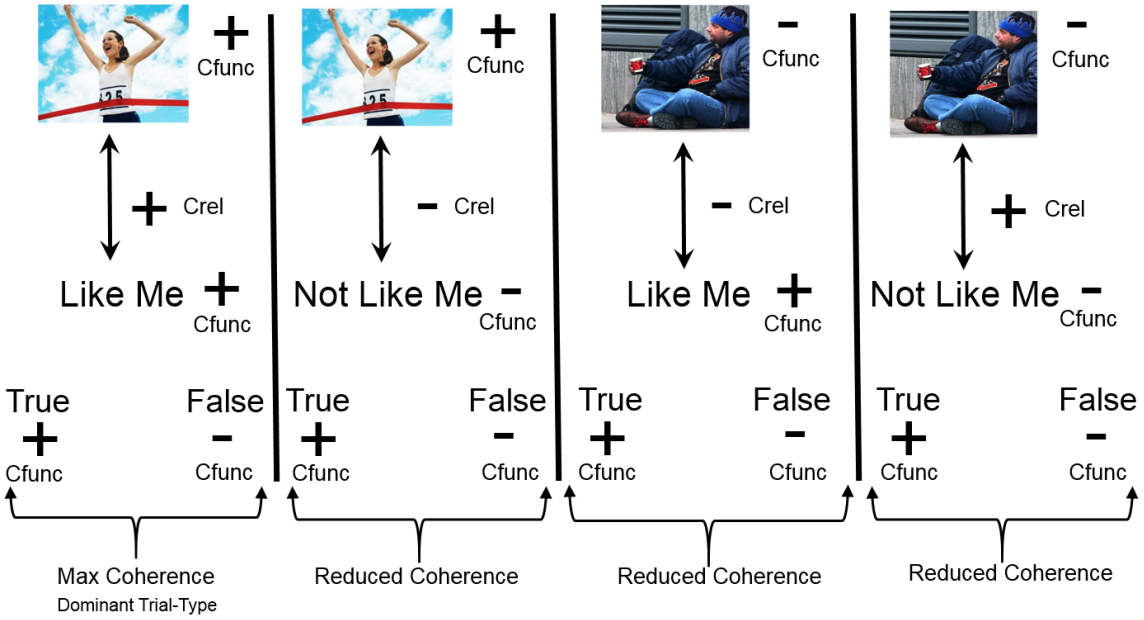
+ Cfunc
+ Crel
Like Me + Cfunc
True + Cfunc
False - Cfunc
+ Cfunc
- Crel
Not Like Me - Cfunc
True + Cfunc
False - Cfunc
- Cfunc
- Crel
Like Me + Cfunc
True + Cfunc
False - Cfunc
- Cfunc
+ Crel
Not Like Me - Cfunc
True + Cfunc
False - Cfunc
Max Coherence
Dominant Trial-Type
Reduced Coherence
Reduced Coherence
Reduced Coherence

Fig. 4

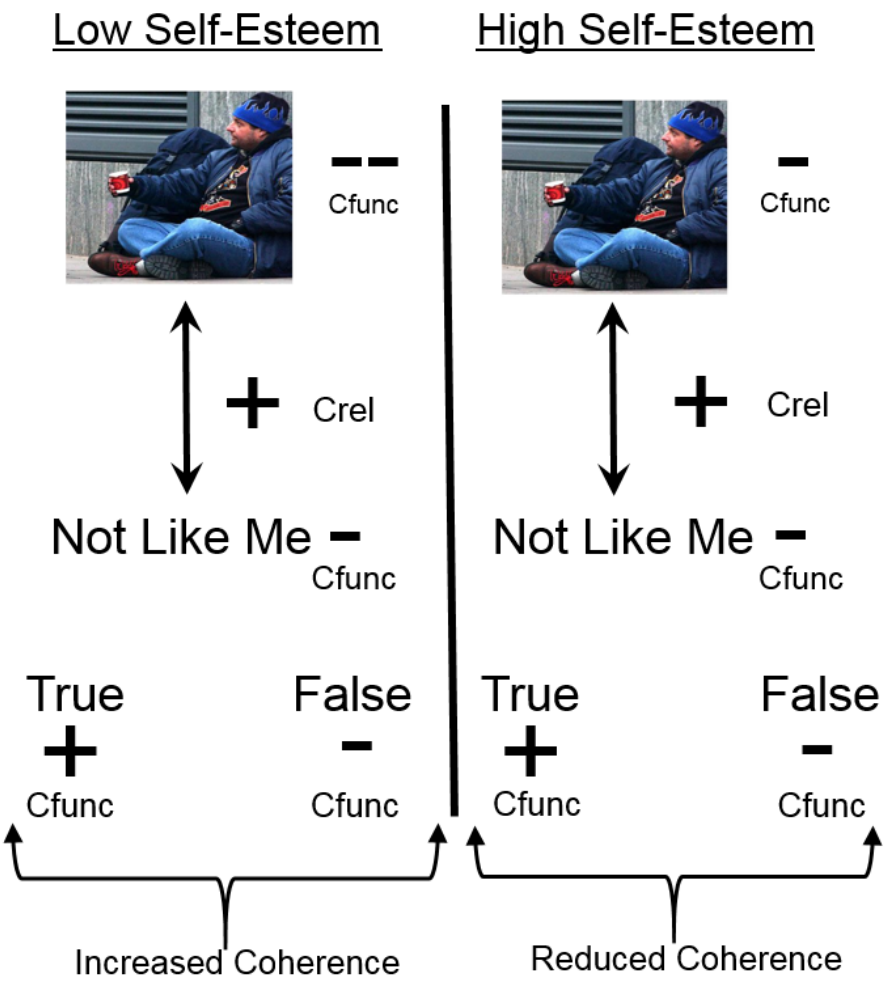
Low Self-Esteem
High Self-Esteem
-- Cfunc
- Cfunc
+ Crel
+ Crel
Not Like Me - Cfunc
Not Like Me - Cfunc
True + Cfunc
False - Cfunc
True + Cfunc
False - Cfunc
Increased Coherence
Reduced Coherence

Fig. 5