

Predicting and Influencing the Single Trial-Type Dominance Effect: The First Study

Martin Finn, Dermot Barnes-Holmes, Ciara McEnteggart & Deirdre Kavanagh

Experimental, Clinical and Health Psychology, Ghent University, Ghent, Belgium

Correspondence should be addressed to Martin Finn, Experimental, Clinical and Health

Psychology, Ghent University, Ghent, Belgium, Martin.Finn@UGent.be

Acknowledgements

The data for the current manuscript was collected, and prepared with the support of the FWO

Type I Odysseus Programme at Ghent University, Belgium.

Predicting and Influencing the Single Trial-Type Dominance Effect: The First Study**Abstract**

A recently published article reported a particular pattern of responding that has been observed on the implicit relational assessment procedure (IRAP), referred to as a single-trial-type-dominance-effect (STTDE; Finn, Barnes-Holmes, & McEnteggart, 2018). To account for the phenomenon, the differential arbitrarily applicable relational responding effects (DAARRE) model of IRAP performance was proposed. The DAARRE model predicts the STTDE in terms of an overlap in the functional properties of the label, target, and response-option stimuli presented within an IRAP. The current paper presents an initial attempt at engineering a STTDE within an experimental session. Forty participants were exposed to a series of training procedures and IRAPs. Critically, the training procedures consisted of a series of trials that aimed to establish a “True” function for a picture stimulus that was subsequently presented in the IRAP; participants were then exposed to an IRAP in which participants were required to respond “True” on a specific trial-type that presented that picture. Consistent with the DAARRE model, the STTDE emerged for the predicted trial-type, with analyses at both the group and individual-participant level supporting this conclusion. The implications of the findings for future research on analyzing the dynamics of arbitrarily applicable relational responding are discussed.

Key words: RFT, IRAP, Crel, Cfunc, DAARRE

Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001) is a modern functional contextual account of human language and cognition. The account is rooted in the Skinnerian tradition and draws heavily on Sidman's work on stimulus equivalence (e.g., Sidman, 1971, Sidman, et al., 1982; see Sidman, 1994, for a book-length review). According to RFT the core unit of human language and cognition involves framing events relationally, or in other words responding to one stimulus in terms of another stimulus. A relational frame is defined as a generalized pattern of relational responding possessing the properties of mutual entailment (if $A = B$, then $B = A$), combinatorial mutual entailment (if $A = B$, and $B = C$, then $C = A$ and $A = C$), and the transformation of functions. The distinction between relational entailment and the transformation of functions is critically important in RFT because it distinguishes between the act of relating stimuli in an "abstract sense" from the impact of that relating on the functions of those stimuli. The classic demonstration of the transformation of functions thus involves establishing a frame of equivalence among three stimuli ($A=B=C$) and then establishing a specific function for one of the stimuli (you are told that A tastes like your favorite chocolate bar) and then observing that the other stimuli within the frame also acquire that function in the absence of direct training (B and C acquire at least some of the functions of your favorite chocolate bar). As such, RFT as an account is very much focused on how language and cognition impacts on our moment-to-moment "real-life" interactions with the natural environment rather than being a theory focused only on abstract or logical reasoning.

The distinction between relational entailment and transformations of stimulus functions indicates that each instance of relational framing occurs under two types of contextual control. One kind of contextual control (i.e., Crel) specifies the particular type of relation defining the relational response, while the other kind of contextual control (i.e., Cfunc) specifies the particular behavioral functions that are transformed in accordance with the relational response. In any instance of relational framing as an act in context, the stimuli

may be seen as possessing entailing or Crel properties and transformation-of-function or Cfunc properties. Thus, for example, the word "coffee" may have numerous entailment functions, or semantic meanings, such as "drink," "tea," and "Starbucks," and additional functions that occur, in part due to the entailed relations, such as actually going to Starbucks when a friend suggests having a coffee. The critically important distinction between Crel and Cfunc contextual control thus separates the two key properties of the stimuli that participate in relational frames.

The distinction within RFT between the Crel and Cfunc properties of stimuli is well established in the literature (Dymond & Barnes, 1994; Dougher, Augustson, Markham, Greenway & Wulfert, 1994), but empirical analyses of the distinction have typically been limited to studies exploring the derived transfer or transformation of functions, as described above (e.g., Perez, Fidalgo, Kovac, & Nico, 2015; Perez, et al., 2017). More recently, however, the importance of the distinction has become increasingly apparent. In particular, the distinction helps to explain subtle but reliable patterns of behavior observed with a methodology that was designed to capture the dynamics of arbitrarily applicable relational responding (AARRing) or relational framing "in flight."

In targeting the dynamics of AARRing, methodologies for measuring or indicating the relative strength of a previously established pattern of AARRing was required. Most, if not all, of the widely used methodologies employed within RFT research up until the mid-2000s were largely dichotomous (e.g., Dymond & Barnes, 1995). Specifically, the methods tended to involve procedures for training and testing specific patterns of AARRing to determine if they were present or absent in a behavioral repertoire, rather than indicating other properties of that pattern, such as the relative probability of one pattern versus another. One methodology that was developed to address the issue of relative probabilities in AARRing was the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes,

Stewart, & Boles, 2010). A typical IRAP presents a series of stimulus-stimulus relations with two response options, and requires opposing patterns of responses across blocks of trials, such that the relationship confirmed in one block is denied in the opposite block. For example, a Flowers-and-Insects IRAP might present the word “Flower” or “Insect” at the top of the screen, examples of positive and negative words (e.g., “Pleasant”, “Healthy”, “Happy”, “Unpleasant”, “Death”, “Pain”) in the middle of the screen, and the response options “True” and “False” at the bottom of the screen. The IRAP program combines these stimuli to produce four trial-types; *Flower-Positive*, *Flower-Negative*, *Insect-Positive*, and *Insect-Negative*. During one block of trials participants must select the response option “True” for *Flower-Positive* and *Insect-Negative* trials, and “False” for *Flower-Negative* and *Insect-Positive* trials. For most English speakers the assumption would be that this pattern of responding would be consistent with their verbal histories. During the opposing block of trials, the pattern is reversed, and participants must now select “False” for *Flower-Positive* and *Insect-Negative* trials, and “True” for *Flower-Negative* and *Insect-Positive* trials (assumed to be history-inconsistent for most English speakers). Comparing the response latencies across blocks of trials is taken to indicate the probability of one pattern versus the other. The general assumption is that shorter response latencies indicate higher response probabilities (e.g., all things being equal one might predict responding “True” more quickly than “False” on the *Flower-Positive* trial-type).

Although a large body of IRAP research has tended to use the IRAP as a measure of so called implicit cognition or attitudes (see Vahey, Nicholson, & Barnes-Holmes, 2015, for a clinically relevant meta-analysis), more recently some RFT-based research has focused on variables impacting IRAP performances such as rules or instructions (Finn, Barnes-Holmes, Hussey, & Graddy, 2016), the properties of the stimulus categories employed (O’Shea, Watson, & Brown, 2016), and the properties of the response options (Maloney & Barnes-

Holmes, 2016). In one series of studies a persistent pattern of IRAP effects termed a Single-Trial-Type-Dominance-Effect (STTDE) was noted (Finn, Barnes-Holmes, & McEntegart, 2018). This pattern involved significant differences in magnitude between trial-types that shared the response option “True” during history-consistent blocks of trials. Specifically, participants who completed a *Shapes-and-Colors* IRAP (see below) consistently produced significantly larger IRAP effects for the *Color-Color* than the *Shape-Shape* trial-type. This pattern emerged, and has been reported in other IRAP research (e.g., Kavanagh, Barnes-Holmes, Barnes-Holmes, McEntegart, & Finn, 2018), despite the fact that the same response option is required within each block of IRAP trials. These results were not readily accounted for by an existing model of IRAP performances, and so the authors proposed a new model, the Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) model.

The DAARRE model incorporates variables beyond the particular stimulus-stimulus relation (between label and target) presented within each trial-type of the IRAP. Thus, if a particular relationship between stimuli in an IRAP remains constant, the IRAP effects may still differ due to the properties of the stimuli that extend beyond the stimulus-stimulus relation. Consider the example of the *Shapes-and-Colors* IRAP from Finn et al. (2018), in which the effect for the *color-color* trial-type was larger than for the *shape-shape* trial-type. An illustration of this effect in the DAARRE model is presented in Figure 1. The actual IRAP presented the labels “Color” or “Shape” at the top of the screen with examples of colors and shapes as targets in the center of the screen (e.g., blue, red, orange, square, rectangle, circle).

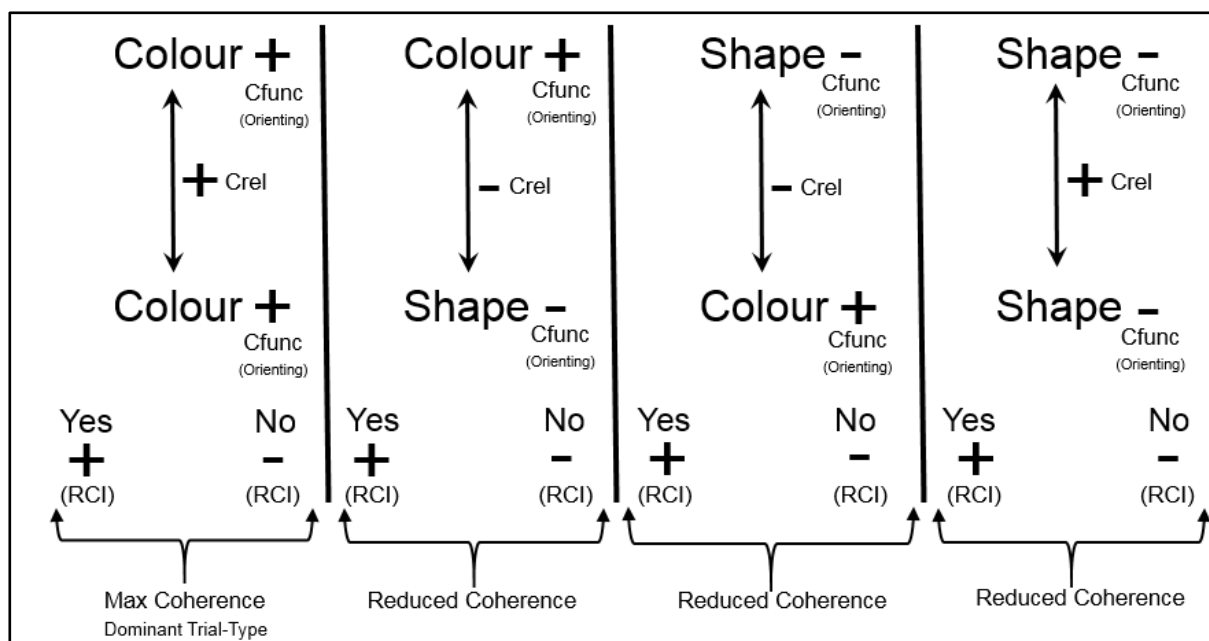


Fig. 1 A DAARRE model representation of the *Shapes-and-Colors* IRAP.

In natural language, “blue” *is a* “color” in the same way that “square” *is a* “shape”. Thus, in Figure 1 the relationships between these pairs of stimuli are labelled in the same way (i.e., with a plus sign). The authors (i.e., Finn, et al., 2018) suggested that the difference in the size of the effect for the *color-color* and *shape-shape* trial-types (the STTDE) may be explained, in part, by the fact that in natural language color words occur with far greater frequency than do shape words (Kuleers, Diependaele, & Brysbaert, 2010). It is assumed that these differences in frequency in natural language are likely to have produced differences in the functional properties of color words and shape words. Critically, therefore, participants may orient more strongly towards the color words than the shape words. To reflect the relative dominance of color words over shape words in natural language, in Figure 1 color words are labelled with a positive symbol (+) and shape words are labelled with a negative symbol (-). Furthermore, the functional properties of the two response options are also not equal – specifically, in the context of the IRAP confirmatory words, such as “True,” likely possess stronger orienting functions than dis-confirmatory words, such as “False”; frequently

referred to as a “positivity bias” (see Dodds et al., 2015). Words such as these (i.e., “True”, and “False”), may serve to indicate the coherence or incoherence of a particular relational network rather than specify the relationship between stimuli as is the case with Crels (Maloney & Barnes-Holmes, 2016, Hayes, Stewart, & McElwee, 2017; see Hayes, Stewart, & McElwee, 2016, for an applied example). Thus, these words have come to be referred to as Relational Coherence Indicators (RCIs).

According to the DAARRE model, the functions of the stimuli, the relationships between them, and the response options, interact within and across the trial-types and blocks of trials. That is, responding to a trial in an IRAP involves responding appropriately to the label stimulus, the target stimulus, the relationship between them, and selecting the appropriate response option for that block of trials¹. The responses to each of these elements are defined, in part, by the functional properties of the other stimuli in the IRAP context (e.g., in the context of the IRAP colors possess stronger orienting functions than shapes, and the response option “True” possesses stronger orienting functions than the response option “False”). As a result, in an IRAP the responses to each element of a trial (i.e., to the label stimulus, the target stimulus, and the relationship between them) will overlap with the orienting properties of the response option to a greater or lesser extent. In the case of the *Shapes-and-Colors* IRAP, this means that confirming that colors are colors involves a higher level of functional overlap across the labels, targets and response options (4 plus signs), than confirming that shapes are shapes (2 plus and 2 minus signs). In other words, “differential trial-type effects may be explained by the extent to which the Cfunc and Crel properties cohere with the RCI properties of the response options across blocks of trials” (Finn, et al., 2018, p. 21).

¹ Here, the word “appropriate(ly)” simply implies responding in accordance with the prevailing contingencies in the natural environment.

The interpretation of the STTDE offered by Finn et al. (2018) was consistent with the finding of their third experiment. Specifically, the STTDE failed to emerge when Shapes and Colors as stimulus categories were replaced by Forks and Spoons, because the latter were not strongly differentiated in terms of orienting functions (i.e., they occur with relatively equal frequency in natural language; Keuleers, et al., 2010). This effect is illustrated in terms of the DAARRE model in Figure 2, in which the lack of a naturally dominant stimulus category is indicated by labelling each of the label and target stimuli in the same manner (i.e., +/-). Insofar as the properties of the labels and targets do not differ significantly, the trial-type effects for the *Fork-Fork* and *Spoon-Spoon* trial-types should not differ significantly, and as indicated above this is what was reported by Finn et al. (Experiment 3).

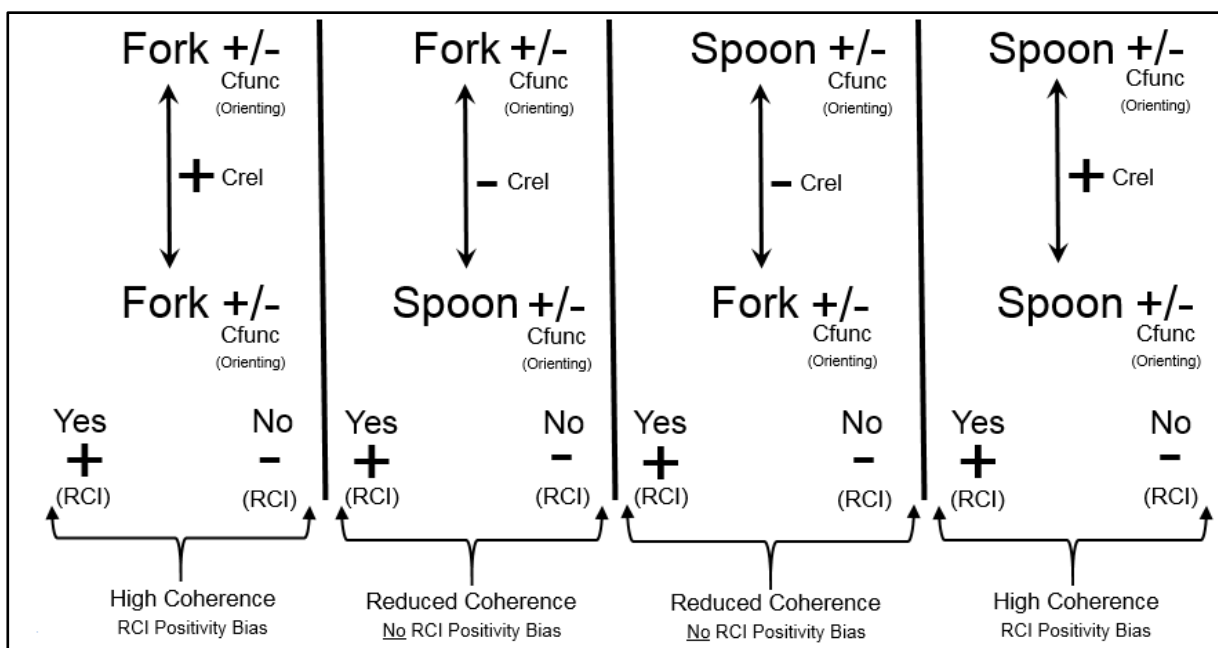


Fig. 2 A DAARRE model representation of the *Forks-and-Spoons* IRAP.

At this point, the DAARRE model has been used to predict the STTDE based on likely pre-experimental histories of the participants (i.e., differential word frequencies in natural language). At the time of writing, however, no published study had attempted to engineer (i.e., predict-and-influence) the STTDE by providing appropriate behavioral

histories *within* the experimental context itself. The current study constitutes the first attempt to achieve this outcome. Specifically, we sought to produce opposing STTDEs for two groups of participants within a single experimental session by establishing “True” functions for pairs of categories of stimuli that occur with relatively equal frequency in natural language (e.g., batteries and cans; boxes and pencils, etc.). This involved pre-training one of the stimuli from each pair by requiring participants to respond “True” whenever an exemplar of that category appeared on screen (e.g., battery-“True”; box-“True”).

The training thus aimed to produce a “positive” or “confirming” function for the stimulus (i.e., we aimed to change the orienting functions of the stimulus), and that stimulus would then act as a label stimulus in a subsequent IRAP. If a positive function was established for the stimulus it would then dominate over the other untrained label stimulus. This would in turn increase the level of functional overlap between the elements making up the trials bearing this newly “positive” label stimulus and the orienting properties of the response options (relative to the trials that bore label stimuli for which no function had been established experimentally). The critical question we asked was would this IRAP produce a STTDE effect for the trial-type that presented the stimulus category that had been pre-trained with “True”. This was assessed by exposing participants to an IRAP in which the category that had been trained to “True” was inserted into the IRAP. One group of participants were trained to produce the STTDE in one direction, and a second group were trained to produce the STTDE in the opposite direction. We did not include a no-training condition in the current study because Finn et al. (2018, Experiment 3) had already shown the absence of a STTDE when no training was provided for stimuli that did not differ in terms of orienting or any other known positivity/negativity functions (i.e., forks and spoons). In a sense, therefore, the current study involved “backward engineering” from the *Forks-and-Spoons* IRAP to the *Shapes-and-Colors* IRAP.

Method

Participants

Forty participants with an age range of 17 to 46 ($M = 24.3$ years, 28 females and 12 males) participated in the study in exchange for payment of €15. Participants volunteered for the study via Ghent University's SONA system, the vast majority of which were students attending the University. Participants were matched into pairs (described later), being allocated to their role within each pair based on the sequence in which participants volunteered on the SONA system. Ethical approval was granted by the ethics committee of the Faculty of Psychology and Educational Sciences at Ghent University.

Procedure

Stage 1: Familiarization IRAP. Upon arrival at the experimental session, participants were seated in front of a computer monitor and were asked to sign the informed consent form. The experimenter then initiated the IRAP program, which was presented on a standard computer. The GO-IRAP software was used to present stimuli and instructions to participants and record responses (available for download from <https://GO-RFT.com>). Nine IRAPs were employed in the current study (described here as one familiarization IRAP and eight Test IRAPs). The familiarization IRAP was used to create a broadly similar experimental history for each participant before completing the critical Test IRAPs (i.e., it was not used as a basis for including/excluding participants from the subsequent stages of the experiment).²

The experimenter described what participants would see on each trial, namely, a label stimulus, a target stimulus, and two response options. Participants were informed that their

² Finn et al. (2018) found that the STTDE appeared to be relatively stronger for participants with extended histories of completing latency-based measures (see *General Discussion* of Finn et al. for a post-hoc explanation of this moderating variable). To ensure that all participants in the current study had a minimum level of experience with the IRAP, they completed a familiarization IRAP before the experimental manipulation (the stimuli from the familiarization IRAP were never used in subsequent training and testing).

goal was first, to figure out how to respond to the various combinations of pictures and words that the program would present, and second, to respond as quickly and as accurately as they could. At this point, the experimenter invited participants to place one finger on the “D” key, one finger on the “K” key, and once they were ready to begin, to press the spacebar, and begin to respond to the trials.

During the familiarization IRAP, the label stimuli were pictures of tables and pictures of chairs, the target stimuli were the words “Table”, and “Chair”, while the response options were “True” and “False”. The IRAP program combined the labels and targets to produce four trial-types; *Table-Table*; *Table-Chair*; *Chair-Table*; and *Chair-Chair*. Eight examples of each trial-type were presented in each IRAP block. During the first block (history-consistent), participants were required to select the response option indicating “True” in response to the *Table-Table* and *Chair-Chair* trial-types, and to select the response option indicating “False” for the remaining two trial-types. The response options themselves alternated position across trials from left to right at the bottom of the screen such that they did not remain in the same location for more than three successive trials. If at any point participants selected the incorrect response for a trial, all stimuli remained in place and the program displayed a red “X” beneath the target stimulus. Only selecting the correct response option resulted in the removal of labels, targets, and the red “X”, and the initiation of a 400ms inter-trial interval. To encourage participants to respond quickly a red “!” would appear on screen if no response had been selected within 2000ms of the beginning of the trial.

When participants had completed a block of trials, the program displayed their mean latency and accuracy for that block, and instructed participants to press the spacebar to continue. Doing so presented participants with the following instruction “The previously correct and incorrect answers have been reversed. Press the spacebar to continue”. The pattern of responding for the second block of trials, a history-inconsistent block, was the reverse of

that required for the first block of trials. In this case, progressing from trial to trial was contingent on selecting “False” for *Table-Table* and *Chair-Chair* trial-types, and selecting “True” for the remaining trial-types. Aside from this difference, all other parameters remained the same across blocks of trials. If participants failed to achieve the pre-specified latency and accuracy criteria of at least 80% accuracy and a mean latency of less than 2000ms in either of the first pair of blocks, they were exposed to another pair of blocks. This sequence continued until the latency and accuracy criteria had been achieved during both a history-consistent block and history-inconsistent block within a block-pair, at which point participants proceeded to the test phase. If participants failed to achieve the criteria across four successive block-pairs, the IRAP software was programmed to terminate the session (note, however, that no participants failed to achieve the criteria within the allotted four pairs of blocks). The test phase involved three pairs of blocks and was procedurally similar to the practice phase, except no latency or accuracy criteria were applied to the test block performances. Accuracy and latency feedback were presented on screen between each block in order to encourage participants to maintain the performance criteria (one participant failed to maintain the practice criteria across two pairs of test blocks, but the participant continued to the next stage of the experiment because they were deemed to be familiar with the IRAP). When the test phase was completed, the program instructed participants to alert the researcher who terminated the IRAP program.

Stage 2: Function Training and Test IRAP. Following the familiarization IRAP, participants completed a sequence of four Function Training Tasks (described below) and four Test IRAPs (i.e., Function Training Task – IRAP – Function Training Task – IRAP – Function Training Task – IRAP – Function Training Task – IRAP). Each trial within the Function Training Task presented a picture from one of two categories (e.g., a picture of a battery) on-screen along with the response options “True” and “False” (see Table 1).

Participants were required to select the response option “True” for one of kind of picture (i.e., the Target) and to select the response option “False” for the other kind of picture (i.e., the Non-target). Each of the four tasks was followed by a corresponding IRAP (see Table 1). Each IRAP presented pictures of two different everyday items as label stimuli, and the Dutch words for these items as target stimuli. The target stimuli for each of these sets is presented in the final column of Table 1. All stimulus sets were presented with the response options “True” and “False”.

Participants were first split into pairs. One participant was randomly exposed to four of the eight Function Training Tasks using stimulus sets listed in the Group 1 column of Table 1 (e.g., Battery-Nail, Bag-Vase, Locker-Backpack, Socks-Towel). The second participant in that pair was also exposed to four Function Training Tasks using the corresponding stimulus sets listed in the Group 2 column of Table 1 (i.e., Nail-Can, Vase-Torch, Backpack-Glove, Towel - Ticket). In each case, participants were required to select ‘True’ when presented with the target stimulus and ‘False’ when presented with the non-target stimulus. Critically, both participants in each pair were trained to select ‘True’ in the presence of different target stimuli, which later appeared in the Test IRAP (e.g., Participant A selected ‘True’ for Battery and ‘False’ for Nail, whereas Participant B selected ‘True’ for Nail and ‘False’ for Can).

After each Function Training Task, both participants were exposed to a corresponding Test IRAP. The critical point here was that if a participant had been trained to respond ‘True’ to the picture of batteries in the Function Training Task, then a picture of a battery would appear as one of the label stimuli in the corresponding IRAP. The other label stimulus (pictures of cans) that was used in the IRAP was ‘novel’ in that it had not been presented in the Function Training Task. The IRAP for the other participant was the same, except the label stimulus for which they had been trained to select ‘True’ was pictures of cans, and the novel

stimulus were pictures of batteries. The other seven possible combinations of stimuli that were employed for the Function Training Tasks and IRAPs are presented in Table 1.

Table 1

Category labels for the picture stimuli employed in the Function Training Tasks and related Test IRAPs.

Function Training Tasks		Test IRAP Label stimuli
Group 1	Group 2	
Target – Non-target	Target – Non-target	
Battery - Nail	Can – Nail	Battery – Can
Bag - Vase	Torch – Vase	Bag – Torch
Box - Scarf	Pencil – Scarf	Box – Pencil
Fork - Comb	Spoon – Comb	Fork – Spoon
Hook - Chalk	Tie – Chalk	Hook – Tie
Locker - Backpack	Glove - Backpack	Locker – Glove
Pan - Ladder	Button – Ladder	Pan – Button
Socks - Ticket	Towel - Ticket	Socks – Towel

Note. Each row presents the Function Training Task stimuli for each group and the label stimuli in the IRAP that followed the Function Training Task for each group. During the Function Training Task participants were trained to respond ‘True’ to a picture (i.e., the Target stimulus in the Function Training Task), and this picture was subsequently presented as one of the two label stimuli in the Test IRAP. Each pair of participants were exposed to four of the eight stimulus sets listed above; for each pair the stimulus sets were allocated randomly.

Function Training Task. The Function Training Task was presented by PsychoPy version 1.85.4 (Peirce, 2007). The task comprised of four blocks of 36 trials (a total of 144 trials), with each block presenting a target stimulus (e.g., “Battery”) and a non-target stimulus

(in this case, “Nail”) an equal number of times (i.e., 18 trials each per block presented in random order).

The first screen of the Function Training Task summarized the task as follows:

In this task you will be presented with a series of screens.

On each screen there will be a picture, and two response options. You are required to select the appropriate response option on each screen. Failing to do so will result in a red X appearing on screen.

If you have not responded to a screen after 2000 milliseconds an exclamation mark will appear. Try to avoid the exclamation mark.

Your goal is to respond as accurately and quickly as possible.
Press the spacebar to continue.

On each trial there was a picture stimulus at the top center of the screen, and two response options at the bottom left and right of the screen. The response options alternated position from left to right of the screen from trial to trial such that they appeared an even number of times in each location. Selecting the incorrect response option produced a red “X” in the middle of the screen, whereas selecting the correct response option resulted in the disappearance of the picture stimulus, and the initiation of the subsequent trial after an inter-trial interval of 400ms. All stimuli remained on-screen until a correct response was emitted. To encourage participants to respond quickly a red “!” appeared if no response had been made 2000ms after the appearance of the picture stimulus. Upon completing a block of 36 trials, the task displayed the mean latency and accuracy for that block, and encouraged participants to improve their score. The task proceeded in this manner until four blocks of 36 trials had been completed. The experimenter then exited the program and initiated the corresponding Test IRAP.

Test IRAP. Each IRAP was similar to the familiarization IRAP with two modifications. First, the stimuli employed corresponded to those presented in the preceding

Function Training Task (see Table 1). Second, the test phase consisted of two pairs of test blocks (three pairs are often used in IRAP studies but this was reduced to two pairs because participants were required to complete five separate IRAPs within a single session). Each test IRAP presented practice blocks. Participants were required to successfully complete the practice phase (i.e., accuracy of more than 80% correct and an average time to correct response of less than 2000ms) before they were allowed to progress to the test blocks. When this IRAP had been completed, the next Function Training Task was presented, again followed by a corresponding IRAP. This sequence was repeated until each participant had completed each Function Training Task and corresponding Test IRAP with four sets of stimuli, thus ending the experiment.

Results

Familiarization IRAP

All 40 participants met the performance criteria during the practice blocks of the Familiarization IRAP; one failed to maintain the criteria during the test blocks but continued to the remaining stages of the experiment because the function of the familiarization IRAP had been fulfilled (i.e., to ensure each participant had a minimal level of exposure to the IRAP). Given that performance on the Familiarization IRAP was not the focus of the current research the data from that IRAP are not presented here.

Function Training Task

As noted previously, there were no performance criteria applied to the Function Training Task. However, all but one participant responded at or above 88% accuracy on all four Function Training Tasks; the remaining one participant responded at 80% correct, and only on one of four Tasks. Mean accuracy calculated across the four Function Training Tasks for the two groups were, Group 1 = 95.9% and Group 2 = 95.59%. An independent *t*-test indicated the difference was non-significant, $t = .46$, $p = .64$, $df = 38$). Mean latency scores

(time in milliseconds from stimulus onset to correct response) calculated across the Four Function Training Tasks for the two groups were, Group 1 = 661ms and Group 2 = 757ms. An independent *t*-test indicated the difference was significant, $t = 2.32$, $p < .05$, $df = 38$). While this difference is statistically significant, the two groups were treated no differently during the Experiment. We assume, therefore, that the observed difference is the result of random variation and draw no further conclusions. The mean reaction time for each Function Training Task completed by each participant in each group is displayed in Figure 5 in Appendix B.

Test IRAPs

For each of the four Test IRAPs, an average *D*-IRAP score was produced for the four trial-types in a manner consistent with the description provided by Finn et al. (2018)³. These calculations were as follows: 1. Only data from test block-pairs with a mean latency of 2000ms or less and accuracy of 80% were retained⁴. 2. A mean latency was calculated for each trial-type in each test block in each IRAP. 3. A raw difference score was produced for each trial-type in each test block-pair of each IRAP by subtracting the mean latency for that trial-type in the history-consistent block of trials from the latency for that trial-type in the history-inconsistent block of trials. 4. A standard deviation was calculated for each trial-type from the latencies in both blocks within each test block-pair of each IRAP. 5. A *D*-IRAP score was produced for each trial-type in each test block-pair of each IRAP by dividing the

³ Because the IRAPs that followed each iteration of the Function Training Task presented only two test block-pairs, the *D*-IRAP scores comprised the average of these two pairs of test blocks. In a small number of cases ($N = 11$) where participants did not maintain the performance criteria across a test block-pair, a *D*-IRAP score was produced from one test block-pair. It should be stressed, however, that each participant completed 4 separate IRAPs, each with two test block-pairs (i.e., 8 test block-pairs in total). The effects reported in the current study were thus generated from a minimum of 5 test block-pairs for each participant. Therefore, the effects reported here are based on more data points than is typical in IRAP research. The data from the first test IRAP completed by participant 2 were excluded from all analyses due to a translation error in the stimulus set. Participant 2 highlighted the error, and the stimuli were removed from the stimulus pool and were thus not presented to any other participant.

⁴ Secondary analyses employing the performance criteria of mean latency of less than 2000ms and accuracy of greater than 75% are presented in Appendix A.

raw difference scores produced in step 3 by the corresponding standard deviation produced in step 4 (e.g., the raw difference score for a trial-type in test block-pair 2 of IRAP 3 was divided by the standard deviation of the latencies for that trial-type 1 in test block-pair 2 or IRAP 3).

6. An average *D*-IRAP score was produced for each IRAP by averaging the *D*-IRAP scores for the test block-pairs in each IRAP that were produced in step 5. 7. A single aggregate *D*-IRAP score was produced for each trial-type for each participant by averaging the *D*-IRAP scores produced in step 6.

The foregoing calculations resulted in four aggregate *D*-IRAP scores for each participant. The four trial-types are identified in Table 2 based on whether the label stimulus for the trial-type was a target picture for Group 1 (e.g., batteries, boxes, and forks), or Group 2 (e.g., cans, spoons, and gloves), and whether the label (picture) stimulus matched the category descriptor. The intervention in the current study targeted trial-types requiring the response option “True”, therefore trial-types where labels and targets matched are referred to as the Group 1 Targeted Trial-Type, and the Group 2 Targeted Trial-Type. The means and standard errors of the *D*-IRAP scores for each of the four trial-types for the two groups are presented in Table 2.

Table 2. The means and standard errors of the *D*-IRAP scores for each of the four trial-types for the two groups.

	Group 1 Targeted Trial-Type (Descriptor Match)	Group 1 Picture (Descriptor Mismatch)	Group 2 Picture (Descriptor Mismatch)	Group 2 Targeted Trial-Type (Descriptor Match)
Group 1	.54 (.06)	.15 (.07)	.14 (.08)	.41 (.08)
Group 2	.33 (.05)	.14 (.05)	.13 (.06)	.44 (.05)

The data from the two trial-types during which participants were required to respond “False” on history consistent trials were excluded from the inferential statistical analyses

because no specific theoretically driven prediction was made regarding them. The two trial-types of interest were those that required the response option “True” during history consistent blocks of trials.

Group level analyses. The four averaged *D*-IRAP scores were all positive indicating that in general participants responded more quickly when confirming, rather than denying, that the picture stimulus and the word were coordinate (e.g., responding “True” rather than “False” when presented with a picture of a battery and the word battery). The critical effect, however, was the difference in the size of the effects across the two targeted trial-types (with matching descriptors) for each group. Specifically, if the Function Training Task produced the predicted effect, then the size of the *D*-IRAP score for the Group 1 participants should be larger for the Group 1 stimuli than for the Group 2 stimuli, and the opposite should be the case for the Group 2 participants (a larger *D*-IRAP score for the Group 2 relative to the Group 1 stimuli). This is indeed the pattern observed in Table 2. Participants in Group 1 produced a larger effect for the trial-type bearing Group 1 targeted stimuli ($M = .54$) than for the trial-type bearing Group 2 targeted stimuli ($M = .41$). Participants in Group 2, meanwhile, produced a larger effect for the trial-type bearing Group 2 targeted stimuli ($M = .44$) than for the trial-type bearing Group 1 targeted stimuli ($M = .33$). A mixed 2x2 repeated measures Analysis of Variance (ANOVA) with trial-type and group as factors, confirmed that the critical interaction effect was statistically significant, $F(1, 38) = 9.00, p < .01, \eta_p^2 = .19$ (both of the main effects were non-significant, $ps > .2$).

The overall IRAP performances did not differ between the groups (note Table 2, and the absence of a main effect). For both groups, the largest trial-type effect was produced on the trial-type targeted by the Function Training Task intervention. The absence of a baseline measure prevents a definitive conclusion from being drawn about the pattern of effects for *a single group*. However, in the interaction effect each group acts as an active control for the

treatment received by the other group. That is, the relative dominance of the trial-types requiring the response option “True” in history-consistent blocks of trials should be in the opposite direction for both groups. The observed patterns of effects were consistent with this prediction and are unlikely to have been obtained by chance ($p < .01$). This indicates that the Function Training Task did influence which trial-type produced the largest effect for both groups. Additional analyses of the relationship between performance during the Function Training Task and the STTDE are presented in Appendix B.

Individual level analyses. The foregoing analysis provides supportive evidence for the successful manipulation of the STTDE at the group level. To explore the extent to which this manipulation was successful at the individual participant level we conducted additional analyses. For each participant, the difference between the two *D*-IRAP scores was calculated and these are presented in Figure 3. Scores above zero indicate that a participant produced an aggregate difference between *D*-IRAP scores that was greater for the targeted trial-type (i.e., contained label stimuli that had previously been trained to “True”) relative to the non-targeted trial-type (i.e., contained label stimuli that had *not* been trained to “True”). Scores below zero indicate the opposite effect. Figure 3 shows that 28 out of the 40 participants produced effects in the predicted direction (i.e., greater than zero). Assuming that scores above zero and scores below zero are equally likely by chance, a binomial probability test indicates that observing 28 out of 40 effects in the predicted direction is statistically significant ($B(40, 0.5)$ where $N \geq 28$, $p < .01$). The effects that deviated from zero to the greatest extent were in the predicted direction. Thus, the group effect presented above arose from relatively large differences produced by a relatively large number of participants. That is, there were 16 cases where the aggregate difference in trial-type *D*-IRAP scores exceeded .2 in the predicted direction, compared to six cases where the aggregate difference was less than -.2.

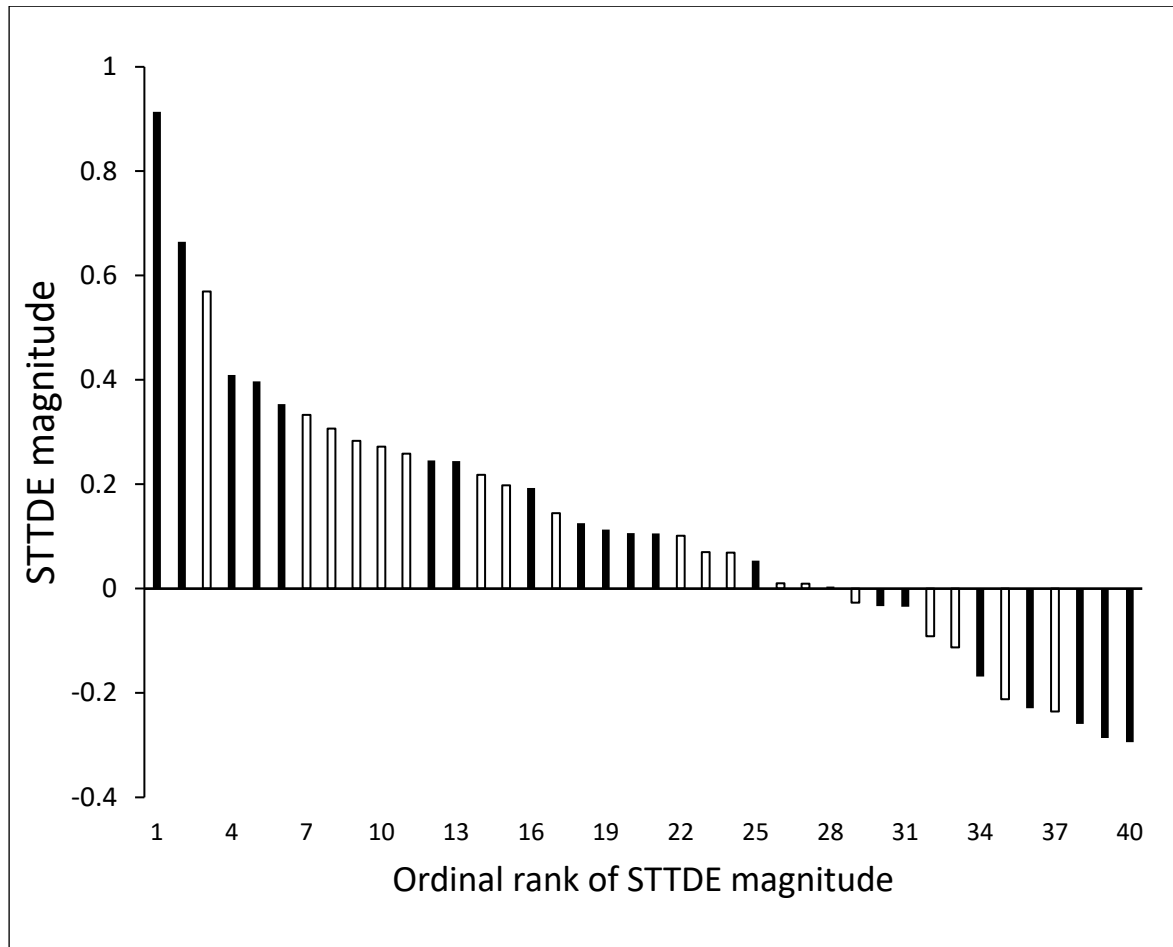


Fig. 3 The aggregate difference between the Group 1 targeted trial-type and the Group 2 targeted trial-type for each participant of the 40 participants. The x-axis, from left-to-right, represents the ordinal rank of aggregate differences between the trial-types based on the extent to which they were in the predicted direction (i.e., 1 indicates the largest effect in the predicted direction, whereas 40 indicates the largest effect that was in the unpredicted direction).

Discussion

The results from the current study provide evidence that a STTDE may be influenced within an experimental session. That is, manipulating the Cfunc properties of a stimulus, by establishing a “True” function, appears to produce a STTDE in an IRAP that presents this stimulus as a label. At a group level, there were differences in the differential trial-type effects generated by the two groups of participants exposed to the two sets of stimuli. Furthermore, at the individual level the majority of participants produced STTDEs that were consistent with

predictions derived from the DAARRE model. The current findings thus support the argument that a relatively comprehensive explanation of the patterns of *D*-IRAP effects will likely involve appealing to the *Crel* and *Cfunc* properties of the stimuli presented in an IRAP.

At this point, it should be acknowledged that the STTDE could be seen as directly relevant to effects that have been reported previously in the “mainstream” psychological literature (e.g., salience asymmetry, Rothermund & Wentura, 2004; the polarity principle, Proctor & Cho, 2006; positivity framing biases, O’Shea, et al., 2016). We agree that at the level of the basic effect there is certainly a large degree of overlap among these effects. However, the purpose of the current research is to develop a behavior-analytic account or explanation of the STTDE and related phenomena. Whereas mainstream (cognitive) models, broadly speaking, explain the processing and retrieval of information by a cognitive system that occurs during the completion of a procedure, the DAARRE model appeals to a single behavioral process (i.e., mutual entailment and the transformation of functions). The current study is thus part of an on-going research program that is seeking to develop and refine the DAARRE model as a way of explaining patterns of AARRing, and to facilitate greater prediction-and-influence of such patterns.

Although the current data could be seen as reasonably compelling, it remains the case that 12 of the 40 participants produced negative scores (in Figure 3). As pointed out in the Introduction, however, it is important to understand that the properties of the stimuli employed in the current study would be determined, in part, by unknown pre-experimental histories. That is, the STTDE is a multivariate phenomenon. Thus, a particular trial-type may have produced a relatively strong STTDE in a negative direction, which may be highly resistant to experimental manipulation using a training procedure that lasted no more than a few minutes. We attempted to circumvent this possibility by requiring participants to complete several training tasks, and thus one might be surprised to find that 12 of the

participants “failed” to show the effect. On balance, in the absence of a baseline measure of the STTDE, it is not possible to determine the precise impact that the training tasks had on the IRAP performances. Or to put it another way, it is possible that a participant who produced a negative score in Figure 3 may have produced an even more negative score in the absence of the training procedure. In this context, it is important to note that the entire distribution of scores as illustrated in Figure 3 appear to have been shifted in the predicted direction. Indeed, the largest positive score relative to zero (0.9 approx.) was approximately three times greater than the largest negative score relative to zero (-0.3 approx.). If the training had no effect on the STTDE then one would expect a relatively even distribution around zero, and if it failed to have an effect on some of the participants a larger number of relatively stronger scores in a negative direction should have been observed. However, the predicted STTDE was shown by 28 of the 40 participants, a result that would be observed, by chance alone, in less than 1 in 100 instances.

At this point, it seems important to note that the current study should be seen as simply the beginning of what will necessarily be a protracted program of RFT research on the dynamics of arbitrarily applicable relational responding. In this context, we recognize that there are many issues that will require additional experimental analyses to refine the conceptual analysis we are developing. For example, it could be argued that the manner in which we established a “True” function for the various stimuli presented within the IRAP in the current study employed a conditional discrimination procedure (participants picked “True” in the presence of a picture), which may be seen as primarily involving Crel control. Although this may well be the case, the DAARRE model only predicts the subsequent emergence of the STTDE if the function of the picture (a Cfunc property) had been transformed by the training. It is thus important to understand that the distinction between Crel and Cfunc properties should be seen as relatively “fuzzy” in much the same way that the

distinction between the eliciting and reinforcing functions of a stimulus are not separable in an absolute sense (Epstein, 2012). Of course, only time will tell if the distinction between Crel and Cfunc properties proves to be increasingly pragmatically useful in subsequent experimental analyses, but that may be seen as a core part of the protracted program of RFT research that we are currently engaged in.

In going forward, one of the main issues that will need to be addressed is the extent to which it is possible to manipulate DAARRE model variables within participants. For example, a set of relatively novel stimuli could be inserted into an IRAP and a baseline measure of responding to those stimuli could be recorded. Subsequently, the functions of the stimuli, and/or the relationships between them, could be targeted with various interventions that were designed to produce STTDEs in one direction or the other. Such research may help to determine the extent to which targeting relations versus functions, or both, produces weaker or stronger STTDEs. This would test and develop the DAARRE model and also enhance our understanding of the dynamics of AARRing.

In closing, we wish to emphasize a relatively unique feature of the current study in that it involves a shift towards single-participant research using the IRAP. This shift involves recognizing the IRAP as a measure of the dynamics of arbitrarily applicable relational responding, rather than as a measure of so called implicit attitudes or cognition. Our view of the IRAP as a measure of these dynamics has promoted a focus on the variables impacting on those dynamics (e.g., Finn, et al., 2016, 2018; Maloney & Barnes-Holmes, 2016). This emerging program of research has allowed us to target a subtle feature of IRAP performance, the STTDE; an effect that is based on differences between differences in response latencies, the latter of which (i.e., response latencies) are often seen as an inherently “noisy” metric. A feature of the current research, therefore, is that it appears to show some promise in using what many consider to be a measure that is best used only at the group level of analysis for

individual participant research. When addressing the analysis of relational framing “in flight” the relative strength of Crel and Cfunc properties in relational networks becomes a primary focus, which requires methodologies that allow us to analyze complex human behaviors at the individual level. The current study should thus be seen as a first step in using the IRAP to conduct relatively precise experimental analyses of the dynamics of AARRing with individual participants.

Compliance with Ethical Standards

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

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional 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.

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*, 527-542. doi: 10.1007/BF03395726
- Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., ... & Megerdooian, K. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*(8), 2389-2394. doi: 10.1073/pnas.1411678112
- 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. doi: 10.1901/jeab.1994.62-331
- Dymond, S., & Barnes, D. (1994). A transfer of self-discrimination response functions through equivalence relations. *Journal of the Experimental Analysis of Behavior*, *62*(2), 251-267. doi: 10.1901/jeab.1994.62-251
- 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. doi: 10.1901/jeab.1995.64-163
- Epstein, R. (2012). Crying babies. *Journal of Contextual Behavioral Science*, *1*(1-2), 43-48. doi: 10.1016/j.jcbs.2012.05.001
- Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules. *The Psychological Record*, *66*, 309-321. doi:10.1007/s40732-016-0173-4
- Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the Single-Trial-Type-Dominance-Effect in the IRAP: Developing a Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) 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: Kluwer Academic.
- Hayes, J., Stewart, I., & McElwee, J. (2016). Assessing and training young children in same and different relations using the relational evaluation procedure (REP). *The Psychological Record*, *66*(4), 547-561. doi: 10.1007/s40732-016-0191-2
- Hayes, J., Stewart, I., & McElwee, J. (2017). Children's answering of yes-no questions: A review of research including particular consideration of the relational evaluation procedure. *Behavioral Development Bulletin*, *22*(1), 173. doi: 10.1037/bdb0000027
- Kavanagh, D., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Finn, M. (2018). Exploring differential trial-type effects and the impact of a talk-aloud procedure on deictic

- relational responding on the IRAP. *The Psychological Record*, 68(2), 163-176. doi: 10.1007/s40732-018-0276-1
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, 1, 174. doi: 10.3389/fpsyg.2010.00174
- Maloney, E., & Barnes-Holmes, D. (2016). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: The role of relational contextual cues versus relational coherence indicators as response options. *The Psychological Record*, 66, 1-9. doi: 10.1007/s40732-016-0180-5
- O'Shea, B., Watson, D. G., & Brown, G. D. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*, 28(2), 158. doi: 10.1037/pas0000172
- Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, 162(1-2), 8-13. doi: 10.1016/j.jneumeth.2006.11.017
- Perez, W. F., Fidalgo, A. P., Kovac, R., & Nico, Y. C. (2015). The transfer of Cfunc contextual control through equivalence relations. *Journal of the Experimental Analysis of Behavior*, 103(3), 511-523. doi: 10.1002/jeab.150
- Perez, W. F., Kovac, R., Nico, Y. C., Caro, D. M., Fidalgo, A. P., Linares, I., de Almeida, J.H., & de Rose, J. C. (2017). The transfer of Crel contextual control (same, opposite, less than, more than) through equivalence relations. *Journal of the Experimental Analysis of Behavior*, 108(3), 318-334. doi: 10.1002/jeab.284
- Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: a general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, 132(3), 416. doi: 10.1037/0033-2909.132.3.416
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: dissociating salience from associations. *Journal of Experimental Psychology: General*, 133(2), 139. doi: 10.1037/0096-3445.133.2.139
- Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech, Language, and Hearing Research*, 14(1), 5-13. doi: 10.1044/jshr.1401.05
- Sidman, M. (1994). Stimulus equivalence: A research story. *Boston: Authors Cooperative*.
- Sidman, M., Rauzin, R., Lazar, R., Cunningham, S., Tailby, W., & Carrigan, P. (1982). A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children. *Journal of the Experimental Analysis of Behavior*, 37(1), 23-44. doi: 10.1901/jeab.1982.37-23
- 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

Appendix A

Secondary Analyses

The following analyses were conducted on data from test IRAPs in which performances in each block of a test block-pairs was at least 75% accurate and a mean response latency of 2000ms or less. This resulted in the exclusion of four of the 318 test block-pairs completed by participants, as opposed to the exclusion of 14 test block-pairs as per the original analyses. Both sets of statistical analyses yield similar results.

Secondary analysis: The means and standard errors of the *D*-IRAP scores for each of the four trial-types for the two groups.

	Group 1 Targeted Trial-Type (Descriptor Match)	Group 1 Picture (Descriptor Mismatch)	Group 2 Picture (Descriptor Mismatch)	Group 2 Targeted Trial-Type (Descriptor Match)
Group 1	.57 (.06)	.15 (.07)	.14 (.08)	.43 (.08)
Group 2	.32 (.05)	.14 (.05)	.13 (.06)	.44 (.05)

Secondary analysis: Mixed 4x2 repeated measures Analysis of Variance (ANOVA) with trial-type and group as factors.

	Df	Sum Sq	Mean Sq	F value	<i>p</i>
Group	1	0.276	0.276	2.699	.109
Residuals	38	3.885	0.102		
Trial-Type	1	0.003	0.003	0.092	.76
Trial- Type*Group	1	0.371	0.371	10.56	.002
Residuals	38	.1.334			

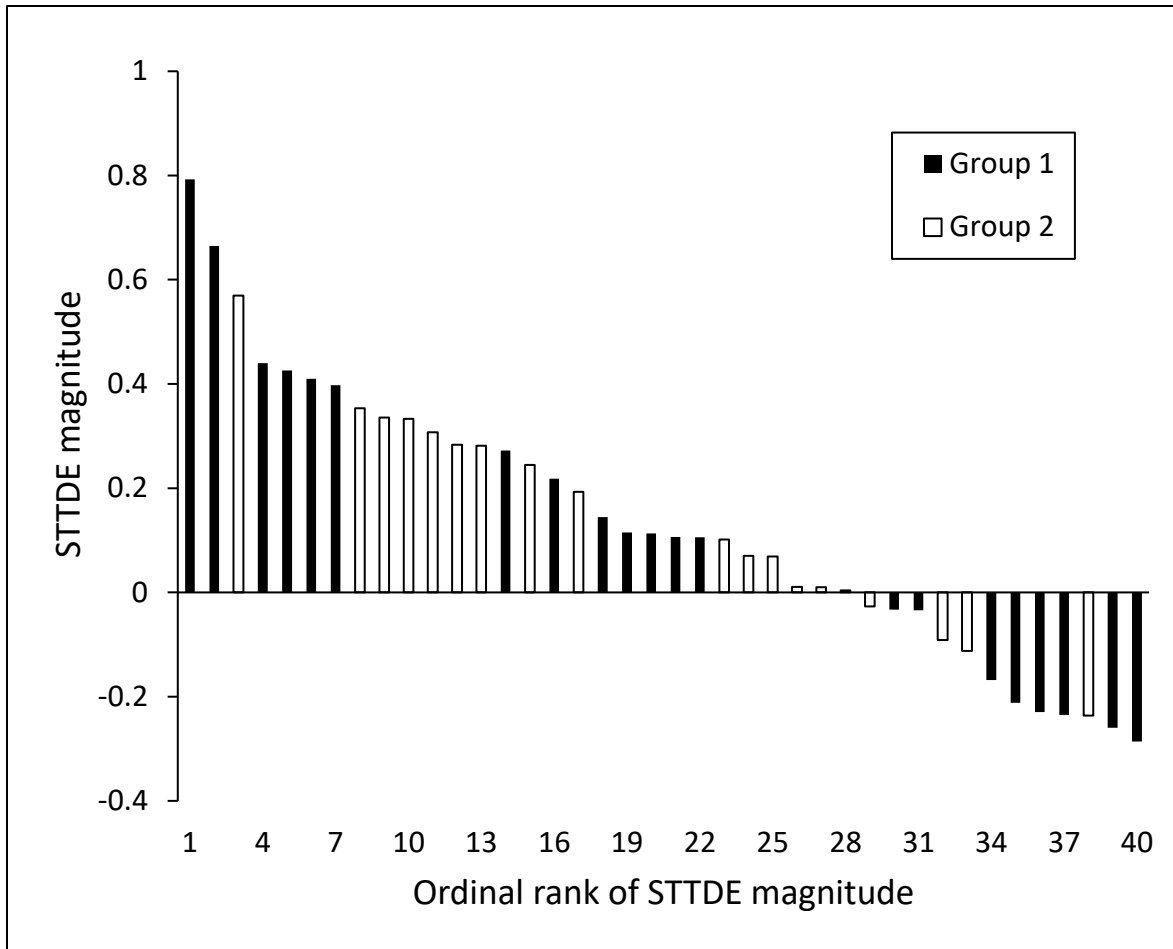


Fig 4. The equivalent to Figure 3 from the secondary analysis. The aggregate difference between the Group 1 targeted trial-type and the Group 2 targeted trial-type for each participant of the 40 participants. The x-axis, from left-to-right, represents the ordinal rank of aggregate differences between the trial-types based on the extent to which they were in the predicted direction (i.e., 1 indicates the largest effect in the predicted direction, whereas 40 indicates the largest effect that was in the unpredicted direction).

Appendix B

Function Training Task and STTDE

To assess the impact of performance during the Function Training Task on the STTDE dominance effect a regression analysis was performed. The variables included in the regression were the mean reaction time across the final two blocks of each Function Training Task and the magnitude of the STTDE from the subsequent IRAP. One of the 158 Function Training Tasks was excluded from the regression analysis as it was an extreme outlier more than five standard deviations above the mean. The regression was not statistically significant ($p = .2$, adjusted $r^2 = 0.004$). A scatterplot with the line of best fit is presented in Figure 5 (below). The absence of a relationship between performance on the Function Training Task and the size of the STTDE can be explained by the *D*-IRAP algorithm controlling for individual differences in response latency (see Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010).

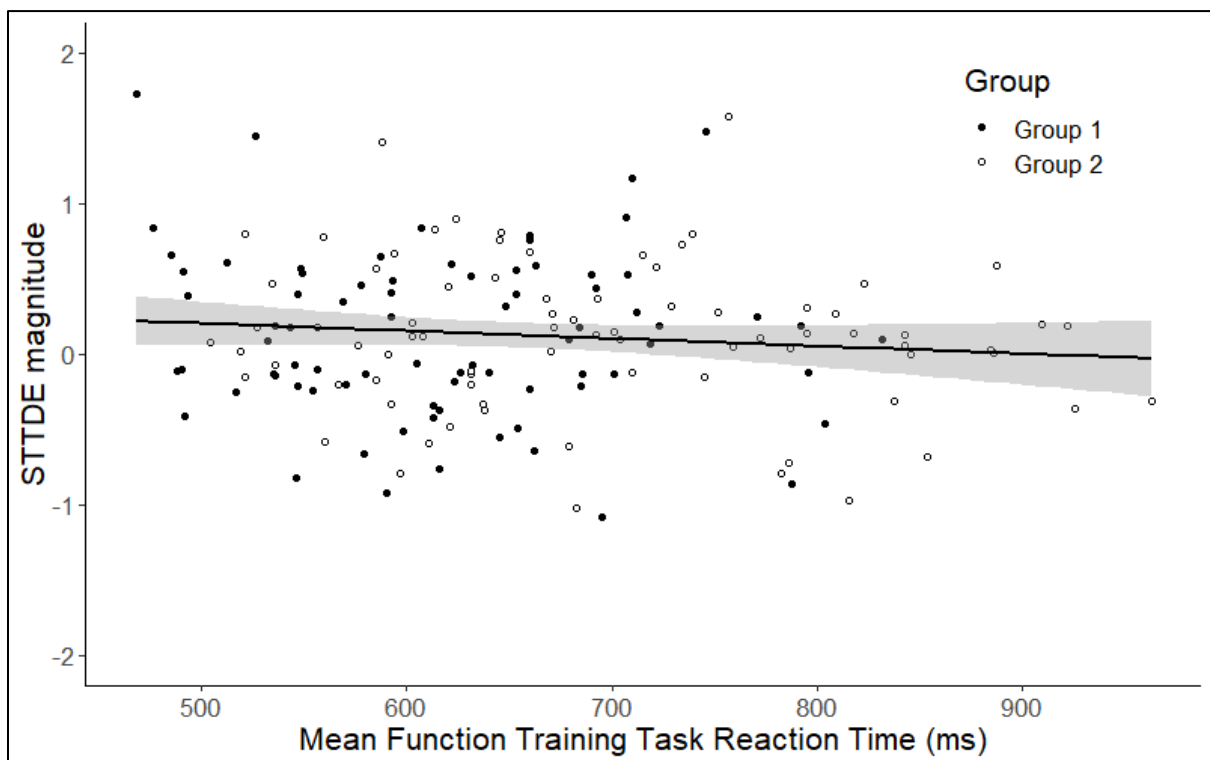


Fig 5. The relationship between latency during the Function Training Task and the Single-Trial-Type-Dominance-Effect. The line of best fit is displayed in black, with the 95% confidence interval shaded in grey.