Exploring the Behavioral Dynamics of the Implicit Relational Assessment Procedure: The Role of Relational Contextual Cues versus Relational Coherence Indicators as Response Options

Emma Maloney

Department of Psychology, Maynooth University, Ireland

Dermot Barnes-Holmes

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

Corresponding Author:

Dermot Barnes-Holmes

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan 2

B-9000 Ghent

Belgium

Email: Dermot.Barnes-Holmes@ugent.be

Authors' Note This article was prepared with the support of an Odysseus Group 1 grant awarded to the second author by the Flanders Science Foundation (FWO). Correspondence concerning this article should be sent to Dermot.Barnes-Holmes@ugent.be

Abstract

Purpose The current study examined the role of Relational Contextual Cues (Crels) versus Relational Coherence Indicators (RCIs) as response options in the Implicit Relational Assessment procedure (IRAP).

Method Fifty-two university undergraduate participants successfully completed two consecutive IRAPs. Both IRAPs were similar except for the response options employed. The Crels "Similar" and "Different" served as response options for one IRAP with the RCIs "True" and "False" as response options for the other. The order in which the two different IRAPs were completed was counterbalanced across participants.

Results Although the two types of response options yielded similar effects for the participants' first exposures to the IRAPs, differences emerged during the second exposures. In addition, one of the four trial-types from the IRAP appeared to be particularly sensitive to the Crel-RCI manipulation and the order in which the two types of IRAP blocks were presented (consistent-first versus inconsistent-first with natural verbal relations). The findings highlight the complex behavioural dynamics that may be involved in IRAP performances, and suggest that even seemingly trivial components of the procedure require systematic analysis.

Key words: Relational frame theory, contextual cues, relational coherence indicators, human adults

Exploring the Behavioral Dynamics of the Implicit Relational Assessment Procedure: The Role of Relational Contextual Cues versus Relational Coherence Indicators as Response Options

In recent decades, behavior-analytic researchers have directed increasing attention towards the study of human language and cognition, with a substantive amount of this work being conducted under the rubric of Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). The theory argues that the core units of human language involve generalized relational operants, or relational frames, and a growing body of evidence has provided support for the account (see Hughes & Barnes-Holmes, in press-a, in press-b, for extensive reviews). Much of the early work in RFT focused on demonstrating the emergence of relational frames under tightly controlled experimental conditions. That is, participants were typically trained, using differential reinforcement contingencies, to relate specific stimuli to each other under particular types of contextual control and were then tested, in absence of programmed reinforcement, to determine if predictable relational responses would emerge. A basic example of a relevant study might be as follows. An individual might first be trained to respond to an abstract shape as functionally equivalent to the relation of "sameness". Subsequently, the person would then be trained in the presence of the sameness cue to relate a nonsense word such as CUG to another nonsense word such as VEK and to relate VEK to a third nonsense word such as ZID. The critical test phase then involved presenting CUG and ZID in the context of the sameness cue to determine if participants would respond to these two indirectly related stimuli as being the same. If such a relational response emerged in the absence of direct reinforcement, prompting or further instruction, the generalized relational operant (or relational frame) of coordination was deemed to have emerged. Numerous patterns of relational framing were studied using this basic research strategy, focusing on relations such as opposite, difference (Steele & Hayes, 1991), more-than and less-than

(Dymond & Barnes, 1995), before and after (O'Hora, Barnes-Holmes, Roche, & Smeets, 2004) and the research provided increasing support for the basic tenets of RFT.

Although this early work was critical in testing the basic concepts of RFT other research directions have emerged in recent years. One particularly active area of research has involved the development of a procedure that was designed to measure or assess the strength of generalized relational responses that had been established in the natural environment, rather than simply assessing relational responses that were established in the laboratory. Specifically, the Implicit Relational Assessment Procedure (IRAP) was developed to assess relational framing in natural language. In the seminal study in this area (Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008), participants were presented with a task that involved presenting one of two label stimuli, "Pleasant" or "Unpleasant", at the top of a computer screen and one of a number of positively or negatively valenced target stimuli (e.g., "Love", "Happy", "War", "Hate") in the centre of the screen. On each trial of the IRAP participants were required to choose between one of two response options, "Similar" and "Opposite". On some blocks of trials participants were required to respond in a manner that was deemed consistent with natural language (e.g., choosing "Similar" when "Pleasant" and "Love" appeared on screen) and on other blocks of trials responding in a manner deemed inconsistent with natural language was required (e.g., choosing "Opposite" when Unpleasant and "Hate" appeared). The basic metric, or so called IRAP effect, involved calculating the difference in response latencies between consistent versus inconsistent blocks of trials, with the basic prediction that participants would respond more quickly during consistent relative to inconsistent blocks.

The prediction was up-held in this first study and an RFT-based account of the IRAP effect, known as the Relational Elaboration and Coherence (REC) model, was proposed (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010; Hughes, Barnes-Holmes, & Vahey,

2012). The basic concept behind this model is that the IRAP effect reflects the relative probabilities in brief and immediate relational responses (BIRRs) that exist in the natural environment, with higher probabilities being reflected in lower response latencies. Thus, for example, the average English speaker should choose "Similar" more rapidly, in a time-pressured context, than "Opposite" when presented with the label "Pleasant" and the target "Love". The REC model contrasts BIRRs with extended and elaborated relational responses (EERRs), which are seen as responses that occur more slowly and typically involve greater levels of relational complexity than BIRRs. An important feature of the REC model is that as responses become increasingly EERR-like it becomes more difficult to predict exactly what impact they will have on the IRAP. In other words, all things being equal, the IRAP was designed to capture relational framing that occurs "in flight" or under time pressure, and thus it will not function as a relatively reliable and valid measure of non-BIRR-like responding.

Over the past five or six years the IRAP has been developed, refined and used successfully to assess relational responding across a range of psychological domains, particularly in the area of clinical psychology (Vahey, Nicholson, & Barnes-Holmes, 2015; see also Golijani-Moghaddam, Hart & Dawson, 2013). Although the findings emerging from the use of the IRAP are indeed encouraging it is important to recognise that many features or properties of the measure do need to be subjected to careful systematic empirical analysis. Even seemingly obvious or simple questions remain unanswered or unexplored and the study presented in the current article constitutes an effort to begin and stimulate this programme of research.

At this stage, it is worth noting that the early IRAP studies often involved using contextual cues for specific relations as the response options. As noted above, for example, Barnes-Holmes, et al. (2008) employed the words "Similar" and "Opposite". According to RFT words such as these may be defined as Crels. A large number of IRAP studies, however,

have also employed other types of response options, the two most common being “True” and “False” (e.g., Kosnes, Whelan, O’Donovan, & McHugh, 2013; Nicholson, McCourt, & Barnes-Holmes, 2013). Thus a participant would be required to choose “True” when “Pleasant” and “Love” were presented on consistent blocks but “False” on inconsistent blocks. The implicit but untested assumption here was that the nature of the response options would have little if any impact on the IRAP effects that emerged. Perhaps somewhat surprisingly, this basic assumption has never been tested. The current study was designed to address this gap in the literature. Participants were exposed to two IRAPs, using stimuli similar to those employed by Barnes-Holmes, et al (2008). One IRAP employed the response options “Similar” and “Different” and the other IRAP employed the response options “True” and “False”. The order in which the two IRAPs were presented was counterbalanced across participants. Given that the research was largely exploratory no specific predictions were made.

Before continuing, it is important to note that although the research was exploratory it did not lack a theoretical basis. As mentioned above, the response options “Similar” and “Different” may be considered Crels by RFT. However, the terms “True” and “False” would not typically be defined as such. Rather, such terms are often defined as indicating or referring to relational coherence (see Hayes & Barnes-Holmes, 1997). According to RFT, a pattern of relational responding may be deemed as coherent when it “makes sense” or is deemed to be true in some way. For example, the statement “mice are smaller than elephants” coheres with wider patterns of relational responding in natural language and is thus deemed to be a true statement; the opposite (elephants are smaller than mice) does not cohere with natural language practices and would thus be considered a false statement. Terms such as “true” and “false“, therefore, may be defined as relational coherence indicators or RCIs and should be distinguished from Crels, such as “Similar” and “Different”. In the current article

we will make a clear distinction between Crels and RCIs. Furthermore, we will suggest, if only tentatively, that the former may be seen as involving a "lower level" of relational responding relative to the latter. Or to put it another way, once some basic level of Crel control has been established in an individual's verbal repertoire only then is it possible to respond with appropriate RCIs. For example, an individual can only determine if a relation is true or false if the relation itself has first been established (e.g., the truth or falsity of the relation "X is the same as Y" can only be determined if the Crel "same" possesses the appropriate relational functions for that individual). The current study therefore provided us with an opportunity to determine if the distinction we are making between Crels and RCIs would be reflected or captured in some way by the performances that emerged across the two IRAPs to which the participants were exposed.

Method

Participants

Fifty-two undergraduate students attending Maynooth University completed the study. They were aged 17 to 42 years (M = 21.9), were native English speakers, and had normal or corrected-to-normal vision. Having obtained informed consent, participants were randomly allocated to one of four experimental conditions (described subsequently). No financial payment or other inducements were offered for participation in the study. The research was conducted in accordance with the ethical guidelines of the Department of Psychology at Maynooth University.

Materials and Stimuli

The IRAPs were delivered via a computer program, which controlled the presentation of stimuli and recorded all responses. On each trial, one of two label stimuli ("Pleasant" or "Unpleasant") and a single positively or negatively valenced target stimulus was presented on screen. Two response options were presented in the lower left- and right-hand corners of the

screen. For one IRAP the response options were Crels (“Similar” and “Different”) and for the other IRAP they were RCIs (“True” and “False”); hereafter the former IRAP will be referred to as the SD-IRAP and the latter as the TF-IRAP. The target stimuli consisted of synonyms for “Pleasant” and “Unpleasant”. The six synonyms for “Pleasant” were “Good,” “Positive,” “Nice,” “Likeable,” “Lovely” and “Wonderful”; the six synonyms for “Unpleasant” were “Bad,” “Negative,” “Nasty,” “Unlikeable,” “Horrible” and “Awful”. The target stimuli were independently rated along a 7-point scale by a random sample of thirty students from Maynooth University, with a score of 1 representing “Very Unpleasant” and 7 representing “Very Pleasant.” The positively valenced target stimuli were rated with a mean score of 6, and the negatively valenced target stimuli were rated with a mean score of 2. All participants rated each of the positively valenced words as more pleasant or less unpleasant than each of the negatively valenced words.

Procedure

The experiment consisted of two phases. In each phase, the participant was required to complete a single IRAP. The two IRAPs were similar, except for the two response options that were employed (i.e., “Similar” and “Different” for one IRAP and “True” and “False” for the other IRAP). The order in which the two different IRAPs were presented was counterbalanced across participants. Each participant completed both IRAPs on an individual basis in a small quiet room with the door closed. In all cases, the researcher interacted directly with participants only during instructional phases of the task; during these phases a set of general instructions was presented, which described the task and how to complete it.

Each IRAP consisted of a minimum of eight blocks of 24 trials, with a minimum of two practice blocks followed by a fixed set of six test blocks. Consistent with common practice in IRAP studies, progression from the practice to the test blocks required each participant to reach two pre-determined performance criteria (>= 80% correct responses with

a median response latency <=2000ms). If participants failed to meet the performance criteria on one or both of the first pair of practice blocks a message appeared stating that they had failed to reach the criteria and invited them to try again; the message also reassured participants that it was a difficult task and normally required considerable practice to master.

Up to a total of four pairs of practice blocks were presented in this manner. If the performance criteria were achieved after any of these pairs of practice blocks the program progressed immediately to the fixed set of six test blocks. Specific performance criteria were not required to progress across the test blocks but performance feedback (see below) was presented after each block to encourage participants to maintain the accuracy and latency criteria. If participants failed to reach the performance criteria across all four pairs of practice blocks they were thanked and invited to return on another day for a second attempt to complete the practice phase. Twelve participants returned for a second attempt and completed the study successfully.

On each trial of the IRAP, four stimulus words appeared on screen simultaneously; a label stimulus ("Pleasant" or "Unpleasant"), a target stimulus (e.g., "Positive," "Negative," etc.) and the two response options ("Similar" and "Different" or "True" and "False"). Within each block, the label and target stimuli were presented quasi-randomly across trials, with the constraint that each label stimulus appeared once with each of the six target words across the 24 trials. This 2x2 cross-over of label with target stimuli thus yields four IRAP trial-types, which may be denoted in the current study as; (i) Pleasant-Positive, (ii) Pleasant-Negative, (iii) Unpleasant-Positive, and (iv) Unpleasant-Negative.

The left-right positions of the two response options alternated quasi-randomly across trials, with the constraint that they did not appear in the same positions across more than three consecutive trials. The phrases "PRESS 'd' FOR" and "PRESS 'k' FOR" appeared directly above the two response options. Thus, on one trial, a participant might be required to press

the “d” key for “Similar” (or “True”) and “k” for “Different” (or “False”) but on another trial the reverse would apply (“d” for “Different”/”False” and “k” for “Similar”/”True”).

Consistent with previous IRAP research, each block of trials (both practice and test) required a pattern of responding that was deemed either consistent or inconsistent with the natural verbal relations found within the relevant language community in which the participants resided (in this case a predominately English speaking community). Blocks of trials that were deemed consistent with natural verbal relations thus required the following response pattern across the four trial-types for each IRAP; (i) Pleasant-Positive-*Similar/True;* (ii) Pleasant-Negative-*Different/False;* (iii) Unpleasant-Positive-*Different/False;* (iv) Unpleasant-Negative-*Similar/True.* The response pattern required for blocks of trials deemed inconsistent with natural verbal relations was orthogonal to the consistent pattern; (i) Pleasant-Positive-*Different/False;* (ii) Pleasant-Negative-*Similar/True;* (iii) Unpleasant-Positive-*Similar/True;* (iv) Unpleasant-Negative-*Different/False.*

The blocks of IRAP trials were presented in two sequences. In one sequence all of the odd-numbered blocks (1, 3, 5, etc.) required a pattern of responding that was consistent with natural verbal relations and all even-numbered blocks (2, 4, 6, etc.) required the orthogonal (inconsistent) pattern (hereafter, these two patterns will simply be denoted “consistent” versus “inconsistent”). For the other block sequence, all odd-numbered blocks required the inconsistent response pattern and all even-numbered blocks required the consistent pattern. These two block sequences were counterbalanced across participants for both IRAPs. Thus half of the participants commenced both IRAPs with a consistent block and the other half commenced with an inconsistent block. With the manipulation of the Crel versus RCI response options, there were four separate groups: SD-IRAP-first/consistent-block-first; SD-IRAP-first/inconsistent-block-first; TF-IRAP-first/consistent-block-first; TF-IRAP-first/inconsistent-block-first.

On each IRAP trial, participants were required to select one of the two response options by pressing the appropriate response key (‘d’ or ‘k’); all other computer keys were disabled. If participants chose the response option deemed correct for that block of trials all stimuli were removed from the screen for an interval of 400-ms before the next trial was presented. If participants chose the response option deemed “incorrect” for that block of trials a red “X” appeared mid-screen directly below the target stimulus, and remained there until the correct response option was chosen. When the correct response was emitted, the program progressed to the 400-ms interval followed by the next trial. If a participant failed to respond within 2000ms on any trial, an exclamation mark appeared in red towards the bottom centre of the screen. The exclamation mark remained there until the participant emitted a correct or incorrect response.

Brief on-screen instructions were presented by the IRAP programs before each block of trials. These instructions informed the participant that the upcoming block of trials was a either a practice or test block. For practice blocks, the instructions stated that participants were to “Try to avoid the red ‘X’ on every question.” For test blocks, this was changed to “Please try to get as many right as possible.” A rule relating to the subsequent block of trials was also presented on screen. For consistent blocks this rule read “Pleasant words are positive. Unpleasant words are negative”; for inconsistent blocks, the rule altered to “Pleasant words are negative. Unpleasant words are positive.” Participants pressed the space bar to initiate each block of trials. Feedback was presented on screen following each block of practice and test trials. This feedback detailed the median response latency for that block and the percentage of correct responses. Upon completion of all six test blocks, a message appeared on-screen requesting that the participant report to the researcher.

Results

Data Preparation

The primary datum from the IRAP was response latency, defined as the time in *ms* that elapsed between the onset of a trial and the input of a correct response by the participant. If participants failed to maintain the accuracy (=> 80%) and/or latency criteria (<= 2000 ms) in any of the six test blocks their entire data set was removed from subsequent analyses. The response latency data for each participant were transformed by the IRAP program into normalised indices of response latency differences between consistent and inconsistent blocks of the IRAP trials, yielding *D*-IRAP scores for each of the four trial-types (see Barnes-Holmes, et al., 2010, for a detailed description of this data transformation process). Following data transformation, positive *D*-IRAP values indicated that responses, on average, were faster during blocks of trials that required responding in a manner that was consistent (e.g., Pleasant-Love-True) rather than inconsistent (e.g., Pleasant-Love-False) with natural verbal relations. Negative *D*-IRAP values indicated the opposite response pattern (i.e., more rapid responding on inconsistent than consistent blocks).

Mean Analyses

The mean *D*-IRAP scores for each of the four trial-types for each IRAP were entered into a 4x2x2x2 mixed repeated measures analysis of variance (ANOVA). Trial-type (*Pleasant-Positive; Pleasant-Negative; Unpleasant-Positive; Unpleasant-Negative*) and response-options (True/False versus Similar/Different) were entered as within-participant variables. Block-order (consistent-first versus inconsistent-first) and response-option-order (True/False-first versus Similar-Different-first) were entered as between-participant variables. Three significant effects emerged from the analyses, a main effect for trial-type, $F(3, 144) = 25.85$, $p < .0001$, $\eta p2 = .35$, and two interaction effects; a two-way interaction between response-options and response-option-order $F(1, 48) = 4.68$, $p = .03$, $\eta p2 = .09$, and a three-way interaction between IRAP trial type, response-option-order and block-order: $F(3, 144) =$

2.94, $p = .03$, $\eta p2 = .06$. The two interaction effects were explored separately with a series of follow-up analyses.

The two-way interaction: Response-option by response-option-order. The mean *D*-IRAP scores from the two-way interaction are presented in Table 1. Descriptively, the impact of the two different response options was minimal across the first exposure to the IRAP. The second exposure led to a reduction in the *D*-IRAP scores for both response option conditions, but the reduction was considerably more pronounced for the participants who completed the first IRAP using Crels (Similar/Different) rather than RCIs (True/False). An independent *t*-test confirmed that the two groups did not differ significantly across their first exposures to the IRAP, $t = .11$, $n = 24$, $p = .91$, and a second indicated that the difference remained insignificant for the second exposure, $t = -.81$, $n = 24$, $p = .42$. Critically, however, two paired *t*-tests indicated that the difference in the overall *D*-IRAP scores between the first and second exposures was significant for the group who completed the first IRAP using Crels, $t = -2.15$, $n = 25$, $p = .04$, but was not significant for the group who completed the first IRAP using RCIs, $t = 1.03$, $n = 25$, $p = .31$. Four one-sample *t*-tests indicated that three of the *D*-IRAP effects were significantly different from zero: True/False-First, $t = 4.66$, $n = 25$, $p < .0001$; Similar/Different-First, $t = 4.65$, $n = 25$, $p < .0001$; Similar/Different-Second, $t = 2.45$, $n = 25$, $p = .02$ (remaining $p > .1$). The inferential statistics therefore confirmed the descriptive analyses by indicating that the reduction in the overall *D*-IRAP effect was significant when shifting from Crel to RCI response options, and the effect did not remain significantly different from zero during the second exposure. In contrast, shifting from RCI to Crel response options did not produce a significant reduction in the *D*-IRAP effect and it remained significantly different from zero during the second exposure.

Table 1

Overall mean D-IRAP effects for each IRAP exposure divided according to Crel versus RCI response options (standard errors appear in parentheses)

Response Options	First IRAP	Second IRAP
True/False	.22 (.04)*	.09 (.04)
Similar/Different	.21 (.04)*	.15 (.05)*

*$p < .05$

Three way interaction: Trial type by block-order by response-option-order

The significant three way interaction between IRAP trial type, block-order and response-option-order was explored by conducting four separate between groups 2x2 ANOVAs (i.e., one ANOVA for each trial-type) with the IRAP scores collapsed across the Crel and RCI response options. Only the ANOVA for the *Unpleasant-Positive* trial-type yielded a (marginally) significant effect, indicating an interaction between response-option-order and block-order, $F(1, 48) = 3.48$, $p = .07$, $\eta 2 = .07$ (all *ps* > .12 for the remaining three ANOVAs). The nature of the interaction is presented in Table 2, which shows that the group who were presented with the *Crel* response options in their first IRAP were not influenced by the order in which the blocks were presented. In contrast, the group who were presented with the RCI response options first produced a negative IRAP effect in the consistent-first condition and a positive effect in the inconsistent-first condition. Follow-up *t*-tests indicated that the groups who received the True/False RCI first differed significantly across the block-order conditions, $t = -2.40$, $df = 24$, $p = .02$. The difference between the two response-option conditions for the inconsistent-first block-order condition approached significance, $t = -1.8$, $df = 23$, $p = .09$. The two remaining follow-up *t*-tests were non-significant (*ps* > .3). Four one-sample *t*-tests indicated that the effect for the group who commenced with the RCI response options and an inconsistent block was significantly different from zero, $t = 2.22$, $df = 12$, p

< .05 (remaining *ps* > .2). Overall, therefore, the inferential statistics supported the pattern of differences presented in Table 2.

Table 2

Mean D-IRAP effects for the Unpleasant-Positive trial-type divided according to the order in which the Crel versus RCI response options were presented and the block order variable (standard errors appear in parentheses)

	Response Option Order	
Block Order	Similar/Different First	True/False First
Consistent First	-.01 (.08)	-.13 (.10)
Inconsistent First	-.03 (.08)	.16 (.07)

Discussion

The results of this preliminary and exploratory study revealed that the type of response options that are inserted into an IRAP may impact upon the direction and strength of the effects produced by the measure. Specifically, a two-way interaction indicated that when participants first completed an IRAP using the Crels "Similar" and "Different", and then completed a second IRAP using the RCIs "True" and "False", there was a significant reduction in the effect and it also became non-significant (from zero); this was not the case for the participants who completed the RCI-IRAP first and the Crel-IRAP second. Thus, although the two types of response options yielded similar effects for the participants' first exposures to the IRAPs, differences emerged during the second exposures. This finding clearly indicates that Crels versus RCIs within an IRAP should not be considered functionally equivalent, and as such the use of different types of response options within the IRAP and its derivatives requires careful and systematic analysis.

The impact of the response options on the IRAP performances was further complicated by a three-way interaction with trial-type, block-order and the order in which the response options were presented. Follow-up analyses suggested that one particular trial-type

was driving this effect. Specifically, for the *Unpleasant-Positive* trial-type the IRAP effects did not differ between the two block-order conditions for participants who started with the Crel-IRAP but there was a large divergence between the effects for the participants who started with the RCI-IRAP. Once again, therefore, the results of the current study underscore the need to be cautious in assuming that apparently trivial differences across IRAPs make little if any difference in the effects produced by the measure. Clearly, the type of response option employed does make a difference and its impact may be quite complex in that it interacts with other features of the IRAP itself. But how might we interpret or explain the effects observed in the current study?

The two-way interaction effect may be relatively straightforward to interpret. Towards the end of the introduction we suggested that Crel control may be seen as involving a "lower level" of relational responding than that involved in RCI control. In other words, the truth or falsity of a relation can only be determined if the relation itself has first been established in an individual's verbal repertoire. If we accept this conceptual analysis, then we might explain the two-way interaction effect as follows. Imagine that a participant was asked to complete the Crel-IRAP first. It seems likely that the participant will learn to identify the relation between the label and target stimulus on each trial and respond accordingly (choosing "Similar" or "Different" depending on the label-target relation). If the participant is then presented with the RCI-IRAP, it seems likely that she would continue to respond to the relations between the label and target stimuli as similar or different on each trial and then respond to that relational response itself as true or false. In this sense, shifting from the Crel- to the RCI-IRAP involved a "step up" in complexity from one relational response per trial to two responses. Given the highly time-constrained context of the IRAP this increase in relational complexity may have reduced the extent to which responding on the second IRAP could be seen as involving BIRRs (brief and immediate relational responses).

Imagine, however, if a participant was asked to complete the RCI-IRAP first before moving to the Crel-IRAP. In this case, it seems likely that she would learn to respond *directly* to the relationship between label and target stimuli as true or false, rather than first responding with "similar" or "different" and then determining whether the relation was "true" or "false." If the participant was then asked to complete the Crel-IRAP, as suggested above the task requires a lower level of relational complexity and thus responding at a higher level first would seem largely unnecessary. Imagine, for example, that the label-target pair *Pleasant-Love* was presented on a specific trial with "Similar" and "Different" as response options. Once the relation is identified as *similar* there would little if any motivation to then produce the additional RCI "True" before pressing the "Similar" response key. In this sense, therefore, the shift from RCI- to Crel-IRAP does not involve any increase in relational complexity (only one relational response is required on each trial across both IRAPs), and as such the BIRR-like property of the relational responding may remain in-tact from one IRAP to the next.

But how might we explain the three-way interaction, which appeared to be driven by one particular trial-type (*Unpleasant-Positive*). One possible explanation involves recognising a potentially important role for the rules or instructions that were presented to participants before each test block of the IRAP. Before each block of consistent trials (i.e., consistent with natural language) participants were instructed "Pleasant is positive and unpleasant is negative" whereas before each block of inconsistent trials the instruction read "Pleasant is negative and unpleasant is positive." Critically, there was no counterbalancing of the two elements of each instruction across participants. That is, the first part of the instruction always specified whether "pleasant" was "positive" or "negative" and the second part whether "unpleasant" was "negative" or "positive." Furthermore, the two instructions may be seen as coordinating with the four trial-types as follows: The first part of the

consistent instruction coordinated with the *Pleasant-Positive* trial-type, with the second part coordinating with the *Unpleasant-Negative* trial-type; the first part of the inconsistent instruction coordinated with the *Pleasant-Negative* trial-type, with the second part coordinating with the *Unpleasant-Positive* trial-type. Recent findings from our research group have suggested that instructional variables such as these may impact quite dramatically upon the size and direction of effects that emerge from IRAP performances (Finn, Barnes-Holmes, & Hussey, 2015). That is, larger history-consistent IRAP effects may be observed for those trial-types that cohere with natural language and coordinate with the first part of the instruction or rule that is presented to participants before each block of the IRAP.

In the case of the current research it may be that such instructional effects also interact with the type of response options that are employed (i.e., Crels versus RCIs). The critical point here is that the trial-type that appeared to drive the three-way interaction was the trial-type that "coordinated" with the second part of the instruction that did not cohere with natural language (i.e., "Unpleasant is positive"). Or to put it another way, the trial-type that was least well instructed in terms of both natural-language coherence and instructional sequence appeared to be impacted most by the use of Crels versus RCIs as response options. In terms of the REC model, it may be that the relational responding that occurred with this trial-type was the least BIRR-like of the three and thus was most susceptible to the impact of other moderating variables.

Of course, the foregoing explanation is highly speculative and must remain so until the interacting effects of IRAP trial-types, instructional variables, response-options, and IRAP block-order effects are analyzed systematically. Given the complexity and dynamic nature of the variables involved, this work will be time consuming and difficult to conduct but we will only gain a sophisticated level of understanding of how the IRAP and its derivatives work from conducting the necessary experimental analyses. Although certainly

preliminary and largely exploratory it is our hope that the current findings will serve to motivate other researchers to begin to investigate the complex behavioral dynamics that underlie the so called IRAP effect.

Compliance with Ethical Standards

Conflict of Interest: Emma Maloney declares that she has no conflict of interest. Dermot Barnes-Holmes declares that he has no conflict of interest.

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

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

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.

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,* 497–515.

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,* 163-184.

Finn, M., Barnes-Holmes, D. & Hussey, I. (2015). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: The impact of detailed, minimal, and response-option-focused rules. *Manuscript submitted for publication.*

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2,* 105-119.

Hayes, S. C. & Barnes, D. (1997). Analysing derived stimulus relations requires more than the concept of stimulus class. *Journal of Experimental Analysis of Behavior. 68,* 235-244.

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

Hughes, S., & Barnes-Holmes, D. (in press-a). Relational frame theory: The basic account. In R. Zettle, S. C. Hayes, D. Barnes-Holmes, & T. Biglan (Eds.), *Handbook of Contextual Behavioral Science.* New York: Wiley-Blackwell.

Hughes, S., & Barnes-Holmes, D. (in press-b). Relational frame theory: Implications for the study of human language and cognition. In R. Zettle, S. C. Hayes, D. Barnes-Holmes, & T. Biglan (Eds.), *Handbook of Contextual Behavioral Science.* New York: Wiley-Blackwell.

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

Kosnes, L., Whelan, R., O'Donovan, A. & McHugh, L. A. (2012). Implicit measurement of positive and negative future thinking as a predictor of depressive symptoms and hopelessness. *Consciousness and Cognition, 22,* 898-912.

Nicholson, E., McCourt, A. & Barnes-Holmes, D. (2013). The Implicit Relational Assessment Procedure (IRAP) as a measure of obsessive beliefs in relation to disgust. *Journal of Contextual Behavioral Science, 2,* 23-30.

O'Hora, D., Barnes-Holmes, D., Roche, B., & Smeets, P. M. (2004). Derived relational networks and control by novel instructions: A possible model of generative verbal responding. *The Psychological Record, 54,* 437-460.

Steele, D. & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56,* 519-555.

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.