The IRAP as a Measure of Implicit Cognition: A Case of Frankenstein's Monster

The IRAP as a measure of implicit cognition: A case of Frankenstein's monster

Dermot Barnes-Holmes[1] & Colin Harte[2, 3]

[1]School of Psychology, Ulster University, Northern Ireland

[2]Departmento de Psicologia, Universidade Federal de São Carlos, Brazil

[3]Paradigma – Centro de Ciências e Technologia do Comportamento, Brazil

Acknowledgements: Preparation of this manuscript was supported by a research fellowship awarded to the second author by the São Paulo Research Foundation (FAPESP; Grant #2019/24210-0).

Abstract

The implicit relational assessment procedure (IRAP) was initially developed as a way to assess the strength and probability of natural verbal relations, as defined within relational frame theory (RFT), and was conceptually rooted within the behavior-analytic tradition. However, the IRAP quickly became employed primarily as a measure of implicit cognition, more in line with mainstream psychology than behavior analysis. In doing so, research using the IRAP increasingly relied on group designs, employed ill-defined mainstream psychological terms, and focused on correlational analyses with traditional psychometry. While perhaps beneficial to the study of implicit cognition, this approach could be argued to have limited the IRAP's utility in behavior analyses of human language and cognition. In the current article we will reflect on this suggestion, on the IRAPs place and current use in the field of behavior analysis, and on its potential future within behavioral psychology in light of recent conceptual and empirical advances in RFT. In doing so, it is hoped that the measure may be refined into a better understood, more precise, functional-analytic tool.

KEYWORDS: IRAP; RFT; VERBAL RELATIONS; FUNCTIONAL ANALYTIC; IMPLICIT COGNITION

In Mary Shelley's classic novel, 'Frankenstein' (1818), we are presented with the case of a doctor who creates a living monster by successfully piecing together and reanimating body parts from different people. However, not long after the monster has been brought to life does he become Dr. Frankenstein's nemesis and ultimately leads to their joint demise. In one sense, this tale seems like an appropriate metaphor for the creation of the implicit relational assessment procedure (IRAP), a behavioral measure brought to life by piecing together parts of different tools such as the implicit association test (IAT; Greenwald, et al., 1998) and the relation evaluation procedure (see Barnes-Holmes, et al. 2010). The original purpose of the IRAP was to assess the strength and probability of natural verbal relations, as defined within relational frame theory (RFT; Hayes, et al., 2001), and was conceptually rooted firmly within the behavior-analytic tradition. However, as was the case with Dr. Frankenstein's monster, the creator of the IRAP seemingly lost control of his creation as the procedure became almost exclusively employed as a measure of implicit cognition. In the current article we will reflect on the history of the IRAP, its place and use in the field currently, and its potential future within behavior analysis (and behavior science more generally) in light of recent conceptual and empirical developments in RFT (see Barnes-Holmes et al., 2020, 2021; Barnes-Holmes & Harte, 2022). In doing so, we hope that this story will not end in the same way that Shelley's did. Rather we hope that the IRAP, unlike Frankenstein's monster, will be tamed and refined into a better understood, more precise, functional-analytic tool.

Before continuing, we understand that some of our colleagues who have contributed to the IRAP literature, may be somewhat alarmed by what we are presenting here. We should be clear that we are not denying there may be considerable value in what the IRAP has contributed toward the study of implicit cognition. However, in the service of intellectual honesty, we believe it is important to 'call it how we see it', so to speak, and argue that as a measure for advancing a behavior-analytic account of human language and cognition, the

IRAP has until relatively recently fallen short. Thus, the current article simply aims to highlight this argument and present an example of what the IRAP may contribute as a tool for conducting experimental analyses of behavior in the domain of human language and cognition. To quote Shelley, “learn from me, if not by my precepts, at least by my example” (p. 58).

The IRAP as a Mainstream Measure

Numerous methods for assessing so-called implicit attitudes have been developed over the years such as the IAT, the Go/No-Go Association Task (GNAT), evaluative priming and the Extrinsic Affective Simon Task (EAST). Critically, however, each of these methods may be considered relatively indirect measures because they target associations (in memory) rather than verbal relations. Associations are considered to be bi-directional activations (e.g, flower-good, insect-bad), but attitudes and beliefs are seen as involving verbal relations (e.g, flower better than insect, insect worse than flower) or indeed several relations (e.g., insect if not a butterfly is bad). In contrast to the IAT and other associative measures, the IRAP was designed specifically to assess natural verbal relations.

Associative measures typically involve participants responding to pairs of associated stimuli in some way (e.g., pressing a specific key when “flower” and “good” appear together). In contrast, on each trial of the IRAP participants are typically asked to confirm or deny a specific verbal relation (or set of relations) between two label and target stimuli within a short response window. That is, rather than simply associating stimuli, participants must confirm or disconfirm the truth value of a specific verbal relation or proposition (see Hughes et al., 2011, for a detailed argument). Corrective feedback is also presented on the IRAP which, in general, is implemented in a manner consistent with participants' pre-experimental learning histories for half of the blocks of trials and inconsistent with this history for the other half. For example, an IRAP assessing verbal relations in the context of race might present a picture of a person at the top of the screen (e.g., a picture of a white person or black person) as a label stimulus, a word or phrase in the centre of the screen (e.g., 'safe' or 'dangerous') as a target stimulus, and the response options 'True' and 'False' on the bottom of the screen. Half of the trials would require participants to respond 'True' when presented with the picture of the white person and the word 'safe' and a black person and the word 'dangerous,' while the other half would require opposite responding (e.g., 'False' in the presence of the white person and 'safe' and 'True' in the presence of the black person and 'dangerous'). The basic logic is that participants should tend to respond more quickly to relations that are consistent versus inconsistent with their pre-experimental learning histories. The difference between history consistent and history inconsistent responding is referred to as the IRAP effect and is used as a measure of the strength of this relational responding.

The IRAP's focus on confirming versus disconfirming specific truth values rendered it, at least in principle, a method for assessing (the strength of) natural verbal relations rather than as a test of so-called implicit cognition (Barnes-Holmes et al., 2008). Despite this conceptual starting point, however, the descriptor "implicit" was nonetheless added to the name for a number of reasons: 1. The IAT was a source of inspiration for the IRAP; 2. There appeared to be some potential for the IRAP to function as a test of so-called implicit cognition; 3. The name "I-rap" was quite catchy and reflected what the test required -- rapid verbal responding. And as we now know, although the IRAP did not start out as a measure of implicit cognition, it did seem to have relative success in becoming one (Vahey, et al., 2015). From a behavior-analytic perspective, however, combining an ill-defined domain (i.e., implicit cognition) with a "measure" (the IRAP) that was not well understood in a functional-analytic abstractive manner, would lead us down an intellectual blind alley. We elaborate on this argument below.

Over the years, IRAP studies increasingly relied upon group designs and employed ill-defined mainstream psychological terms (e.g., attitudes, self-esteem, prejudice). Furthermore, the validity of the measure was largely assessed using traditional nomothetic mainstream methods (e.g., determining if the measure predicted or correlated with other implicit or psychometric instruments; see Hofmann et al., 2021, for a recent challenge to the use of traditional nomothetic psychometry in psychological science). This "mainstream" method of employing the IRAP quickly became its most dominant application, with a meta-analysis conducted in 2015 (Vahey et al.) revealing that it compared favourably with a range of other mainstream measures of implicit cognition (in the clinical domain). Although the IRAP seemed to be performing well in this regard, the fact remained that using the instrument in this classic mainstream psychological manner failed to provide a functional-analytic abstractive account of the behaviors produced by the IRAP itself. As such, the research achieved little in terms of meeting the analytic goals of a behavior-analytic science – the prediction and influence of behavior (with precision, scope and depth) that its initial conceptualisation strived to achieve. Thus, although the IRAP was masquerading, to some extent, as a behavior-analytic tool, it was not being used as such.

The IRAP as a Behavior-Analytic Tool: A Brief Review of Some Relevant Research.

Exploring the functional properties of the IRAP: Identifying "contaminating" variables. We have argued that the IRAP became widely used as a type of mainstream measure of implicit cognition, but it is also important to describe how it may be used in ways that speak more directly to a behavior-analytic research agenda. Interestingly, the potential dangers of relying on the IRAP as a measure of implicit cognition, rather than a functional-analytic one, were identified over 10 years ago. Specifically, Barnes-Holmes et al. (2010) highlighted the potential sensitivity of the IRAP to verbal relations that extended beyond those that were being targeted in a typical study of implicit cognition. Specifically, the authors warned that procedural variables, such as a tendency to respond quicker with 'True' than 'False', may interact with the assessment of implicit attitudes or biases:

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

This sensitivity to non-targeted verbal relations presented a clear problem when using the IRAP to assess implicit cognitive processes, independent of what may be seen as spurious variables (O'Shea, et al., 2015). On the other hand, treating the IRAP as a context for exploring the functional-analytic properties of verbal relations renders this "problem" largely irrelevant (i.e., because an IRAP performance is not seen as a proxy for underlying mental events).

The foregoing argument is perhaps best illustrated in a study by Maloney and Barnes-Holmes (2016), who found that IRAP effects were differentially impacted by the type of response options employed. Specifically, the study involved using the response options, "Same" and "Different" (defined as Crels, or relational cues, within RFT) in one IRAP and "True" and "False" (defined as relational coherence indicators, RCIs) in another. Interestingly, the sequence in which these two types of response options were used across the two IRAPs appeared to produce significantly different response patterns. As such, this study served to support the point made five years earlier by Barnes-Holmes and colleagues (2010) that the response options 'True' and 'False' may indeed impact upon performance on the IRAP in perhaps subtle and complex ways. In light of this and other findings (covered below) it became apparent that a more functional-analytic approach to the IRAP was required.

Not long thereafter, further studies began to highlight other variables to which the IRAP seemed sensitive. For example, Finn et al. (2016) found that the type of instructions provided to participants about how to complete the IRAP differentially impacted the strength and direction of the effects produced. Specifically, participants in this study were presented with an IRAP that presented names of colors (e.g., 'red', 'blue') and shapes (e.g., 'square', 'circle') as label and target stimuli and were provided with rules that varied in terms of the level of detail pertaining to the relational network being assessed. Specifically, some participants received rules that were quite detailed (e.g., "Respond as if shapes are shapes and colors are colors"), others received rules that were more general but specified pre-experimentally established relations (e.g., "Respond correctly to the stimuli"), while others received a rule that was also general but did not specify pre-experimentally established relations (e.g., "Please respond as if true is consistent and false is inconsistent").

The results demonstrated that the level of detail presented in the rule dramatically impacted upon the size and direction of the IRAP effect produced, and that this effect may be in part moderated by the order in which the blocks were presented (i.e, history-consistent blocks presented first versus history-inconsistent presented first). Related work by Finn et al. (2018) subsequently found that the amount of past experience participants had with completing latency-based measures in general (e.g., the IAT, priming tasks, stroop tests) also dramatically impacted performance on the IRAP. That is, the more experience participants had with completing latency based measures, the larger the IRAP effects. Interestingly, the authors also reported that implementing a read-aloud procedure seemed to reduce this difference (i.e., in the read-aloud condition participants were asked to report out loud the on-screen IRAP stimuli and the emitted response on each trial). Similar findings for the read-aloud procedure were reported by Kavanagh et al. (2018) but in the context of deictic (i.e., self-other) relations.

Exploring the functional properties of the IRAP: Differential trial-type effects. Apart from identifying "contaminating" variables involved in IRAP performances, a specific pattern of effects was increasingly being observed on the IRAP in many studies conducted across a range of domains. One such effect showed a response pattern in which one of the IRAP trial-types was consistently larger than the other three. This effect was referred to as the single trial-type dominance effect (STTDE; Finn et al., 2018). Critically, this pattern seemed difficult to explain in terms of the response options alone because two of the trial-types involve selecting the same option within each block. Consider, for example, the shapes and colors IRAP described above. In this IRAP participants must choose "True" on both color-color and shape-shape trials during history-consistent blocks (False must be chosen on history-inconsistent blocks). However, the trial-type effect for the color-color trial-type has been found to be significantly larger than the effect for the shape-shape trial-type (Finn, et al., 2018). Given this finding, and other similar results, a new model was proposed; the Differential Arbitrarily Applicable Relational Responding Effects (DAARRE; Finn et al., 2018) model. According to the model, it is important to distinguish between the relational (Crel) and functional (Cfunc) properties of stimuli. For RFT, the Crel property refers to the symbolic (arbitrarily applicable relational) functions, or more informally the "cold" semantic meaning of a stimulus (e.g., the word "spider" is semantically related to a range of eight-legged arthropods). The Cfunc property refers to the non-semantic behavioral functions of the stimulus, or more informally its "hot" attentional, emotional, or motivative effects (e.g., the word "spider" may evoke a mild aversive reaction in a listener). Within RFT, it has been argued that these two classes of functional properties do not "exist" independently but serve to co-define each other (see Dymond & Barnes, 1994). The critical point is that the DAARRE model proposes that differential trial-type effects, such as the STTDE, may be explained by the extent to which the Crel and Cfunc properties of the stimuli overlap or cohere with specific properties of the two response options.

For illustrative purposes, consider a study reported by Bortoloti et al. (2019), which found a STTDE using abstract stimuli that had been related via arbitrarily applicable relations to other stimuli with specific Cfunc properties. Specifically, the abstract stimuli had been derived as equivalent to pictures of emotional faces. The pictures of the emotional faces and the abstract stimuli were then inserted into an IRAP. The basic DAARRE model as it applies to this particular IRAP is presented in Figure 1. The model identifies three key sources of behavioral influence: (1) the relationship between the label and target stimuli (Crels); (2) the evoking (or valence) functions of the label and target stimuli (Cfuncs); and (3) the Cfunc properties of the two RCIs (e.g., “True” and “False”). The two critical trial-types in this context were Happy symbol-Happy face and Negative symbol-Negative Face because participants were required to press “True” during symbol-face consistent blocks and “False” during symbol-face inconsistent blocks. As noted above, these trial-types are described as critical here because both trial-types require that participants respond with the same RCI within each block of trials. A tendency to respond “True” more quickly than “False” cannot, therefore, explain any difference in the size of the two trial-type effects. Interestingly, the participants tended to respond “True” more quickly than “False” on the Happy-Happy trial-type, but in the Negative-Negative trial-type they responded "True" and "False" with almost equal latencies. In effect, a single-trial-type-dominance effect was observed.

INSERT FIGURE 1 HERE

The DAARRE model explains this dominance effect by assuming that the pictures of the happy faces and equivalent symbol, possessed relatively positive evoking functions, whereas the pictures of negative faces and equivalent symbols, possessed relatively negative evoking functions. Specifically, the STTDE is explained by the extent to which the Cfunc and Crel properties cohere with the RCI properties of the response options across blocks of trials. Note that the Cfunc and Crel properties for the happy symbol-happy face trial-type are all labelled with plus signs; in addition, the RCI that is deemed correct for history-consistent trials is also labelled with a plus sign (the only instance of four plus signs in the diagram). According to the model, therefore, this trial-type may be considered as maximally coherent during history-consistent trials. In contrast, during history-inconsistent trials, where participants have to respond "False" for the happy symbol-happy face trial-type, there is no coherence between the required RCI (minus sign) and the properties of the Cfuncs and the Crel (all plus signs). According to the DAARRE model, this stark contrast in levels of coherence across blocks of consistent and inconsistent trials serves to produce a relatively large IRAP effect.

Now consider the Negative symbol-Negative face trial-type, which requires that participants choose the same RCI as the Happy symbol-Happy face trial-type during history-consistent trials, but here the property of the RCI (plus sign) does not cohere with the Cfunc properties of the label and target stimuli (both minus signs). During history-inconsistent trials the RCI does cohere with the Cfunc properties (minus signs) but not with the Crel property (plus sign). Thus, the differences in coherence between history-consistent and history-inconsistent trials across these two trial-types is not equal (i.e., the difference is greater for the Happy symbol-Happy face trial-type) and thus explains, at least in part, the single-trial-type-dominance-effect. What the DAARRE model provides, therefore, is the potential for a relatively precise analysis of the functional properties that are at play when participants are required to complete an IRAP. Of course, the model will likely need to be adapted and refined as new data emerge in years to come. However, thus far other researchers within behavior analysis seem to be having increased success in analysing patterns of IRAP effects in light of the DAARRE model (e.g., Bortoloti et al., 2020; Finn et al., 2019; Pidgeon et al., 2020; Pinto et al., 2020, Schmidt et al., 2021).

Extending the use of the IRAP as a behavior-analytic tool. In the last number of years, researchers have also begun to use the IRAP as a context for both training and testing verbal relations. For example, Leech et al. (2018) employed pictures of spiders and pets to establish fearful and pleasant stimulus functions for arbitrary shapes within a training version of the IRAP. When used as a context for training, IRAP blocks do not alternate between history-consistent and history-inconsistent responding but rather simply require responding in whatever manner is being trained to specific accuracy and latency criteria. In the Leech et al. study, pictures of spiders and pets were trained to pictures of two geometric shapes, a circle and a square (i.e., Spider-Circle-Similar; Spider-Square-Different; Pet-Square-Similar; Pet-Circle-Different). After training these mutually entailed relations, researchers then assessed the transformation of fear functions using a 'traditional' format IRAP. Specifically, the circle and square shapes were presented as label stimuli with words as targets that referred to negative (e.g., "I hate it") and positive (e.g., "I like it ") reactions. A transformation of functions was observed for the two geometric shapes, in that participants showed IRAP effects that were consistent with the previously trained relations (e.g., they responded more quickly with Circle-"I hate it"- Yes, than with Circle-"I hate it"- No). In this example, the circle had acquired some of the evoking properties of spiders.

Interestingly, a follow up study by Leech and colleagues (2020) attempted to replicate this result but at the level of combinatorial entailment (by adding a middle node to the trained network; e.g., Spider-Circle-Similar/Circle-Bem-Similar) but failed to find any evidence of the transformation of stimulus functions (Experiment 1). This result suggested a potential boundary condition for using the IRAP in this way. That is, a transformation of functions seemed to occur at the level of mutual entailment (Leech et al., 2018) but not with combinatorially entailed relations. In Experiment 2 of the 2020 study, however, the authors manipulated levels of derivation and coherence of the trained and derived relations. Critically, the results in this case showed that manipulating derivation by providing participants with an opportunity to respond in accordance with the derived relations in the absence of differential feedback (e.g., Spider-Bem-Similar) produced a reliable transformation of functions. Interestingly, increasing coherence (i.e., by increasing the number of trained but not derived relational responses) failed to produce the transformation effects observed in the "derivation" condition. These studies, and indeed others (e.g., Gomes et al., 2019), thus demonstrated the potential utility of the IRAP to train and/or test transformations of stimulus functions while also highlighting the importance of ongoing functional analyses of doing so at different levels of relational complexity (i.e., mutual versus combinatorial entailment).

The impact of levels of derivation and coherence have also been explored, in a series of studies by Harte and colleagues, using the training IRAP in the context of persistent rule-following. The basic preparation involved first training and/or testing novel derived relations using the training IRAP (e.g., train A=B=C, test A=C), and then manipulating the level of derivation and/or coherence of the novel network, similar to the research reported by Leech and colleagues. Subsequently, specific stimuli from the network were inserted into a rule for responding on a matching task. Initially, reinforcement contingencies supported responding in accordance with the rule (involving a derived relation) before subsequently reversing so that rule-consistent responding was now punished. Results of one of these studies (Harte et al., 2018) showed that participants were more likely to continue following the rule in the face of reversed reinforcement contingencies when the IRAP-trained network (or part of it) involved 15 blocks, rather than 1 block, of relational training. The reader should note that although the authors referred to this manipulation as involving derivation, it could also be viewed as involving coherence (see Harte et al., 2021, p. 224). Indeed, in a subsequent study reported by Harte, et al. (2020), the researchers attempted to manipulate coherence directly via the presence versus absence of performance feedback, with the finding that rule-persistence was greater when feedback was provided during the IRAP training and testing.

Follow-up research by Harte, Barnes-Holmes, Barnes-Holmes, et al. (2021) attempted to manipulate both coherence and derivation of the IRAP-trained network. In this study, the researchers made a clearer distinction between coherence and derivation than in the earlier works. That is, while coherence was again manipulated via the presence and absence of performance feedback, the derivation manipulation involved differential opportunities to derive the critical relation that would be inserted into the rule. The results showed that increasing the coherence, by providing differential feedback when "testing" the IRAP-generated network, appeared to reduce the impact of the derivation manipulation on persistent rule-following.

In other related research, Harte, Barnes-Holmes, Moreira, et al. (2021) used a training IRAP to assess the extent to which flexibility in reversing derived relations would subsequently control participant responding on a broadly similar contingency-switching rule-following task. Specifically, researchers first used a training IRAP to establish a relational network involving two combinatorially entailed relations (train A1=B1=C1, test A1=C1 and train A2=B2=C2, test A2=C2). Next, the researchers reversed the B and C relations in the network (A1=B1=C2 and A2=B2=C1) and tested the newly reversed derived relations (e.g., A1=C2 and A2=C1). All participants successfully reversed the derived relations across three experiments, a performance that matching-to-sample (MTS) procedures have previously been limited in producing (e.g., Pilgrim & Galizio, 1995). The authors suggested that the success in doing so with the IRAP versus MTS may be due, in part, to the four trial-type structure of the IRAP, which requires that all relations, confirmatory and disconfirmatory, are explicitly trained. For example, participants might be trained to confirm that A is the same as B, but also disconfirm that A is different to B. With respect to rule-persistence, the results showed that the networks generated by the IRAP training and testing controlled responding on the rule-following task when the task contingencies initially cohered with the network as opposed to when the task contingencies were immediately in opposition with the network. Overall, therefore, these studies suggest that the IRAP, in both traditional and training formats, provides a useful context for manipulating and assessing full relational networks and their impact on numerous domains (e.g., persistent rule-following, fear and avoidance responding).

Summary and conclusion. As has been described above, attempting to analyze the functional properties of the behaviors typically observed on the IRAP has helped to reveal at least some of the important controlling variables involved when individuals respond, quickly and accurately, in accordance with relational networks. Specifically, it now seems important to consider both the Cfunc and Crel properties of the stimuli presented within an IRAP, including the controlling functions of the two response options. That is, individuals are not just sensitive to the relations between label and target stimuli, but also to the functional overlap between the properties of these stimuli and the response options. In addition, this work served to highlight that simply focusing on the effect sizes of individual trial-types, as a measure of so-called implicit cognition, fails to fully reveal the complex and rich behavioral dynamics that may be explored using the IRAP as an experimental tool.

One Example of How the IRAP Might be Used in Future Experimental Analyses

Having considered a range of recent studies that have employed the IRAP as an experimental context for exploring the dynamics of derived relational responding (under time pressure), it seems useful to explain how using the IRAP in this way may contribute towards relatively precise experimental analyses in another research area. Specifically, it has been argued recently (Barnes-Holmes et al., 2021; Harte & Barnes-Holmes, 2021) that the IRAP could be used to provide experimental analogs of the concepts of fusion and defusion, which have been widely employed in the literature on acceptance and commitment therapy (ACT; Hayes, et al., 2012). The term fusion is used to refer to those instances in which individuals find it difficult to distance themselves from their own emotionally-driven psychological content; the term defusion is in a sense the antonym and refers to instances in which individuals can distance themselves from such content. ACT therapists will typically focus on helping clients to defuse from emotionally-driven psychological reactions that are seen as undermining the achievement of valued personal goals.

Obviously, the foregoing definitions of the concepts of fusion and defusion, while useful therapeutically, may not yield readily to relatively precise experimental analyses (Barnes-Holmes et al., 2016). Recently, however, it has been argued that it may be useful to begin to develop experimental analyses of these terms by drawing on the relative dominance of Cfunc versus Crel stimulus properties, as measured with the IRAP (Barnes-Holmes et al., 2021; Harte & Barnes-Holmes, 2021). Allow us to elaborate.

As argued previously, the DAARRE model interpretation of the STTDE is taken to indicate that the Cfunc properties (i.e., orienting and/or evoking) of the stimuli involved are strongly influencing the IRAP performance. Alternatively, a pattern of IRAP effects within which no trial-type dominates significantly over the others (i.e., all four trial-types are more or less equal) may indicate that the Cfunc properties of the stimuli have limited impact (e.g., Finn et al., 2019). In this case, the IRAP performance may be driven largely by responding to the relations between the label and target stimuli, with a limited role for their Cfunc properties. Recently, this distinction in IRAP performances has been described in terms of the relative dominance of the Cfunc versus Crel properties. Critically, some researchers have argued that this relative dominance effect may provide the beginnings of an experimental analog of what ACT researchers and therapists refer to with the terms fusion and defusion. Specifically, when Cfunc properties dominate on the IRAP this suggests that the orienting and/or evoking responses produced by the stimuli are impacting upon the "pure" relational responding; if the stimulus relations (Crels) between the label and target stimuli were dominating the performance then broadly similar effect sizes should be observed across the trial-types (see Figure 2). In other words, when Cfunc properties dominate an IRAP performance it suggests that the individual finds it difficult to respond to the stimuli in a purely relational manner (i.e., is fused with those stimulus properties). When the Crel properties dominate, the opposite is the case in that the individual finds it easier to relate the stimuli without being unduly influenced by their non-relational (i.e., Cfunc) properties (i.e., is defused from those stimulus properties).

INSERT FIGURE 2 HERE

Although highly speculative, these two patterns of responding on the IRAP might help to provide a relatively precise experimental analysis of the distinction between fusion and defusion, and the behavioral processes involved. While no direct empirical evidence is yet available to support or contest this suggestion, a research programme is currently underway in Brazil that aims to empirically test and develop this analysis. In broad terms, this will first involve replicating a STTDE and subsequently assessing whether defusion-based interventions reduce the differential trial-type effect, such that no trial-type significantly dominates over another. In other words, the general strategy will involve first replicating a Cfunc dominated "fused" pattern of responding on the IRAP and demonstrating that responding to the same IRAP following a defusion intervention produces a relatively Crel dominated "defused" pattern. As such, the foregoing project provides just one example of the possible avenues that an increasingly precise, bottom-up, functional-analytic view of IRAP effects can generate in other research areas.

Conclusion

In Shelley's classic novel, the tale ends tragically with the demise of both Dr. Frankenstein and his creation. We hope that this article perhaps represents an alternative ending when it comes to the tale of the IRAP within behavior analysis. While this particular 'monster' may not yet be fully tamed, it does seem like the task is well underway as increasingly sophisticated functional analyses of the patterns of behaviour produced on the measure are being conducted and refined. Analysing the subtleties and dynamics involved in these patterns have provided some critical insights concerning, for example, the complexities involved in responding in accordance with relational networks (under time-pressure), and some of the potential variables involved in training and manipulating networks to assess their impact on other behaviors, such as fear/avoidance responding and persistent rule-following. Approaching the IRAP in this way appears to provide a means of avoiding at least some of the pitfalls encountered when employing the instrument as a simple "mainstream" measure of implicit cognition. Thus, while there is undoubtedly still a long way to go, we hope that the current article has illustrated the power of keeping functional analyses and behavioral processes centre stage in the use of the IRAP. Of course, the benefits and longevity of such analyses will ultimately be decided by the wider research community. However, perhaps this re-focused direction may mark the beginning of a different outcome for the IRAP 'monster'; one in which it can contribute in a meaningful way to the ongoing experimental analysis of human language and cognition within the behavior-analytic tradition.

Statements and Declarations

The authors declare that they have no conflicts of interest.

References

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

Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2020). Updating RFT (more field than frame) and its implications for process-based therapy. *The Psychological Record, 70,* 605-624. https://doi.org/10.1007/s40732-019-00372-3

Barnes-Holmes, D., Barnes-Holmes, Y., McEnteggart, C., & Harte, C. (2021). Back to the future with an up-dated version of RFT: More field than frame? *Perspectivas em Análise do Comportamento, 12*(1). https://doi.org/10.18761/PAC.2021.v12RFT.03

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. https://doi.org/10.1007/BF03395726

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*, 497–516. https://doi.org/10.1007/BF03395634

Barnes-Holmes, D. & Harte, C. (2022). Relational Frame Theory 20 years on: The Odysseus Voyage and Beyond. *Journal of the Experimental Analysis of Behavior.* Advanced Online Publication. https://doi.org/10.1002/jeab.733

Barnes-Holmes, Y. Hussey, I., McEnteggart, C., Barnes-Holmes, D., & Foody, M. (2016). Scientific ambition: The relationship between relational frame theory and middle-level terms in acceptance and commitment therapy. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 365–382). Wiley Blackwell.

Bortoloti, R., de Almeida, R.V., de Almeida, J.H., & de Rose, J.C. (2020). A commentary on the dynamics of arbitrarily applicable relational responding involving positively valenced stimuli and its implications for the IRAP research. *The Psychological Record, 71*, 481-486. https://doi.org/ 10.1007/s40732-020-00413-2

Bortoloti, R., de Almeida, R. V., de Almeida, J. H., & de Rose, J. C. (2019). Emotional faces in symbolic relations: A happiness superiority effect involving the equivalence paradigm. *Frontiers in Psychology, 10,* 1–12. https://doi.org/10.3389/fpsyg.2019.00954.

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. https://doi.org/10.1901/jeab.1994.62-251

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*(2), 309–321. https://doi.org/10.1007/s40732-016-0173-4

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

Finn, M., Barnes-Holmes, D., McEnteggart, C., & Kavanagh, D. (2019). Predicting and influencing the single trial-type dominance effect. *The Psychological Record, 69*(3), 425–435. https://doi.org/10.1007/s40732-019-00347-4

Gomes, C., Perez, W., de Almeida, J., Ribeiro, A., de Rose, J., & Barnes-Holmes, D. (2020). Assessing a derived transformation of functions using the implicit relational assessment procedure under three motivative conditions. *The Psychological Record, 69*, 487–497. https://doi.org/10.1007/s40732-019-00353-6

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480. https://doi.org/10.1037//0022-3514.74.6.1464

Harte, C. & Barnes-Holmes, D. (2021). A primer on relational frame theory (RFT). In M.P. Twohig, M.E. Levin, & J.M. Peterson (Eds.). *The Oxford Handbook of Acceptance and Commitment Therapy.* Oxford University Press. https://doi.org/ 10.1093/oxfordhb/9780197550076.013.4

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2018). The impact of high versus low levels of derivation for mutually and combinatorially entailed relations on persistent rule-following. *Behavioral Processes, 157,* 36-46. https://doi.org/10.1016/j.beproc.2018.08.005

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2021). Exploring the impact of coherence (through the presence versus absence of feedback) and levels of derivation on persistent rule-following. *Learning and Behavior*, *49,* 222-239. https://doi.org/10.3758/s13420-020-00438-1

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., McEnteggart, C., Gys, J., & Hassler, C. (2020). Exploring the potential impact of relational coherence on persistent rule-following: The first study. *Learning and Behavior, 48,* 373-391. https://doi.org/10.3758/s13420-019-00399-0

Harte, C., Barnes-Holmes, D., Moreira, M., de Almeida, J.H., Aparecida-Passarelli, D., & de Rose, J.C. (2021). Exploring a Training IRAP as a single participant context for analyzing reversed derived relations and persistent rule-following. *Journal of the Experimental Analysis of Behavior, 115*(2), 460-480. https://doi.org/10.1002/jeab.671

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

Hayes, S. C., Strosahl, K.D., & Wilson, K.G. (2012). *Acceptance and Commitment Therapy: The process and practice of mindful change* (2nd Edition). Guilford Press.

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

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

Leech, A. & Barnes-Holmes, D. (2020). Training and testing for a transformation of fear and avoidance functions via combinatorial entailment using the implicit relational assessment procedure (IRAP): Further exploratory analyses. *Behavioral Processes, 172,* 104027. https://doi.org/10.1016/j.beproc.2019.104027

Leech, A., Bouyrden, J., Bruijsten, N., Barnes-Holmes, D., & McEnteggart, C. (2018). Training and testing for a transformation of fear and avoidance functions via combinatorial entailment using the implicit relational assessment procedure

(IRAP): The first study. *Behavioral Processes, 157,* 24-35. http://doi.org/10.1016/j.beproc.2018.08.012

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*, 395–403. http://doi.org/10.1007/s40732-016-0180-5

O'Shea, (2015). B., Watson, D.G., & Brown, G.D.A. (2015). Measuring implicit attitudes: A positive framing bias flaw in the implicit relational assessment procedure. *Psychological Assessment, 28*(2), 158-170. https//:doi.org/ 10.1037/pas0000172

Pidgeon, A., McEnteggart, C., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2021). Four self-related IRAPs: Analyzing and interpreting effects in light of the DAARRE model. *The Psychological Record, 71,* 397-409. https//:doi.org/10.1007/s40732-020–00428-9

Pilgrim, C. & Galizio, M. (1995). Reversal of baseline relations and stimulus equivalence: I. Adults. *Journal of the Experimental Analysis of Behavior, 63*, 225-238. https//:doi.org/10.1901/jeab.1995.63-225

Pinto, J. A. R., de Almeida, R. V., & Bortoloti, R. (2020). The stimulus' orienting function may play an important role in IRAP performance: Supportive evidence from an eye-tracking study of brands. *The Psychological Record, 70*, 257–266. http://doi.org/10.1007/s40732-020-00378-2

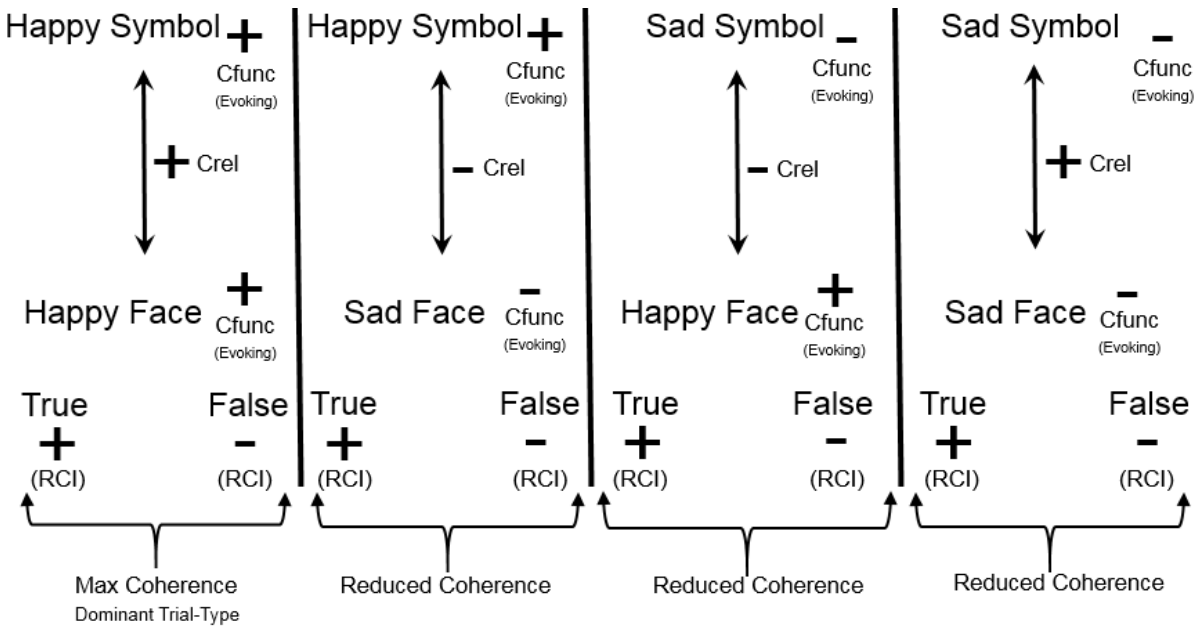
Schmidt, M., de Rose, J.C, Bortoloti, R. (2021). Relating, orienting and evoking functions in an IRAP study involving emotional pictographs (emojis) used in electronic messages. *Journal of Contextual Behavioral Science, 21*, 80-87. https://doi.org/10.1016/j.jcbs.2021.06.005

Shelley, M. (1818). *Frankenstein.* Penguin Classics.

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

Figure 1

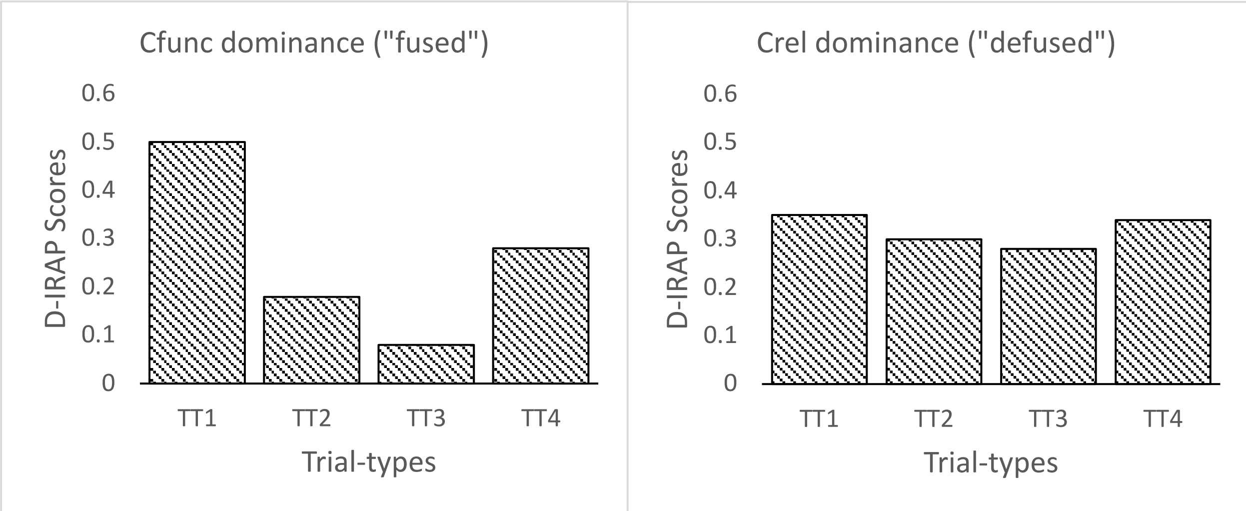
The DAARRE model as it might apply to the emotional faces IRAP from Bartoloti et al. (2019).



Note. The plus and minus symbols refer to the relative positivity of the Cfuncs (for each label and target), the Crels (between each label and target) and the RCIs. This relative positivity should be considered in the context of all other Cfuncs, Crels and RCIs in this stimulus set.

Figure 2

Hypothetical IRAP response patterns illustrating Cfunc (left panel) versus Crel (right panel) dominance as a potential analog for "fusion" and "defusion" respectively.



Note. Left-hand panel: A pattern of differential IRAP trial-type effects indicative of the STTDE and the dominance of the Cfunc properties of the stimuli involved. This general pattern may indicate "fusion" with the orienting and/or evoking properties of the stimuli. Right-hand panel: A pattern of IRAP effects that do not differ significantly from one another, indicative of the dominance of the Crel properties of the stimuli involved. This general pattern may indicate "defusion" from the orienting and/or evoking properties of the stimuli.