PRE PROOF VERSION – PLEASE DO NOT QUOTE THIS VERSION DIRECTLY

A primer on relational frame theory (RFT)

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

[1]Department of Psychology, Federal University of São Carlos, Brazil

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

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

Abstract

Both relational frame theory (RFT) and acceptance and commitment therapy (ACT) are based on the assumption that the evolution of human language (as derived relational responding) creates the potential for a form of psychological suffering unique to the human species. Furthermore, it has often been argued that RFT provides the basic science foundation for ACT. The current chapter will not dwell on these features of RFT but will focus instead on providing an up-to-date summary of the theory itself. Specifically, an historical and contemporary overview of RFT is presented, along with the details of recent ongoing efforts to advance the theory as a general behaviour-analytic account of human language and cognition. In doing so, the chapter endeavours to provide a modern vision of how RFT may continue to connect with ACT in the years to come.

KEYWORDS: RFT; ACT; DERIVED RELATIONAL RESPONDING; HDML; ROE-M

Historically, relational frame theory (RFT; Hayes, Barnes-Holmes, et al., 2001) has been seen as providing the basic science foundation for ACT by offering a detailed and empirically supported account of human language and cognition. The basic idea behind both RFT and ACT is that the evolution of human language, conceptualized as derived relational responding, creates the potential for a type of psychological suffering that is largely unique to humans. The purpose of the current chapter is not to focus on how RFT accounts for human psychological suffering, but rather to provide an up-to-date summary of the theory itself. The general aim, therefore, is to help readers of the current volume to contextualize, and better understand, any references that are made to RFT in other more clinically-focused chapters.

Writings on RFT are numerous and widespread, with several hundred published empirical studies. The theory is over 30 years old and its concepts appear to have stood the test of time, debate, and experimental scrutiny (e.g., see Hughes & Barnes-Holmes, 2016a, 2016b for recent reviews; but see Kissi, et al., 2017). As a result, it is thus possible to argue that RFT offers a relatively adequate, functional-analytic (behavioural) account of human language and cognition, while of course remaining very much a work in progress. In this chapter, we will first cover the historical background of RFT within behaviour analysis, to its emergence as a behavioural theory of human language and cognition. We will then provide a detailed overview of the core concepts of RFT, in the context of what appear to be important recent developments in the theory, both empirically and conceptually. We also reflect briefly on how these recent advances may connect to clinically relevant issues within the ACT literature. The current chapter thus aims to present an overview of RFT, and recent ongoing efforts to advance the theory, as an account of human language and cognition (Barnes-Holmes et al., 2016; Barnes-Holmes et al., 2017; Barnes-Holmes et al., 2018; Harte, Barnes-Holmes, Barnes-Holmes, & Kissi, 2020; Barnes-Holmes et al., 2020; Barnes-Holmes et al., in press). Incorporating these recent advances and developments in RFT into the current chapter was deemed to be important for the present volume because it will provide readers with a modern vision of how RFT may continue to connect with ACT over the coming years.

Historical Background to the Emergence of RFT in Behaviour Analysis

During the earlier part of the mid-20th century, behaviourism may have been seen as quite dominant, and indeed there is some substance to this view. Behaviourism, however, is a very broad term. One form of behaviourism, radical behaviourism, most closely associated with B.F. Skinner, could be seen as surviving to this day, although paradoxically it is perhaps seen as most closely associated with the demise of behavioural psychology. Specifically, it was Noam Chomsky's review (1959) of B.F. Skinner's book Verbal Behaviour (1957) that is often seen as marking the failure of behaviour analysis to provide an adequate account of human language. While there may be some truth to this historical narrative, it fails to recognise the fact that Skinnerian behaviourism has indeed survived and continues to work on many features of human language and cognition.

The first serious attempt within the school of radical behaviourism to develop an account of human language was provided by Skinner (1957). Although Chomsky's review raised at least some legitimate concerns with regard to Skinner's work, what is less well-known is that almost 10 years later, Skinner proposed another concept that was directly relevant to the study of human language and cognition. Specifically, he suggested that human problem-solving drew heavily upon a type of behaviour he referred to as rule-governed behaviour or instructional control (1966). In doing so, Skinner recognised that verbally-able humans frequently solved problems, not through direct contact with reinforcement contingencies that shaped-up appropriate behaviour through trial-and-error, but through the selection of verbal statements about the world and how to interact with it. Thus, for example, a child could learn to avoid eating a toxic berry by following a rule provided by a caregiver rather than having to eat the berry and experience sickness and risk even death. Skinner thus introduced the idea that a complete understanding of human psychology would require dealing with the extent to which human language created a type of learning pathway not shared with nonhuman animals.

Indeed, it was only five years later that another major figure in behaviour analysis, Murray Sidman, reported an effect that highlighted another way in which human learning may differ dramatically from that of other animals (1971). At the time, Sidman was attempting to develop procedures for teaching basic reading skills to an individual with severe learning disabilities. Specifically, Sidman and colleagues taught the individual to match 20 spoken words to 20 pictures, and to 20 printed words, over more than 15 hours across 4 weeks. At the end of this time, and to the surprise of the researchers, the individual spontaneously matched the 20 printed words to the pictures and vice versa in the absence of direct reinforcement for doing so. That is, reinforcing a subset of relational (reading) responses produced a number of emergent or unreinforced matching behaviours. Interestingly, these untaught or emergent matching or relating responses were discovered by Sidman and colleagues in the context of attempting to teach basic reading skills, and thus were clearly relevant to human language. Subsequently, the phenomenon that Sidman had revealed came to be known generally as the study of stimulus equivalence relations (see Sidman, 1994, for a book length review of the early history of this research program).

While the concept of the equivalence relation was refined over the years that followed, it was not until the 1980’s that a more rigorous and formalised account was presented (Sidman & Tailby, 1982; Sidman, 1986). Specifically, it was argued that the phenomenon comprised three formal properties, all shown in the absence of direct reinforcement: reflexivity, symmetry, and transitivity. Reflexivity required that each stimulus is conditionally related to itself (e.g., if A then A). In more concrete terms, given a picture of a dog, this picture will be chosen from an array containing a picture of the dog and other options (e.g., pictures of a cat and an apple). Symmetry required that the relation between stimuli is reversible (e.g., if A=B then B=A). Or in more concrete terms, for example, if a child is presented with the written word "dog" and taught to choose a picture of a dog, then the child should also readily choose the written word "dog" in the presence of the picture of the "dog". Finally, transitivity required that a relation between two stimuli (e.g., A=B), combined with a relation between one of those stimuli and a novel stimulus (e.g., A=C), so that the relations B=C and C=B readily emerged. To again apply this example, imagine that, as above, a child was presented with a picture of a dog and was taught to pick the written word 'dog', and also the written word 'woof'. Subsequently, the child may spontaneously match the written word 'dog' with the word 'woof' and the word 'woof' with the written word 'dog'. When such a pattern of responses emerged, the participating stimuli were said to form an equivalence class or relation. Crucially, these emergent, untrained responses were demonstrated with relative ease in humans but were largely absent (or at best extremely weak) in nonhumans (e.g. Sidman et al., 1982; Dugdale and Lowe, 2000; see also Zentall et al., 2014, and related commentaries in Dougher et al., 2014, indicating that clear evidence that stimulus equivalence, as defined by Sidman, has yet to be observed in non-humans species).

The phenomenon of stimulus equivalence thus raised two key but related issues. The first was the fact that it was difficult to explain in terms of direct reinforcement contingencies because previously unreinforced matching responses "emerged" during testing or probe trials. Second, there appeared to be some link between stimulus equivalence and human language (because it was discovered when teaching basic reading skills, and non-humans had failed to demonstrate clear evidence of equivalence responding). In attempting to reconcile these two issues, Sidman, et al. (1982) suggested that equivalence may be a basic stimulus function unique to humans and thus provided an explanation for human language (or at least symbolic relations) itself. In contrast, other researchers suggested that human language, and in particular naming, provided the basis for stimulus equivalence (Horne and Lowe, 1996).

A third alternative explanation for the emergent properties of equivalence relations was proposed within an account that came to be known as relational frame theory (RFT). Specifically, Hayes (1991) argued that the relating behaviour observed in emergent equivalence responding could be considered a class of generalized operant behaviour (i.e., equivalence responding was essentially learned during early language acquisition, and thus equivalence and symbolic relations were, functionally, synonymous). Furthermore, Hayes argued that a wide variety of these classes of generalized operants were possible and he referred to these as relational frames. In effect, during the course of early language learning human children were taught to respond in accordance with relational frames, such as opposite, difference, comparison (e.g., bigger versus smaller than), and so on, and thus a wide variety of derived relational responses should be possible. We will now turn to a description of the core concepts of RFT and its extension beyond stimulus equivalence as a basis for the complexities of human language and cognition.

RFT: Core concepts and technical explanation.

Just as Sidman had proposed that there were distinct properties involved in the equivalence relation, RFT posits three basic properties involved in the relational frame. Unlike the properties involved in stimulus equivalence, however, the properties involved in the relational frame are inherently more generic, because they need to reflect the numerous different generalized patterns of derived stimulus relating that are possible from an RFT perspective (i.e., not just frames of coordination/equivalence but also of opposition, comparison, difference, hierarchy, etc.). For example, the frame of opposition differs from coordination in that two opposite relations yield a derived relation of coordination, not opposite (e.g., “chilly” and “cold” are both opposite to hot, but coordinate with each other).

The frame of comparison includes many examples, but in the abstract may be represented using “more” and “less” signs (e.g., if A > B and B > C, then A > C and C < A). Comparison is also one of those frames that may yield “unspecified” relations when presented in abstract form. For example, if A > B and A > C then the relation between B and C remains unspecified; B and C could be more or less than each other or indeed the same. Note, however, that according to RFT, correctly deriving that the relation between B and C remains unspecified is a “correct” derived response in this instance.

Other relational frames, such as hierarchy, are perhaps best considered to be complex relational networks, rather than basic or simple frames. As such, a hierarchical network may involve “containment”, “coordination” and “difference”. For example, the term “fruit” *contains* all fruits, but dividing fruits into “citrus” and “non-citrus” involves establishing one frame of *coordination* among all citrus fruits and a separate frame of coordination among all non-citrus fruits, and a frame of *difference* between the two categories. If you then divide the two categories (citrus and non-citrus) into hard versus soft-skinned fruits the resulting “frame” of hierarchy seems more like a network than a “basic” or simple frame composed of just three relata. That is, the superordinate category “fruit” is at the top of the hierarchy, with citrus and non-citrus at the next level down, and then below that level are the four categories citrus (hard- and soft-skinned) and non-citrus (hard- and soft-skinned).

The most basic or simple relational network (or frame) is defined as a *generalized* (i.e., arbitrarily applicable) pattern of relational responding possessing the properties of *mutual entailment*, *combinatorial entailment*, and the *transformation of stimulus functions*. Mutual entailment is the most basic form of derived relational responding and marks the beginning of *symbolic* language development (Lipkens et al., 1993). It requires that the relation between two stimuli are related, *bidrectionally,* in a very specific way. For example, if A is *more than* B, then this relation mutually entails that B is *less than* A. In more concrete terms, if a child is taught that a car costs *more than* a bike, then the child may derive (i.e. without further information, instruction, prompting or reinforcement) that a bike costs *less than* a car.

The second property, combinatorial entailment, refers to the novel relations that emerge among and between stimuli when three or more stimuli are related. For example, if A is the opposite to B (mutually entailing that B is the opposite to A) and B is the opposite to C (mutually entailing that C is the opposite to B) then the derived relations A is the same as C and C is the same as A may emerge. In more concrete terms, imagine a child is taught that “wrong” is the opposite of “right” and that “right” is the opposite of “mícheart” (Irish for wrong). Again, in the absence of further instruction or prompting, etc. the child may derive that “mícheart” is the same as (coordinated with) “wrong”. As noted above the terms, mutual and combinatorial (entailment) are used within RFT, rather than symmetry and transitivity, respectively, because the former are not bound or limited to derived relations in which the individual elements simply become substitutable or equivalent to each other.

The third and final core property of a relational frame (or basic network) is the transformation of stimulus functions; that is, the change in the functions of one stimulus participating in a frame, which results in spontaneous changes in the functions of other stimuli in the frame. Critically these transformations of function occur in the absence of direct reinforcement, instruction or prompting. This third defining property of a relational frame thus highlights that symbolic relations in human language are involved in stimuli gaining, losing, or changing (i.e., transforming) their psychological properties. 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. Although not considered a core property within stimulus equivalence, a trans*fer* of functions was recognized in stimulus equivalence research. The classic demonstration involves establishing an equivalence class composed of three or more stimuli (e.g., A=B=C=D), establishing a specific function for at least one of the stimuli (e.g., pairing A with an unpleasant taste or smell), and then observing that the other stimuli within the class also acquire that function in the absence of direct training (B and C and D acquire at least some of the unpleasant taste or smell functions).

The term trans*formation* of function (rather than transfer) is employed within RFT because the functions of stimuli participating in relations other than equivalence/coordination do not transfer from one stimulus to another; rather the functions of the other stimuli in the frame are changed or *transformed* in accordance with the entailment properties. That is, the same function does not necessarily emerge among all participating stimuli within the frame - the nature of the transformation of stimulus functions depends on the specific relations involved (e.g., Dymond & Barnes, 1995). Imagine, for example, a situation in which a child has been bitten by a relatively small dog. The child later learns that a neighbour has just bought a very large dog. Based on the transformation of fear functions, in accordance with the frame of comparison (in this case, smaller/larger), it is possible that the neighbour's larger dog will evoke even greater fear and avoidance than the smaller dog that actually bit the child in the first place (see Dougher et al., 2007 for relevant experimental evidence).

In making a distinction between entailment and transformation of functions, RFT stipulates that these properties are under separate classes of contextual control. Specifically, entailment is determined by Crel contextual cues (i.e. controlling the type of relation) and the transformation of function is determined by Cfunc contextual cues (i.e controlling the specific behavioural functions produced during the act of relating). Specifying these types of contextual control is essential in determining how entailment and transformation effects combine in any given instance of what RFT refers to as arbitrarily applicable relational responding (AARRing). For example, if a friend told you that their new pet dog was called "Bongo" then the phrase "called" could function as a Crel for coordination (between the word 'Bongo' and your friend's new pet dog). If your friend then says "Bongo is really friendly" then the phrase "really friendly" may function as a Cfunc for actualizing some of the functions of a "friendly" dog (tail wagging, bouncy, safe, etc). Of course, "tail wagging" and other phrases may also be *entailed* with the events to which they refer. However, in this example we are highlighting their Cfunc properties to illustrate how RFT uses the defining properties of a frame (both Crel and Cfunc contextual control) to describe how verbal stimuli produce their effects in the natural environment of the wider verbal community.

We have just provided a description of the core properties of relational framing, which is seen as providing a behavioural unit of analysis for studying human language and cognition. On balance, RFT is not a "nativist" theory of language, in the sense that AARRing is deemed to be learned behaviour. That is, RFT aims to provide an explanation for the establishment of different classes of relational operants or AARR, and their combination into increasingly complex networks of relations. For illustrative purposes consider one of the most basic classes of AARR, naming. Young children may learn to point or look at a specific object upon hearing the name for that object, and they may also learn to produce the spoken name for the object. Across multiple exemplars of coordinating multiple objects and their names across many contexts, the operant class of coordination is established such that direct learning is no longer required in the presence of novel objects. That is, derived relating (coordination in this naming example) is established in the child's behavioural repertoire. For example, if the child is subsequently shown a novel object and is told its name, the child may subsequently name the object without having to be trained to do so. That is, once the generalized relational response of coordinating objects and their names is established, simply hearing the name for a novel object may "spontaneously" generate the appropriate naming response. Crucially, when this pattern of relational responding has been established, the generalized relational response may then be applied to any stimuli, given appropriate contextual cues (e.g., “this is a”). We will return to the important issue of learning histories involved in AARR later in the chapter.

According to an updated version of RFT, the ability to learn to AARR emerges from the evolution of highly cooperative behaviours within the human species (Hayes & Sanford, 2014). This updated version of RFT still maintains that AARRing, for any given individual, involves years of increasingly complex interactions with the wider verbal community within which the individual resides. However, a more detailed treatment of the phylogenic and ontogenic origins of AARRing is now emerging in the RFT literature (e.g., Hayes, et al., 2017), and we will incorporate these developments into the current chapter.

RFT: The role of human cooperation in the evolution of AARR

As a behavioural account of human language and cognition, RFT traditionally focused on the learning experiences that occur within the lifetime of the individual. This focus is understandable because the theory has been driven by a pragmatic concern with predicting-and-influencing human language and cognition itself in clinical, educational, and wider social settings. On balance, it has always been recognized that the ability to acquire the relational operants identified by RFT, with relative ease, is likely to have emerged from a particular evolutionary history, but until recently work in this area has been limited (e.g., Hayes & Sanford, 2014; Wilson et al., 2014).

Wilson (2007) summarized human evolution as the “three C’s”: cognition, culture, and cooperation. While all three of these were considered in early renditions of RFT, it appears that cooperation was somewhat underplayed, if not largely ignored. According to the first book-length treatment of RFT, Hayes, Barnes-Holmes, et al. (2001) suggested that mutual entailment (the bi-directional relational responding that may occur between two stimuli) in a listener could enhance or support avoidance of predators even if entailment was not yet present as part of a vocal or speaking repertoire. In addition, it was argued that this small difference could generate a group of listeners who could then reinforce mutually entailed responses in a speaking/vocal repertoire. Upon reflection, this account relies heavily on the evolution of mutual entailment as an adaptation of cognition in listening responses, and then spreads to speaking or vocal responses, thereby leading to increased social cooperation throughout the wider group or culture. In contrast to this account, Hayes and Sanford (2014) argued that it is more evolutionarily viable to assume that human cooperation was the primary driver in the evolution of mutual entailing, rather than the other way around. Indeed, as Hayes and Sanford point out, there is a wealth of empirical data supporting the argument that human cooperation was established by multilevel selection of cooperation itself, because it provided advantages for human group competition, which occurred alongside the cultural suppression of individual selfishness.

From an RFT perspective, a critical feature of human cooperation involves pointing and grunting, for example, which provided humans with highly important behavioural skills, such as social referencing and joint attention. These skills, it is argued, increased the likelihood that more advanced forms of cooperation, involving the emission of specific vocal sounds, would be selected or reinforced, as is the case with young children. For example, if a young child orients to a care-giver and then towards a toy while emitting a vocal sound (e.g., “eh”), and tries to reach for the toy, the care-giver may reinforce this cooperative act by giving the toy to the child. In the words of Hayes and Sanford (2014): “The entire exchange will build cooperation, perspective taking, and joint attention as patterns that are maintained within the group because it is a functionally useful communication exchange. If we unpack this highly likely sequence it means that in the context of high levels of cooperation, and adequate skills in joint attention, social referencing, and perspective taking, any characteristic vocalization in the presence of a desired object would likely lead to reinforced instances of symmetry or mutual entailment" (p. 122).

According to an updated version of RFT, which focuses on cooperation as a key driver of derived relational responding itself (AARR), the critical behavioural history does not begin with speaking or even simply listening (in a manner which involves "understanding" what was said in a symbolic sense). Rather it begins with mutually entailed orienting (Barnes-Holmes & Sivaraman, 2020), which we will argue is a key precursor for establishing AARRing. A potential "marker" for this type of orienting, which appears to be unique to the human species, is bi-directional orienting, characterized by a child orienting (e.g., looking at) back and forth between a care-giver and an object or stimulus that the care-giver is oriented towards[1]. Mutually entailed orienting should therefore be seen as a type of trans-generational behaviour (i.e., a class of behaviour that stretches across ontogeny and phylogeny) and is selected by reinforcement contingencies operating within the lifetime of the individual. In this sense, an up-dated version of RFT seeks increased scope in terms of linking directly with a modern evolutionary science (e.g., Wilson, et al., 2007), which argues that evolution operates at multiple levels (e.g., genetic, cellular, symbolic and cultural). As noted above, therefore, the critical behavioural history for AARR does not begin with listening and speaking (with understanding), it starts with one of the most basic of human cooperative acts (i.e., mutually entailed orienting). Mutually entailed orienting provides the infant with an opportunity to continue interacting with the caregiver as a dyad, which likely

[1] Mutually entailed orienting obviously involves orienting responses, but these occur as part of a uniquely human cooperative act between two or more individuals. Orienting *per se* remains a relatively basic response to any event that functions as a stimulus (or roughly speaking is simply noticed). Indeed, strictly speaking a stimulus cannot be defined as a stimulus without some orienting property. Thus, a young infant may show a startle response (and start to cry) if it hears a loud "unexpected" bang but this may occur when the child is alone and thus not engaging in what we are labelling mutually entailed orienting (because it is not part of a cooperative act). An important caveat, however, is that later we will refer to orienting as one of the core properties of a new generic unit of analysis that is emerging in an up-dated RFT. As such, orienting within this unit does not necessarily involve mutually entailed orienting but the establishment of orienting "inside" the unit necessarily involved a history of mutually entailed orienting.

serves as a reinforcer for continuing to engage in such acts of cooperation (i.e., gradually creating a dynamical feedback loop between cooperation and AARR).

The critical importance of mutually entailed orienting cannot be underestimated because it allows caregivers to establish appetitive and aversive evoking functions for stimuli in the child's environment. Once a caregiver and an infant are engaging in mutually entailed orienting, the caregiver can now orient the child towards a particular stimulus and encourage the child to approach "safe" and avoid "dangerous" stimuli. Mutually entailed orienting may thus also involve establishing specific orienting and evoking functions for particular stimuli. For example, if a caregiver shouts loudly when the child approaches a dangerous stimulus (e.g., an insect with a powerful venom) that stimulus will likely acquire strong orienting and (aversive) evoking properties for the child. Furthermore, when an infant engages in mutually entailed orienting, even items that are simply oriented towards by the caregiver, without issuing any sort of warning signal, may acquire relatively positive evoking (approach) functions for the infant. Mutually entailed orienting is thus more accurately labelled mutually entailed orienting and evoking. As a listening repertoire then develops, mutually entailing and evoking functions for particular stimuli become related, in an arbitrarily applicable manner, to specific sounds (i.e., words). Gradually, therefore, a new response unit involving relating, orienting, and evoking is established for the child. In an updated version of RFT we refer to this response unit as the ROE (pronounced "row", which is an acronym for relating, orienting and evoking); we will return to this conceptual unit of analysis later in the current chapter. At this point, however, it is important to understand that the term mutually entailed (orienting and evoking) serves to highlight that such cooperative acts occur in parallel with establishing a basic listener repertoire (e.g., a caregiver rarely engages a child in orienting and evoking without also emitting language-appropriate sounds, such as "Look, it's teddy", when orienting the child towards a toy teddy-bear).

We should emphasize that mutually entailed orienting/evoking are not simply new terms for mutual eye gaze, joint attention and social referencing. The latter have no "technical weight" within RFT itself, and thus by introducing these new concepts (mutually entailed orienting/evoking), an up-dated version of RFT seeks to establish explanatory depth. As noted above, the new concepts link the behavioural account of human language and cognition more directly to the evolution science argument that human cooperation drove, at least initially, the evolution of human language and cognition itself. In addition, the concepts of mutual eye gaze, joint attention, and social referencing are relatively topographical (e.g., all three behaviours involve mutual eye contact between two individuals). The term mutually entailed orienting/evoking aims to establish a functional-analytic-abstractive quality to the conceptual analysis of the behavioural topographies usually associated with the terms, joint attention and social referencing (and perspective-taking more generally).

To appreciate the point being made here, imagine a dog owner trained his dog to fetch an object and bring it to him by pointing at it, or even simply gazing at it, and shouting "fetch". One might argue that this interaction was clearly cooperative and involved at least some element of joint attention (and perhaps even social referencing), because the dog and its owner both needed to attend to the same object for the dog to fetch it. In addition, there have been studies showing that some dogs can follow human pointing to locations where food was hidden (e.g., Hare et al., 1998). According to an up-dated version of RFT, however, these interactions would not be defined as mutually entailed orienting, for the dog, unless it was functioning as part of an ontogenic and phylogenic history for AARR for that animal. Or to put it another way, if the cooperative interaction is part of an evolutionary history that leads to the establishment of AARR (and ROEing) for the dog, then the dog could be considered as engaging in mutually entailed orienting; if there is little or no evidence of AARR in the dog's behavioral repertoire in the past (as a species) or in the future as an individual organism, then the term mutually entailed orienting/evoking should not be applied to the dog's behaviour in this example of joint attention (or social referencing).

In emphasizing the importance of cooperation as a driver of AARR, and introducing the concept of mutually entailed orienting, the potential origins of contextual control over the transformation of functions becomes apparent. When an infant engages in mutually entailed orienting, even items that are simply oriented towards by the caregiver, without issuing any sort of danger-warning signal, may become more valuable than other items in the environment and acquire relatively positive evoking (approach) functions for the infant. It is important to note that during the acts of cooperation involved in mutually entailed orienting/evoking, the caregiver does not necessarily become appetitive or aversive as a consequence of their reactions to the pleasurable and dangerous items in the environment. This may be the case at first -- for example, if a child pulls away from or aggresses towards a caregiver when they shout at the child as a warning not to approach a dangerous object. However, an infant quickly learns to respond to the objects as being appetitive or aversive, and not the caregiver. In effect, mutually-entailed orienting and evoking between the care-giver and numerous stimuli serves to establish the care-giver as a stimulus that transforms the functions of novel stimuli and events in the environment while maintaining generally appetitive functions for the care-giver. In effect, a care-giver's actions or behaviors may transform the functions of a novel stimulus, but the care-giver appears to function as a context for limiting the transformation of functions to that stimulus.

This control (or limiting) over the transformation of functions could be seen as the basis for Cfunc control in RFT generally. This type of contextual control is seen in RFT as critical in selecting the specific functions that are transformed in any act of relating. For example, when an older child learns to relate the written word "chocolate" to actual chocolate they rarely attempt to eat the written word. Thus, it can be seen that the early cooperative acts involved in mutually entailed orienting and evoking in a sense provide the basis for the more sophisticated types of contextual control that are required as derived relational responding involving arbitrary stimuli is established in the child's listening and speaking repertoires.

As mutually entailed listening and speaking are established through ongoing interactions between the child and its caregivers, extended cooperation further facilitates the adaptation of the species, by allowing for more complex adaptations of the functional units, such as combinatorial entailment. This increasing complexity in derived relational responding involves the use of symbols and the ability to problem-solve in the natural and social environment. According to this updated version of RFT, therefore, cooperation facilitates more useful forms of cognition, rather than cognition producing more useful forms of cooperation, although it is important to appreciate that the relationship is likely non-linear and dynamical (i.e., cooperation generates increasingly advanced cognition, which in turn feeds back into generating increasingly complex forms of cooperation).

The relational development of increasingly complex patterns of AARR

Once the generic response unit of AARR (i.e., the ROE) is established, it allows for the evolution of increasingly complex relational responding inside the ROE, such as relational networking, the relating of relations (e.g., analogy and metaphor), and the relating of entire relational networks to other relational networks (e.g., extracting common themes from different narratives). An updated version of RFT has proposed a new multi-level framework for conceptualizing this increasing complexity in relational responding in terms of five levels of relational development; (i) mutually entailing, (ii) combinatorial entailing, (iii) relational networking, (iv) relating relations, and (v) relating relational networks.

Before considering this framework in greater detail, it is important to consider the generic RFT explanation for the establishment of different classes of relational operants or AARR, known as relational frames, and their combination into increasingly complex networks of relations. Imagine, for example, the wider verbal community directly reinforces a young child for pointing to or looking at a household pet such as a rabbit upon hearing the word ‘rabbit’ and/or the rabbit’s name (e.g., Roger). The child is also directly reinforced for producing other appropriate naming responses such as saying “rabbit” or “Roger” when this pet is observed, or in response to appropriate contextual cues such as “what is the rabbit's name?” or “what is this?”. Across multiple exemplars of coordinating multiple other stimuli with their names in multiple other contexts, the operant class of coordination comes to be established such that direct reinforcement is no longer required in the presence of novel stimuli. That is, derived coordination is established in the child’s behavioural repertoire. For example, if the child is subsequently shown a picture of a kangaroo alongside the written word “kangaroo” and is told its name, upon being presented with a relevant picture or the word, the child may then say “That’s a kangaroo!” in the absence of prompting or direct reinforcement. That is, once the generalized relational response of coordinating pictorial stimuli, spoken stimuli, and written words is established, directly reinforcing a subset of the relating behaviours “spontaneously” generates the complete set. Crucially, when this pattern of relational responding has been established, the generalized relational response may then be applied to any stimuli given appropriate contextual cues (e.g., “is”).

In the same way that derived coordination responding was established above in the presence of appropriate contextual cues (Crel; e.g., “is a” to specify the relationship between a rabbit and “Roger”), other cues such as “smaller than” or “faster than” would be established across multiple exemplars to specify other patterns of relational frames. Across time, this generalized derived relational responding becomes arbitrarily applicable -- the relating is not based solely on the physical or formal relations between and among the stimuli, but on additional contextual cues that determine the appropriate relational responses (again, in the absence of direct reinforcement).

For example, someone can abstractly say and understand that “a pig is bigger than a centipede” (which of course it is), despite the word “pig” being physically smaller and audibly shorter. Thus, the relationship between the two stimuli becomes arbitrarily applicable and is no longer determined by length or other physical characteristics. Subsequently, following a sufficient number of relevant exemplars to establish appropriate patterns of relational frames, you could be told that “A is bigger than B” and thus respond that “B must be smaller than A” without any knowledge of what A and B actually are.

Early research in RFT demonstrated a number of distinct patterns of AARR or relational frames. These patterns included: coordination (or sameness; e.g., Carr et al., 2000; Dunne et al., 2014; Luciano et al., 2007), distinction (or difference; e.g., Dunne et al., 2014; Roche & Barnes, 1997; Steele & Hayes, 1991), opposition (e.g., Barnes-Holmes, et al., 2004; Dunne et al., 2014), comparison (e.g., Barnes-Holmes, et al., 2004; Berens & Hayes, 2007; Dunne et al., 2014), temporality (e.g., O’Hora et al., 2004; O’Hora et al., 2005), hierarchy (e.g., Foody et al., 2013; Gil et al., 2012; Griffee & Dougher, 2002; Slattery, & Stewart, 2014), and deictics (or perspective taking; e.g., Barnes-Holmes, 2001; McHugh et al., 2004; McHugh et al., 2007). In addition, some early studies demonstrated the transformation of functions (as described previously) in accordance with specific relational frames (e.g., Dougher et al., 2007; Dymond & Barnes, 1995; Roche & Barnes, 1997). Furthermore, research demonstrated relational framing could be shown with numerous experimental preparations, thus indicating that the phenomenon was not tied specifically to any particular experimental procedure. And finally, and indeed critically, empirical evidence emerged to support the argument that exposure to multiple exemplars appeared to be essential in establishing specific frames (e.g., Barnes-Holmes et al., 2004; Lipkens et al., 1993; Luciano et al., 2007). Thus, the argument that relational framing could be thought of as a generalized relational operant (i.e., established by appropriate multiple exemplars) gained considerable traction (see Barnes-Holmes & Barnes-Holmes, 2000; Healy, Barnes-Holmes, & Smeets, 2000).

Complex relational networking. According to RFT, the combination of relational frames into increasingly complex relational networks helps to explain scaling up to complex levels of human language and cognition, such as rule following and analogical reasoning. For RFT, a rule or instruction can be thought of as a network of relational frames, typically involving temporal and coordination relations accompanied by appropriate contextual cues that transform specific behavioural functions within the network (Barnes-Holmes, et al., 2001). Take, for example, the instruction "if the alarm clock rings, then get out of bed". This simple rule involves frames of coordination between the words "alarm clock", "rings", and "get out of bed" and the physical alarm clock, the sound it makes when it rings, and the action of getting out of bed. The words "if" and "then" function here as contextual cues for establishing a temporal relation between the sound and the act of getting up (i.e., sound before getting up). Insofar as one actually gets up when the alarm clock rings, the functions of the sound itself have, in principle, been transformed by the network such that it now controls this specific behaviour in this context. This conceptual analysis of rules as complex relational networks has also been successfully modelled in the lab (e.g., O'Hora et al., 2004, 2014). Excessive reliance on rules at the expense of contact with direct environmental contingencies has been at the core of the ACT explanation for human psychological suffering since the conception of the approach (Hayes et al., 1999). And while little experimental work has explored the complexities involved in excessive rule-following as derived relational networks, recent research has successfully begun to do so (see Harte, Barnes-Holmes, Barnes-Holmes, & Kissi, 2020). We will return to this later in the current chapter.

Relating relations and relating relational networks. In scaling up in complexity again, other advanced levels of human language and cognition, such as metaphorical and analogical reasoning, may be readily explained by RFT (Barnes et al., 1997). For example, consider the simple analogy “a hammer is to a mallet as a comb is to a brush”. In this case, hammer and mallet are coordinated, as are comb and brush (via the cue “is to”). Furthermore, a coordination relation connects both of these coordination relations via the cue “as” (see Stewart & Barnes-Holmes, 2004, for a review of empirical work in this area). The example involves the relating of relations because the four relata (comb, brush, mallet, and hammer) do not “collapse” into a single relational network but involve relating one relation to another. In this sense, relating relations appears to involve responding relationally to one’s own relational responding; that is, coordinating the hammer-mallet relational response with the comb-brush relational response. Critically, this level of relational responding likely involves deictic relational responding (see next paragraph). At an even more advanced level of AARR, RFT proposes the relating of complex relational networks to other complex relational networks (Hayes, Gifford, et al., 2001). Empirical research in this area is somewhat limited (Ruiz & Luciano, 2011), but highly advanced verbal abilities such as complex problem solving and comparisons of extended narratives would likely involve this level of relating.

Perspective-taking and deictic relations. A considerable body of conceptual and empirical research has been conducted on deictic relational responding in RFT, which is seen to be critical for the emergence of a verbal self, and perspective taking (mentioned previously). In an up-dated version of RFT, mutually entailed orienting would be seen as providing a critically important historical context for the gradual emergence of a verbal self. Specifically, it involves cooperation between two separate individuals -- the infant and care-giver -- while the caregiver utters sounds (words) that later come to participate in arbitrary relations with the infant (e.g., the child’s name), the caregiver (e.g., “Daddy”) and the stimulus they are orienting towards (e.g. “teddy”). For example, a father might pick up a toy teddy, orient their child towards the teddy (i.e., hold the teddy in front of the child) and ask, "would you like daddy to give you the teddy?" Initially, of course, the words in the question have no symbolic functions for the infant, but this example of mutually entailed orienting is a critical part of the history that serves to establish those symbolic functions across thousands of such cooperative episodes in the child's first months and years of life. As the words in these types of questions gradually acquire their appropriate symbolic functions, and the ROE as a generic response unit becomes established, deictic relating (see below) may then emerge.

For RFT, three core relations are involved in deictic relating (Barnes-Holmes, 2001): the interpersonal relation, I-You, the spatial relation, Here-There, and temporal relation, Now-Then. These three types of relations combine into the basic or simplest deictic relational frame, which involves locating oneself in time and space relative to another individual. The core idea is that as children learn to respond in accordance with these deictic relations, they are essentially learning to relate the self to others in the context of particular times and spaces. For example, imagine a very young child being asked "What did you have for breakfast at home this morning?" while they are eating lunch in a restaurant later that day with their family. If the child responds simply by referring to what, for example, their sister is currently eating, they may be corrected and told "No, that's what your sister is now eating here. What did you eat earlier at home for breakfast?" Ongoing refinement of the three deictic relations in this way thus allows the child to respond appropriately to questions about their own behaviour in relation to others, as it occurs in specific times and places (e.g., McHugh et al., 2004).

Deictic relational responding is viewed as being relatively advanced because it involves learning to respond to one's own relational responding. As noted previously, this level of relational responding is likely involved in relating relations, and certainly in relating entire relational networks to other relational networks. In simple terms, a child would find it difficult to relate two separate relational responses if they could not "locate" those relational responses in a specific time and space. Indeed, this basic argument has been elaborated recently by Kavanagh et al. (2019) in their presentation of an RFT interpretation of the classic false belief perspective-taking task. We will return to this issue later in the chapter.

A Hyper-Dimensional Multi-Level (HDML) framework for conceptualizing the dynamics of AARR

As noted previously, an up-dated version of RFT proposes five key levels of behavioural development. In addition, the updated version of the theory emphasizes the dynamic nature of the relating activity that may occur along four dimensions. These dimensions are coherence, complexity, derivation, and flexibility. Each level of the framework intersects with the four dimensions, thus yielding 20 units of analysis (see Table 1; the reader should note that the Table also illustrates how the ROE fits into the framework, which will be covered later in the current chapter).

INSERT TABLE 1 HERE

Coherence refers to the extent to which a pattern of derived relational responding coheres or is consistent with previously established patterns of such responding. For instance, if an individual is told that a dog is smaller than a bear, and is then told that a bear is larger than a dog, the second statement would likely be deemed coherent with the first. In this case, coherence would be high because the overall pattern ($A<B = B>A$) coheres with the manner in which such verbal relations have been established by the wider verbal community (e.g., there are few instances in which an English-speaking listener would reinforce, or not correct, the statement, “if A is bigger than B, then B is bigger than A”).

Complexity refers to the level of detail or density of a particular pattern of derived relational responding. For example, a mutually entailed relation of coordination may be seen as less complex than a mutually entailed relation of comparison, because the former involves only one type of relation (e.g., if A is the same as B, then B is the same as A), whereas the latter involves two types (if A is larger than B, then B is smaller than A).

Derivation refers to the extent to which a particular pattern of derived relational responding has been "practiced" or emitted in the past. Each time a relation is derived, its derivation reduces because it acquires its own history that extends beyond the derivation that is made from the "baseline" relation. Imagine, for example, that an individual learns that bears are larger than dogs, and thus derives that dogs are smaller than bears. The first time that the 'a dog is smaller than a bear' relation is derived, it is derived "directly" from the 'a bear is larger than a dog' baseline relation. However, as the individual subsequently continues to relate dogs as smaller than bears, that relational response gradually acquires its own history, rendering it less and less derived from the original baseline relation (irrespective of whether or not it is reinforced directly).

Flexibility refers to the extent to which a given instance of derived relational responding may be modified by current contextual variables. As a simple example, imagine a young child who is asked to respond with the wrong answer to the question, "Which is bigger, a bear or a dog?" The quicker the child responds with "dog", the more flexible the relational responding (see O'Toole & Barnes-Holmes, 2009). Of course, flexibility is always context dependent and thus if the child had been told previously not to give a wrong answer when asked to do so, it would be difficult to use the production of a correct or wrong answer as an indication of flexibility.

The levels of relational development and the dimensions along which they may vary have been formalized recently within a Hyper-Dimensional Multi-Level (HDML; Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2020) framework for conceptualising and analyzing the dynamics involved in AARR (see Table 1). As noted earlier, an updated version of RFT proposes that most if not all human psychological events involve the ROE.

As an illustrative example, a mutually entailed relation (e.g., "hornets are dangerous") may be conceptualized as varying in coherence, complexity, derivation, and flexibility. In general terms, the relation between hornets and danger may be relatively high in coherence if the statement coheres with similar assertions (e.g., "a small number of hornet stings can kill"); relatively low in complexity if understanding the statement involves a limited number of other relational responses (e.g., the words "hornet" and "dangerous" are directly related to actual hornets and danger); relatively low in derivation (e.g., if similar statements have been heard many times in the past); and low in flexibility (e.g., if it is difficult to modify or "challenge" the perceived truth of the statement). Critically, this relational activity is seen to interact in a non-linear and dynamical manner with the orienting and evoking functions of stimulating events for humans as they navigate their environments. For example, the statement ("hornets are dangerous") may increase orienting and (aversive) evoking functions for hornets if the statement is uttered just before entering an area where they are commonly found. This updated RFT framework for conceptualizing the dynamical interplay among relating, orienting, and evoking (i.e., ROEing) has been defined as hyper-dimensional and multi-level (i.e., the HDML framework; Barnes-Holmes et al., 2020)[2].

A graphical representation of the HDML is presented in Table 1. Each cell of the grid, which shows the intersection between the five levels and four dimensions, contains an inverted 'T' with a third dashed line representing motivating variables (see below). This symbol represents the orienting and evoking functions that may occur within each of the 20 functional-analytic abstractive units of relating. Conceptually, orienting is seen as lying on a continuum, on the vertical axis, from complete absence (0) to strongest orienting response

[2] As noted in footnote 1, the orienting property of the ROE does not necessarily involve mutually entailed orienting. Nevertheless, the ROE itself necessarily involved a history of mutually entailed orienting, and thus the orienting property of the ROE is necessarily determined, in a strictly functional-analytic sense, by a history of mutually entailed orienting.

possible (1). Evoking is seen as lying on a continuum, on the horizontal axis, from the strongest aversive response possible (-1) to the strongest appetitive response possible (+1) with 0 representing the absence of either an aversive or appetitive reaction. Again, it is important to emphasize the inseparable, interactive and non-linear nature of the relating, orienting, and evoking (ROEing) that the HDML aims to capture. Returning to the "hornets are dangerous" example above. Imagine that you are shown some pictures of hornets and told that they are quite dangerous just before you enter a forest where they are commonly found (a relational event). As a result, orienting towards (or noticing) any insect that resembles a hornet and reacting aversively towards it may be more likely as you make your way through the forest. In contrast, imagine that you are provided with no warning about hornets before you enter the forest. You may be less likely to orient aversively towards a hornet, should you come across one, but may still engage in some level of relating, such as emitting the simple self-generated rule (i.e., relational network), "That looks quite nasty, I'll keep my distance." In essence, the concept of the ROE is designed to capture the constant, dynamical, and non-linear nature of the core unit of responding that characterizes human psychological events.

As noted above, the current version of the HDML includes an inverted 'T' with a dashed line representing motivating variables. Strictly speaking, motivating variables are not generic response functions (similar to orienting and evoking), but rather constitute a ubiquitous property of all psychological events that impact upon the ROE itself. Hence, motivating is represented with the broken line, which is scaled from 0 to 1, indicating the putative strength of a variable(s) that impacts upon orienting and/or evoking functions in some specific manner. In this sense, the influence of motivating variables is always inferred through changes in measures of orienting and/or evoking functions. We have chosen to include motivating variables in this manner in light of a recent study (Gomes, et al., 2020) that reported the impact of three different motivating conditions upon a measure that appears to be sensitive to orienting and evoking functions (see below). The critical point is that the concept of the ROE, as articulated above (and in previous publications), remains largely unchanged. Nevertheless, the inclusion of a motivating variable indicates that motivating variables are always at play in co-determining the relative values of the orienting and evoking functions within each of the 20 units of analysis contained within the HDML framework. In recognizing the importance of motivating variables, the ROE acronym has been modified to ROE-M (pronounced "roam").

From an updated RFT perspective, the set of relational abilities, and associated orienting and evoking functions contained within the ROE-M, evolved into complex forms of communication and problem-solving in only a few thousand years. Indeed, as argued above, the ability to engage in ROE-Ming appears to be a defining characteristic of the human species, and allows us to predict and influence our environment in increasingly sophisticated and powerful ways. From this perspective, once ROE-Ming evolves, the natural environment becomes thick and rich with stimuli that are symbolic, rather than direct-acting, as they appear to be for non-human species. For example, symbolic stimuli can be used to form new meanings and to construct new realities detached from direct experience (e.g., fiction, poetry, metaphor). As such, the transmission of behaviours, from one individual to another and from one generation to the next, is increased dramatically. This ultimately leads to greater variation in behaviour and the potential for the acquisition of new behaviours that serve to increase survival at multiple levels -- individuals, groups, and species.

Updating RFT: Some recent empirical advances

The current chapter has presented the historical background to RFT, its core descriptive and explanatory concepts and also more recent conceptual developments in the ongoing updating of the theory. At this point it seems important to consider some of the more recent empirical research that connects directly with the conceptual developments we have considered above. To this end, we will briefly present research that focused on (i) orienting functions, (ii) evoking functions, (iii) motivating variables, (iv) relational networks (as rules), and (v) relating relational networks (in perspective-taking).

Orienting functions. The potential importance of recognizing the role of orienting functions in dealing with the dynamics of AARR first became apparent in recent research reported by Finn et al. (2018). Finn and colleagues conducted a study using a procedure that had been developed in an effort to measure the relative strength or probability of AARR. This procedure is known as the implicit relational assessment procedure (IRAP). The IRAP is a computer-based programme that requires participants to respond quickly and accurately to specific stimuli deemed to be either consistent or inconsistent with participants' pre-experimentally established learning histories. On each trial, participants are required to choose one of two response options (e.g., True and False), indicating the relation between the label (presented at the top of the screen) and target (presented in the centre of the screen) stimulus. On some blocks of trials, participants are required to respond in a manner coherent with their pre-experimental learning histories, while on other blocks they are required to respond in a manner incoherent with these histories. The general assumption that underpinned early IRAP research was that, all things being equal, relational responding should be quicker and more accurate across blocks of trials that require relational responding that is coherent with a participant's learning history than on blocks that require responding that does not cohere with that history. For example, an IRAP might present the word 'flowers' or the word 'insects' as label stimuli at the top of the screen, positive or negative adjectives as target stimuli in the centre of the screen, and the response options True and False on the bottom left- and right-hand side of the screen. During some blocks of trials participants would be required to respond in a history-consistent manner (i.e., choosing "True" on *Flowers-Positive* and *Insects-Negative* trials and "False" on *Flowers-Negative* and

Insects-Positive trials), while on other blocks of trials the opposite response pattern would be required (e.g., responding "False" on a *Flowers-Positive* trial).

The typical IRAP may be conceptualized as comprising four separate trial-types involving a 2x2 cross-over of the label and target stimuli. Following on from the pleasant-flowers example above, the trial-types could be summarized as (i) *Flowers-Pleasant,* (ii) *Flowers-Unpleasant,* (iii) *Insects-Pleasant,* and (iv) *Insects-Unpleasant.* The primary datum from the IRAP is response latency, measured in milliseconds, and is defined as the time that elapses from the onset of stimulus presentation on each trial to the emission of a correct response. And as noted above, the basic assumption was that participants would produce response biases in which the size of the IRAP effects (the difference score between latencies/accuracies on history-coherent versus history-incoherent trials) indicates the probability of responding in the natural environment. Thus, for example, one would expect an individual who "loved" flowers and "hated" insects to produce relatively large IRAP effects in the predicted direction.

Since the IRAP's development, many studies have demonstrated its utility in measuring response biases in a range of areas and domains such as, age (e.g., Cullen et al., 2009), gender (e.g., Cartwright et al., 2017), race (e.g., Barnes-Holmes et al., 2010), religion (e.g., Hughes et al., 2017), and forensics (e.g., Dawson et al., 2009). The measure has also been used to predict racial group membership (Power et al., 2017) and parental smoking status (Cagney et al., 2017) over and above that of standard self-report measures. Finally, a meta-analysis of clinically-related IRAP studies reported a relatively high level of predictive validity (Vahey et al., 2015).

As mentioned above, a core assumption of early IRAP research was that responding would be faster and more accurate when the procedure required response patterns that were coherent with pre-existing patterns of AARR, than when it required patterns that were incoherent. On this basis, a simple assumption would be that the IRAP effects for all four trial-types should be roughly equal and in the same direction. However, this simple assumption did not always turn out to be the case (e.g., Finn et al., 2016; O'Shea et al., 2016). For instance, Finn et al. (2018) employed what they called a 'shapes and colours' IRAP, and although all of the effects were generally in the predicted direction, the effect for the *colour-colour* trial-type was significantly larger than for the other three trial-types (*colour-shape, shape-colour,* and *shape-shape*; see Figure 1). The smaller effect sizes for the *colour-shape* and *shape-colour* trial-types could be explained by the fact that responding during history-coherent blocks of trials required choosing False rather than True. More specifically, if there was an inherent response bias towards confirming rather than disconfirming relations, then reduced effect sizes would be expected when False was the correct response option (i.e., for an incoherent relation, such *colour-shape*). This explanation could not be used, however, to account for a larger effect for the *colour-colour* relative to the *shape-shape* trial-type, because they both involved responding "True" during history-coherent blocks. In grappling with an explanation for why this latter trial-type difference emerged, the authors of the study argued that the colour words employed within the IRAP occurred with higher frequency in natural language relative to the shape words (Keuleers et al., 2010). As such, it was possible that the colour words produced a stronger orienting response than their shape counterparts because the concept of colour, and colour words in general, were simply more salient than shapes, for the average participant (because colour words are used far more often in everyday discourse).

INSERT FIGURE 1 HERE

In developing a formal explanation for the differential trial-type effects described above, Finn et al. (2018) proposed the differential arbitrarily applicable relational responding effects (DAARRE) model (pronounced "dare"). According to this model, the differential trial-type effects may be explained by the extent to which the *Cfunc* and *Crel* properties of the stimuli contained within an IRAP cohere with specific properties of the response options across blocks of trials. As mentioned earlier in the chapter, for RFT, each instance of relating occurs under two types of contextual control. One kind specifies the particular type of relation defining the relational response (Crel), and the other specifies the particular behavioural functions that are transformed in accordance with the response (Cfunc). The reader should also note that response options, such as “True” and “False”, are referred to as relational coherence indicators (RCIs) because they are often used to indicate the coherence or incoherence between the label and target stimuli that are presented within an IRAP (see Maloney & Barnes-Holmes, 2016, for a detailed treatment of RCIs). A visual representation of the basic DAARRE model, as it applies to the Shapes-and-Colours IRAP, is presented in Figure 2.

INSERT FIGURE 2 HERE

Three key sources of behavioural influence are highlighted: (1) the relation between the label and target stimuli (Crels); (2) the orienting functions of the label and target stimuli (Cfuncs); and (3) the coherence functions of the two RCIs (e.g., “True” and “False”). As mentioned above, the two critical trial-types here were *Colour-Colour* and *Shape-Shape*. As can be readily observed, the Cfunc property for Colours is labelled as positive and the Cfunc property for Shapes is labelled as negative. This is in line with the suggestion above that, based on differential frequencies in natural language, colour-related stimuli likely possess stronger orienting functions relative to shape-related stimuli (the negative labelling for shapes should not be taken to specify a negative orienting function but simply an orienting function that is relatively weaker to that of colours). Relations between the label and target stimuli are labelled with plus or minus signs to indicate the extent to which they do or do not cohere, based on the participants’ relevant history. Thus, the colour-colour relation is labelled with a plus sign (i.e., coherence) whereas the colour-shape relation is labelled with a minus sign (i.e., incoherence). Finally, the two response options are similarly labelled to indicate their functions as either coherence or incoherence indicators. In the current example, "True" (+) would typically be used in natural language to indicate coherence and "False" (-) to indicate incoherence.

To appreciate the DAARRE model explanation for the differential trial-type effects, consider first the *Colour-Colour* trial-type and note that the Crel and Cfunc properties are all labelled with plus signs. Additionally, the RCI that is deemed correct on history-coherent trials is also labelled with a plus sign; this is the only trial-type that involves four plus signs. During history-coherent trials, therefore, this trial-type may be considered maximally coherent[3]. By contrast, during incoherent trials there is no coherence between the properties of the Crel, Cfuncs (all plus signs) and the required RCI (minus sign). According to the DAARRE model, this clear contrast in levels of coherence across blocks of trials results in a relatively large IRAP effect. Now consider the *Shape-Shape* trial-type. During history-coherent trials, participants are required to choose the same RCI as is required for the *Colour-Colour* trial-type, 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-incoherent trials,

[3] The term "coherent" is being used here in a manner that is consistent with the general definition provided earlier (i.e., "the extent to which a pattern of derived relational responding [involving both Crel and Cfunc properties] coheres or is consistent with previously established patterns of such responding"). The term does *not*, therefore, apply simply to Crel properties (i.e., A=B coheres with B=A), but also applies to the Cfunc properties (including RCIs). In the context of the shapes-and-colours IRAP, the *colour-colour* trial-type is considered maximally coherent because all of the critical responses during a history-coherent block involve relatively strong "confirmatory" responses. We are assuming here that most, if not all, participants would be subject to a general confirmation bias effect (e.g., Nickerson, 1998). For RFT such a bias is based on a history that involves a higher frequency of "confirming responses" for stimuli and events that are functionally similar (rather than functionally dissimilar). In principle, it would be possible to manipulate coherence within the current IRAP. Imagine, for example, if participants were exposed to a shapes-and-colours IRAP that presented the *shape-shape* trial-type far more frequently, across numerous sessions, than the other three trial-types. Given such a history, the pattern of Crel and Cfunc properties occurring for this trial-type would likely increase in coherence (thus over-riding the standard confirmation bias effect) because there would be greater functional overlap between the response pattern on this trial-type and the dominant pattern observed during previous sessions for that IRAP. The reader should note that this all-embracing functional definition of coherence is required when the ROE-M is defined as the generic unit of analysis involving Crel *and* Cfunc properties (as co-determined by motivational variables).

however, the RCI coheres with the Cfunc properties but not with the Crel property (plus sign). Thus, the differences in coherence between history-coherent and history-incoherent trials across these two trial-types is not equal (i.e., the difference is greater for the *Colour-Colour* trial-type). Finally, as becomes apparent from inspecting Figure 2 for the remaining two trial-types (*Colour-Shape* and *Shape-Colour*), the differences in coherence across history-coherent and history-incoherent blocks is reduced relative to the *Colour-Colour* trial-type (two plus signs relative to four), thus again explaining, at least in part, the dominance of the colour-colour trial-type over the other three. Subsequent studies have provided additional experimental support for this DAARRE model explanation in terms of orienting functions (see Finn et al., Experiment 3, 2018; Finn et al., 2019; Pinto et al., 2020).

Evoking functions. The capacity for functions to transform across stimuli and evoke appetitive or aversive responses has long been recognised within RFT and behaviour-analysis more generally, particularly with regard to explaining the ubiquity of human psychological distress (e.g., Dougher et al., 2007; Luciano et al., 2013, 2014). However, fully appreciating the potential importance of the role of evoking functions in explaining the dynamics of AARRing became particularly apparent in research reported by Leech and colleagues (2016, 2017). In both of these studies, participants responded on IRAPs that involved ‘cute’ puppies or kittens versus aggressive-looking spiders as label stimuli, while target stimuli involved approach (e.g., “I can pick it up”) versus avoidance (“I need to get away”) descriptors. Response options were once again “True” and “False” RCIs.

The results of both studies were generally in accordance with what one might expect (i.e., positive response biases on the two pet trial-types and a negative bias on one of the spider trial-types). However, the response pattern on the *Spider-Approach* trial-type was in the opposite direction to a common-sense prediction. Specifically, assuming that participants would not readily approach spiders in the natural environment one would likely anticipate a negative response bias on the *Spider-Approach* trial-type; counter-intuitively, however, participants tended to press “True” more quickly than “False”. On balance, this response bias did correlate significantly with participant performances on a behavioural task, which involved approaching a live spider. That is, the stronger the tendency to respond “True” more quickly than “False” (for *Spider-Approach*), the more likely participants were to approach an actual live spider. Thus, although the direction of the response pattern on this trial-type may appear somewhat counter-intuitive, it predicted *actual* behaviour. How might we explain this outcome?

Conceptually, it is possible that two separate Cfunc properties (i.e., orienting *and* evoking) were involved in determining participants' responses. To appreciate this suggestion, consider the stimuli involved within the IRAP. First, the pictures of spiders could be seen as potentially dangerous or threatening stimuli, and thus may likely possess strong orienting and aversive evoking functions, relative to the pet pictures. In contrast, the cute and cuddly-looking pet stimuli, would likely possess relatively strong appetitive evoking functions (but perhaps relatively weaker orienting functions due to their lack of threat or danger). Additionally, the approach and avoidance descriptors may not possess orienting functions that differ dramatically from each other, but the evoking functions they possessed would differ (i.e., avoidance = aversive, approach = appetitive).

According to a DAARRE model interpretation, therefore, the orienting functions of spiders dominated over the evoking functions for participants relatively low in self-reported spider fear (i.e., because spiders were not particularly aversive or appetitive for these participants). In contrast, for participants who were relatively high in self-reported spider fear, the (aversive) evoking functions may have dominated over the orienting functions (because spiders were seen as highly threatening). If this was indeed the case, choosing “True” more quickly than “False” would be highly coherent for low-fear participants, but less so for the high-fear participants (note that participants were from a normative sample and thus the relative differences in levels of spider fear would not be particularly extreme). To fully appreciate this argument, consider the DAARRE interpretation of the *Spider-Approach* trial-type illustrated in Figure 3.

INSERT FIGURE 3

This figure indicates that the Crel between spiders and approach is negative (i.e., most people would not report readily approaching spiders). A correct response on history-coherent trial-types, therefore, would be “False”. However, within the wider context of the IRAP, a relatively strong spider *orienting* function is likely established for the low spider fear participants, while a relatively strong aversive *evoking* function is likely established for the high spider fear participants. Thus, for the low-fear participants, the dominating Cfunc property for spiders (orienting) is positive as is the Cfunc property for the approach descriptor (evoking), both of which cohere with the positive (“True”) RCI. For the high-fear individuals, however, the dominating evoking Cfunc for spiders is negative but positive for the approach target stimulus. Thus, one of the Cfunc properties coheres with the positive “True” RCI while the other coheres with the negative “False” RCI. If the foregoing (albeit post-hoc) interpretation is correct, it would explain why performance on this trial-type appears to predict actual approach towards a live spider, although the overall direction of the effect is in a counter-intuitive direction (i.e., the latter effect is explained by the fact that the sample was normative).

Assessing the Cfunc properties of the label stimuli within an IRAP. The foregoing material on the DAARRE model draws heavily on the assumption that the Cfunc properties of the label and target stimuli play an important role in determining the types of effects that are observed with the IRAP. At the time of writing, however, direct experimental evidence for the impact of Cfunc properties for label and/or target stimuli was absent. That is, no published research had attempted to determine the impact of the Cfunc properties of, for example, the label stimuli independent of the target stimuli. However, very recent unpublished research has indicated that it is possible to examine the differential impact of the label stimuli (independently of the targets) using a modified IRAP combined with measures of neural activity (electroencephalograms [EEG]; Leech, 2020). Specifically, the research involved developing what is called a sequential IRAP in which the label stimulus on each trial is presented before the target stimulus. EEG signals are then recorded from the presentation of the label stimuli (i.e., before the target is presented). In Experiment 7 reported by Leech, pictures of pets and spiders were presented to participants. Critically, the difference in the EEG signals between pictures of pets and spiders interacted with laterality (i.e., whether the signal was recorded from the left or right side of the cortex) and whether the IRAP block required a history-coherent or history-incoherent response. And this three-way interaction effect was observed within 300ms of the label stimulus being presented on each trial (i.e., before the target stimulus was presented). In other words, it was possible to identify the impact of a Cfunc property for the label stimuli within an IRAP independently of the target stimuli. Admittedly, this research is very new and will need to be replicated, but it does suggest that IRAP effects involve complex clusters of 'interactants' that will require systematic analyses to better understand its functional-analytic properties and how it might be used to further explore human language and cognition within a behaviour-analytic framework.

The impact of motivating variables on IRAP performances. As noted previously, the ROE acronym has been modified (the ROE-M) to highlight the ubiquity of motivating variables influencing orienting and evoking functions. Indeed, the important role of motivating variables has long been recognized in behavior analysis generally (e.g., Skinner, 1953, 1957; Michael, 1993, 2007) and also in RFT with the concept of augmenting. On balance, the latter concept (augmenting) is more specific to rule-governed behaviour per se and could be considered a so-called middle-level term (see below). In any case, the impact of three different motivating conditions on an IRAP performance was recently reported in a study that used drops of pepper sauce to increase the size of appetitive functions for water related stimuli presented within an IRAP (Gomez, et al., 2020). Specifically, when two drops of pepper sauce were ingested by participants, the size of the IRAP effect for the "water-positive" trial-type increased dramatically relative to a group of participants who did not ingest any pepper drops or ingested only a single drop of pepper. In effect, the evoking (appetitive) functions of the water stimuli appeared to increase when a motivating variable for access to water was manipulated. It therefore seems wise to assume that such motivating variables are part of the behavioral field of interactants that are always involved in determining the properties of any given instance of ROEing (hence the ROE is more appropriately labelled the ROE-M).

Relevance to ACT and middle-level concepts. At this point, the DAARRE model interpretation of at least some IRAP effects could be seen as becoming so abstract that its connection to clinical psychology and human psychological suffering has been completely lost. On balance, we would argue that the effects and conceptual analyses we have considered here could be directly relevant to one or more of the well-known ACT–based middle-level terms (Barnes-Holmes et al., in press). Let us consider the concept of defusion, for example. Perhaps, the relative dominance of Cfunc over Crel properties described in the current chapter, as measured within the IRAP, could provide a bottom-up approach to this middle-level concept. To appreciate this analysis, consider the pattern of trial-type effects illustrated in Figure 4 (left-hand side). As argued previously, the DAARRE model interpretation of this effect is taken to indicate that the Cfunc properties of the label, target, and RCIs strongly influence IRAP performance. Perhaps these differential trial-type effects could be seen as evidence for “fusion” with the Cfunc properties of the stimuli because they produce differential trial-type effects (i.e., the orienting and evoking functions of the stimuli partly determine the response patterns). In contrast, consider the right-hand side of Figure 4. In this case, all four IRAP trial-type effects are more or less even; critically, no trial-type particularly dominates over the other. In this case, therefore, the Cfunc properties of the stimuli could be seen as having relatively limited impact on the IRAP performance. Or more informally, the Cfunc properties fail to create high levels of “fusion” because the participant is simply relating the stimuli (responding to their Crel properties) without being unduly influenced by their Cfunc properties. In principle, therefore, these two patterns of responding on the IRAP might provide a relatively precise experimental analysis of the distinction between fusion and defusion. Of course, this conceptual analysis is purely speculative and will require systematic experimental analysis.

INSERT FIGURE 4 HERE

Relational networks (as rules). The importance of rule-governed behaviour as a distinct feature of human language and cognition has long been acknowledged within RFT and behaviour-analysis more generally. In the 1970’s and 80’s, a plethora of experimental research emerged exploring this concept and the extent to which rule-governance led to insensitivity to direct contingencies of reinforcement (e.g., Lowe et al., 1983; Shimoff et al., 1986). Excessive rule-following has also been at the heart of ACT’s conceptual understanding of human psychological suffering. The basic argument is that the human propensity to engage in rule-governed behaviour undermines sensitivity to direct contingencies of reinforcement, and excessive rule-following in this regard is thus a critical feature in human psychological suffering (see Baruch et al., 2007; McAuliffe et al., 2014, for examples of experimental research exploring this suggestion in clinical samples; but see also Kissi et al., 2020, for a recent review).

Within RFT, the conceptual analysis of rules as derived relational networks (as laid out earlier in the chapter) gained empirical support from two studies that successfully modelled rules as derived relational networks in the experimental lab (O'Hora et al., 2004, 2014). However, while the link between rule-following and AARRing was clearly evident, other experimental research integrating these concepts was somewhat lacking until relatively recently (see Harte, Barnes-Holmes, Barnes-Holmes, & Kissi, 2020, for an extended discussion). Specifically, a series of studies have been published over the past number of years that have conducted experimental analyses of the impact of derived relations, varying across multiple dimensions within the HDML framework, on persistent rule-following (see also Monestes et al., 2017).

In an initial study, Harte et al., (2017) provided some participants with a direct rule, which did not involve responding in accordance with a derived relation generated within the experiment. That is, they were simply instructed to choose the *least like* comparison stimulus in a MTS task (see below). In contrast, other participants were presented with the same instruction but the phrase "least like" was replaced with a nonsense word that participated in a derived equivalence relation with that phrase. Or more informally, participants were required to treat the nonsense word as equivalent to the phrase "least like". The critical question that the researchers sought to address was: would there be any difference in persistent rule-following between these two groups of participants? For example, would participants who had been provided with a "direct" rule (one that did not require any derivation within the experiment) show higher levels of persistent rule-following than participants who were required to (partly) derive the meaning of the rule.

As noted above, persistent rule-following was assessed by presenting participants with a MTS task. Each trial on the task involved presenting a sample shape at the top of the screen and three comparison shapes at the bottom of the screen, each varying in terms of their similarity to the sample shape (i.e., one shape was clearly the most like the sample shape, one shape was quite like the sample shape but with more variations, and one shape was completely different to the sample shape with few or no overlapping similarities). The instruction or rule that participants received for responding on this task was initially consistent with feedback contingencies for responding on the task (i.e., the rule asked the participant to choose the least like comparison and points were awarded to participants whenever they chose this comparison). After a certain number of MTS trials, however, the feedback contingencies were reversed, and now points were awarded for choosing the most similar comparison. The researchers took a number of measures of the extent to which participants persisted with following the rule or instruction, even though doing so ceased to produce points on the MTS task. In general, the results showed that participants who had been provided with a direct rule persisted for longer than those who were provided with a rule that required deriving a relation within the experiment, but only when they had at least 100 opportunities (Experiment 2) to follow the rule before the contingencies reversed.

In a subsequent study, Harte et al. (2018) explored the impact of level of derivation on rule persistence using a similar contingency switching MTS task as above (i.e., task contingencies initially supported the derived rule but later reversed). In this study all rules provided to participants required a novel derivation within the experiment (i.e., there was no direct rule as in the previous study). In one condition, participants had many opportunities to make this novel derivation (low derivation), while in a second condition participants had relatively few opportunities (high derivation). The impact of mutual and combinatorial entailment was also explored in this study. That is, in one experiment, participants were required to derive that ‘least like’ was equivalent to a nonsense word (i.e., A=B) many or few times, while in a second experiment, participants were required to derive that ‘least like’ was equivalent to a nonsense word through a middle node (i.e., A=B=C), also many or few times.

This relation was then inserted into the rule for responding on the contingency switching MTS task. In broad terms, the results indicated that lower levels of derivation produced greater persistence in rule-following at both mutually and combinatorially entailed levels. That is, the more opportunities participants had to derive the rule, the more they persisted with rule-following when the MTS task contingencies no longer cohered with the rule.

In another two studies, researchers explored the impact of manipulating the coherence of the derived rule (Harte, Barnes-Holmes, Barnes-Holmes, McEnteggart, et al., 2020; Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, 2020). In both studies, participants were again provided with rules that required a novel derivation within the experiment, but the coherence of these rules was manipulated through the provision or non-provision of performance feedback for deriving the relations between the nonsense word and key phrase ‘least like’. Level of derivation was also manipulated, but across studies. That is, within each experiment all participants had the same number of opportunities to derive the relation, but across experiments they had relatively more or less opportunities (e.g., 5 blocks of training trials in one experiment versus only 1 block of training trials in another experiment). In other words, *within* each experiment, feedback was manipulated and level of derivation remained constant, but *across* each experiment level of derivation varied. Results showed that feedback for deriving significantly impacted upon rule persistence within the experiments in which derivation was high (i.e., fewer opportunities to derive), but not when derivation was relatively low (i.e., more opportunities to derive). That is, it seemed that the less derived the rule became, the less impact feedback for deriving had on participants’ MTS rule following. However, when the rule was relatively high in derivation, feedback significantly impacted upon MTS rule persistence.

The foregoing experimental analyses of rule persistence highlight what appear to be relatively subtle and complex effects. Specifically, the research has involved exploring the impact of coherence and derivation, for both mutually and combinatorially entailed relations (contained within a rule or network), on persistent rule-following. The findings indicate that persistent rule-following may be influenced by variables identified within the HDML framework and thus there is a clear need to continue to explore the impact of these types of variables if persistent or excessive rule-following is to be better understood. Indeed, this work will be particularly important in advancing our understanding of how excessive or inflexible rule following plays a key role in human psychological suffering, as has been long argued in the ACT literature. For example, in providing a better understanding of the variables involved in persistent rule-following, in laboratory studies, it should be possible to develop increasingly sophisticated assessment and treatment models of human psychological suffering in which excessive rule-following is implicated (see Harte, et al., 2017).

Relating relational networks (in perspective taking). As noted earlier, the ability to perspective-take (i.e., through the deictic relations of I-You, Here-There and Now-Then) has been implicated in the development of the verbal self. Specifically, these relations are thought to be critical in the ability to locate oneself in time and space relative to others. While a considerable body of research on perspective taking as deictic relational responding exists within RFT (e.g., Barnes-Holmes, 2001; McHugh et al., 2004; McHugh, et al., 2007), more complex conceptual analyses of this concept have only recently been offered. For example, as mentioned earlier in the chapter, Kavanagh et al. (2019) recently presented a conceptual analysis of a classic false belief perspective-taking task, wrought directly from the HDML. Although there is no experimental evidence yet available to support this interpretation, it seems wise to present it here as an example of potential future RFT analyses. In doing so we hope that the apparent precision and specificity that the HDML may offer will be fully appreciated, particularly when attempting to articulate the relationship between experimental RFT analyses and high level, clinically relevant concepts such as perspective-taking.

A large body of research has emerged over the years, particularly within mainstream psychology, that has focused on the development of different types of perspective-taking, emphasising in particular children with a diagnosis of autism spectrum disorder (ASD; e.g., Boucher, 2012) and adults with specific disorders, such as schizophrenia and borderline personality disorder (BPD; e.g., Németh et al., 2018). In broad terms, groups with these diagnostic labels tend to perform poorly on perspective-taking tasks such as false belief tasks relative to typical controls, but the literature indicates that performances vary widely depending on the nature of the tasks that are employed.

False belief refers to assumptions made about (a) another person's false beliefs and/or (b) another person's assumptions about beliefs held by a third party (Boucher, 2012). Perspective taking tasks (e.g., the Sally-Anne test; the Deceptive Container task; Unexpected Transfer task) aimed at assessing the first type of false belief typically involves presenting a child, for example, with a scenario in which they are asked a question about a known false belief that differs from the child's own belief (e.g., that there is a glove in a box despite the child's belief that there is actually a scarf in the box). The second type of false belief is typically assessed through presenting the child with a scenario involving a 'change in location' element to determine whether they understand that someone can hold a false belief about someone else's belief. For example, this can involve first telling the child that two people are given something to share (e.g., a chocolate bar), and that both then leave the item in a specific location (e.g., a cupboard). The child is then told that shortly afterwards, one of the two people moved the item (e.g., to a rucksack). The child is then asked where the other person thinks the item is, the correct answer being in the original location (i.e., the cupboard). It is this second type of false belief that is the focus of the conceptual functional analysis presented by Kavanagh et al. (2019), to which we now turn.

Kavanagh et al. (2019) first suggested that in order to respond successfully on a false belief task, which likely involves responding at the highest level of relational development (relating relational networks), there are a number of critical relational precursors that would already need to be firmly established within the individual's behavioural repertoire. First, a number of basic relational frames (i.e., coordination, distinction, and temporality) would be required to be in place, therefore involving the first two levels of relational development as specified within the HDML framework (mutual entailment and relational framing). These basic patterns of relational responding would also need to be high in coherence (i.e., consistent with many other past and current instances of responding in accordance with these patterns), relatively high in complexity (i.e., subject to multiple sources of contextual control), low in derivation (i.e., have relatively extended histories), and low in flexibility (i.e., should persist in the absence of supporting contextual variables, such as reinforcement, and in contexts that could undermine such responding, such as "mild" punishment).

Second, the three core deictic relations (I-You; Here-There; Now-Then) would also need to be firmly established within the individual's repertoire. The authors suggest that while these frames would naturally be located at the second level of relational development, if well established, they would likely also participate within larger relational networks. Responding in accordance with these relations would therefore additionally include this third level of relational development. Furthermore, as is the case for the frames of coordination, distinction, and temporality, the deictic frames would similarly need to be high in coherence and complexity, and low in levels of derivation and flexibility. Responding in accordance with these relations at these dimensional levels would likely be crucial in order for the individual to "locate" relevant relational responses in a specific time and space.

Finally, the authors suggest that frames of causality (i.e., 'if-then') would also be required, again at the first three levels of relational development with the same dimensional requirements specified above for the other relevant precursor frames. These ‘if-then’ frames would likely participate in complex relational networks with the deictic frames so that the individual could successfully derive such things as ‘if you and I both see something occur, then you and I both know that something has occurred’. This type of derived relational responding would likely be essential given the causal and temporal nature of the false belief task.

Assuming the relevant precursors are sufficiently established in the individual's history, Kavanagh and colleagues (2019) suggested that the ability to understand and successfully engage with the false belief task likely involves both (a) relating relations, and (b) relating entire relational networks to other entire relational networks. A graphical representation of the suggested relational responding involved in correctly identifying the unexpected location aspect of the classic False Belief task is presented in Figure 5. The reader is first invited to examine the left-hand side of Figure 5. This indicates that, at Time 1, both the self and other observe a hat being placed into a box. Based on this observation and the relational precursors detailed above, the self can conclude that both self and other know that the box contains a hat. The right-hand side of Figure 5 indicates that, at Time 2, the self observes the hat being replaced with a glove when the other is not present to see this happen. Once again, based on this observation and the relational precursors described above, the self can conclude that only they know that there is now a glove, not a hat, in the box. The double-headed arrow linking both Times 1 and 2 indicates that correct responding requires that the self relate the two networks as distinct in terms of what each knows after Time 2. Crucially, if the self simply reported that the other does not know what is in the box after Time 2, that would indicate responding at the level of relating relations. If, however, the self reported that the other thinks that the box contains a hat, that would require the relating of relations at Time 2 to the relating of relations at Time 1. In other words, the self must understand that what the other thinks at Time 2 is still what they knew at Time 1.This would involve relating relational networks by relating relations to relating relations at a second point in time (a combination of the final two levels within the HDML).

INSERT FIGURE 5 HERE

The foregoing conceptual analysis, while speculative, reveals how complex even a relatively simple task like the false belief task appears to be and why young children often fail to complete the task successfully. The individual differences in levels of coherence, complexity, derivation, and flexibility among the relational precursors discussed above could also help to explain, at least in part, why the literature contains such wide variation in the ages at which false belief tasks can be solved correctly, and why performances vary widely depending on the variation of the task that is presented (see Kavanagh et al., 2019 for a recent review). Again, despite the speculative nature of this analysis in the absence of experimental testing, we present it here so that the reader can appreciate the potential precision offered by the HDML, as an example of cutting-edge RFT-based analyses of complex behaviours. The critical point, of course, is that such analyses might be of use in helping applied researchers and practitioners alike to tackle deficits in perspective-taking when they are identified in specific clinical populations.

Concluding Comments

The main aim of this chapter on RFT was to help readers of the current volume to contextualize any references that are made to the theory in other, more clinically-focused chapters, and to appreciate how RFT may connect with ACT or ACT related work. We began by exploring the historical roots of the theory, moving then to its conceptual and methodological foundations, and presenting some of its most recent conceptual and empirical advances. Reflected within this, we hope, is the fact that RFT, as a functional-analytic account of human language and cognition, has not simply stood still in the 20 years since publication of the seminal volume (Hayes, Barnes-Holmes, et al., 2001). For example, recent advances have helped researchers to ask increasingly sophisticated questions, such as (i) what is involved in a simple perspective-taking task, relationally speaking; (ii) what potential precursors are necessary for this type of relating; (iii) how can patterns of IRAP effects potentially be used to interpret concepts like fusion and defusion; (iv) what variables are important in excessive rule-governance, and how are these variables relevant to human psychological suffering? Of course, the work we have outlined here has only begun to scratch the surface, and thus the recent conceptual and empirical work we have presented here should be seen as an ongoing work-in-progress.

In presenting these recent developments in RFT, we acknowledge the relatively rapid emergence of new RFT terms and concepts, including the DAARRE model, the HDML framework, the concept of mutually entailed orienting/evoking, and also the concept of the ROE-M. One could question the need for these new terms or concepts, or at least such a rapid pace of development. On balance, it is important to emphasize that these developments emerged directly from experimental research, and not solely from abstract theorizing. In time, alternative terms and concepts that allow for greater precision, scope, and depth may emerge, but that is exactly what we mean when we argue that RFT, and particularly the more recent empirical and conceptual developments, should be seen as a work in progress. Indeed, adopting this view seems important now as ever in taking RFT forward as a modern behaviour-analytic account of human language and cognition, and in endeavouring to create “a science more adequate to the challenges of the human condition” (Hayes et al., 2012, pp. 2).

References

Barnes, D., Hegarty, N., & Smeets, P.M. (1997). Relating equivalence relations to equivalence relations: A relational framing model of complex human functioning. *The Analysis of Verbal Behavior, 14,* 57-83. https://doi.org/10.1007/BF03392916

Barnes-Holmes, Y. (2001). *Analysing relational frames: Studying language and cognition in young children* (Unpublished doctoral thesis). National University of Ireland Maynooth.

Barnes-Holmes, D., & Barnes-Holmes, Y. (2000). Explaining complex behaviour: Two perspectives on the concept of generalised operant classes. The Psychological Record, 50, 251-265. https://doi.org/10.1007/BF03395355

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I. & Luciano, C. (2016). Relational frame theory: Finding its historical and philosophical roots and reflecting upon its future development: an introduction to part II. In R.D. Zettle, S.C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds), *The Wiley handbook of Contextual Behavioural Science* (pp. 117-128). Wiley-Blackwell.

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From IRAP and REC model to a multi-dimensional multi-level framework for analysing the dynamics of arbitrarily applicable relational responding. *Journal of Contextual Behavioural Science, 6*(4), 473-483. https://doi.org/10.1016/j.jcbs.2017.08.001

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

Barnes-Holmes, D., Barnes-Holmes, Y., McEnteggart, C., & Harte, C. (In press). Back to the future with an updated version of RFT: More field than frame? *Brazilian Journal of Behaviour Analysis.*

Barnes-Holmes, Y., Barnes-Holmes, D., Smeets, P. M., Strand, P., & Friman, P. (2004). Establishing relational responding in accordance with more-than and less-than as generalized operant behavior in young children. *International Journal of Psychology and Psychological Therapy, 4*, 531-558.

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

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., O'Hora, D., Roche, B., Hayes, S.C., Bissett, R.T., & Lyddy, F. (2001). *Understanding and verbal regulation*. In S.C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), Relational frame theory: A post Skinnerian account of human language and cognition (pp 103-117). Plenum.

Barnes-Holmes, D. & Sivaraman, M. (2020, August 14). Updating RFT: cooperation came first, the ROE as a unit of analysis, and engineering prosocial behaviour. https://science.abainternational.org/up-dating-rft-cooperation-came-first-the-roe-as-a-unit-of-analysis-and-engineering-prosocial-behavior/louise-mchughucd-ie/

Baruch, D. E., Kanter, J. W., Busch, A. M., Richardson, J. V., & Barnes-Holmes, D. (2007). The differential effect of instructions on dysphoric and nondysphoric persons. *The Psychological Record, 57*, 543–554. https://doi.org/10.1007/BF03395594.

Berens, N. M., & Hayes, S. C. (2007). Arbitrarily applicable comparative relations: Experimental Evidence for relational operants. *Journal of Applied Behavior Analysis, 40*, 45-71. https://doi.org/10.1901/jaba.2007.7-06

Boucher, J. (2012). Putting theory of mind in its place: Psychological explanations of the socio-emotional-communicative impairments in autistic spectrum disorder. *Autism, 16*, 226-246. https://doi.org/10.1177/1362361311430403

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

Carr, D., Wilkinson, K. M., Blackman, D., & McIlvane, W. J. (2000). Equivalence classes in individuals with minimal verbal repertoires. *Journal of the Experimental Analysis of Behavior, 74*, 101-115. https://doi.org/10.1901/jeab.2000.74-101

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

Chomsky, N. (1959). A review of B.F. Skinner's Verbal Behavior. *Language. 35*(1), 26-58. https://doi.org/10.2307/411334

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

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

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

Dougher, M.J., Twohig, M.P., & Madden, G.J. (Eds.). (2014). Stimulus-stimulus relations [Special Issue]. *Journal of the Experimental Analysis of Behaviour, 101*(1), 130-170.

Dugdale, N. & Lowe, C.F. (2000). Testing for symmetry in the conditional discriminations of language trained chimpanzees. *Journal of the Experimental Analysis of Behaviour, 73,* 5-22. https://doi.org/10.1901/jeab.2000.73-5.

Dunne, S., Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D. & MURPHY, C. (2014). Facilitating Repertoires of Coordination, Opposition Distinction, and Comparison in Young Children with Autism. *Behavioural Development Bulletin, 19*(2), 37-47. https://doi.org/10.1037/h0100576

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

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/10.10 07/s40732-017-0262-z

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2019). Predicting and Influencing the Single Trial-Type Dominance Effect. *The Psychological Record, 69*(3), 425-435. https://doi.org/10.1007/s40732-019-0347-4

Foody, M. Barnes-Holmes, Y., Barnes-Holmes, D., & Luciano, C. (2013). An Empirical Investigation of Hierarchical versus Distinction Relations in a Self-based ACT Exercise. *International Journal of Psychology and Psychological Therapy, 13*(3), 373-388.

Gil, E., Luciano, C., Ruiz, F. J., & Valdivia-Salas, V. (2012). A Preliminary Demonstration of Transformation of Functions through Hierarchical Relations. *International Journal of Psychology and Psychological Therapy, 12*(1), 1-19.

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

Griffee, K. & Dougher, M.J. (2002). Contextual control of stimulus generalisation and stimulus equivalence in hierarchical categorisation. *Journal of the Experimental Analysis of Behaviour, 78*(3), 433-447. https://doi.org/10.1901/jeab.2002.78-433

Hare, B., Call, J., & Tomasello, M. (1998). Communication of food location between food and dog (canis familaris). Evolution of Communication, 2(1), 137-159. https://doi.org/10.1075/eoc.2.1.06har

Harte, C., Barnes-Holmes, D, Barnes-Holmes, Y., & Kissi, A. (2020). The study of rule-governed behavior and derived stimulus relations: Bridging the gap. *Perspectives on Behavior Science, 43,* 361-385. https://doi.org/10.1007/s40614-020-00256-w

Harte, C., Barnes-Holmes, Y., Barnes-Holmes, D., & McEnteggart, C. (2017). Persistent rule-following in the face of reversed reinforcement contingencies: The differential impact of direct versus derived rules. *Behaviour Modification, 41*(6), 743-763. https://doi.org/10.1177/0145445517715871.

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. *Behavioural Processes, 157,* 36-46. https://doi.org/10.1016/j.beproc.2018.08.005.

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2020). Exploring the impact of coherence (through the presence versus absence of feedback) and levels of derivation on persistent rule-following. *Learning and Behavior*. Advanced online publication. 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

Hayes, S.C. (1991). A relational control theory of stimulus equivalence. In L.J. Hayes and P.N. Chase (Eds.), *Dialogues on Verbal Behavior* (pp. 19-40). Context Press.

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., Barnes-Holmes, D., & Wilson, K. (2012). Contextual behavioural science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioural Science, 1*(1-2), 1-16. https://doi.org/10.1016/j.jcbs.2012.09.004

Hayes, S.C., Gifford, E.V., Townsend, R.C., Jr., & Barnes-Holmes, D. (2001). Thinking, problem-solving, and pragmatic verbal analysis. In S.C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post Skinnerian account of human language and cognition* (pp 87-101). Plenum.

Hayes, S.C. & Sanford, B.T. (2014) Cooperation came first: Evolution and human language and cognition. *Journal of the Experimental Analysis of Behaviour, 101*(1), 112-129. https://doi.org/10.1002/jeab.64

Hayes, S.C., Sanford, B.T., Chin, F.T. (2017). Carrying the baton: Evolution science and contextual behavioural analysis of human language and cognition. *Journal of Contextual Behavioural Science, 6(*3), 314-328. https://doi.org/10.1016/j.jcbs.2017.01.002

Hayes, S. C., Strosahl, K., & Wilson, K.G. (1999). *Acceptance and Commitment Therapy: An experiential approach to behavior change.* Guilford Press.

Healy, O., Barnes-Holmes, D., & Smeets, P.M. (2000). Derived relational responding as generalised operant behaviour. Journal of the Experimental Analysis of Behaviour, 74(2), 207-227. https://doi.org/10.1901/jeab.2000.74-207.

Horne, P.J., & Lowe, C.F. (1996). On the origins of naming and other symbolic behaviour, *Journal of the Experimental Analysis of Behaviour, 65*, 185-241. https://doi.org/10.1901/jeab.1996.65-185

Hughes, S. & Barnes-Holmes, D. (2016a). Relational frame theory: The basic account. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 129-178). John Wiley & Sons, Ltd.

Hughes, S. & Barnes-Holmes, D. (2016b). Relational frame theory: Implications for the study of human language and cognition. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes,

& A. Biglan (Eds.), *The Wiley handbook of contextual behavioral science* (pp. 129-178). John Wiley & Sons, Ltd.

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

Kavanagh, D., Barnes-Holmes, Y., & Barnes-Holmes, D. (2019). The study of perspective taking: Contributions from mainstream psychology and behaviour analysis. *The Psychological Record.* Advanced online publication. https://doi.org/10.1007/s40732-019-00356-3

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. https://doi.org/10.3389/fpsyg.2010.00174.

Kissi, A., Harte, C., Hughes, S., De Houwer, J., & Crombez, G. (2020). The rule-based insensitivity effect: A systematic review. *Peer J 8:e9496.* https://doi.org/10.7717/peerj.9496

Kissi, A., Hughes, S., Mertens, G., Barnes-Holmes, D., De Houwer, J., & Crombez, G. (2017). A systematic review of pliance, tracking, and augmenting. *Behaviour Modification, 41*(5), 683-707. https://doi.org/10.1177/0145445517693811

Leech, A. (2020). *Analyzing the functional independence of the derived transfer of fear and the derived transfer of avoidance responses* [Doctoral dissertation, Ghent University].

Leech, A., Barnes-Holmes, D., & Madden, L. (2016). The implicit relational assessment procedure (IRAP) as a measure of spider fear, avoidance, and approach. *The Psychological Record, 66*, 337-349. https://doi.org/10.1007/s40732-016-0176-1

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

Lipkens, Hayes, S.C., & Hayes, L.J. (1993). Longitudinal study of derived stimulus relations in an infant. *Journal of Experimental Child Psychology, 56,* 201-239. https://doi.org/10.1006/jecp.1993.1032

Lowe, C. F., Beasty, A., & Bentall, R. P. (1983). The role of verbal behavior in human learning: Infant performance on fixed interval schedules. *Journal of the Experimental Analysis of Behaviour, 39*, 157–164. https://doi.org/10.1901/jeab.1983.39-157

Luciano, C., Gómez-Becerra, I., & Rodríguez-Valverde, M. (2007). The Role of multiple-exemplar training and Naming in Establishing Derived Equivalence in an Infant. *Journal of Experimental Analysis of Behaviour, 87*, 349-365. https://doi.org/10.1901/jeab.2007.08-06

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodríguez-Valverde, M., Barnes-Holmes, D., Dougher, M. J., Cabello, F., Sanchez, V., Barnes-Holmes, Y., & Gutierrez, G. (2013). Extinction of aversive conditioned fear: Does it alter avoidant responding? *Journal of Contextual Behavioural Science, 2*, 120-134. http://doi.org/10.1016/j.jcbs.2013.05.001

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodríguez-Valverde, M., Barnes-Holmes, D., Dougher, M. J., Lopez-Lopez, J. C., Barnes-Holmes, Y., & Gutierrez-Martinez, G.

(2014). Effects of an acceptance/diffusion intervention on experimentally induced generalised avoidance: A laboratory demonstration. *Journal of the Experimental Analysis of Behaviour, 101,* 94-111. http://doi.org/10.1002/jeab.68

Maloney, E., & Barnes-Holmes, D. (2016). Exploring the behavioural 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

McAuliffe, D., Hughes, S., & Barnes-Holmes, D. (2014). The dark-side of rule governed behavior: An experimental analysis of problematic rule-following in an adolescent population with depressive symptomatology. *Behaviour Modification, 38*(4), 587–613. http://doi.org/10.1177/0145445514521630.

McHugh, L., Barnes-Holmes, Y., & Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *The Psychological Record, 54*, 115-144. http://doi.org/10.1007/BF03395465

McHugh, L., Barnes-Holmes, Y., Barnes-Holmes, D., Whelan, R., & Stewart, I. (2007). Knowing me, knowing you: deictic complexity in false-belief understanding. *The Psychological Record, 57*, 533-542. http://doi.org/10.1007/BF03395593

Micheal, J. (1993). Establishing operations. *The Behaviour Analyst, 16*(2), 191-206. http://doi.org/0.1007/BF03392623

Michael, J. (2007). Motivating operations. In O.J. Cooper, T.E. Heron, W.L. Heward (Eds), *Applied Behaviour Analysis* (2nd ed.), 374-391. Merrill Prentice Hall.

Monestes, J. L., Villatte, M., Stewart, I., & Loas, G. (2014). Rule-based insensitivity and delusion maintenance in schizophrenia. *The Psychological Record, 64*(2), 329–338. http://doi.org/10.1007/s40732-014- 0029-8

Németh, N., Mátrai, P., Hegyi, P., Czéh, B., Czopf, L., Hussain, A., Pammer, J., Szabo, I., Solymar, M., Kiss., L/. Hartmann, P., Szilagyi, A.L., Kiss, Z., Simon, M. (2018). Theory of mind disturbances in borderline personality disorder: A meta-analysis. *Psychiatry Research, 270*, 143-153. https://doi.org/10.1016/j.psychres.2018.08.049

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175-220. http://doi.org/10.1037/1089-2680.2.2.175

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. http://doi.org/10.1007/BF03395484

O'Hora, D., Barnes-Holmes, D., & Stewart, I. (2014). Antecedent and consequential control of derived instruction-following. *Journal of the Experimental Analysis of Behaviour, 102* (1), 66-85. http://doi.org/10.1002/jeab.95.

O'Hora, D., Pelaez, M., Barnes-Holmes, D., & Amesty, L. (2005). Derived relational responding and human language: Evidence from the WAIS-III. The Psychological Record, 55, 155-174.

O'Shea, B.A., Watson, D.G., & Brown, G. (2016). Measuring implicit attitudes: A positive framing bias flaw in the implicit relational assessment procedure (IRAP). *Psychological Assessment, 28*(2), 159-170. http://doi.org/10.1037/pas0000172

O'Toole, C. & Barnes-Holmes, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligence test: The importance of relational flexibility. *The Psychological Record, 59*, 119-132. http://doi.org/10.1007/BF03395652

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

Power, P., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y. (2017). Exploring racial bias in a European country with a recent history of immigration of black Africans. *The Psychological Record, 65*, 365-375. http://doi.org/10.1007/s40732-017-0223-6

Roche, B., & Barnes, D. (1997). A transformation of respondently conditioned stimulus function in accordance with arbitrarily applicable relations. *Journal of the Experimental Analysis of Behaviour, 67*, 275-300. http://doi.org/10.1901/jeab.1997.67-275

Ruiz, F.J. & Luciano, C. (2011). Cross-domain analogies as relating derived relations among two separate relational networks. *Journal of the Experimental Analysis of Behaviour, 95,* 369-385. http://doi.org/10.1901/jeab.2011.95-369

Shimoff, E., Matthews, B. A., & Catania, A. C. (1986). Human operant performance: Sensitivity and pseudosensitivity to contingencies. *Journal of the Experimental Analysis of Behaviour, 46,* 149–157. http://doi.org/10.1901/jeab.1986.46-149

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

Sidman, M. (1986). Functional analysis of emergent verbal classes. In T. Thompson and M.E. Zeiler, (Eds.), *Analysis and Integration of Behavioural Units* (pp. 213-245). Laurence Erlbaum Associates.

Sidman, M. (1994). *Equivalence relations and behaviour: A research story*. Boston, MA: 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 Behaviour, 37*, 23–44. https://doi.org/ 10.1901/jeab.1982.37-23

Sidman, M. & Tailby, W. (1982). Conditional discrimination versus matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behaviour, 37*, 5-22. https://doi.org/10.1901/jeab.1982.37-5

Skinner, B.F. (1953). *Science and human behaviour.* Macmillan.

Skinner, B.F. (1957) *Verbal Behavior*. Appelton-Century-Crofts.

Skinner, B.F. (1966). An operant analysis of problem solving. In B. Kleinmuntz (Ed.), Problem-Solving: Research, Method, and Theory, (pp. 225-257). Wiley.

Slattery, B. & Stewart, I. (2014). Hierarchical classification as relational framing. *Journal of the Experimental Analysis of Behaviour, 101*(1), 61-75. https://doi.org/10.1002/jeab.63

Steele, D. L., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behaviour, 56*, 519–555. https://doi.org/10.1901/jeab.1991.56- 519.

Stewart, I. & Barnes-Holmes, D. (2004). Relational frame theory and analogical reasoning: Empirical investigations. *International Journal of Psychology and Psychological Therapy,* 4, 241-262.

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 Behaviour Therapy and Experimental Psychiatry, 48,* 59-65. https://doi.org/10.1016/j.jbtep.2015.01.004.

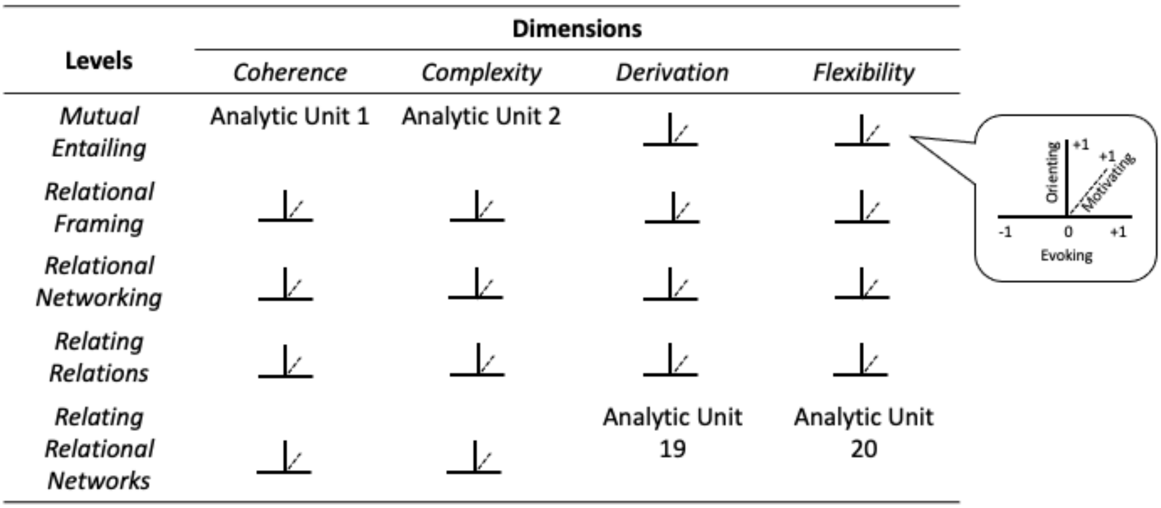
Wilson, D.S. (2007). *Evolution for everyone: How Darwin's theory can change the way we think about our lives.* Delacorte Press.

Wilson, D.S., Hayes, S.C., Biglan, A., & Embry, D.D. (2014). Evolving the future: toward a science of intentional change. Behavioural and Brain Sciences, 37(4), 395-416. https://doi.org/10.1017/S0140525X13001593

Zentall, T.R., Wasserman, E.A., & Urcuioli, P.J. (2014). Associative concept learning in animals. *Journal of the Experimental Analysis of Behaviour, 101*(1), 130-15. https://doi.org/10.1002/jeab.55

Table 1

A visual representation of the Hyper-Dimensional, Multi-Level (HDML) framework



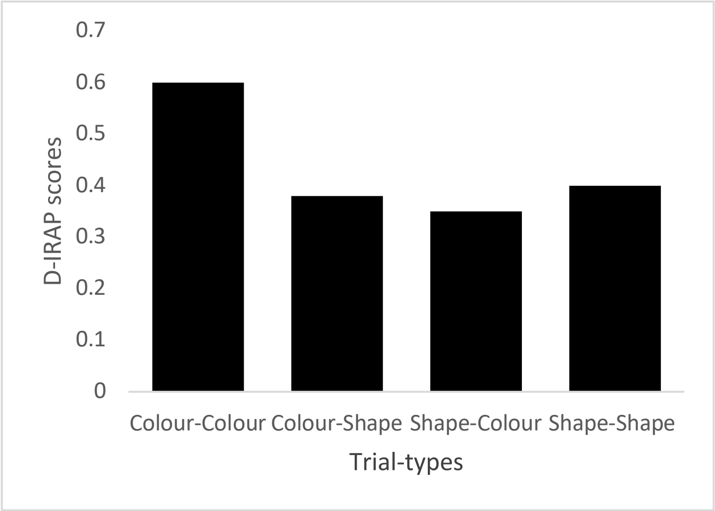
Levels	Dimensions			
	Coherence	*Complexity*	*Derivation*	*Flexibility*
Mutual Entailing	Analytic Unit 1	Analytic Unit 2	⊥	⊥
Relational Framing	⊥	⊥	⊥	⊥
Relational Networking	⊥	⊥	⊥	⊥
Relating Relations	⊥	⊥	⊥	⊥
Relating Relational Networks	⊥	⊥	Analytic Unit 19	Analytic Unit 20

Orienting
+1
+1
Motivating
-1
0
+1
Evoking

Note. 20 intersections between the five levels and four dimensions of arbitrarily applicable relational responding, combined with orienting and evoking functions, and motivating variables. Note that motivating is represented by a broken line because its impact is inferred based on changes in orienting and evoking functions. Overall, this table aims to capture the dynamic nature of AARRing (i.e., relating, orienting, evoking, and motivating; the ROE-M).

Figure 1

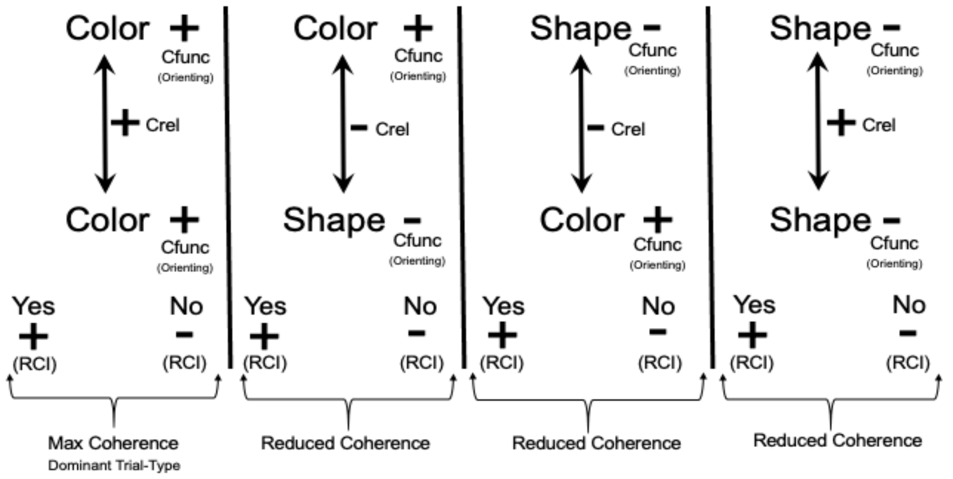
General pattern of trial-type effects produced by the Shapes-and-Colours IRAP



Note. In the Finn et al. (2018) study, participants were divided into two experimental groups based on experience with the IRAP (but this is not represented in the current figure).

Figure 2

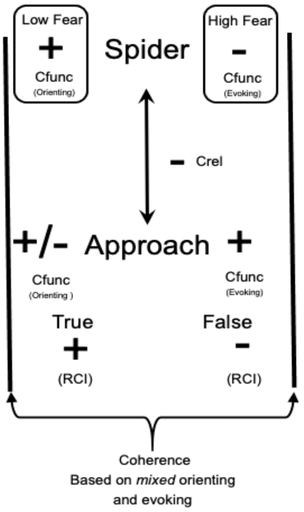
A DAARRE model analysis of the Shapes and Colours IRAP



Note. The '+' and '-' refer to the relative positivity of the transformation of function property (Cfunc) for each label and target stimulus, the entailment property between them (Crel), and the relational coherence indicator (RCI).

Figure 3

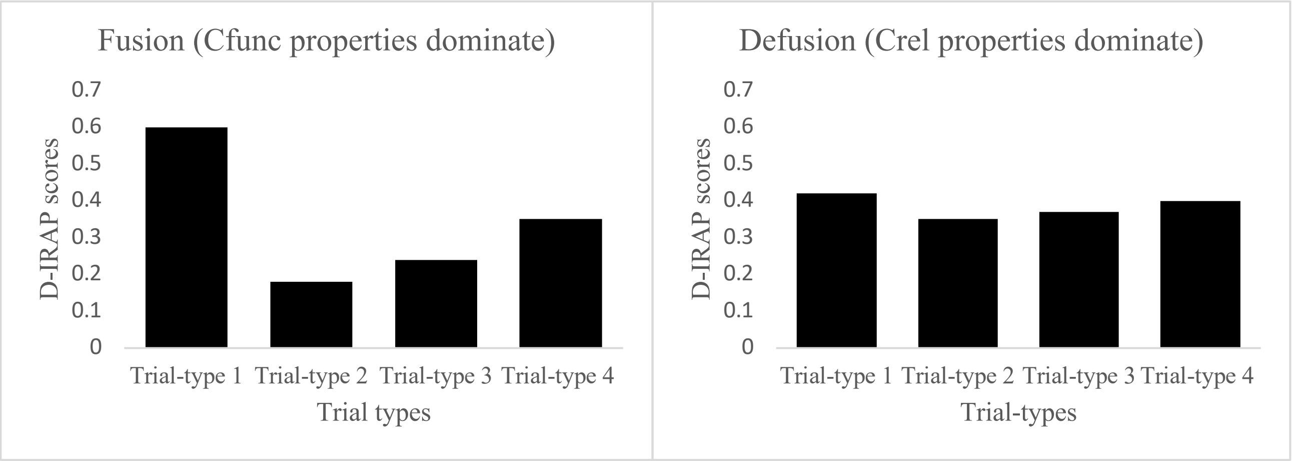
DAARRE model analysis of the Spider-Approach trial-type for low and high spider fear participants



Note. The figure illustrates the Cfuncs that most likely dominate for individuals who are low (orienting; left-hand side) versus high (evoking; right-hand side) spider fear. The "+/-" symbol indicates the assumption that the orienting functions of the "approach" relative to the "avoidance" adjectives would not differ dramatically within this particular IRAP, but the evoking functions for "approach" would be positive relative to "avoidance".

Figure 4

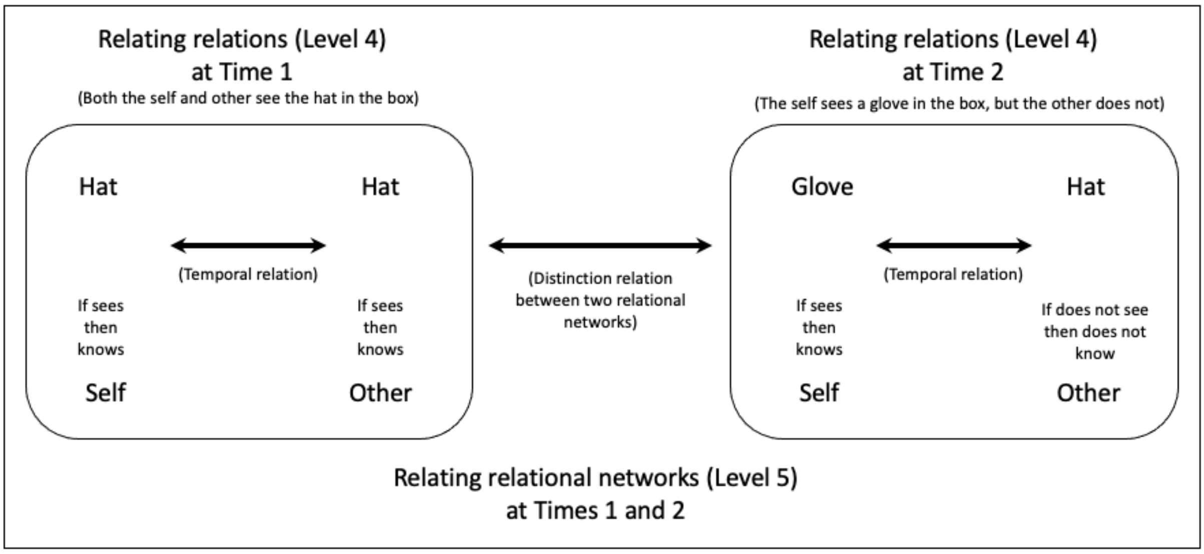
Hypothetical data illustrating the interpretation of differential trial-type effects on the IRAP



Note. On the left-hand panel, the Cfunc properties of the stimuli dominate over Crel properties, potentially modelling “fusion”. On the right-hand panel, the Crel properties of the stimuli dominate over the Cfunc properties, potentially modelling “defusion”.

Figure 5

Visual representation of the functional analyses provided for responding correctly on a false belief task



Relating relations (Level 4) at Time 1

(Both the self and other see the hat in the box)

Relating relations (Level 4) at Time 2

(The self sees a glove in the box, but the other does not)

Relating relational networks (Level 5) at Times 1 and 2