Chapter 8
Function over Form:
A Behavioral Approach to Implicit Attitudes

Anthony G. O'Reilly
National University of Ireland – Maynooth, Ireland

Bryan Roche
National University of Ireland – Maynooth, Ireland

Aoife Cartwright
National University of Ireland – Maynooth, Ireland

ABSTRACT

Research surrounding the construct of "implicit attitudes" and the various methodologies for measuring that construct is currently founded on the social cognitive paradigm. However, no robust and agreed upon theoretical framework has emerged from this paradigm, despite the widespread adoption of implicit testing methodologies and their associated theoretical assumptions. The current chapter outlines a functional approach to implicit testing, describing research stemming from Relational Frame Theory that was developed in parallel with the emergence of the IAT, and arguing for the benefits of connecting these two strands of research to improve the understanding of attitude behaviors and create better understood implicit testing methodologies. The chapter concludes with descriptions of two examples of such methodologies: the IRAP and the FAST.

INTRODUCTION

Form: The Implicit Attitude Construct

The central pillar of the cognitive paradigm is the position that mental representations mediate how information is perceived, processed, analyzed, stored in the brain, and that these representations precede behavior in the chain of cause and effect. These mental representations are conceptualized in terms of hypothetical constructs infered from observable behavior, which are in turn thought to explain that behavior. In the field of Social Psychology, there is no more ubiquitous construct than the Attitude.

Despite its central place in explaining human social behavior (Allport, 1935), there is no universally agreed upon definition of what precisely

DOI: 10.4018/978-1-4666-6599-6.ch008

is represented by an attitude. There is, however, broad agreement on the general form of the attitude construct. An attitude is usually defined as being a combination of cognitive (i.e. propositional) and affective evaluations of an object with a variable strength (Olson & Kendrick, 2008). An attitude is thought to be stored in the mind as a set of *associations* between the attitude object and these evaluations, with attitude strength being a function of the relative accessibility of these associations. When activated, attitudes predispose the person towards a favorable or unfavorable behavioral response towards the attitude object (see Olson & Zanna, 1993, for a review). Attitudes can be formed in relation to any given object including other individuals (e.g., McConnell, Rydell, Strain, & Mackie, 2008), social groups (e.g., Dovidio, Kawakami, & Gaertner, 2002), or even abstract concepts (e.g., nationality; Devos & Banaji, 2005). Stereotypes, prejudice, self-esteem, general positive or negative evaluations and biases all fall under the umbrella of "attitude". As such, attitude research has historically been a key topic in social cognitive research, and it is likely to continue to be so for many years to come. However, attitude research underwent a minor revolution in the last years of the 20th century, with the introduction of the concept of *implicit attitudes*.

Greenwald and Banajii's 1995 paper simultaneously introduced the concept of implicit attitudes and a methodology designed to detect this new hypothetical construct – The Implicit Association Test (IAT). Drawing upon research in implicit memory, the authors described how past experience can influence present attitudes and the responses mediated by those attitudes without conscious awareness. Central to the new concept was the suggestion that some experiences lead to the formation of evaluative associations (i.e., attitudes) that were not readily accessible by introspection and whose influence on behavior is outwith the control of the subject. The implicit attitude construct was said to explain why self-reported attitudes were not reliable predictors of behavior. More specifically, the behavior was considered to be mediated by implicit attitudes, which in turn can be defined as "the introspectively unidentified (or inaccurately identified) trace of past experience that mediates [favorable or unfavorable feeling, thought or action towards social objects]" (Greenwald & Banaji, 1995, p.8).

The seminal IAT experiment used a "known groups" paradigm, presenting participants with flower names (e.g. Tulip), "insect" names (e.g. Spider), pleasant words (e.g. Love) and unpleasant words (e.g. Ugly) and required participants to categorize them by means of a key press. (Greenwald, McGhee & Schwartz, 1998). In the first (consistent) condition, the same response key was assigned to both "flower" and "pleasant" words, while the other response key was assigned to "insect" and "unpleasant" words. In the second (inconsistent) condition, one response key was assigned to "unpleasant" words and "flower" words, and the other to "pleasant" and "insect" words. The researchers found (as expected) that reaction times were shorter in the consistent condition than in the inconsistent condition. In line with the pre-experimental assumptions outlined above, Greenwald and colleagues stated that the IAT effect (i.e., the difference in response times between the two conditions) was indicative of the existence of an implicit attitude construct in which flowers were associated with positive evaluations and insects with negative associations. The magnitude of the difference between the normed reaction times in each condition is taken to be an indicator of the *associative strength* between the category of interest and a positive/negative attribute (e.g., Hoffman, Gawronski, Geschwnder, Le, & Schmidt, 2005). The implicit associations so measured by the IAT are assumed to be relatively stable, trait-like cognitive associations that are existent objects in the individual (Nosek & Hanson, 2008).

Despite the widespread adoption of the IAT methodology and the tacit acceptance of its theoretical assumptions, Greenwald and his colleague

are "theory uncommitted" (Greenwalk, Nosek, Banaji & Klauer, 2005) with regard to the precise structure of the mental associations the IAT purports to measure, and with regard to precisely what mental construct causes the IAT effect. The main empircal evidence for Greenwald's account of implicit attitudes (which informed the creation of the IAT) comes from the IAT effect itself, which it turn relies on the associative assumption within the implicit cognition theory in order to explain the effect. This position represents a form of circular reasoning, rendering the IAT and implicit social cognition research more generally lacking in a unaminously agreed-upon testable model than can explain the IAT effect (Fiedler, Messner, & Blumke, 2006; Gavin, Roche & Ruiz, 2008; Roche, O'Reilly, Gavin, Ruiz, & Aranciba, 2012). There is widespread concern in the social cognitive field (e.g., Blanton & Jaccard 2006; De Houwer, 2009; Fazio & Olson, 2003; Gawronsk, Lebel, Peters & Banse, 2009; Hughes et al, 2011; Rothermund & Wentura, 2004) regarding the "deplorable disconnect between basic research on the mechanisms underlying implicit measures and the somewhat wider reception of research using these methods" (Gawronski, Lebel & Peters, 2007).

It is precisely that "deplorable disconnect" that this chapter aims to address. Using behavioral manifestations such as response latencies of assumed cognitive structures, such as unconscious bias, to infer the existence of those very constructs is the source of much of the conceptual ambuguity and disagreement surrounding the theory underlying implicit attitude research (De Houwer, 2011). When the sources of an effect are hypothethical and conceptual controlling the outcomes of tests for those processes becomes a haphazard endeavour, in which changes in experimental methodology are dictated by a disputed or ambiguous point of theory. Functionally oriented approaches to the analysis of behavior aim their scrutiny directly at the underlying mechanisms of behavior. A behavior-analytic account of implicit attitudes and implicit testing methodologies can explain from the bottom up, in well defined technical terms, each and every apsect of an implicit test performance, as well as identifying the sources of historical and environmental control over test outcomes. This approach yields a far less speculative account of the processes involved, which in turn leads to more robust theory.

Jan DeHouwer (2011) has outlined the benefits of integrating functionally oriented research into the cognitive paradigm, leading to what he calls a functional-cognitive framework. DeHouwer argued that the functional approach is useful to the social cognitivist in that it provides useful and actionable information about the environmental causes of behavior and the environmental variables that can be experimentally manipulated to produce or alter a behavioral effect (such as the IAT effect) without requiring any reference to mental constructs as causal events. This information allows the cognitively oriented psychologist to make more informed inferences about the mental constructs assumed to mediate such behavioral effects by eliminating *a priori* assumptions about the processes underpinning those constructs and by providing clear information about the input to those mental processes. The functional approach informs the cognitive approach as to the facts (behaviors) that need to be accounted for with mental explanations, without reference to those mental explanations themselves. This is a pragmatic approach which puts aside the philosophical differences regarding the ontological status of mental representations in favor of developing a research program with strong theoretical underpinnings on the process level, leaving the debate regarding the neccesity of mental representations as causal objects to a future in which the body of evidence is more complete.

Function: The Stimulus Equivalence/ Relational Frame Theory Account

As it happens, there is already a wealth of functionally-oriented behavior-analytic work examining

the same phenomena that inspired the Implicit Social Cognition revolution and the development of the IAT. This work has been proceeding in parrallel with the development of the the IAT, with the two threads only rarely making contact. The following section will detail the history of the behavior-analytic research that informs this functional account of implicit attitudes, before bringing the threads together by offering a functional account of the Implicit Association Test and of Implicit Attitudes.

*Relational Frame Theory (*RFT: Hayes, Barnes-Holmes & Roche, 2001) emerged from the behavior-analytic literature around the same time that Implicit Social Cognition research began to take off in the social-cognitive mainstream, and while the methodologies and theoretical/ philosophical underpinnings of the work are quite different, there are many interesting parallels in the subject matter of both fields. As RFT will likely be unfamiliar to many readers, the authors will now describe its basic premises and development with regard to implicit attitude phenomena in detail.

The RFT approach is a modern behavioral account of human language and cognition that has proved able to tackle questions long thought to be out of reach of the behavior-analytic approach. RFT has made a multitude of advances in the behavioral understanding of such areas of cognition as analogical reasoning (Carpenter, Smeets & Barnes-Holmes, 2003; Stewart & Barnes-Holmes 2004; Stewart, Barnes-Holmes & Weil, 2009), assessment and training of intellectual skills (e.g., Cassidy, Roche & Hayes, 2011; O'Toole & Barnes-Holmes, 2009), in applied developmental and clinical arenas (McHugh, Barnes-Holmes & Barnes-Holmes, 2004; Rehfeldt, Dillen, Ziomek & Kowalchuck, 2007; Villatte, Monestes, McHugh, Freixa i Baque & Loas, 2010a, 2010b; Weil, Hayes & Capurro, 2011) and generative verbal behavior in developmentally delayed children (e.g., Heagle & Rehfeldt, 2006; Moran, Stewart, McElwee & Ming, 2010; Murphy & Barnes-Holmes 2009). In addition to its success in these areas, the authors believe that RFT can be a powerful explanatory tool for researchers interested in implicit attitudes.

Relational Frame Theory offers the attitude researcher an account of language and cognition that is founded upon a single core process – relational framing. This process is precisely articulated in a bottom up account that specifies the interactions necessary between organism and environment (see below) for this behavioral process to emerge. The bottom-up, functional-analytic approach taken by RFT builds incrementally from elementary core processes, but retains the explanatory power necessary to tackle more complex forms of behavior with the same precise terminology. The relational framing account of language has the potential to shed light on the fundamental processes which underlie attitude constructs, allowing for a more nuanced theoretical account of their formation and change, and to guide research into methodologies which might more accurately and reliably measure attitudes.

Much of the research that underpins current thinking on relational framing has grown from the discovery of the phenomenon of stimulus equivalence over forty years ago. The way in which this simple phenomenon was studied and built upon exemplifies the functional approach.

Murray Sidman's (1971) investigations into this phenomenon began while conducting research into the behavior of participants who experienced difficulty reading, writing, and speaking. He used a Matching-to-Sample (MTS) procedure, a conditional discrimination procedure in which two stimuli (let's call them B1 and B2) are presented as response options to be discriminated between and the correct response is determined by the presence of the conditional stimulus (let's call it A1). In effect, an "if-then" relation is established for the child, such as "If A1 is present, select B1 rather than B2". Or, "If A2 is present, select B2, rather than B1". In Sidman's study each trial involved the presentation of a sample stimulus, either a picture of the object to be named (e.g., a picture of a cat), a word (e.g., "cat"), or an

auditory stimulus (e.g., the word "cat" spoken aloud to the participant). In matching tests, the participants were required to choose the correct comparison stimulus (a picture or a word) from an array of eight choices (pictures or words). In oral naming tests, the participant was required to name the sample stimulus aloud.

Using the matching to sample procedure, the researchers taught the first participant to match the spoken word samples to the correct written word. Without any further direct training, the participant was then able to match written words to pictures (and vice versa) and to name the written words. This emergent behavior caused significant excitement for Sidman and his collaborators (Sidman, 1982).

The vital finding was that teaching two sets of conditional discriminations caused novel behaviors to emerge without direct training. This early glimpse of the phenomenon caused Sidman to focus his research on defining stimulus equivalence and establishing the necessary and sufficient historical and current conditions required to produce and test it in the laboratory.

Stimulus equivalence was defined procedurally (Sidman, 1982) as responding that displays the properties of reflexivity, transitivity, and symmetry. When a verbally able human participant is trained in (at least) two conditional discriminations (e.g., given A1 pick B1 and not B2, given A2 pick B2 and not B1, given B1 pick C1 and not C2, and given B2 pick C2 and not C1) the participant will behave in ways that have not been reinforced by the experimenter. Specifically, when a participant is presented with A1, she will pick A1, matching each stimulus with itself (reflexivity). When presented with B1, she will pick A1, reversing the direction of the trained relation (symmetry). When presented with C1, she will select A1, deriving the untrained identity relation between the stimuli that were never paired (transitivity).

The generativity and stimulus substitutability characteristics of Stimulus Equivalence suggested a strong link between stimulus equivalence and language. With the emergence of this new phenomenon, behavior analysts began developing a new model of verbal behavior involving derived or emergent stimulus relations (For a full review of the evidence linking the stimulus equivalence phenomenon to language the reader is referred to Dymond and Roche, 2013). Expanding their explanation beyond simple stimulus equivalence into more varied forms of stimulus relations, behavior analysts developed Relational Frame Theory, which in turn informs the functional account of implicit testing and implicit attitudes described later in this chapter.

The Relational Frame Theory account of language and cognition draws upon and elaborates the stimulus equivalence phenomenon. As well as being able to discriminate (i.e., detect and respond to) specific stimuli, organisms are also capable of responding to relations *between* stimuli such as similarity, difference, distance, greater than, and so on. These forms of responding are known collectively as *relational responding* or *relational framing*. Nonverbal organisms are capable of learning to respond to such *formal* relations as size and distance, via traditional learning processes such as operant conditioning (see Reese, 1968). Verbal organisms, however, display the unique ability to respond to *arbitrary* stimulus relations such as oppositeness, value and time, that are not tied to the formal properties of the stimuli involved (e.g., a small coin can be worth *more than* a large coin). This form of responding is called *arbitrarily applied relational responding* (AARR; Hayes et al., 2001).

As seen in stimulus equivalence, verbal organisms can derive equivalences between stimuli that have never been explicitly matched. The different forms of *relational responding* (difference, opposition, greater than, less than, etc) can also be derived without explicit training (e.g., Dymond & Barnes, 1995; Lipkens, Hayes, & Hayes, 1993; Roche & Barnes, 1996). This leads us to *derived*

relational responding (DRR), perhaps the core behavior underpinning language and cognition. The following paragraphs describe the specific properties of DRR.

Mutual entailment means that if A is related to B, then B is related to A in a complimentary fashion. For example, if A is opposite to B, then B is opposite to A. If A is more than B, then B is less than A. *Combinatorial entailment* occurs when three or more stimuli are related. If A is opposite to B, and B is opposite to C, then the relation that is derived between A and C is one of equivalence, because both are opposite to B. Combinatorial entailment refers to the reciprocal relationships that exist between two stimuli as mediated by other intermediary stimuli (Blackledge, 2003).

The RFT account of attitudes begins to come into focus with the introduction of one final feature -*Transformation of function*. Expressed simply, transformation of function refers to the well documented fact that when two stimuli are related, the psychological functions of each stimulus changes the functions of the other, according to how the stimuli are related (see Dymond & Rehfeldt, 2000, for a review). For instance, if an individual shows a preference for a particular soft-drink labelled using a specific term, and that term in turn participates in a derived (i.e., untrained) relation with other sitmuli, a similar preference will also be shown for novel drinks labelled using those other stimili (Barnes-Holmes, Keane, Barnes-Holmes & Smeets, 2000; Smeets & Barnes-Holmes, 2003). In effect, the psychological response functions established for the original label by virtue of its being directly related to a preferred drink (i.e., salivation, appetitive thoughts, increased behavioral probability of drinking the liquid) transform the response functions of all indirectly related stimuli (e.g., other related labels and drinks).

A critical feature of the RFT approach to language is that the derivation of stimulus relations and the transformation of stimulus functions is entirely controlled by historical events and is predictable where control over trained stimulus relations is possible. Specifically, AARR emerges from the process of operant conditioning involved in such tasks as simply learning to name objects. For example, a child is trained to name an object out loud when presented with the object, and to orient towards the object when the object's name is spoken. After a number of object-name and name-object relations are trained, the generalized operant response class of "naming" is established in the presence of appropriate contextual cues such as the word "is", as employed in the utterances "This **is** your shoe" or "What **is** that?" The history of responding establishes the specific contextual cues for "naming", a form of relational responding (equivalence). Likewise, other forms of relational responding, both basic and derived, are established through multiple exemplar training under the control of environmental cues (usually words). In effect, no further process at the psychological level needs to be appealed to in order to understand how humans come to derive relations.

The ability of an individual to derive stimulus relations and transform stimulus functions accordingly need not be explained by such phenomena as "propositions" or insight". Rather, the process of arbitrarily applicable relational responding explains those phenomena. This is a radical departure in conceptual terms from the mainstream view that behavior is ultimately controlled from the inside out. However, this perspective brings with it a remarkable and parsimonious explanatory power that requires no mental representational constructs or the attendant obligation to study these rather than the original behavioral phenomenon of interest. Researchers can focus on researching the specific histories that lead to the emergence of the verbal behaviors associated with the attitude construct. The following section details the RFT model of attitudes, and the methodologies that emerged from that model parallel to the IAT and other social-cognitive implicit tests.

Measuring Attitudes as Histories of Verbal Behavior

The phenomena of transfer of function and relational responding combine to inform the RFT model of attitudes. From the behavior-analytic perspective, an attitude may now be conceived of as a history of derived and explicitly reinforced stimulus relations according to which the functions of events are transformed (e.g., Grey & Barnes, 1996). As such, an attitude can be thought of as a verbal event (or series thereof) which emerges from our interactions with others and with our environment across our personal learning history. Thus, an attitude is established and maintained through a history of both explicitly reinforced relations and untrained derived relations between verbal stimuli. The remote social contingencies which support these networks of relational responses can be represented by the verbal practices of the wider community (i.e., the culture with which the participant interacts). For instance, rules, norms, mores, and taboos all constitute forms of verbal contingency that specify relations between stimuli (e.g., "sex" and "dirty") in often complex and subtle (i.e., indirect) ways. A participant's past participation in a verbal environment (i.e., a culture) provides many hundreds of training exemplars that establish complex derived relational networks through which functions of stimuli may be transformed, and this can explain the emergence of apparently untrained or indirectly trained responses and attitudes. This occurs because of the way in which the various terms were framed relationally in language, whether explicitly, or in turn by further derived relations (e.g., innuendo, jokes; see Roche, Barnes-Holmes & Barnes-Holmes, 2002).

In an early example of a behavior-analytic study of attitudes, Kohlenberg, Hayes, and Hayes (1991) investigated a model of social stereotyping based on verbal control over equivalence classes. They suggested, for example, that in the sentence "the woman complained and complained", the word "woman" may serve as a contextual cue for relating the word "complain" to the words "*nag*" or "*bitch*". In a structurally similar sentence "the man complained and complained", the word "man" serves as a contextual cue which occasions relating "complain" with "*assertive*" or "*forceful*". In Experiment 1 of their study, participants were exposed to training to form six four-member equivalence classes using nonsense syllables as stimuli. The members of each class were contingent on a contextual cue – either a male or female name. That is, when a male name was present, responding to A1-B1-C1 and D1 as equivalent was reinforced, and when a female name was on screen, responding to A1-B1-C3 and D3 was reinforced. The experimenters then tested for derived equivalence relations using novel male and female names as contextual cues. Participants related the novel male and female names according to the trained derived relations despite having no experimental history of responding to those names. In effect this demonstrated that control over the trained equivalence classes had generalized through classes of by pre-existing contextual cues (i.e., gender classifications). The researchers had in effect demonstrated a means by which socially established classes based on gender could be assessed subtly and indirectly without engendering social desirability.

Experiment 2 brought the process under further control by generating pre-experimental equivalence classes of nonsense syllables, which were then used as the contextual cues in place of male and female names. Thus, the researchers modeled the process shown in experiment 1 with entirely laboratory-controlled histories (whereas Male and Female names would have had extensive pre-experimental histories which the experimenters could not know the details of). These experiments showed that contextual control could transfer without direct training from one stimulus to other stimuli with which it shares equivalence class membership. This process can be studied in a highly controlled manner, thus shedding light on

the types of histories which could produce social stereotyping behavior without any need for an appeal to an explanatory mental construct.

In another early study, Grey and Barnes (1996) suggested that a negative attitude towards normal heterosexual interactions can be seen as responding in accordance with an equivalence relation between normal opposite-sex adults and descriptive terms such as 'disgusting'. Their empirical study demonstrated a transformation of a trained attitudinal or evaluative response from one member of an equivalence class to the other members of the equivalence class. These researchers provided participants with the necessary conditional discrimination training to form the following three derived equivalence relations; A1-B1-C1, A2-B2-C2, and A3-B3-C3, using nonsense syllables as stimuli. Participants then viewed the video contents of VHS videocassettes that were clearly labeled with the A1 and A2 stimuli. One of the cassettes contained sexual/romantic scenes, while the other contained religiously themed scenes. Subsequently, participants were asked to categorize four novel videocassettes, each labeled as B1, C1, B2 or C2. They were given no information about these cassettes, but they categorized them according to the derived equivalence classes. That is, participants classified the B1 and C1 cassettes in the same way as the A1 cassette, and the B2 and C2 cassette in the same way as the A2 cassette. In effect, the study demonstrated the sexual and religious evaluative functions of the A-labeled cassettes and the derived relations in which the A stimuli participated, transformed the functions of the C-labeled cassettes, such that these were responded to as sexual or religious as appropriate. This study demonstrated a process by which evaluative responses (i.e. attitudes) can be indirectly trained and measured.

Dixon, Dymond, Rehfeldt, Roche, & Zlomke (2003) applied the transfer-of-function model to the understanding of attitude changes to Muslim men in the aftermath of the September 11th terrorist attacks in the USA. They suggested that on hearing of the terrorist (A) attack a white American male may instantly experience feelings of rage (B). The media may claim that Terrorists (A) are responsible for these acts, and pictures of these Terrorists may be shown on the television (C). In effect, an A-B and A-C relation has been established by direct media reports, but a derived relations account explains how images of the terrorists themselves may now come to elicit feelings of hate or rage. This occurs by virtue of a transfer of response function across the stimuli in the newly created relation (i.e., B to C). As the most salient features of the terrorists are their race, religion, and country of origin these feelings of hate and rage towards the terrorists begin to transfer to other persons sharing the same skin colour, religion, and country of origin because of a formal similarity between them and the terrorists (i.e., simple stimulus generalization). That is, innocent Muslims of a Middle Eastern descent are now responded to as equivalent to the A, B and C stimuli and the feelings of hate and rage felt by the hypothetical American may now have transferred beyond the terrorists themselves to all middle Eastern people. This process was modeled in the laboratory by Dixon et al. (2003).

The above studies showed the potential of an RFT paradigm to detail the processes involved in the establishment of behaviors that we might refer to as attitudes. RFT and stimulus equivalence researchers were able to leverage derived relational responding processes to experimentally study attitudes and to use these processes as part of a theoretical account, but one challenge still remained- to develop a method which would allow a researcher to detect those histories. This breakthrough occurred when researchers (Watt, Keenan, Barnes, & Cairns, 1991) began to investigate how a particular history of verbal behavior might interfere with the formation of new relational responses – and as such, reveal itself to the experimenter.

The Methodological Breakthrough: The Watt et al. Paradigm

In their seminal study, Watt et al. (1991) used a simple stimulus equivalence paradigm in which participants were trained to relate stimuli with strong socially established functions in ways that were inconsistent with their social history. Specifically, they took advantage of the fact that people in Northern Ireland often respond to names as indicative of religious background, and utilized stimuli representative of Catholic and Protestant names and symbols.

A three-phase experimental procedure was employed. Participants were first exposed to a matching to sample procedure comprising the presentation of either a nonsense syllable or a first and last name at the top of the screen (the "sample" stimulus). Three "comparison" stimuli were displayed separately below. Participants were instructed to select a comparison stimulus by pressing a corresponding key. Training comprised of one of three Catholic names being randomly chosen to serve as the sample stimulus. Beneath this, three nonsense syllables served as comparison stimuli and were arranged in a random order across the screen. Participants were required to select the correct comparison in the presence of the sample stimulus (A-B Relations). The second stage trained B-C relations. Here, the sample stimuli were selected from the list of nonsense syllables, and the comparison stimuli were selected from the list of Protestant symbols. Feedback was provided on all trials during Stage 1. During stage 2, corrective feedback was presented following 50% of responses. The stimulus combinations described in Stage 1 were all presented in random order during this condition. Each stimulus combination was presented twice and participants were required to meet a 100% criterion.

Stage 3 of the Watt et al. procedure involved Equivalence Testing. For this stage, no corrective feedback was provided. Ten presentations of each of the stimulus combinations from Stage 1 were randomly presented. Interspersed with these were ten presentations each of six other stimulus combinations. Each of the three Protestant symbols served as sample stimuli and two of the Catholic names served as comparison stimuli. An additional Protestant name was included as a comparison stimulus for each of these three combinations of sample and comparison stimuli. The results of the Watt et al. study showed that during equivalence testing, all of the English participants correctly matched the Catholic names with the Protestant symbols, but 12 of the 19 Northern Irish participants chose a novel Protestant name in the presence of the Protestant symbols, thereby failing to derive the equivalence relations that the procedure usually occasions. These findings strongly suggested that the social contingencies operating in Northern Ireland interfered with the establishment of equivalence relations in the laboratory. The equivalence test required Northern Irish participants to juxtapose names and symbols in a manner that was counter-cultural for this group of participants. As such, the Watt et al. procedure had hit upon a basic methodology for inferring the social histories of participants, without using a direct questioning approach. It was, in effect, an early behavior-analytic implicit test.

Subsequent studies supported the suggestion made by Watt et al. that social history interferes with the formation of equivalence classes and that this phenomenon could be used to indirectly assess participants' personal and social histories. For instance, in a study on gender identity, Moxon and Keenan (1993) found that participants had more difficulty forming equivalence classes when the classes included female names and stereotypic male occupations. Leslie, Tierney, Robinson, Keenan, Watt, and Barnes (1993) also employed the Watt et al. procedure in a study designed to differentiate between anxious and non-anxious participants. They found that anxious participants had more difficulty in matching pleasant-state adjectives to threatening situation descriptors than did control participants. In another study,

Merwin and Wilson (2005) required participants to form equivalence classes between self-referential terms and negative evaluations, and in a second procedure, between those terms and positive evaluations. Participants who reported high distress and low esteem made significantly more errors when required to match "self" terms with positive items. Plaud (1995) showed that participants required significantly more training to derive equivalence classes made up of aversive stimuli (snake-related words) than classes made up of innocuous stimuli (flower related words). Importantly, the increase in the amount of training required to establish the equivalence classes correlated with self-reported fear of snakes.

These studies supported the assertion that pre-experimental functions of stimuli could interfere with the formation of equivalence classes in the laboratory. However, in order to gain complete experimental control over this effect it was also necessary for researchers to create their own stimuli and to establish psychological functions for those stimuli using respondent and operant learning methods. Only in this way could the process by which individual words (i.e., stimuli with conditioned evaluative functions) interfere with class formation (derived or trained) be fully understood in functional, rather than theoretical, terms

The first study to address this issue was reported by Roche, Barnes, and Smeets (1997). The experimenters trained participants on a matching to sample procedure that formed two three-member equivalence classes using nonsense syllables as stimuli (i.e. A1-B1-C1 and A2-B2-C2, where the A-C linkages are derived). The authors then paired two of the stimuli (A1 and C2) with sexually explicit film clips and two other stimuli (A2 and C1) with nonsexual film clips, establishing conflicting sexual response functions for stimuli that had been trained in the laboratory as equivalent. When re-exposed to the equivalence testing, participants reproduced the original equivalence relations (A1-C1 and A2-C2) and not the newly created and incongruous functional stimulus classes (A1-C2 and A2-C1). In a second experiment, Roche et al. firstly established the sexual and nonsexual functions for the A1/C2 and A2 / C2 pair respectively, and then presented participants with the matching-to-sample procedure. In this case, participants matched stimuli based on their conditioned sexual/non-sexual functions (i.e., A1 with C2 and A2 with C1) rather than forming the equivalence classes required by the matching to sample procedure. This demonstrated that once an equivalence class is formed, it is difficult to disrupt with succeeding functional relations, and that the reverse is also true; it is difficult for participants to form equivalence relations when they are incongruous with existing functional relations.

This issue was further examined by Tyndall, Roche and James (2004). They established two functional classes of stimuli; Six S+ stimuli (responding to the stimulus was reinforced) and Six S- stimuli (responding away from the stimuli was reinforced). In matching to sample training, participants were trained to form two three-member equivalence classes using four different combinations of S+/S- stimuli. Participants required more training to establish two distinct equivalence classes from amongst 6 S+ stimuli (i.e., stimuli sharing the same response function) than from amongst 6 S- stimuli (i.e., without a shared response function). Further, participants found it easier to form equivalence classes when they were required to separate S+ and S- stimuli than when they were required to form classes which mixed S+ and S- stimuli.

While in the 2004 paper, the stimulus functions were emotionally neutral, Tyndall, Roche & James 2009 later studied the process using emotive stimuli. They created these stimuli by pairing six stimuli with aversive images and six further stimuli with neutral images, as an analog of everyday evaluative experiences often studied in attitude research. They then tested for the formation of those functional classes. In equivalence training and testing, the authors found that

participants required significantly more training to establish two three-member equivalence classes from amongst the six aversive stimuli than from amongst the six neutral stimuli. In effect, Tyndall et al. (2004, 2009) more clearly demonstrated that learning tasks involving the formation of new stimulus relations can be used as indicators of the pre-existing relations between those stimuli.

The studies above were conducted at the same time that the IAT emerged as the preeminent Implicit Test. However, this research thread was developed independently from the social cognitive paradigm, with little or no contact between the two. In recent years, however, RFT researchers have begun to apply the lessons gleaned from this research to modeling performances on the IAT in terms of relational responding, and to develop testing methodologies that utilize insights from both threads of research. In the following section, the authors will outline a behavioral model of the IAT, before moving on to detail two new implicit testing methodologies which are built upon functional foundations.

Connecting the Threads: A Functional Account of the Implicit Association Test

The Watt et al. procedure showed that a functional process-based account of attitudes could be harnessed to create a subtle test for the verbal histories that underpin the attitude construct, at least from a behavioral perspective. However, the Watt et al. procedure was developed prior to the IAT and the widespread acceptance of the social cognitive / associationistic approach to attitudes. Their definition of attitudes did not make contact with the yet-to-emerge phenomenon of implicit attitudes. Indeed, from a behavioral perspective, the term "implicit" may not merit a unique account. It likely refers only to the relative speed of a response and therefore the relative probability that the response was mediated by further verbal behavior, usually private. The processes of producing verbal responses that have been mediated by further private verbal responses (such as "I had better not say that" is) what behavior analysts refer to as thinking (or more technically behavior-behavior relations). It does not in itself constitute a special process different to other forms of relational responding or simple speaking. In effect, we take the word "implicit" to refer to the observation that the contingencies controlling the relevant responding are not discriminable by the participant to a sufficient degree to allow a conscious (i.e., verbally-mediated) alteration of the response so that it is relationally consistent with anything other than the stimulus presented during the implicit testing task. Put simply, the participant would require more time during implicit test trials to observe the stimulus, respond to it privately, discriminate that response privately along with the controlling source as a specific prior history of trained or derived stimulus associations, discriminate that this history is socially undesirable, and then alter the response consciously in a manner referred to loosely as an example of social desirability. Rather, under time pressure the participant likely simply responds directly to the stimulus on screen, without additional mediating sequences of private verbal behavior.

From within a functional Relational Frame approach, the IAT is viewed as a measure of an individual's verbal history vis-à-vis an assessment of their ability to form stimulus relations under time constraints. IAT effects are conceived in terms of participants' fluency with the relevant verbal categories and their degree of experience at juxtaposing members of those verbal categories. For instance, an individual who has many dealings with people of a specific race, and has encountered both pleasant and unpleasant individuals from this racial group, will likely find it easy to juxtapose racial and evaluative terms in an IAT according to the test rules across the two test blocks. Such an individual will show no IAT effect (i.e., no response time or accuracy differential across the test blocks). On the other hand, if they have expe-

rienced mostly unpleasant individuals from one racial group or other, the juxtaposition of response rules across the IAT blocks will likely expose a fluency differential across those two blocks (i.e., an IAT effect; see the behavioral account of Roche, Ruiz, O'Riordan & Hand, 2005).

The behavioral model of the IAT was tested empirically by Gavin et al. (2008) using nonsense syllables as stimuli and experimentally produced derived relations between them as laboratory analogs of verbal relations between words in the vernacular. Two equivalence relations were established in the usual manner, leading to the two classes of nonsense syllables, labeled here as A1-B1-C1 and A2-B2-C2, where the A-C relations were derived, not reinforced. An IAT-type test was then administered to measure participants' ability to respond in the same way to common class member pairs (e.g., A1 and C1) compared to cross-class pairs (e.g., A1 and C2). Not surprisingly, more errors were made in responding under rule conditions in which a common response was required for incompatible, compared to compatible stimuli. In effect, typical IAT outcomes were generated using only directly manipulated variables and laboratory created stimuli, and in purely functional terms.

These entirely laboratory produced IAT effects were subsequently shown to be manipulable vis-à-vis reversals of some of the baseline relations underlying the derived equivalence relations (Ridgeway et al., 2010). Such findings strengthen the behavior-analytic position that the IAT test format is sensitive to a participant's history of relating the test stimuli, perhaps even including histories of stimulus relating that a participant would wish to conceal. However, they also shed new light on the fundamental processes that may entirely underpin the IAT effect and demonstrate that informative research can be done on implicit test processes from a functional perspective, without necessary recourse to associationistic explanatory mechanisms. For the cognitive researcher, these types of studies provide useful information that can inform the ongoing debate as to the form of the implicit attitude construct, opening up a wider array of explanations that go beyond the underlying associative narrative that informs most theorizing on the subject (Hughes et al., 2011). Further, this research thread has birthed two new functionally oriented implicit tests which are built on the insights gained through functional research, meaning that their core processes are well articulated and supported by a wealth of basic research.

Functional Implicit Tests

The Implicit Relational Assessment Procedure

The behavioral methodologies described thus far were focused largely on identifying behavioral processes and were prototypical in nature rather than designed for use in real world studies of attitudes. In recent years, two distinct behavioral implicit tests have been developed for real world application. The first of these is the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2006). This test is in many ways procedurally similar to the IAT. However, each trial of the IRAP displays two stimuli (rather than just one) on screen (e.g., "Child" and "Sexual") along with a contextual cue that specifies the relation between the two stimuli (e.g., "Same" or "Opposite."). The participant is required to respond quickly to this resulting relational statement ("Child SAME Sexual" or "Child OPPOSITE Sexual") with a key press that corresponds to one of two response options (e.g., for "TRUE" press z, for "FALSE" press m). Feedback is only presented if the response in incorrect as defined by the block rules (a red X is displayed) or if the response latency is lower than the stated criterion (the words "too slow" are displayed). Like the IAT, trials in the IRAP are organized into blocks. In one type of block, the responses defined

as correct are those that are *consistent* with social norms (e.g. Child OPPOSITE Sexual – TRUE = Correct) while the other block requires responses that are *inconsistent* with social norms.

A recent IRAP study by Dawson and Barnes-Holmes (2009) tested for pre-existing child-sex stimulus relations in a child sex offender population. Sixteen participants who had been convicted of a contact sexual offence against a child (the offender group) and sixteen male non-offenders recruited from a college population (the control group) completed an IRAP procedure and a Cognitive Distortion Scale (CDS; Gannon, 2009). The IRAP stimuli consisted of two category labels ("Child" and "Adult") and two sets of target stimuli ("sexual" words and "non-sexual" words). During the consistent blocks, participants were required to respond with "true" to the relations "Adult – Sexual" and "Child – Nonsexual" while in the inconsistent blocks participants were required to respond in the opposite way. The IRAP successfully detected a difference in response time differentials across the control group and the offender group. Furthermore, the IRAP was able to identify the specific relation on which the two groups differed. Specifically, on Child-Sexual trials, the offender group did not show a significant IRAP effect, responding equally quickly to Child-Sexual-False and Child-Sexual-True, whereas non-offenders tended to show a reaction time differential across these blocks.

Dermot Barnes-Holmes and colleagues have devised a behavior-analytic and functional account of the IRAP and related implicit test effects based on Relational Frame Theory, which they call the Relational Elaboration Coherence Model (REC; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). It is a more formally elaborated account than the loose working definition of implicitness provided above and first proposed by Roche et al. (2005). According to the REC model, each individual trial on the IRAP produces an immediate and brief response to the relation presented before the participant presses a response key. The probability of this initial response is a function of participants' verbal and non-verbal history with the stimuli and the current contextual cues (i.e., the relational stimulus, such as the word *opposite* on screen). The most probable response will likely be emitted first. If this immediate response coheres with the response required by the current IRAP trial rule, then the response latency will be lower. If the required response is in opposition with the participants' immediate relational response, then the correct (as defined by the contingencies of the specific trial) response will be emitted more slowly. Across multiple trials, the average latency on inconsistent trials will be higher than the average latency for consistent trials.

The foregoing provides a basic explanation for the IRAP effect, but the REC model extends the explanation to account for why implicit measures and explicit questionnaire methods so often diverge in their results. More specifically, when completing questionnaires or other so called "explicit" measures of attitudes, the participant is under little time pressure and can therefore engage in complex and extended relational responding (i.e., thinking) which allows them to produce a response which is coherent with other responses in their behavioral repertoire (see Barnes-Holmes, Hayes & Dymond, 2001) such as; "it is wrong to categorize children as sexual." It is also possible under these circumstances to produce a response that coheres with the social expectations of others. However, when exposed to the IRAP (or other implicit test) procedure, participants are under significant pressure to respond quickly (commonly within 2000 ms) and, therefore have little time to engage in the elaborate private relational responding necessary to produce alternative socially desirable responses. In effect, the most likely responses under time constraint conditions are those that are immediate and brief and therefore direct measures of history, unmediated by local relational activity.

The IRAP has a well worked out functional explanatory account accompanying its procedure,

and addresses the concerns of the experimental analyst of behavior in how it approaches the matter of implicit testing. Nevertheless, the test procedure is sufficiently intensive for the participant that rapid administrations of the test are not possible. This is because the test consists of a practice block, several re-administrations of test blocks, and a relatively high rate of subject attrition due to a failure to reach fluency criteria within the test itself. However, a further behavior-analytic test has been developed to address this and other concerns. Specifically the Function Acquisition Speed Test (FAST; O'Reilly, Roche, Ruiz, Tyndall & Gavin, 2012; O'Reilly, Roche, Ryan & Campion, 2013) was developed within a functional research paradigm but with the express purpose of being easy and fast to administer. The test also addresses the problem of the relative nature of biases and preferences as indexed by the IAT and the IRAP. Before this issue is addressed, however, a brief outline of the FAST methodology is merited.

The Function Acquisition Speed Test (FAST)

The FAST requires participants to complete a number of simple discrimination blocks with minimal instructions, and no instructions at all relating to appropriate responses. Each trial presents a single stimulus, and participants are required to learn, via trial and error and corrective feedback, whether to respond with a "left" key press (e.g., press "a" or press "z") or a "right" key press (e.g. press "j" or press "m"). For two of the stimuli, a "left" key press is reinforced, while for the other two, a "right" key press is reinforced. Participants are required to continue the block until their responses are fluent; that is both rapid and accurate as defined by the production of a specific succession of correct responses (usually 10). As a result there is no predetermined block length. Blocks are completed when response fluency has been achieved. Feedback is presented following all trials and there is no error correction procedure. Response speed is constrained not by instruction but by a finite response window (usually 3s). There are no instructions or response rules, other than to produce as many correct responses as possible. The metric of interest is the differential in the number of trials required to reach response fluency across two critical test blocks.

A FAST contains two "test blocks", each of which uses the same four stimuli. Two stimuli are the stimuli of interest, suspected to be related in the participant's history. The other two are novel, unrelated stimuli. In this regard the FAST is not unlike a single target IAT. In the "consistent" block, the same response is reinforced (e.g., press z) for both of the target stimuli of interest, while another response (e.g., press m) is reinforced for the novel innocuous stimuli. These responses are consistent with the participant's learning history, and so should quickly result in stable, high rate responding by the participant. In the "inconsistent block", the reinforced responses are inconsistent with the participant's learning history insofar as different responses ("press z" and "press m") are required for each of two (thought-to-be-related) target stimuli. In effect, the juxtaposition of current and past reinforcement contingencies during the inconsistent block functions as a type of learning *disrupter*. The current learning contingencies need to overcome the "behavioral inertia" of the pre-experimental contingencies that maintain a specific stimulus relation in order for fluent responding to be produced (i.e., learning). In simple terms, the "inconsistent" response classes are more difficult for participants to acquire. The differences in rates of response class acquisition across the two test blocks is used to index the pre-existing "strength" of the relation under investigation.

Critically, the FAST also contains a "baseline" block, in which all four stimuli are novel and unrelated (e.g., nonsense syllables). The purpose of this block is to measure the rate of acquisition of an arbitrary response class under test conditions but without any interfering effect from pre-experimental contingencies. The number

of trials to criterion on the baseline block places differences in acquisition speed across the main FAST blocks in context. For example, for a participant who completes such tasks very quickly, a small difference in acquisition rates between the inconsistent and consistent test blocks is more meaningful that the same raw fluency difference is for a participant with large baseline trial requirements. This relationship between differences in learning speed across critical FAST blocks, and baseline acquisition speed is the basis of a Strength of Relation index used to quantify the "strength" of a pre-existing relation between the stimuli of interest.

The FAST has been shown to be capable of detecting relations between laboratory created stimuli and relations (O'Reilly et al., 2012) as well as derived relations (O'Reilly et al, 2013). It has also been used with real world stimuli to measure categorization of teenaged females as sexual (see Roche et al., 2012), but is still in early stages of development.

One exciting feature of the FAST worth outlining at this stage relates to its circumnavigation of the problem of relativism in inferences regarding attitudinal biases. Specifically, the IAT provides only a relative measure of association strength (De Houwer, 2002, Greenwald, Nosek & Banaji, 2004). That is, an IAT trial involves both responding towards one category while responding away from the other, and thus it is impossible to determine which of these response forms is at work in controlling the participant's correct response rate differentials across test blocks, or the extent to which one or the other contributes to the total effect. More specifically, given the result of a race IAT, only the question; "Is White preferred to Black" is answered. Even though the bias may be reliably recorded it remains unclear whether or not the bias is characterized by a participant having a history of pro-white responding or a history of anti-black responding. In effect, the structure of the IAT (and the IRAP) requires that the attitude being studied be stated in symmetrical terms. While this provides a useful index of whether an attitude is present for a participant, the specific pattern of relational responding that characterizes the attitude may be crucial to the development of interventions to change those attitudes.

CONCLUSION

While there remains much to be done in the development of the FAST, its development and that of the IRAP illustrate that behavior analysis has much to offer the field of implicit testing by way of ground-up functional accounts of the behavioral processes relevant to attitude measurement and in the development of appropriate methodologies to this end. The functional account of attitudes detailed in this chapter led directly to these new methodologies, and the understanding of the processes leveraged by these tests means that strong predictions can be made and empirically tested, thus rapidly advancing knowledge with regard to the conditions under which attitudes, implicit and otherwise, are formed and maintained, and how they can be changed. The functional approach offers the hope of providing an account that renders explicit those very processes social cognitivists refer to as implicit.

REFERENCES

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of Social Psychology*. Worcester, MA: Clark University Press.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*, 527–542.

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record*, *58*, 497–516.

Barnes-Holmes, D., Hayes, S. C., & Dymond, S. (2001). Self and self-directed rules. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of language and cognition* (pp. 119–139). New York: Plenum Press.

Barnes-Holmes, D., Keane, J., Barnes-Holmes, Y., & Smeets, P. M. (2000). A derived transfer of emotive functions as a means of establishing differential preferences for soft drinks. *The Psychological Record*, *50*, 493–511.

Blackledge, J. T. (2003). An introduction to relational frame theory: Basics and applications. *Behavior Analyst Today*, *3*(4), 421–433.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192–212. doi:10.1016/j.jesp.2005.07.003

Carpenter, F., Smeets, P. M., & Barnes-Holmes, D. (2003). Equivalence-equivalence as a model of analogy: Further analyses. *The Psychological Record*, *53*, 349–371.

Cassidy, S., Roche, B., & Hayes, S. C. (2011). A relational frame training intervention to raise Intelligence Quotients: A pilot study. *The Psychological Record*.

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. *Sexual Abuse*, *21*(1), 57–75. PMID:19218478

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, *50*(2), 77–85. doi:10.1026//1618-3169.50.2.77 PMID:12693192

De Houwer, J. (2006a). Using the Implicit Association Test does not rules out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, 1–12.

De Houwer, J. (2006b). What are implicit measures and why are we using them. In The handbook of implicit cognition and addiction. Academic Press.

De Houwer, J. (2011). Why the Cognitive Approach in Psychology Would Profit From a Functional Approach and Vice Versa (2011). *Perspectives on Psychological Science*, *6*(2), 202–209. doi:10.1177/1745691611400238

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology*, *43*(6), 972–978. doi:10.1016/j.jesp.2006.10.007

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Theoretical claims necessitate basic research: Reply to Gawronski, Lebel, Peters, and Banse (2009) and Nosek and Greenwald (2009). *Psychological Bulletin*, *135*(3), 377–379. doi:10.1037/a0015328 PMID:19379021

Devos, T., & Banaji, M. R. (2005). American= white? *Journal of Personality and Social Psychology*, *88*(3), 447–466. doi:10.1037/0022-3514.88.3.447 PMID:15740439

Dixon, M. R., Dymond, S., Rehfeldt, R. A., Roche, B., & Zlomke, K. R. (2003). Terrorism and relational frame theory. *Behavior and Social Issues*, *12*(2), 129. doi:10.5210/bsi.v12i2.40

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity & Ethnic Minority Psychology*, *8*(2), 88–102. doi:10.1037/1099-9809.8.2.88 PMID:11987594

Dymond, S., & Barnes, D. (1995). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more-than, and less-than. *Journal of the Experimental Analysis of Behavior*, *64*(2), 163–184. doi:10.1901/jeab.1995.64-163 PMID:16812766

Dymond, S., & Rehfeldt, R. A. (2000). Understanding complex behavior: The transformation of stimulus functions. *The Behavior Analyst*, *23*(2), 239. PMID:22478349

Dymond, S., & Roche, B. (Eds.). (2013). *Advances in relational frame theory: Research and application*. Oakland, CA: New Harbinger.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. doi:10.1037/0022-3514.69.6.1013 PMID:8531054

Fazio, R. H., & Olson, M. A. (2003). Attitudes: Foundations, functions, and consequences. In M. A. Hogg & J. Cooper (Eds.), *The handbook of social psychology* (pp. 139–160). London: Sage.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the 'I', the 'A', and the 'T': A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*(1), 74–147. doi:10.1080/10463280600681248

Gavin, A., Roche, B., & Ruiz, M. (2008). *A behaviorally modifed and functionally transparent Implicit Association Test to measure levels of culturally appropriate sexual categorization of children among a sample of normal males and females*. Academic Press.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181–193. doi:10.1111/j.1745-6916.2007.00036.x

Gawronski, B., LeBel, E. P., Peters, K. R., & Banse, R. (2009). Methodological issues in the validation of implicit measures: Comment on De Houwer, Teige-Mocigemba, Spruyt, and Moors (2009). *Psychological Bulletin*, *135*(3), 369–372. doi:10.1037/a0014820 PMID:19379019

Greenwald, A., Nosek, B., & Banaji, M. (2003). Understanding and using the Implicit Association Test: I: An improved scoring algorithm. *Journal of Personality and Social Psychology*.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, *102*(1), 4–27. doi:10.1037/0033-295X.102.1.4 PMID:7878162

Greenwald, A. G., McGhee, D. E., Schwartz, J. L. K., & Attitudes, M. I. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality*, *74*(6), 1464–1480. PMID:9654756

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). *Validity of the salience asymmetry interpretation of the implicit association test: Comment on Rothermund and Wentura (2004)*. Academic Press.

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

Heagle, A. I., & Rehfeldt, R. A. (2006). Teaching perspective-taking skills to typically developing children through derived relational responding. *Journal of Early and Intensive Behavior Intervention*, *3*, 1–34.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2004). *A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures*. Unpublished manuscript, University of Trier, Germany.

Houwer, J. D. (2002). The Implicit Association Test as a tool for studying dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behavior Therapy and Experimental Psychiatry*, *33*(2), 115–133. doi:10.1016/S0005-7916(02)00024-1 PMID:12472175

Hughes, S., & Barnes-Holmes, D. (2011). The Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives. *The Psychological Record*, *61*, 465–496.

Kohlenberg, B. S., Hayes, S. C., & Hayes, L. J. (1991). The transfer of contextual control over equivalence classes through equivalence classes: A possible model of social stereotyping. *Journal of the Experimental Analysis of Behavior*, *56*(3), 505–518. doi:10.1901/jeab.1991.56-505 PMID:1774542

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV: Procedures and validity. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 59–102). New York: Guilford Press.

Leslie, J. C., Tierney, K. J., Robinson, C. P., & Keenan, M. et al. (1993). Differences between clinically anxious and non-anxious subjects in a stimulus equivalence training task involving threat words. *The Psychological Record*, *43*, 153–161.

Lipkens, R., Hayes, S. C., & Hayes, L. J. (1993). Longitudinal study of the development of derived relations in an infant. *Journal of Experimental Child Psychology*, *56*(2), 201–239. doi:10.1006/jecp.1993.1032 PMID:8245768

Lowe, C. F. (1979). Determinants of human operant behavior. In M. D. Zeiler & P. Harzem (Eds.), Advances in analysis of behavior: Reinforcement and the Organization of behavior (vol. 1, pp. 159-192). Chichester, UK: Wiley.

McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, *94*(5), 792–807. doi:10.1037/0022-3514.94.5.792 PMID:18444739

McHugh, L., Barnes-Holmes, Y., & Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *The Psychological Record*, *54*(1), 115–144.

Merwin, R. M., & Wilson, K. G. (2005). Preliminary findings on the effects of self-referring and evaluative stimuli on stimulus equivalence class formation. *The Psychological Record*, *55*, 561–575.

Moran, L., Stewart, I., McElwee, J., & Ming, S. (2010). Brief report: The training and assessment of relational precursors and abilities (TARPA): A preliminary analysis. *Journal of Autism and Developmental Disorders*, *40*(9), 1149–1153. doi:10.1007/s10803-010-0968-0 PMID:20151185

Moxon, P., Keenan, M., & Hine, L. (1993). Gender-role stereotyping and stimulus equivalence. *The Psychological Record*, *43*(3), 381–393.

Murphy, C., & Barnes-Holmes, D. (2009). Establishing derived manding for specific amounts with three children: An attempt at synthesizing Skinner's *Verbal Behavior* with relational frame theory. *The Psychological Record*, *59*, 75–91.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social cognition, 19*(6), 625–666. Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*(4), 553–594. doi:10.1080/02699930701438186

O'Reilly, A., Roche, B., Ruiz, M., Ryan, A., & Campion, G. (in press). A Function Acquisition Speed Test for equivalence relations (FASTER). *The Psychological Record*.

O'Reilly, A., Roche, B., Ruiz, M. R., Tyndall, I., & Gavin, A. (2012). The Function Acquisition Speed Test (FAST): A behavior-analytic implicit test for assessing stimulus relations. *The Psychological Record*, *62*, 507–528.

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.

O'Toole, C., Barnes-Holmes, D., & Smyth, S. (2007). A derived transfer of functions and the Implicit Association Test. *Journal of the Experimental Analysis of Behavior*, *88*(2), 263–283. doi:10.1901/jeab.2007.76-06 PMID:17970419

Olsen, J. M., & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology*, *44*(1), 117–154. doi:10.1146/annurev.ps.44.020193.001001

Olson, M. A., & Kendrick, R. V. (2008). Origins of attitudes. *Attitudes and Persuasion*, 111-130.

Plaud, J. J. (1995). The generalized expectancy bias: An explanatory enigma. *Behavioral and Brain Sciences*, *18*(02), 311–312. doi:10.1017/S0140525X00038644

Reese, H. W. (1968). *The perception of stimulus relations: Discrimination learning and transposition*. New York: Academic Press.

Rehfeldt, R. A., Dillen, J. E., Ziomek, M. M., & Kowalchuk, R. K. (2007). Assessing Relational Learning Deficits in Perspective-Taking in Children with High-Functioning Autism Spectrum Disorder. *The Psychological Record*, *57*, 23.

Ridgeway, I., Roche, B., Gavin, A., & Ruiz, M. R. (2010). Establishing and eliminating IAT effects in the laboratory: Extending a behavioral model of the Implicit Association Test. *European Journal of Behavior Analysis*, *11*, 133–150.

Roche, B., & Barnes, D. (1996). Arbitrarily applicable relational responding and sexual categorization: A critical test of the derived difference relation. *The Psychological Record*, *46*, 451–475.

Roche, B., Barnes, D., & Smeets, P. M. (1997). Incongruous stimulus pairing and conditional discrimination training: Effects on relational responding. *Journal of the Experimental Analysis of Behavior*, *68*(2), 143–160. doi:10.1901/jeab.1997.68-143 PMID:16812852

Roche, B., Barnes-Holmes, Y., Barnes-Holmes, D., Stewart, I., & O'Hora, D. (2002). Relational Frame Theory: A new paradigm for the analysis of social behavior. *The Behavior Analyst*, *25*, 75–91. PMID:22478379

Roche, B., O'Reilly, A., Gavin, A., Ruiz, M., & Arancibia, G. (2012). Using behavior-analytic implicit tests to assess sexual interests among normal and sex-offender populations. *Socioaffective Neuroscience and Psychology*, *2*(0), 17335. doi:10.3402/snp.v2i0.17335 PMID:24693346

Roche, B., O'Riordan, M., Ruiz, M., & Hand, K. (2005). A relational frame approach to the psychological assessment of sex offenders. In M. Taylor & E. Quayle (Eds.), *Viewing Child Pornography on the Internet: Understanding the Offence, managing the Offender, and Helping the Victims* (pp. 109–125). Dorset: Russell House Publishing.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology. General*, *133*(2), 139–165. doi:10.1037/0096-3445.133.2.139 PMID:15149248

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research*, *14*(1), 5–13. doi:10.1044/jshr.1401.05 PMID:5550631

Sidman, M. (1986). Functional analysis of emergent verbal classes. In T. Thompson & M. E. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 213–245). Hillsdale, NJ: Laurence Erlbaum Associates.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston: Authors Cooperative.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. doi:10.1901/jeab.1982.37-5 PMID:7057129

Skinner, B. F. (1957). *Verbal Behavior*. Acton, MA: Copley Publishing Group. doi:10.1037/11256-000

Smeets, P. M., & Barnes-Holmes, D. (2003). Children's emergent preferences for soft drinks: Stimulus-equivalence and transfer. *Journal of Economic Psychology*, *24*(5), 603–618. doi:10.1016/S0167-4870(03)00004-7

Stewart, I., Barnes-Holmes, D., & Roche, B. (2004). A Functional-Analytic Model of Analogy Using the Relational Evaluation Procedure. *The Psychological Record*, *54*, 531.

Stewart, I., Barnes-Holmes, D., & Weil, T. (2009). *Training analogical reasoning as relational responding. Derived relational responding: Applications for learners with autism and other developmental disabilities*. Oakland, CA: New Harbinger.

Tyndall, I. T., Roche, B., & James, J. E. (2004). The relation between stimulus function and equivalence class formation. *Journal of the Experimental Analysis of Behavior*, *81*(3), 257–266. doi:10.1901/jeab.2004.81-257 PMID:15357509

Tyndall, I. T., Roche, B., & James, J. E. (2009). The interfering effect of emotional stimulus functions on stimulus equivalence class formation: Implications for the understanding and treatment of anxiety. *European Journal of Behavior Analysis*, *10*, 121–140.

Villatte, M., Monest ès, J. L., McHugh, L., Frei xa i Baqu é, E., & Loas, G. (2010b). Assessing perspective taking in schizophrenia using relational frame theory. *The Psychological Record*, *60*, 413–436.

Villatte, M., Monestès, J.-L., McHugh, L., Freixa i Baqué, E., & Loas, G. (2010a). Adopting the perspective of another in belief attribution: Contribution of relational frame theory to the understanding of impairments in schizophrenia. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(2), 125–134. doi:10.1016/j.jbtep.2009.11.004 PMID:20034611

Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorisation and stimulus equivalence. *The Psychological Record*, *41*, 33–50.

Weil, T. M., Hayes, S. C., & Capurro, P. (2011). Establishing a deictic relational repertoire in young children. *The Psychological Record*, *61*, 5.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. doi:10.1037/0022-3514.72.2.262 PMID:9107001

KEY TERMS AND DEFINITIONS

Attitude: An attitude is usually defined as being a combination of cognitive (i.e. propositional) and affective evaluations of an object with a variable strength.

Conditional Discrimination: A discrimination between stimuli, in which the reinforcement of responding to one stimulus (e.g., a choice from an array) is conditional upon the presence of a further stimulus (i.e., the conditional stimulus).

Construct: A hypothetical construct is an explanatory variable that is not directly observable, but which nevertheless has heuristic value. For example, the concepts of *intelligence* and *motivation* are used to explain phenomena in psychology, but neither is directly observable or deemed to be extant entities.

Functional Analysis: An analysis of behavior in terms of its antecedents, consequences and over-arching context.

Operant Conditioning: A process by which behavior rate is altered as a function of delivered consequences (appetitive or aversive) that immediately follow responses according to a known schedule.

Relational Responding: Responding to relations *between* stimuli rather than to the stimuli in isolation. Example of relations include; similarity, difference, and temporal sequence.

Stimulus Equivalence: A phenomenon whereby untrained relations between at least three stimuli emerge spontaneously and bear the defining characteristics of reflexivity, symmetry, and transitivity. Reflexivity refers to identity matching (e.g., *A is A*); symmetry refers to functional reversibility (e.g., given *A is B, B is A* is derived spontaneously); and, transitivity refers to the functional equivalence of stimuli related to each other only via a third stimulus (e.g., given *A is B* and *B is C*, *A is C* is derived spontaneously).