Chapter 14: Implicit Cognition and Social Behavior

Dermot Barnes-Holmes, Colin Harte* and Ciara McEnteggart*

Ghent University, Belgium

The human species is composed of relatively weak and slow moving primates and yet we have, in one sense, come to dominate the planet in a very short period of time. Arguably, the propelling force behind our rise in dominance is our ability to learn how to relate stimuli and events in increasingly abstract, or arbitrarily applicable, ways. Irrespective of how this ability evolved so strongly in our species (see Hayes & Sanford, 2014), it appears to lie at the core of human language and cognition (Hayes, Barnes-Holmes, & Roche, 2001). On the one hand, the human ability to engage in abstract relational responding has been the key to our success in developing modern civilization with increasingly sophisticated sciences and technology. On the other hand, abstract relational responding allows humans to categorize other members of their own species into in- and out-groups and to react to members of an out group in a highly negative manner without having a direct negative experience with any member of that group. Such prejudicial behaviors appear to be widespread and common among the human species. Indeed, at the time of writing we had just witnessed populist appeal for literally building a wall along the border between the United States of America and Mexico and the withdrawal of the United Kingdom from the European Union to allow for greater control over its borders. And in the last 100 years of human history we have witnessed two world wars in which millions were slaughtered based, at least in part, on the human ability to categorize outgroups as sufficiently threatening to warrant warfare and in extremist attempts at mass genocide.

The ability to categorize other members of our species into in- and out-groups may have served us well as small groups of hunter-gatherers, because it supported necessary

* These authors contributed equally to the chapter

cooperation within a group against another in competing for potentially limited resources. However, when human groups become increasingly abstract and increase in size (e.g., to nation states) and the level of science and technology allows for the annihilation of literally millions at the press of a button, the evolutionary value of prejudicial behavior, or social categorization, seems completely lost. Indeed, one could argue that in the modern context of a genuinely globalized world there is only one in-group – the entire human species and even more broadly life itself on this planet.

Clearly, the human ability to engage in social categorization is a potentially lethal behavior that we need to study and to understand if we are to predict and influence it in a manner that will serve to protect life on this planet. The science of psychology, and social psychology in particular, has devoted considerable time and effort in tackling this issue (see Tajfel, 1981, for seminal work). Behavior analysis has been less engaged, empirically, in this area but its main progenitor, B. F. Skinner, literally became a house-hold name for publishing books that described how behavioral principles could be used to engineer human cultures in a positive and peaceful direction. Empirical efforts at understanding social categorization within behavior analysis have not been entirely absent, however. As we shall see, relatively early in the study of stimulus equivalence and arbitrarily applicable relational responding attempts were made to create experimental models of human prejudicial behavior. In conducting this early work it is fair to say that there was little appreciation of how rapidly social categorization effects can occur in human behavior. Only relatively recently has the "split-second" nature of such responding come to light. As such, the pervasive and uncontrolled nature of human prejudicial behavior is rendered even more threatening because it seems to occur "involuntarily" (because it is so rapid). The area of research that has focused on this phenomenon has been labelled implicit social cognition, and the current chapter will provide an overview of this work.

Specifically, the current chapter will focus on the research investigating the role of derived relations in the development of social stereotypes, prejudices, and beliefs. In doing so, we will briefly review the large body of literature that has employed a method derived from relational frame theory (RFT; Hayes, et al., 2001), known as the implicit relational assessment procedure (IRAP). The social domains upon which we will focus comprise: national identity; religion; race; gender; sexuality and sexual preferences; age; body image; and smoking as a stigmatized behavior. We will then discuss the existing conceptual attempts to understand these findings in the context of RFT, including a recently proposed model for analyzing IRAP effects -- the differential arbitrarily applicable relational responding effects (DAARRE) model.

The origins of the more recent work on implicit social cognition, from a RFT perspective, actually began in the early 1990s. Specifically, researchers sought to examine social categorization in Northern Ireland, where family names and sectarian symbols are often associated exclusively with either Catholic *or* Protestant communities (Watt, Keenan, Barnes, & Cairns, 1991). The study involved training participants in a series of matching-to-sample tasks that were designed to generate derived equivalence relations between Catholic names and Protestant symbols, that would be inconsistent with the verbal/social histories of participants who resided in Northern Ireland. The results showed that some Northern Irish residents did indeed demonstrate difficulty in forming these equivalence relations, whereas individuals from outside Northern Ireland did not. Numerous studies since have reported broadly similar outcomes in which participants with specific pre-experimental histories appear to show difficulty forming derived relations that are inconsistent with those histories (e.g., Barnes, Lawlor, Smeets, & Roche, 1996; Dixon, Rehfeldt, Zlomke, & Robinson, 2006; Leslie, et al., 1993; Merwin & Wilson, 2005). Interestingly, recent research in this area has also shown that it is possible to undermine these types of effects with appropriate matching-

to-sample training designed to counter racially biased responding in white children (Mizael, de Almeida, Silveira, & de Rose, 2016).

The general strategy of comparing patterns of responding that are consistent versus inconsistent with participants' pre-experimental histories carried through to more recent efforts to develop behavior-analytic procedures that may be used to assess verbal relations. Currently, the most widely used method in this regard is the IRAP (Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010). The IRAP procedure itself presents pairs of stimuli (e.g., words, pictures, statements) on each trial and participants are required to confirm or disconfirm the relation between these pairs within a short response latency window. Corrective feedback is presented after each response. In general, the feedback is designed to be consistent with participants' pre-experimental verbal histories on half of the blocks of trials, and inconsistent on the other half. For example, an IRAP might require responding "True" to a picture of a flower and the word "pleasant" (history consistent) on one block, and "False" (history inconsistent) on another block. The basic logic of the IRAP is that, all things being equal, participants should show a tendency to respond more quickly on history-consistent, relative to history-inconsistent, blocks. This difference in latencies across the two types of blocks is often referred to as the IRAP effect or a positive or negative response bias, depending on whether the effect is above or below zero. It is important to understand that the term IRAP effect, or the concept of response bias, should not be interpreted as a proxy for a mental construct or implicit attitude in a cognitive or social psychological sense. Instead, these terms (response bias) simply denote a tendency to respond in one particular direction over another on the IRAP.

It should be noted that there are a wide range of response-time measures for assessing implicit cognition, such as the implicit association test (IAT), evaluative priming, and the extrinsic affective simon task (EAST). However, all of these other measures emerged from the "mainstream" cognitive tradition, and cannot therefore be attributed to RFT, or behavior

analysis more generally. There is one other task, however, that has emerged within behavior analysis known as the Function Acquisition Speed Test (FAST; O'Reilly, Roche, Ruiz, Tyndall, & Gavin, 2012). The FAST has been used to assess implicit social cognition, but at the time of writing there was only one published study using the measure in this domain. One advantage of the IRAP over many of the other measures is that it allows for a more detailed analysis of the relations being measured. This is made possible through the separation of the measured relations into four individual trial-types. In an IRAP designed to measure racial response biases, for example, the following four trial-types might be presented: White People-Positive-True/False; White People-Negative-True/False; Black People-Positive-True/False; and Black People-Negative-True/False. Therefore, the IRAP permits a functional distinction between, for example, black people as both positive *and* negative, which may be conceptually important in socially sensitive domains. In any case, the current chapter will focus largely on IRAP research because it is clearly contained within the behavior-analytic tradition.

Social Research

National Identity. One of the first published IRAP studies assessed biases towards different nationalities (i.e., Irish, Scottish, American, and African) by native Irish and Irish-American participants (Power, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). Across two experiments, researchers gave participants an IRAP in which the label stimuli "more likeable" and "less likeable" appeared with a target that consisted of two nationalities that were to be compared in terms of likeability on each trial. In Experiment 1, the pairs were "Irish-Scottish", "Scottish-American", "American-African", "Scottish-Irish", "American-Scottish", "African-American". Participants were required to choose either "True" or "False" as response options on each trial. As an example, therefore, the first trial listed above could be read as "Irish are more likeable than Scottish; True or False?" Results indicated a strong "in-group response bias" for Irish over Scottish, and American over African nationalities. That is,

participants tended to respond more quickly when they were required to respond "True" rather than "False" when these two pairs were presented with "More Likeable". In Experiment 2, the target pairs were "American-Irish", "Irish-Scottish", "Scottish-African", "Irish-American", "Scottish-Irish", "African-Scottish" but the participants were Irish-Americans (i.e., American citizens who claimed Irish heritage). Results indicated an in-group response bias for American over Irish, Irish over Scottish, and Scottish over African. In both studies, self-report measures rating preferences of these nationalities diverged from the IRAP biases. For example, the Irish Americans rated Irish as more likeable than Americans (or any other group), but on the IRAP they responded "True" more quickly than "False" when responding to the "More-Likable; American-Irish" trial-type. This study was one of the first to highlight the potential utility of the IRAP in the assessment of potentially sensitive social issues. That is, the IRAP seemed to provide a method for assessing the natural verbal relations at play over and above self-report measures.

Religion. Hughes, Barnes-Holmes, and Smyth (2017) employed the IRAP to assess the biases of Catholics and Protestants in Northern Ireland, a post-conflict context in which sectarian tensions had been the source of violence and discrimination across three decades. Specifically, in a known-groups design, Catholic and Protestant participants in Northern Ireland were presented with one of two label stimuli (i.e., "Catholics" or "Protestants") on each trial with one of six positive (e.g., friendly, safe) or negative (e.g., bad, dishonest) words as target stimuli. The IRAP required responding in a pro-catholic and anti-protestant pattern on some blocks of trials (e.g., pressing a key for "True" when "Safe" appeared with the word "Catholic") and a pro-protestant and anti-catholic pattern on the remaining blocks of trials (e.g., pressing a key for "False" when "Safe" appeared with the word "Catholic"). The IRAP revealed pro-catholic and pro-protestant biases across both groups, however, in-group biases were stronger. For example, the size of the difference in reaction times was larger when

catholic participants were required to confirm rather than deny that catholics were positive, than when they were required to confirm rather than deny that protestants were positive; the opposite was the case for the protestant participants. Similar research was also conducted by Drake et al. (2010) on attitudes toward Christians and Muslims in an undergraduate sample in the United States. Results revealed a distinct positive bias for Christians on both the Christian-Positive and Christian-Negative trial-types, and a *negative* bias for Muslims on Muslim-Negative, but no significant effect was observed on Muslim-Positive. Moreover, Scheel, Roscoe, Scaewe, and Yarbrough (2014) found a broadly similar pro-in-group westerner bias when they assessed attitudes toward Muslims and Westerners in a US-based undergraduate sample.

Race. One of the earliest IRAP studies focused on racial bias and examined the response patterns of white participants toward pictures of black and white individuals (Barnes-Holmes, Murphy, et al., 2010). Specifically, participants were presented with one of two label stimuli (i.e., "Safe" and "Dangerous") on each trial with a picture of a white or black man holding a gun as a target stimulus. The IRAP revealed pro-white and anti-black biases, although the anti-black effect was restricted to one trial-type (i.e., the *Dangerous-Black* trial-type). Indeed, four other studies have also examined racial bias using the IRAP and broadly similar positive in-group biases were found for white participants (Drake et al., 2010, 2015; Power, Harte, Barnes-Holmes & Barnes-Holmes, 2017, a). The most recent of these combined the IRAP for examining racial biases with the measurement of electroencephalograms (EEGs; Power, Harte, Barnes-Holmes, & Barnes-Holmes, 2017, b). The EEG recordings revealed that the event-related potentials (ERPs) for the pro-black trials were more positive than the pro-white trials across six of the frontal sites. Indeed, it is interesting that the differential ERPs patterns observed in the current study were restricted to the frontal sites and that greater positivity was recorded for the IRAP performances that

required responding in a manner that was inconsistent with a white in-group racial bias. Specifically, as the authors point out, the findings are broadly consistent with research in the neuro-cognitive literature indicating that the pre-frontal areas of the cortex may be involved in suppressing emotional reactions (in more primitive areas of the brain) that are deemed to be undesirable in some way; in this case a pro-white/anti-black response.

Gender. The IRAP has also been used across a number of studies in the assessment of gender biases. For example, a study by Cartwright, Hussey, Roche, Dunne, and Murphy (2016) assessed the potential utility of the IRAP for assessing gender binary beliefs. Specifically, participants completed two IRAPs – one of two label stimuli were presented in each IRAP (i.e., "Men" and "Women") on each trial. One IRAP presented a *positive* masculine (e.g., "competitive") or feminine trait (e.g., "nurturing") as target stimuli; the other IRAP presented negative traits (e.g., masculine "aggressive"; feminine "bossy"). Participants also completed a number of self-report measures on sexism, heteronormativity and a hiring task to assess hiring preference. The IRAPs revealed that participants readily paired masculine traits to men (e.g., saying men-competitive-*true*, more easily than false) and feminine traits to women (e.g., saying women-nurturing-*true*, more easily than false), and rejected pairing feminine traits with men (e.g., saying men-nurturing-*false*, more easily than true) and masculine traits with women (e.g., saying women-aggressive-*false*, more easily than true). Interestingly, male traits were evaluated as more hirable across 83% of participants. Similar results were also found in an earlier study using a FAST (Cartwright, Roche, Gogarty, O'Reilly, & Stewart, 2016).

Indeed, similar gender stereotyping biases have also been found using the IRAP concerning gendered house chores (e.g., chopping wood is for men and cooking is for women, Drake et al., 2010), gendered toys in young boys and girls (e.g., dolls are for girls and toy cars

are for boys, Rabelo, Bortoloti, & Souza, 2014), and gendered university disciplines (i.e., men are related to science and arts but women to arts only, Farrell, Cochrane, & McHugh, 2015).

Sexuality and sexual preferences. The IRAP has also been used in the assessment of attitudes toward sexuality and sexual preferences. For example, in a study of sexual orientation, Cullen and Barnes-Holmes (2008) assessed biases among a group of heterosexual and homosexual male participants. The IRAP presented one of two label stimuli (i.e., "Straight" and "Gay") along with a positive stereotypical target for straight people (e.g., "normal", "safe") or a negative stereotypical target for gay people (e.g., "abnormal", "dangerous"). The IRAP revealed pro-gay and straight biases for both groups, however, an anti-gay bias was observed on the gay-negative trial-type among heterosexuals only. Indeed, three further studies have successfully used the IRAP to predict sexual orientation when they asked participants to relate the labels 'straight' or 'gay' with attractiveness and unattractiveness (Ronspies et al., 2015; Timmins, Barnes-Holmes, & Cullen, 2016).

In a study on sexual stereotyping, Scheel, Fischer, McMahon, and Wolf (2011) reported IRAP findings in which women confirmed more quickly than they denied that traits stereotypical of gay men (e.g., "artistic", "feminine") were coordinated with this group; a similar bias was obtained for coordinating straight men with stereotypical traits (e.g., "assertive", "masculine"). And in a study on BDSM practice, students and clinicians who reported no BDSM tendencies produced greater anti-BDSM/pro-normal IRAP effects than those who reported BDSM tendencies (Stockwell, Walker, Echleman, 2010; Stockwell, Hopkins, & Walker, in press).

Age. One of the earlier IRAP studies examined ageist attitudes among young people, and revealed pro-young and anti-old biases in the first experiment (Cullen, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). A second experiment attempted to assess the *malleability* of the IRAP effects. Participants were randomly assigned to either a pro-young or pro-old

condition in which they were exposed to pictures of admired old and disliked young people (pro-old condition) or to pictures of admired young and disliked old people (pro-young condition). Participants completed one IRAP immediately after exposure to the pictures and then to a second identical IRAP 24 hours later. The pro-young condition produced strong pro-young and anti-old effects on both days, whereas the pro-old condition produced relatively weak pro-young effects and relatively strong pro-old effects on both days.

Body image. A substantial body of IRAP research has also emerged in this domain in recent years. One of the first studies was conducted by Roddy, Stewart, and Barnes-Holmes (2010) and results revealed a pro-slim, but not an anti-fat bias. This effect was replicated in another study by Roddy, Stewart, and Barnes-Holmes (2011) which also employed facial electromyography (EMG) to gauge emotional responses of participants while they conducted the IRAP, and both measures found a pro-slim bias. Interestingly, in a similar study by Nolan, Murphy, and Barnes-Holmes (2013), this effect was only found among males. Similarly, in a study that employed only female participants and female stimuli, Exposito, Lopez, and Valverde (2015) found no pro-slim or anti-fat bias, potentially suggesting that women demonstrate less weight related prejudice than men.

Body-image IRAP effects have also been shown to predict body dissatisfaction, body weight, and disordered eating, beyond that of self-report measures (Juarascio, Forman, Timko, & Herbert, 2011). Furthermore, Heider, Spruyt and De Houwer (2015) employed two IRAPs, one assessing actual body image (i.e., I am thin) and one assessing ideal body image (i.e., I want to be thin). Results indicated actual self-thin bias was lower in those with higher levels of body dissatisfaction, whereas ideal self-thin bias was higher in those with higher levels of body dissatisfaction. Similar research on attractiveness, rather than body image, found broadly consistent findings in which pro-attractive biases were observed (Murphy, Hussey, Barnes-Holmes, & Kelly, 2015; Murphy, MacCarthaigh, & Barnes-Holmes, 2014).

Smoker Status. Two published studies have used the IRAP to assess attitudes toward smokers and non-smokers (Cagney, Harte, Barnes-Holmes, Barnes-Holmes, & McEnteggart, in press; Vahey, Boles, & Barnes-Holmes, 2010). In the first of these, Vahey et al. presented an IRAP to adolescents in which the labels 'smoker' and 'non-smoker' appeared with either social acceptance or rejection words (identified from tobacco marketing campaigns), along with *similar* or *opposite* as response options. Results showed that smokers responded more quickly when confirming that smokers were *similar* to social acceptance words, but no difference emerged between acceptance or rejection words among the non-smokers. The second study, by Cagney et al., attempted to conduct a more detailed analysis of attitudes toward smokers using the IRAP in both adult and adolescent smokers and non-smokers, and also explored the impact of parental smoking status. A pro-smoker bias was found with both adults and adolescents smokers, with a neutral bias for non-smokers in both groups. However, adult smokers and non-smokers were more clearly differentiated than the two adolescent groups on the IRAP, whereas the opposite was true in terms of the self-report measures. Post-hoc analyses indicated that non-smokers showed more positive bias scores on the IRAP towards smokers if their parents were smokers (relative to non-smoking parents). Non-smokers also showed more positive bias scores towards non-smokers if their parents did not smoke (relative to smoking parents). Overall, therefore, smoking status and parental smoking status appeared to influence social attitudes towards a socially stigmatized group (i.e., smokers).

Understanding IRAP Effects from a Relational Frame Theory Perspective

As noted at the beginning of the current chapter, the behavior-analytic basis of the IRAP involves comparing two opposing patterns of relational responding, one of which is deemed to be generally consistent with the pre-experimental history of the participant; the opposing pattern, by definition, is not. The so called IRAP effect is therefore deemed to

reflect this difference in pre-experimental relational responding. This is, in essence, the basic assumption underlying the IRAP. Over the years, however, increasingly sophisticated behavior-analytic accounts of the IRAP have emerged, and the remaining half of the current chapter will consider these with a particular focus on social implicit cognition.

The Relational Elaboration and Coherence (REC) Model

The types of effects that have been observed with the IRAP have been referred to as brief and immediate relational responses (BIRRs), in that they are emitted within a short response window of time after the onset of each trial. In contrast, extended and elaborated relational responses (EERRs) are more complex and emitted more slowly and as such occur over a longer period of time (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010; Hughes, Barnes-Holmes, & Vahey, 2012). The distinction between BIRRs and EERRs was conceptualized within the context of the Relational Elaboration and Coherence (REC) model, an initial RFT approach to implicit cognition (Barnes-Holmes, Barnes-Holmes et al., 2010; Hughes, et al., 2012). The basic idea behind the model is that the types of effects observed on the IRAP (and indeed other implicit measures) were because the task forces participants to emit BIRRs rather than EERRs. The latter are generally assumed to be evoked by traditional self-report measures, when participants are not under time-pressure to respond to each item in a questionnaire. The strength or probability of the BIRRs emitted toward experimental stimuli is deemed to be *functionally similar* to responding toward these stimuli in the participant's pre-experimental history, particularly in situations in which individuals have to respond relatively quickly or have little motivation to reflect on how they are responding in that particular moment in time.

Imagine, for example, a white individual who has resided exclusively in white neighborhoods, has no non-white friends or family members, and has been exposed to many media images of black people as violent drug dealers and inner-city gang members. When

presented with an IRAP that displayed pictures of black males carrying guns, according to the REC model, it may be likely that BIRRs confirming that black men are "dangerous" and "criminals" are more probable for this individual than denying such relations, thus revealing an anti-black racial bias (see Barnes-Holmes, Murphy et al., 2010). Indeed, this bias might not be observed if the same individual was asked to rate the same pictures of the black men in the absence of time pressure. In the latter context, there is sufficient time to respond in accordance with an extended and elaborated relationally coherent network (i.e., EERRs), which by definition extends beyond the initial BIRR. In this case, the REC model suggests that the individual may fail to report the initial BIRR based on the additional relational responding, such as "It is wrong to discriminate on the basis of race" and "I am not a racist", etc.. Hence, the IRAP reveals the initial BIRR and a questionnaire reveals the more extended and elaborated EERR.

Limitations to the REC model. In concluding that the IRAP reveals BIRRs rather than EERRs, the REC model assumes that participants respond to each of the four IRAP trial-types in more or less the same manner. This basic assumption has not been upheld empirically, however, and thus there have been recent attempts to develop a more sophisticated RFT-based understanding of the behavioral effects observed with the IRAP. In presenting this more recent account we will focus first on an IRAP that is not relevant to social cognition, but thereafter consider the implications for the social-psychological domain.

Imagine an IRAP that aimed to assess the response probabilities of four well-established verbal relations pertaining to non-socially valenced stimuli, such as shapes and colors. Across trials, the two label stimuli, "Color" and "Shape", could be presented with target words consisting of specific colors ("Red", Green", and "Blue") and shapes ("Square", "Circle", and "Triangle"). As such, the IRAP would involve presenting four different trial-types that could be designated as (i) *Color-Color,* (ii) *Color-Shape,* (iii) *Shape-Color,* and (iv)

Shape-Shape. During such a *Shapes-and-Colors* IRAP, participants would be required to respond in a manner that was consistent with their pre-experimental histories during some blocks of trials; (i) *Color-Color-True;* (ii) *Color-Shape-False;* (iii) *Shape-Color-False;* and (iv) *Shape-Shape-True.* On other blocks of trials the participants would have to respond in a manner that was inconsistent with those histories; (i) *Color-Color-False;* (ii) *Color-Shape-True;* (iii) *Shape-Color-True;* and (iv) *Shape-Shape-False.* Thus, when the four trial-type effects are calculated, by subtracting response latencies for history-consistent from history-inconsistent blocks of trials, one might expect to see four roughly equal trial-type effects. In other words, the difference scores for each of the four trial-types should be broadly similar. Critically, however, the pattern of trial-type difference scores obtained with the IRAP frequently differ across the four trial-types (e.g., Finn, Barnes-Holmes, Hussey, & Graddy, 2016).

The REC model always allowed for the potential impact of the functions of the response options on IRAP performances, in which there may be a bias toward responding 'True' over 'False' for example, and that this interacted with the stimulus relations presented in the IRAP (Barnes-Holmes, Murphy et al., 2010). As such, one might expect to observe larger differences in response latencies for trial-types that required a "True" rather than a "False" response during history-consistent blocks of trials. In the case of the *Shapes-and-Colors* IRAP described above, therefore, larger IRAP effects for the *Color-Color* and *Shape-Shape* trial-types might be observed relative to the remaining two trial-types (i.e., *Color-Shape* and *Shape-Color*). The REC model does *not* predict, however, that the IRAP effects for the *Color-Color* and *Shape-Shape* trial-types will differ (because they both require choosing the same response option within blocks of trials), but in fact our research has shown that they do (e.g., Finn et al., 2016, Experiment 3). Specifically, we have found what we call a "single-trial-type-dominance-effect" for the *Color-Color* trial-type. That is, the size of the difference

score for this trial-type is often significantly larger than for the *Shape-Shape* trial-type. This finding has led us to propose an updated model of the relational responding that we typically observe on the IRAP, which we will briefly outline subsequently. A complete description of the model, and its implications for research using the IRAP is beyond the scope of the current chapter (but see Finn, Barnes-Holmes, & McEnteggart, in press). However, it is important to consider the model here simply to highlight how an ongoing focus on relational responding is continuing to contribute towards a behavior-analytic approach to human social implicit cognition.

In attempting to explain the single-trial-type-dominance-effect for the *Shapes-and-Colors* IRAP, it is important to note that the color words we used in our research tend to occur with higher frequencies in natural language than the shape words (Keuleers, Diependaele, & Brysbaert, 2010). We therefore assume that the color words evoke relatively strong orienting responses relative to the shape words. Or more informally, participants may experience a type of confirmatory response to the color stimuli that is stronger than for the shape stimuli. Such responses may be interpreted as arising from the Cfunc rather than the Crel properties of the stimuli.[†] Critically, a functionally similar confirmatory response may be likely for the "True" relative to the "False" response option (because "True" frequently functions as a confirmatory response in natural language). A high level of functional overlap, or what we define as coherence, thus emerges on the *Color-Color* trial-type among the orienting functions of the label and target stimuli, and the "True" response option. During history-consistent blocks of

[†] According to RFT, many of the functions of stimuli that we encounter in the natural environment may appear to be relatively basic or simple but have acquired those properties due, at least in part, to a history of relational framing. Even a simple tendency to orient more strongly towards one stimulus rather than another in your visual field may be based on relational framing. Identifying the name of your home town or city from a random list of place names may occur more quickly or strongly because it coordinates with other stimuli that control strong orienting functions (e.g., the many highly familiar stimuli that constitute your home town). Such functions may be defined as Cfunc properties because they are examples of specific stimulus functions (i.e., orienting) that are acquired based on, but are separate from, the entailed relations among the relevant stimuli; the latter are labeled Crel properties (see Finn et al., in press).

trials on the IRAP, this coherence also coordinates with the relational responding that is required between the label and target stimuli (e.g., *"Color-Red-True"*). During history-consistent blocks, therefore, the trial-type could be defined as involving a maximum level of coherence because all of the responses to the stimuli, both orienting and relational, are confirmatory. During inconsistent blocks, however, participants are required to choose the "False" response option, which does not cohere with any of the other orienting or relational responses on that trial-type, and this difference in coherence across blocks of trials yields relatively large difference scores. The model we have developed that aims to explain the single-trial-type-dominance-effect, and a range of other effects we have observed with the IRAP, is named the Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) model (pronounced "Dare"). In the following sections we will outline the DAARRE model for the *Shapes-and-Colors* IRAP, and then consider an example of how the model may apply to the results obtained from an IRAP that targeted social cognition (i.e., race).

The DAARRE Model

A core assumption of the DAARRE model is that 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. The reader should 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). The basic DAARRE model as it applies to the *Shapes-and-Colors* IRAP is presented in Figure 1. The model identifies three key sources of behavioral influence: (1) the relationship between the label and target stimuli (labeled as Crels); (2) the orienting functions of the label and target stimuli

(labeled as Cfuncs); and (3) the coherence functions of the two RCIs (e.g., “Yes” and “No”). Consistent with the earlier suggestion that color-related stimuli likely possess stronger orienting functions relative to shape-related stimuli (based on differential frequencies in natural language), the Cfunc property for Colors is labeled as positive and the Cfunc property for Shapes is labeled as negative. The negative labeling for shapes should not be taken to indicate a negative orienting function but simply an orienting function that is weaker than that of colors. The labeling of the relations between the label and target stimuli indicates the extent to which they cohere or do not cohere based on the participants’ relevant history. Thus, a color-color relation is labeled with a plus sign (i.e., coherence) whereas a color-shape relation is labeled with a minus sign (i.e., incoherence). Finally, the two response options are each labeled with a plus or minus sign to indicate their functions as either coherence or incoherence indicators. In the current example, “Yes” (+) would typically be used in natural language to indicate coherence and “No” (-) to indicate incoherence. Note, however, that these and all of the other functions labeled in Figure 1 are behaviorally determined, by the past and current contextual history of the participant, and should not be seen as absolute or inherent in the stimuli themselves.

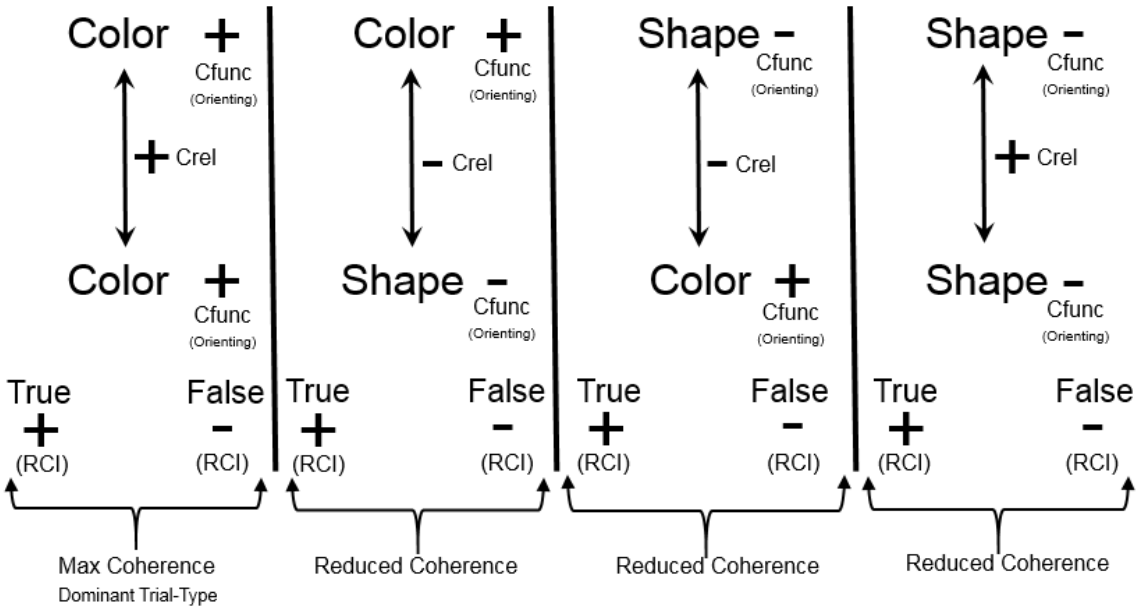

Figure 1. The DAARRE model as it applies to the Shapes-and-Colors stimulus set. The positive and negative labels refer to the relative positivity of the Cfuncs, for each label and target, the relative positivity of the Crels and the relative positivity of the RCIs in the context of the other Cfuncs, Crels and RCIs in that stimulus set.

As can be seen from Figure 1, each trial-type differs in its pattern of Cfuncs and Crels, in terms of plus and minus properties, that define the trial-type for the *Shapes-and-Colors* IRAP. The single-trial-type-dominance-effect for the *Color-Color* trial-type may be explained, as noted above, by the DAARRE model based on the extent to which the Cfunc and Crel properties cohere with the RCI properties of the response options across blocks of trials. To appreciate this explanation, note that the Cfunc and Crel properties for the *Color-Color* trial-type are all labeled with plus signs; in addition, the RCI that is deemed correct for history-consistent trials is also labeled with a plus sign (the only instance of four plus signs in the diagram). In this case, therefore, according to the model this trial-type may be considered as maximally coherent during history-consistent trials. In contrast, during history-inconsistent trials there is no coherence between the required RCI (minus sign) and the properties of the Cfuncs and Crel (all plus signs). According to the DAARRE model, this stark contrast in levels of coherence across blocks of trials serves to produce a relatively large IRAP effect. Now consider the *Shape-Shape* trial-type, which requires that participants choose the same RCI as the *Color-Color* trial-type during history-consistent trials, but here the property of the RCI (plus signs) does *not* cohere with the Cfunc properties of the label and target stimuli (both minus signs). During history-inconsistent trials the RCI *does* cohere with the Cfunc properties (minus signs) but not with the Crel property (plus sign). Thus, the differences in coherence between history-consistent and history-inconsistent trials across these two trial-types is not equal (i.e., the difference is greater for the *Color-Color* trial-type) and thus favors the single-trial-type-dominance-effect (for *Color-Color*). Finally, as becomes apparent from inspecting Figure 1 for the remaining two trial-types (*Color-Shape* and *Shape-Color*) the

differences in coherence across history-consistent and history-inconsistent blocks is reduced relative to the *Color-Color* trial-type (two plus signs relative to four), thus again supporting the single-trial-type-dominance-effect.

At this point, it seems important to consider how the DAARRE model might be used to interpret a single-trial-type-dominance effect that was obtained in an early IRAP study that focused on racial bias. Specifically, Barnes-Holmes, Murphy et al. (2010), reported four trial-type effects for a study that presented pictures of white and black males carrying guns with words related to safety and danger. The two critical trial-types in this context were *Safe-White* and *Dangerous-Black* because participants were required to press "True" during pro-white and "False" during pro-black blocks of trials. The participants (who were all indigenous white Irish individuals) tended to respond "True" more quickly than "False" on these two trial-types. However, the size of the difference for the *Safe-White* trial-type was approximately twice the size of the difference for the *Dangerous-Black* trial-type. In effect, a single-trial-type-dominance effect was observed. One could interpret this result as indicating that participants were simply more certain about the safety of white men carrying guns than they were about the danger of black men carrying weapons. Our recent work with the DAARRE model, however, suggests that we should be more cautious in drawing such a conclusion. For illustrative purposes, consider the following (speculative) interpretation.

Let us assume that the pictures of the white men and the safety words possessed relatively positive evaluative functions, whereas the pictures of black men and danger words possessed relatively negative evaluative functions. If we translate these assumptions into a figure similar to the one we used above for the *Shapes-and-Colors* IRAP, it quickly becomes apparent that the single-trial-dominance effect may have arisen, in part, from differences in coherence across the two trial-types, rather than purely from racially-biased responses (see Figure 2). Specifically, note that the Cfunc and Crel properties for the *Safe-White* trial-type

are all labeled with plus signs; in addition, the relational coherence indicator (RCI) that is deemed correct for pro-white trials is also labeled with a plus sign (the only instance of four plus signs in the diagram). In this case, therefore, according to the model this trial-type may be considered as maximally coherent during pro-white blocks. In contrast, the *Dangerous-Black* trial-type involves a "mixture" of Cfunc and Crel properties; the pictures and words possess negative Cfunc properties but a positive Crel property between them, and participants are required to choose the positive RCI on pro-white blocks. Similar to the *Shapes-and-Colors* IRAP, therefore, the difference between the size of the IRAP effects may be attributed in part to a relative difference in the coherence among the Cfunc, Crel, and RCI properties of the stimuli presented within the IRAP. In making this argument it is important to understand that we are not suggesting that IRAP effects are therefore irrelevant procedural artefacts that do not reflect potentially important behavioral histories with regard to racial differences and other social psychological phenomena. Indeed, the relative difference in coherence that we have just suggested using the DAARRE model requires that participants evaluate the pictures of white men positively and the black men negatively. What the DAARRE model provides, therefore, is the potential to for a more precise analysis of the functional relations that are in play when participants are required to complete an IRAP.

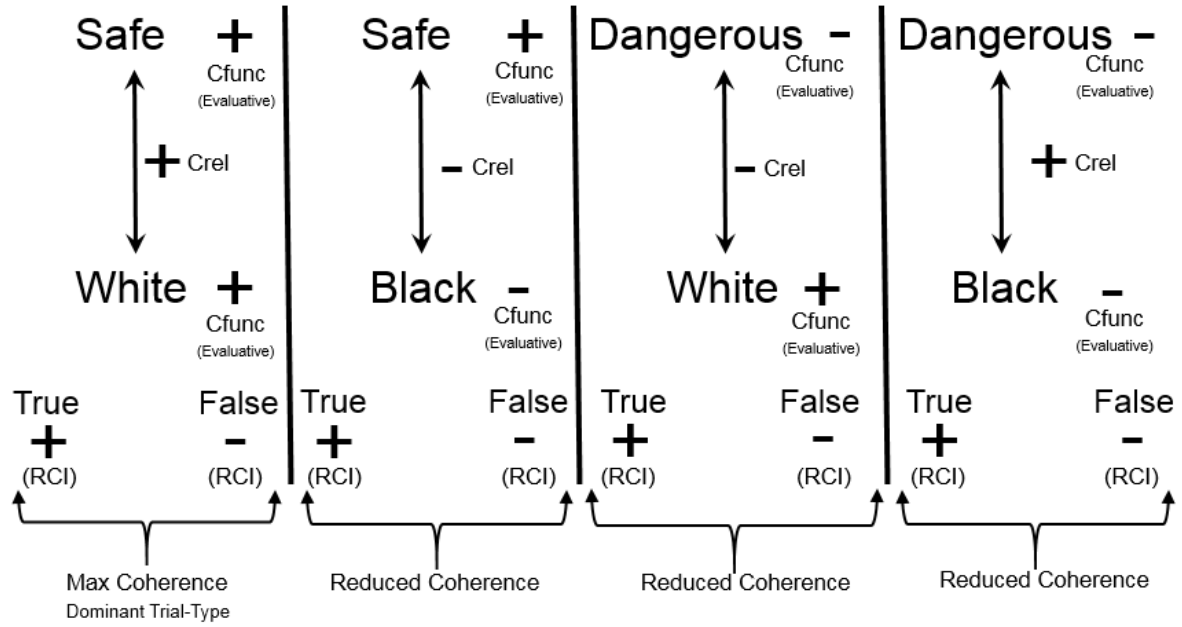

Figure 2. The DAARRE model as it applies to the race stimulus set.

Conclusion

The current chapter aimed to provide an overview of the main body of work on what may be described as implicit social cognition, but from a behavior-analytic perspective. Much of the empirical research has emerged from RFT and a procedure (the IRAP) arising from the theory. Conceptual analyses, such as those provided by the REC and DAARRE models, have also emerged. The research, both empirical and conceptual, has clearly become increasingly sophisticated since the early studies reported in the early 1990s, and it appears our understanding of the relational responding involved is gradually becoming clearer. On balance, there has been very little direct study of the variables that may be used to change or manipulate the types of relational responding that appear to be involved in implicit social cognition. One notable exception was the study reported by Cullen et al. (2009) in which exposure to positive and negative exemplars (of old and young individuals) were shown to impact upon relevant IRAP performances. It seems important, therefore, that future research in this area begins to focus on methods, techniques, and general strategies for changing

implicit social cognition that may be deemed problematic in the natural environment. Indeed, it should be noted that some success has been reported in manipulating IRAP effects, or their relationship to other behavioral measures, in the clinical domain (e.g., Bast & Barnes-Holmes, 2015; Bast, Linares, Gomes, Kovac, & Barnes-Holmes, 2016; Hussey & Barnes-Holmes, 2012; Leech, Barnes-Holmes, & McEnteggart, 2017; Ritzert, Forsyth, Berghoff, Barnes-Holmes, & Nicholson, 2015). Future behavior-analytic research in implicit social cognition is therefore well placed to explore and develop "interventions" for changing the types of relational responding that seem to be involved in implicit cognition (e.g., Mizael et al., 2016). Indeed, at a time when there is increasing populist appeal for building walls and withdrawing from a 40-year European union, there seems to be even greater urgency in the need to tackle the human capacity for categorizing each other in potentially negative and ultimately lethal and toxic ways.

References

Barnes, D., Lawlor, H., Smeets, P. M., & Roche, B. (1996). Stimulus equivalence and academic self-concept among mildly mentally handicapped and nonhandicapped children. *The Psychological Record, 46*, 87-107.

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., 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.

Bast, D. F., & Barnes-Holmes, D. (2015). Priming thoughts of failing versus succeeding and performance on the Implicit Relational Assessment Procedure (IRAP) as a measure

of self-forgiveness. *The Psychological Record*, *65*(4), 667-678. doi: 10.1007/s40732-015-0137-0

Bast, D. F., Linares, I. M. P., Gomes, C., Kovac, R., & Barnes-Holmes, D. (2016). The Implicit Relational Assessment Procedure (IRAP) as a measure of self-forgiveness: the impact of a training history in clinical behavior analysis. *The Psychological Record, 66*(1), 177-190. doi: 10.1007/s40732-016-0162-7

Cagney, S., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & 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. doi: 10.1007/s40732-017-0249-9

Cartwright, A., Hussey, I., Roche, B., Dunne, J., & Murphy, C. (2016). An investigation into the relationship between gender binary and occupational discrimination using the Implicit Relational Assessment Procedure. *The Psychological Record, 67,* 121-130. doi: 10.1007/s40732-016-0212-1

Cartwright, A., Roche, B., Gogarty, M., O'Reilly, A., & Stewart, I. (2016). Using a modified Function Acquisition Speed Test (FAST) for assessing implicit gender stereotypes. *The Psychological Record, 66*(2), 223-233. doi: 10.1007/s40732-016-0164-5

Cullen, C., & Barnes-Holmes, D. (2008). *Implicit pride and prejudice: A heterosexual phenomenon?* In M.A. Morrison & T.G. Morrison (Eds.), *Modern Prejudice* (pp. 195-223). New York, NY: Nova Science.

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,* 591-620.

Dixon, M., Rehfeldt, R. A., Zlomke, K.M., & Robinson, A. (2006). Exploring the development and dismantling of equivalence classes involving terrorist stimuli. *The Psychological Record, 56,* 83-103.

Drake, C.E., Kellum, K.K., Wilson, K.G., Luoma, J.B., Weinstein, J.H., & Adams, C.H. (2010). Examining the Implicit Relational Assessment Procedure: Four preliminary studies. *The Psychological Record, 60,* 81-86.

Drake, C. E., Kramer, S., Sain, T., Swiatek, R., Kohn, K., & Murphy, M. (2015). Exploring the reliability and convergent validity of implicit racial evaluations. *Behavior and Social Issues, 24,* 68-87. doi: 10.5210/bsi.v.24i0.5496

Exposito, P. M., Lopex, M. H., & Valverde, M. R. (2015). Assessment of implicit anti-fat and pro-slim attitudes in young women using the Implicit Relational Assessment Procedure. *International Journal of Psychology and Psychological Therapy, 15*, 17-32.

Farrell, L., Cochrane, A., & McHugh, L. (2015). Exploring attitudes towards gender and science: The advantages of an IRAP approach versus the IAT. *Journal of Contextual Behavioral Science, 4*(2), 121-128. doi: 10.1016/j.jcbs.2015.04.002

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. doi: 10.1007/s40732-016-0173-4

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (in press). Exploring the Single-Trial-Type-Dominance-Effect on the IRAP: Developing a Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) Model. *The Psychological Record.*

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

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

Heider, N., Spruyt, A., & De Houwer, J. (2015). Implicit beliefs about ideal body image predict body image dissatisfaction. *Frontiers in Psychology, 6,* 1402. doi: 10.3389/fpsyg.2015.01402

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

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, *1*(1), 17-38. doi: 10.1016/j.jcbs.2012.09.003

Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice*, *19*(4), 573-582. doi: 10.1016/j.cbpra.2012.03.002

Juarascio, A. S., Forman, E. M., Timko, C. A., & Herbert, J. D. (2011). Implicit internalization of the thin ideal as a predictor of increases in weight, body dissatisfaction, and disordered eating. *Eating Behaviors*, *12*, 207-213. doi: 10.1016/j.eatbeh.2011.04.004

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.

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 Behavior–

Behavior Relation and Its Implications for an Experimental Analysis of Defusion. *The Psychological Record, 67*(3), 387-398. doi: 10.1007/s40732-017-0230-7

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

Maloney, E., & Barnes-Holmes, D. (2016). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: the role of relational contextual cues versus relational coherence indicators as response options. *The Psychological Record, 66*(3), 395-403. doi:10.1007/s40732-016-0180-5

Merwin, I. 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.

Mizael, T.M., de Almeida, J.H., Silveira, C.C., & de Rose, J.C. (2016). Changing racial bias by transfer of functions in equivalence classes. *The Psychological Record, 66*(3), 451-462. doi: 10.1007/s40732-016-0185-0

Murphy, C., Hussey, T., Barnes-Holmes, D., & Kelly, M. (2015). The Implicit Relational Assessment Procedure (IRAP) and attractiveness bias. *Journal of Contextual Behavioral Science, 4*(4), 292-299. doi: 10.1016/j.jcbs.2015.08.001

Murphy, C., MacCarthaigh, S., & Barnes-Holmes, D. (2014). Implicit relational assessment procedure and attractiveness bias: Directionality of bias and influence of gender of participants. *International Journal of Psychology and Psychological Therapy, 14*(3), 333-351.

Nolan, J., Murphy, C., & Barnes-Holmes, D. (2013). Implicit Relational Assessment Procedure and body-weight bias: Influence of gender of participants and targets. *The Psychological Record, 63*, 467-488. doi: 10.11133/j.tpr.2013.63.3.005

O'Reilly, A., Roche, B., Ruiz, M., Tyndall, I., & Gavin, A. (2012). The Function Speed Acquisition Test (FAST): A behavior analytic implicit test for assessing stimulus relations. *The Psychological Record, 62*, 507-528. doi: 10.1007/BF03395817

Power, P., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The Implicit Relational Assessment Procedure (IRAP) as a measure of implicit relative preferences: A first study. *The Psychological Record, 59,* 621-640. doi: 10.1007/BF03395684

Power, P. M., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y. (2017, a). Exploring racial bias in a country with a recent history of immigration of black Africans. *The Psychological Record,* 67*(3)*, 365-375 doi.org/10.1007/s40732-017-0223-6

Power, P. M., Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y. (2017, b). Combining the Implicit Relational Assessment Procedure and the recording of event related potentials in the analysis of racial bias: A preliminary study. *The Psychological Record*, 67(4), 499-506. doi: 10.1007/s40732-017-0252-1

Rabelo, L. Z., Bortoloti, R., & Souza, D. H. (2014). Dolls are for girls and not for boys: Evaluating the appropriateness of the Implicit Relational Assessment Procedure for school-age children. *The Psychological Record, 64*, 71-77. doi: 10.1007/s40732-014-0006-2

Ritzert, T. R., Forsyth, J. P., Berghoff, C. R., Barnes-Holmes, D., & Nicholson, E. (2015). The impact of a cognitive defusion intervention on behavioral and psychological flexibility: An experimental evaluation in a spider fearful non-clinical

sample. *Journal of Contextual Behavioral Science, 4*(2), 112-120. doi: 10.1016/j.jcbs.2015.04.001

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology, 15,* 416-425. doi: 10.1177/1359105309350232

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology, 41,* 688-694. doi: 10.1002/ejsp.839

Ronspies, J., Schmidt, A.F., Melnikova, A., Krumova, R., Zolfagari, A., & Banse, R. (2015). Indirect measurement of sexual orientation: Comparison of the Implicit Relational Assessment Procedure, viewing time, and choice reaction time. *Archives of Sexual Behavior, 44*(5), 1483-1492. doi: 10.1007/s10508-014-0473-1

Scheel, M. H., Fischer, L. A., McMahon, A. J., Wolf, J. E. (2011). The implicit relational assessment procedure (IRAP) as a measure of women's stereotype about gay men. *Current Research in Social Psychology, 18*(2), 11-23.

Scheel, M. H., Roscoe, B. H., Scaewe, V. G., & Yarbrough, C. S. (2014). Attitudes towards muslims are more favorable on a survey than on an Implicit Relational Assessment Procedure (IRAP). *Current Research in Social Psychology, 22*(3), 22-32.

Stockwell, F. M. J., Hopkins, L. S., & Walker, D. J. (in press). Implicit and explicit attitudes toward mainstream and BDSM sexual practices and their relation to interviewer behavior: An analogue study. *The Psychological Record.*

Stockwell, F. M. J., Walker, D. J., & Eshleman, J. W. (2010). Measures of implicit and explicit attitudes toward mainstream and BDSM sexual terms using the IRAP and questionnaire with BDSM/fetish and student participants. *The Psychological Record, 60,* 307-324. doi: 10.1007/BF03395709

Tajfel, H. (1981). *Human groups and social categories: Studies in Social Psychology*. Cambridge: Cambridge University Press.

Timmins, L., Barnes-Holmes, D., & Cullen, C. (2016). Measuring implicit sexual response biases to nude male and female pictures in androphilc and gynephilic men. *Archives of Sexual Behavior, 45,* 829-841. doi: 10.1007/s10508-016-0725-3

Vahey, N., Boles, S., & Barnes-Holmes, D. (2010). Measuring adolescents' smoking-related social identity preferences with the Implicit Relational Assessment Procedure (IRAP) for the first time: A starting point that explains later IRAP evolutions. *International Journal of Psychology and Psychological Therapy*, *10*(3), 453-474.

Watt, A. W., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41,* 371-388.