The Psychological Record, 2010, 60, 527–542

Sketch

A SKETCH OF THE IMPLICIT RELATIONAL ASSESSMENT PROCEDURE (IRAP) AND THE RELATIONAL ELABORATION AND COHERENCE (REC) MODEL

Dermot Barnes-Holmes and Yvonne Barnes-Holmes
National University of Ireland, Maynooth

Ian Stewart
National University of Ireland, Galway

Shawn Boles
Oregon Research Institute

The current article outlines a behavior-analytic approach to the study of so-called implicit attitudes and cognition. The Implicit Relational Assessment Procedure (IRAP), the conceptual basis of which was derived from relational frame theory, is offered as a methodology that may be used in the experimental analysis of implicit attitudes and beliefs. The relational elaboration and coherence (REC) model provides a possible relational-frame account of the findings that have emerged from the IRAP. The article first outlines the research history that led to the development of the IRAP, followed by a description of the method. The REC model and how it explains a range of IRAP data are then considered. The article also outlines how both the IRAP and the REC model overlap with, and differ from, similar research found in the non-behavior-analytic literature.

Key words: Implicit Relational Assessment Procedure, relational frame theory, network, elaboration, coherence, attitudes

Imagine if you were asked, “Do you prefer 7UP or Pepsi?” Assuming that you do have a particular preference, you would likely respond quickly and with little thought or deliberation. Furthermore, your answer would almost certainly correspond to the drink you typically choose. Imagine, however, if you were asked, “Do you prefer white or black people?” Responding to this question seems to differ quite dramatically from the first question. Indeed, the question itself may be considered inappropriate or even insulting

Preparation of this article was supported by funding awarded to the first three authors from the Irish Research Council for Humanities and Social Sciences, under the Research Development Initiative (2008–09).

The authors would like to thank Mark Dixon for his constructive comments on an earlier version of this work.

Correspondence concerning this article should be addressed to Dermot Barnes-Holmes, Department of Psychology, National University of Ireland, Maynooth, Co. Kildare, Ireland. E-mail: Dermot.Barnes-Holmes@nuim.ie

because the answer you provide could be used to judge or label you negatively in some way (e.g., as a racist). Thus you may be inclined to pause for thought, if only briefly, before answering. In addition, your response to the question may not correspond with other aspects of your social behavior—if you answer, for example, "I have no preference," but virtually all of your friends, neighbors, and acquaintances are from your own racial group. In short, the first question may produce an answer that is based almost entirely on an immediate or automatic response to the choice between 7UP and Pepsi, whereas the latter question may generate some thought before an answer is offered (e.g., "Although all of my friends are white, I am not racist, and so I have no preference").

Recognizing this difference in how particular questions may be answered seems to be critically important in the study of attitudes and beliefs and how they relate to actual behavior. If, for example, racially biased actions are better predicted in certain contexts by automatic rather than deliberative responses, it is essential that appropriate methods be developed that can assess such automatic reactions. The current article provides a preliminary exploration (or sketch) of the methodological, empirical, and theoretical development of one such methodology—the Implicit Relational Assessment Procedure, or IRAP[1] (all IRAP articles cited herein are available for download from http://psychology.nuim.ie/IRAP/IRAP_1.shtml).

The IRAP: History and Method

History

The study of human language and cognition has attracted increasing attention from behavior analysts, with some researchers focusing on stimulus equivalence and derived stimulus relations[2] (e.g., Hayes, Barnes-Holmes, & Roche, 2001; Sidman, 1994). Relatively early in this research program, a number of investigators attempted to develop methods for assessing natural verbal relations using procedures that were employed in the study of stimulus equivalence. The basic approach involves training and testing for laboratory-induced equivalence classes that are likely to conflict with specific preexisting verbal relations. Critically, it is predicted that the emergence of laboratory-induced classes will be hindered because they compete with the natural verbal relations.

The first study in this area focused on the topic of sectarian or religious categorization. The study employed a sample of adult participants who resided in Northern Ireland and a group of English participants who

[1] The word *implicit* indicates that the IRAP was designed to measure the probability of automatic responding; this issue is considered in detail in the context of the relational elaboration and coherence (REC) model, which is outlined in the second half of this article.

[2] In a typical study of stimulus equivalence, a series of interrelated conditional discriminations are first reinforced, and then a number of untaught but predictable stimulus relations are seen to emerge in the absence of explicit feedback or verbal instruction. During the training, for example, A–B and B–C matching-to-sample (MTS) responses might be taught. A series of test or probe MTS trials are then presented in which symmetry (B–A, C–B), transitivity (A–C), and combined symmetry and transitivity (C–A) may be observed in the absence of differential reinforcement. If these emergent or untrained patterns of responding occur, the stimuli are said to participate in an equivalence class or derived relation.

did not (Watt, Keenan, Barnes, & Cairns, 1991). In Northern Ireland the verbal community frequently categorizes specific family names and symbols with either the Protestant or Catholic religions (Cairns, 1984), but this verbal practice is rarely found in England. In the Watt et al. study, the initial training involved matching Catholic family names to nonsense syllables and the same nonsense syllables to Protestant symbols, and all participants successfully completed this phase. However, the critical equivalence test involved matching the Catholic names directly to the Protestant symbols, and many of the Northern Irish participants failed this test, but the English participants did not. In effect, the verbal relations previously established within the Northern Irish verbal community appeared to disrupt or retard the formation of laboratory-induced equivalence relations. Since this study was published, the basic effect has been replicated and extended across a range of other content domains, including academic self-concept (Barnes, Lawlor, Smeets, & Roche, 1996), terrorism (Dixon, Rehfeldt, Zlomke, & Robinson, 2006), clinical anxiety (Leslie et al., 1993), and self-esteem (Merwin & Wilson, 2005).

The foregoing approach to assessing natural verbal relations, by pitting those relations against laboratory-induced equivalence classes, provided the conceptual foundation for creating the IRAP. Methodologically, the IRAP drew heavily on earlier work with what is called the relational evaluation procedure (REP).[3] The REP presents participants with a task that requires them to evaluate, or report on, the stimulus relation that is presented on a given trial. For example, two identical shapes might be presented with the relational terms “Same” and “Opposite,” and participants are required to indicate, typically without time pressure, that the relation is “Same” (see O’Hora, Barnes-Holmes, Roche, & Smeets, 2004; O’Hora, Pelaez, Barnes-Holmes, & Amesty, 2005; Stewart, Barnes-Holmes, & Roche, 2002, 2004). Indeed, the importance of the REP in developing the IRAP was such that initially the IRAP was called the IREP. Nevertheless, the IRAP acronym was soon adopted because it can be read as “I rap,” as in “I talk quickly,” which, conceptually, is what the IRAP asks a participant to do.

Like the REP, the IRAP involves presenting specific relational terms (e.g., *similar*, *opposite*, *more*, *less*) so that the properties of the relations among the relevant stimuli can be assessed. Unlike the REP, the IRAP involves asking participants to respond quickly and accurately in ways that are either consistent or inconsistent with their preexperimentally established verbal relations. The basic hypothesis is that average response latencies for a group of participants should be shorter across blocks of consistent trials than across inconsistent trials. In addition, the extent of the observed difference between the trials is assumed to provide an index of the strength of the verbal or relational responses being assessed.

Method

The IRAP is presented on a computer, with detailed instructions typically provided to participants before they commence the IRAP itself. (Software and sample instructions are available from http://psychology.

[3] Research on the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) also provided an important methodological basis for the development of the IRAP (see Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008).

nuim.ie/IRAP/IRAP_1.shtml; access to a web-based version of the procedure is also available at this site; see Figure 1 for a schematic representation of the IRAP.) On each trial of the IRAP one of two label stimuli is presented at the top of the computer screen, with one of two types of target stimuli presented in the center. Participants are required to choose between two response options, which appear at the bottom left and right of the screen, by pressing either the "D" or "K" key; the left–right positions of the response options switch randomly from trial to trial. As an illustrative example, suppose that the two labels are the words "Pleasant" and "Unpleasant," the target stimuli are six positive words (e.g., "Love," "Happy," "Peace") and six negative words (e.g., "Hate," "Sad," "War"), and the two response options are "Similar" and "Opposite." During a block of consistent trials, a response defined as consistent with prevailing verbal contingencies (e.g., choosing "Similar" given "Pleasant" and "Love") clears the screen for 400 ms and then the next trial is presented. If an inconsistent response is emitted (e.g., choosing "Opposite" given "Pleasant" and "Love"), a red *X* appears immediately under the target stimulus. To remove the red *X* and continue to the 400-ms intertrial interval, participants are required to emit the consistent response. In contrast, during inconsistent blocks participants are required to make an inconsistent response in order to progress from one trial to the next (a consistent response produces the red *X*).[4]

The IRAP typically consists of a minimum of two practice blocks and a fixed set of six test blocks.[5] Each block presents the same number of trials, comprised of what are defined as four different trial types. The trial types are created by presenting each label with each of two sets of target words (see Figure 1). Given the previous example, a block of consistent trials thus requires the following pattern of responses: Pleasant–Positive–*Similar*, Pleasant–Negative–*Opposite*, Unpleasant–Positive–*Opposite*, Unpleasant–Negative–*Similar*. A block of inconsistent trials requires the opposite response pattern (Pleasant–Positive–*Opposite*, Pleasant–Negative–*Similar*, etc.). The feedback contingencies are reversed across successive blocks of the IRAP, and thus participants are exposed to an alternating sequence of consistent and inconsistent blocks. The order in which this sequence is presented (consistent followed by inconsistent or inconsistent followed by consistent) is often counterbalanced across participants.[6]

After each block of the IRAP, participants are informed that the previously correct and incorrect answers will be reversed in the next block, thus removing any requirement for trial-and-error learning after the first block. Each IRAP block presents each target stimulus once in the presence of each of the two labels (a minimum of six and a maximum of 12 target stimuli per

[4] At the time of writing, our research group had just begun to explore the use of multiple label stimuli in the IRAP. For example, rather than presenting only "Pleasant" or "Unpleasant" on each trial, other semantically similar labels are presented on other trials, such as "Good" or "Bad." Very early findings suggest that using multiple labels may increase effect sizes.

[5] The number of test blocks presented by the IRAP software can be adjusted to two, four, or six, but at the time of writing our research group had not explored the effect of manipulating this variable.

[6] Thus far, our research has failed to find any important significant effects for the order in which the IRAP blocks are presented (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007; Power, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009; Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009).

label may be entered into the software, thus allowing for a range of 24 to 48 trials per block). The trials are presented quasirandomly, with the typical constraint that none of the four trial types be presented twice in succession.[7] The positioning of the two response options is also quasirandom in that typically they cannot appear in the same left–right position three times in succession.[8]

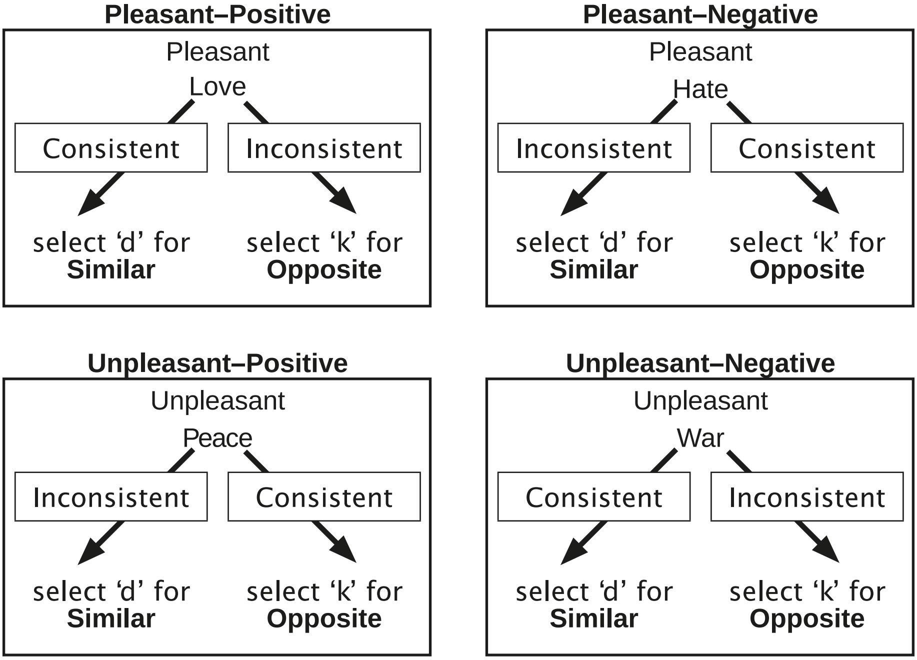

Figure 1. An example of four IRAP trial types. The label ("Pleasant" or "Unpleasant"), target word ("Love," "Hate," "Peace," or "War"), and response options ("Similar" and "Opposite") appear simultaneously on each trial. Arrows with superimposed text boxes indicate which responses are deemed consistent or inconsistent (boxes and arrows do not appear on screen). Selecting the consistent response option during a consistent block, or the inconsistent option during an inconsistent block, clears the screen for 400 ms before the next trial is presented; if the inconsistent option is chosen during a consistent block or the consistent option during an inconsistent block, a red *X* appears on screen until the participant emits the alternative response. In very recent studies, the warning message "Too Slow" appears below the target if a participant fails to respond within a specified latency criterion.

For the practice blocks, participants are informed that it is a practice phase and errors are expected. On-screen feedback is provided after each block, which indicates the percentage of correct responses and median

[7] The number of times a trial type may be repeated can be adjusted by the IRAP software, but at the time of writing our research group had not explored the effect of manipulating this variable.

[8] The number of times the response options may appear in the same position may be changed by the software, but unpublished research by our group indicates that this variable has little impact on IRAP performance.

response latency for that block. Participants are typically required to reach a standard of 80% correct responses[9] and a median response time of less than 2,000 ms.[10] These criteria are used to ensure that participants understand and are complying with the IRAP instructions. If participants fail to achieve the two criteria for either of the two practice blocks, the required standard and the standard of responding they have achieved are presented on the screen, and they are invited to try again. Participants are allowed four attempts to achieve the practice criteria (a total of eight practice blocks), and if they fail to do so, they are thanked and debriefed and their data are discarded.[11] Participants who achieve the practice criteria proceed to the six test blocks.

The procedure for the test blocks is similar to the practice blocks, except that on-screen instructions inform participants that each block is a test and to "go quickly," although making "a few errors is okay." The same alternating sequence employed with the practice blocks is also used with the test blocks. Thus, if a participant is exposed to a consistent–inconsistent sequence during practice, Test Blocks 1, 3, and 5 are consistent and Test Blocks 2, 4, and 6 are inconsistent; if practice involved an inconsistent–consistent sequence, then Test Blocks 1, 3, and 5 are inconsistent and 2, 4, and 6 are consistent. No performance criteria are applied during the test blocks in order to proceed, but if a participant's performance falls below the practice accuracy criterion (e.g., 80%) for any test block, the data for that participant typically are discarded.[12] When all six test blocks have been presented, the IRAP is complete.

The primary datum from the IRAP is response latency, defined as the time in milliseconds (ms) that elapses between the onset of the trial and a correct response emitted by a participant. The response latency data for each participant are typically transformed into D_{IRAP} scores (or IRAP effects) using the D_{IRAP} algorithm, derived from the *D* algorithm developed by Greenwald, Nosek, and Banaji (2003) for the IAT (see also Back, Schmukle, Egloff, & Gutenberg, 2005; Cai, Sriram, Greenwald, & McFarland, 2004; Mierke

[9] This criterion may be reduced to 70% if it becomes apparent that a particular sample of participants cannot achieve the higher criterion (e.g., Vahey et al., 2009).

[10] Initially, our research group typically used a 3,000-ms criterion, but recent research indicates that reducing this to 2,000 ms increases the validity and reliability of the IRAP performance (Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010). On balance, it is important to emphasize that the latency criterion, similar to accuracy, should be adjusted, preferably based on pilot work, to a level appropriate for the population that is being sampled and the stimuli that are being used in the study (e.g., if statements rather than single words are used as labels and/or targets, the latency criterion may need to be 3,000 ms or more to avoid high attrition rates).

[11] In some studies a maximum of three exposures to the practice blocks were presented, but our experience indicates that a maximum of four may be used without incurring fatigue and boredom effects among most participants.

[12] In order to encourage participants to maintain rapid responding during the test blocks, a recent version of the IRAP software has a setting that presents the warning message "Too Slow" on any trial for which a participant fails to respond within the latency criterion. When using this latency feedback our research group typically calculates the mean latency for each of the four trial types across the three consistent and the three inconsistent blocks, and if any of the eight mean latencies exceeds the practice latency criterion (e.g., 2,000 ms), the data for that participant are discarded. Thus, the IRAP test performance must remain within both the accuracy and latency criteria. The importance of both accurate and rapid responding on the IRAP is addressed later in this article.

& Klauer, 2003). The *D* transformation functions to minimize the impact of factors such as age, motor skills, and/or cognitive ability on latency data, allowing researchers to measure differences between groups using a response-latency paradigm with reduced contamination by individual differences associated with extraneous factors[13] (Greenwald et al., 2003). On balance, it is important to note that the data-analytic techniques outlined below are merely a description of the practices that have evolved within a single research group, and as such the D_{IRAP} algorithm should not be seen as prescriptive or necessarily the “best way” to analyze IRAP data.

The steps involved in calculating the D_{IRAP} scores are as follows: (1) Only response-latency data from test blocks are used; (2) latencies above 10,000 ms from the data set are eliminated; (3) all data for a participant are removed if he or she produces more than 10% of test-block trials with latencies less than 300 ms; (4) 12 standard deviations for the four trial types are computed: four from the response latencies from Test Blocks 1 and 2, four from the latencies from Test Blocks 3 and 4, and a further four from Test Blocks 5 and 6; (5) 24 mean latencies for the four trial types in each test block are calculated; (6) difference scores are calculated for each of the four trial types for each pair of test blocks by subtracting the mean latency of the consistent block from the mean latency of the corresponding inconsistent block; (7) each difference score is divided by its corresponding standard deviation from step 4, yielding 12 D_{IRAP} scores, one score for each trial type for each pair of test blocks; (8) four overall trial-type D_{IRAP} scores, or IRAP effects, are calculated by averaging the scores for each trial type across the three pairs of test blocks.[14] The four trial-type scores for each participant are then used to calculate mean D_{IRAP} scores across a group of participants, and these may be presented on a bar graph, such as that shown in Figure 2.

The data presented in Figure 2 are from a pilot study that employed two groups of soccer fans; one group supported the London-based English soccer team, Chelsea, and the other group supported the Spanish team, Barcelona. The IRAP presented the words “Chelsea” and “Barcelona” as labels with six positive target words (e.g., “Great,” “Brilliant,” “Amazing,” etc.) and six negative target words (“Bad,” “Awful,” “Rubbish,” etc.) and the response options “Similar” and “Opposite.” The data indicated predictable

[13] The first IRAP studies (e.g., Barnes-Holmes et al., 2008) did not employ the D_{IRAP} algorithm because individual differences were not being measured. However, our research group often uses the algorithm for two main reasons. First, a recent study has shown that when IRAP difference scores are calculated without using the *D* transformation, they correlate significantly with intelligence (O’Toole & Barnes-Holmes, 2009), which may serve to confound the IRAP measure when factors other than IQ are being assessed. Critically, when the D_{IRAP} transformation was applied to the same data, no significant correlations with IQ were observed (data not reported in the O’Toole & Barnes-Holmes article). Second, using a *D* algorithm facilitates a comparison between the IRAP and the IAT or other implicit measures that use a *D* transformation (e.g., Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart, 2010).

[14] In some of the early IRAP studies the data were analyzed separately for each of the three pairs of test blocks. In general, differences in response latency between consistent and inconsistent blocks did not change significantly across the block pairs, and thus this variable has been ignored in subsequent studies. Furthermore, analyzing block effects seems unwise when examining the data at the level of the individual trial type. Specifically, the number of responses used to calculate a trial-type *D* score from a single pair of test blocks could be a few as 12 (i.e., six responses on each block), which is very low for a reaction-time-based measure.

pro-Chelsea and anti-Barcelona IRAP effects for the Chelsea supporters and the reverse effects for the Barcelona supporters.[15] For example, the mean D_{IRAP} score for the Chelsea–Positive trial type showed that the Chelsea group responded more quickly when "Similar" rather than "Opposite" was the correct response; in contrast, the Barcelona group responded more quickly when "Opposite" rather than "Similar" was correct. Interestingly, the D_{IRAP} scores were larger when the participants responded to the trial types that presented their own rather than the other team as the label. In effect, it appeared that supporters responded more positively toward their own team than they did negatively toward the other team.

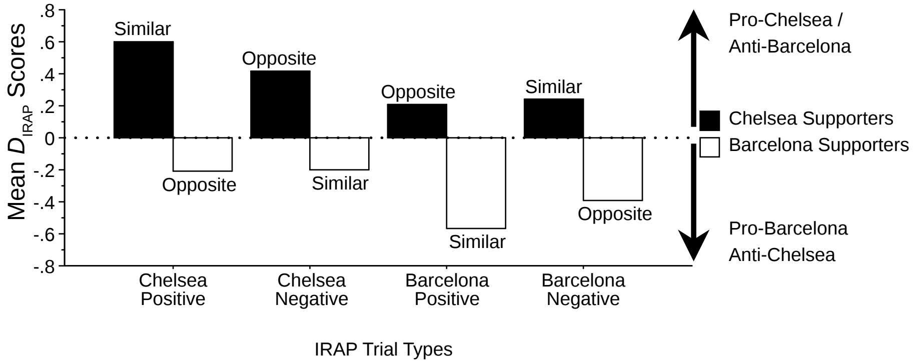

Figure 2. An illustrative example of how the data from four IRAP trial types may be presented on a bar graph. The data are from a pilot study that employed two groups of soccer fans; one group supported the London-based English soccer team, Chelsea, and the other group supported the Spanish team, Barcelona. The IRAP presented the words "Chelsea" and "Barcelona" as labels with six positive and six negative target words and the response options "Similar" and "Opposite." On the graph, positive D_{IRAP} scores indicate pro-Chelsea/anti-Barcelona IRAP effects and negative D_{IRAP} scores reflect pro-Barcelona/anti-Chelsea effects. The zero point indicates no preference. The words "Similar" and "Opposite" indicate which response option was chosen more quickly across the test blocks. For the Chelsea–Positive trial type, for example, the Chelsea supporters chose "Similar" more quickly than "Opposite," but the Barcelona supporters choose "Opposite" more quickly than "Similar."

Analyses of variance (mixed repeated measures) are typically used to test for significant main and interaction effects, and planned one-sample *t* tests are employed to determine if the mean D_{IRAP} scores differ significantly from zero. The internal consistency of the IRAP measures may be assessed by calculating split-half reliability scores for each trial type. In each case, one score for odd trials and one for even trials are calculated, and these are obtained in the same way as for the four original scores except that the algorithm described previously is applied separately to all odd trials and to all even trials. Split-half correlations, applying Spearman-Brown corrections, are then calculated. If explicit measures (e.g., questionnaires or other

[15] When the IRAP is used to examine group differences in this manner, defining the blocks as consistent versus inconsistent per se may be confusing because the consistent blocks for one group will likely be the inconsistent blocks for the other group (and vice versa). In such cases, it seems more appropriate to refer to the blocks as pro-X/anti-Y versus pro-Y/anti-X (see Figure 2 for an example).

rating scales) have been employed, the results of these may be correlated with the D_{IRAP} scores and, if appropriate, regression analyses may be used to determine if the IRAP effects provide incremental predictive validity over the explicit measures.

Although IRAP data may be analyzed at the level of the four trial types, it may simplify the analyses if the data are collapsed across two of the trial types (provided that doing so does not obscure a theoretically important effect). For the results presented in Figure 2, for example, the data could be collapsed across the trial types for each soccer team, yielding a single mean D_{IRAP} score for Chelsea and another for Barcelona (i.e., two IRAP effects for each group of supporters). Indeed, adopting this strategy for these data would likely be acceptable because the to-be-collapsed scores do not differ substantively from each other. Finally, it is worth noting that a single overall D_{IRAP} score (calculated across all four trial types) is sometimes reported, and this measure is also used to calculate split-half reliability and may be employed in correlations and regression analyses with the relevant explicit measures.

In concluding this section, it is important to note that the relationship between IRAP effects and explicit measures is often of critical importance. For example, one approach to establishing the validity of the IRAP is to conduct a “known-groups” analysis, in which IRAP effects are used to predict participants’ self-reports of specific preferences or behaviors. The data presented in Figure 2 provide a simple example of a known-groups test in which the direction of the IRAP effects is consistent with the participants’ self-reported team allegiances (for other examples see Barnes-Holmes, Murtagh, et al., 2010; Barnes-Holmes, Waldron, Barnes-Holmes, & Stewart, 2009; Vahey et al., 2009). On balance, perhaps the most interesting feature of the IRAP (and other implicit measures) is observed when the direction of an IRAP effect diverges from, or appears inconsistent with, an explicit measure (e.g., Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009). Such divergent results are typically observed when “psychologically sensitive” attitudes or beliefs are targeted. For example, white participants may respond positively toward black people on a standard questionnaire of racial prejudice but produce an anti-black IRAP effect (Barnes-Holmes, Murphy, et al., 2010; see also Power et al., 2009). Furthermore, it appears that participants find it difficult to control or “fake” an IRAP effect, even when they are instructed to do so (McKenna et al., 2007). In the following section, a behavior-analytic explanation for these and related findings will be considered.

The Relational Elaboration and Coherence Model: A Relational Frame Theory Explanation for the IRAP Effect

The first IRAP articles provided an outline of a relational-frame interpretation of the IRAP effect (e.g., Barnes-Holmes, Barnes-Holmes, et al., 2006), which is now referred to as the relational elaboration and coherence (REC) model. According to the REC model, specific IRAP trials may produce an immediate and relatively brief relational response before the participant actually presses a response key. The probability of this initial response will often be determined by the verbal and nonverbal history of the participant and current contextual variables. By definition, the most probable immediate response will be emitted first most often, and thus any IRAP trial that requires a key press that coordinates with that immediate response will be emitted relatively quickly; if, however, an IRAP trial requires a key press that opposes the immediate relational response, it may be emitted less quickly. Thus, across multiple trials, the average latency for inconsistent blocks will be longer than for consistent trials. In short, the IRAP effect is based on immediate and brief relational responding, which is made apparent to the researcher when the behavioral system is put under pressure to respond quickly and accurately.

The foregoing interpretation provides a plausible explanation for the basic IRAP effect. But why does the IRAP produce evidence of socially sensitive bias, such as racial stereotyping among white individuals, when explicit measures do not (e.g., Barnes-Holmes, Murphy, et al., 2010)? According to the REC model, IRAP effects indicative of racial bias would likely emerge from exposure to some of the verbal and nonverbal contingencies that operate for white individuals who are raised in a predominately white culture (e.g., portrayal of young black males in the media as violent gang members). In attempting to explain why such contingencies do not produce evidence of stereotyping in self-reports, the REC model assumes that responses to these measures likely reflect relatively elaborate and coherent relational responding. In other words, when asked to express an attitude or belief on a particular issue, it is likely that a person will produce a relational response that coheres with one or more other relational responses in his or her behavioral repertoire (see Barnes-Holmes, Hayes, & Dymond, 2001). Imagine, for example, that a participant is asked to rate pictures of white and black men as “safe” or “dangerous” (apart from race, the pictures are similar), and the two races are rated equally on semantic differential scales. Such relational responses would likely cohere with other relevant relational networks, such as “Apart from race, the pictures are very similar” and “It is wrong to discriminate on the basis of race.” The critical point here is that explicit measures are typically not completed under high time pressure, and thus participants have sufficient time to engage in the extended relational responding that is needed to produce a response that coheres with one or more other relational responses. When exposed to a time-pressured IRAP, however, the impact of a participant’s elaborated relational responding would be absent or significantly reduced because there is insufficient time, on a trial-by-trial basis, to engage in the additional and sometimes complex relational activity that serves to generate a relationally coherent response.

In summary, therefore, the REC model assumes that the IRAP effect, when produced under appropriate time pressure, is driven largely by immediate and relatively brief relational responses, whereas explicit measures reflect extended and coherent relational networks. Or more informally, the IRAP captures spontaneous and automatic evaluations, whereas explicit measures capture more carefully considered reactions. The core of the REC model explanation for the divergence between implicit and explicit measures of psychologically sensitive attitudes rests on the following basic assumption: Immediate or automatic evaluative responses may or may not cohere with subsequent relational responding; when they cohere, implicit and explicit measures will typically converge, but when they do not, the measures will typically diverge.[16] In other words, it is assumed that participants usually "reject" their immediate and brief relational responses (or automatic evaluations) if they do not cohere with their more elaborate and extended relational responding.[17]

The REC model appears to explain at least some of the specific findings obtained with the IRAP. For example, the difficulty that participants experience in faking an IRAP effect is consistent with the model. If an IRAP effect is driven by immediate or spontaneous relational responses that are the result of historical and current contextual variables, such reactions are unlikely to be modified by a simple instruction to "think the opposite."[18] As an aside, faking *has* been observed in unpublished research conducted by our group, but the practice latency criterion was often not maintained during the test blocks. This finding is consistent with the argument that as response latency increases on the IRAP, the "contaminating" effects of elaborated relational responding increasingly impact on the measure. (Note that faking was also associated with reduced accuracy, but this result is difficult to interpret because the critical competition between consistent and inconsistent response patterns is not maintained when accuracy falls.)

A related finding comes from a recent study that showed that reducing the practice latency criterion from 3,000 to 2,000 ms served to increase racial stereotyping effects on the IRAP (Barnes-Holmes, Murphy, et al., 2010). Such a finding again supports the argument that time pressure moderates the contaminating impact of elaborated relational responding (see also Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005, who showed that overcoming bias on implicit measures is reduced when cognitive capacity is diminished). However, the REC model does not predict that decreasing time pressure on the IRAP will necessarily produce increasing convergence with explicit measures. As time pressure decreases, it is difficult to predict exactly what variables will impact on response latency, and thus the potential utility of the measure is lost. Indeed, results from the Barnes-Holmes, Murphy, et al. (2010) study supported this conclusion because the internal reliability of the IRAP decreased as latencies increased.

[16] The term *diverge* is used here to indicate effects that do not go in the same direction (e.g., if a negative racial bias is observed on the IRAP but not on an explicit measure). Note, however, that even when measures diverge in this way they may still correlate positively. For example, individuals who produce high levels of negative racial bias on an IRAP may produce low levels of positive racial bias on an explicit measure, whereas individuals who produce low levels of negative implicit bias may produce high levels of positive explicit bias. Such a pattern would produce overall effects that diverge in direction on a graph but correlate positively.

[17] It should be noted that the REC model does *not* predict that additional relational activity will always produce a positive response in a psychologically sensitive area. For some individuals, additional responding may produce a negative response that coheres with the initial negative evaluation (e.g., "The black man in the photograph looks dangerous *and* it is okay to discriminate on the basis of race"). Alternatively, additional responding may produce a relational response that allows two initially incoherent networks to cohere (e.g., "The black man looks dangerous, *but* it is wrong to discriminate on the basis of race. *However,* the black man in this *particular* photograph does look quite dangerous").

[18] If participants are also asked to engage with exemplars that support "thinking the opposite" (e.g., Cullen, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009), this may alter the probability of immediate relational responding and thus moderate the IRAP effect. Such an outcome, however, is not readily defined as "faking" because immediate responding has actually been modified by the exemplars.

Another finding that the REC model appears to explain was reported by Cullen et al. (2009). Specifically, a negative bias toward old age on the IRAP was found to be malleable using appropriate exemplar training (e.g., viewing pictures of admired older people, such as Nelson Mandela), although the explicit measures were largely unaffected. The REC model assumes that the exemplar training would impact largely on immediate relational responding but less so on extended and coherent relational networks. Thus, the pro-old exemplars may have served to evoke Old–Positive immediate relational responses on the IRAP, but these were subsequently "rejected" during the explicit measures because they did not cohere with elaborated relational responding. More informally, for participants asked to rate older people in terms of "tired" versus "energetic," some of the Old–Positive exemplars may well have "come to mind," but these were rejected as the basis for a rating because they were deemed atypical (e.g., "Nelson Mandela is amazing, but most old people are still less energetic than most young people I know").

There are other IRAP data that the REC model appears to explain (see, e.g., Barnes-Holmes, Murphy, et al., 2010, for an explanation of differential trial-type effects), and indeed many of the findings in the broader literature on implicit attitudes appear to be consistent with the model. Working through the relevant material is beyond the scope of the current article, but it is worth noting that the REC model bears some similarity to the associative–propositional evaluation (APE) model, which has been used successfully to explain a wide range of findings in implicit attitudes research (e.g., Gawronski & Bodenhausen, 2007; Gawronski, LeBel, & Peters, 2007). Similar to the APE model, the REC model assumes that brief, immediate relational responses or automatic evaluations (a) may be discriminated (i.e., implicit attitudes do not necessarily occur at an unconscious level); (b) are sensitive to current contextual factors and are thus not necessarily impervious to social desirability, self-presentation, or other motivational effects; and (c) do not necessarily involve highly stable and long-established responses.[19]

On balance, unlike the APE model, the REC model does not appeal to dual processes (associative and propositional). Rather, the latter appeals to the single process of arbitrarily applicable relational responding, as defined by RFT.[20] Thus, the divergence between implicit and explicit attitudes is explained not by the interplay between associative and propositional processes, but by the extent to which relational responses are elaborated and cohere with each other. Furthermore, the REC model predicts that automatic

[19] Each of these three assumptions follows logically from the REC model's definition of automatic evaluation as inherently behavioral (i.e., as immediate and brief relational responding). According to RFT, (a) a relational response may itself participate in a relational frame and as such is verbally or "consciously" discriminated (see Barnes-Holmes, Hayes, & Dymond, 2001); (b) relational responses are by definition sensitive to current contextual cues (Barnes-Holmes, Hayes, Dymond, & O'Hora, 2001); and (c) relational responding, like any response class, may be more or less stable and involve relatively short or long behavioral histories (Hayes, Fox, et al., 2001).

[20] The REC model is *not* a "single-process" model because, as a behavior-analytic account, it allows for the involvement of other behavioral processes apart from relational framing, such as respondent conditioning and primary stimulus generalization. Strictly speaking, therefore, the REC model is a *multi*-process model, but one in which the difference between implicit and explicit attitudes is not explained by the interaction between distinct psychological processes. Rather, it is the elaboration and coherence involved in the single process of relational framing that provide the core explanation.

evaluations are not restricted to simple associations or activations but may emerge based on a variety of stimulus relations (see Power et al., 2009, for supporting evidence). Only further study, however, will determine if the REC model offers clear advantages over the APE model, and our research group is currently engaged in this work.

Conclusion

The IRAP and the REC model are offered here as one possible way in which behavior-analytic researchers could analyze and explain the functional similarities and differences between responses that are typically described as automatic or implicit versus deliberative. We fully recognize that our contribution thus far only scratches the surface, and a great deal more work is needed in this area. First, additional research is required to assess the reliability and validity of the IRAP across a variety of domains. Second, research should examine the impact of potentially important moderating variables on the IRAP, focusing on both procedural and participant variables. Third, attempts should be made to develop the IRAP into an instrument that may be used effectively with individual participants. Fourth, and perhaps most important, experimental analyses should focus on the behavioral histories and current contextual variables that establish, maintain, or weaken implicit attitudes and moderate their relationships with other verbal and nonverbal behavioral measures. This final area of research would also provide important tests of the REC model and indeed other alternative explanations for implicit attitudes and beliefs.

References

BACK, M., SCHMUKLE, S. C., EGLOFF, B., & GUTENBERG, J. (2005). Measuring task switching ability in the Implicit Association Test. *Experimental Psychology, 52*, 167–179.

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., POWER, P., HAYDEN, E., MILNE, R., & STEWART, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32*, 169–177.

BARNES-HOLMES, D., HAYDEN, E., BARNES-HOLMES, Y., & STEWART, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record, 58*, 497–515.

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 human language and cognition* (pp. 119–140). New York: Plenum.

BARNES-HOLMES, D., HAYES, S. C., DYMOND, S., & O'HORA, D. (2001). Multiple stimulus relations and the transformation of functions. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 51–71). New York: Plenum.

BARNES-HOLMES, D., MURPHY, A., BARNES-HOLMES, Y. & STEWART, I. (2010). The Implicit Relational Assessment Procedure (IRAP): 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*, 57–80.

BARNES-HOLMES, D., MURTAGH, L., BARNES-HOLMES, Y., & STEWART, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes towards meat and vegetables in vegetarians and meat-eaters. *The Psychological Record, 60*, 287–306.

BARNES-HOLMES, D., WALDRON, D., BARNES-HOLMES, Y., & STEWART, I. (2009). Testing the validity of the Implicit Relational Assessment Procedure and the Implicit Association Test: Measuring attitudes toward Dublin and country life in Ireland. *The Psychological Record, 59*, 389–406.

CAI, H., SRIRAM, N., GREENWALD, A. G., & MCFARLAND, S. G. (2004). The Implicit Association Test's D measure can minimize a cognitive skill confound: Comment on McFarland and Crouch (2002). *Social Cognition, 22*, 673–684.

CAIRNS, E. (1984). Social identity in Northern Ireland. *Human Relations, 37*, 1095–1102.

CONREY, F. R., SHERMAN, J. W., GAWRONSKI, B., HUGENBERG, K., & GROOM, C. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469–487.

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.

DAWSON, D. L., BARNES-HOLMES, D., GRESSWELL, D. M., HART, A. J. P., & GORE, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the Implicit Relational Assessment Procedure: A first study. *Sexual Abuse: A Journal of Research and Treatment, 21*, 57–75.

DIXON, M. R., 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.

GAWRONSKI, B., & BODENHAUSEN, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition, 25*, 687–717.

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*, 181–193.

GREENWALD, A. G., MCGHEE, D. E., & SCHWARTZ, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

GREENWALD, A. G., NOSEK, B. A., & BANAJI, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

HAYES, S. C., BARNES-HOLMES, D., & ROCHE, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Plenum.

HAYES, S. C., FOX, E., GIFFORD, E. V., WILSON, K. G., BARNES-HOLMES, D., & HEALY, O. (2001). Derived relational responding as learned behavior. In S. C. Hayes, D. Barnes-Holmes, & B. Roche (Eds.), *Relational frame theory: A post-Skinnerian account of human language and cognition* (pp. 21–49). New York: Plenum.

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.

MCKENNA, I. M., BARNES-HOLMES, D., BARNES-HOLMES, Y., & STEWART, I. (2007). Testing the fakeability of the Implicit Relational Assessment Procedure (IRAP): The first study. *International Journal of Psychology and Psychological Therapy, 7*, 123–138.

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.

MIERKE, J., & KLAUER, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*, 1180–1192.

O'HORA, D., BARNES-HOLMES, D., ROCHE, B., & SMEETS, P. M. (2004). Derived relational networks and control by novel instructions: A possible model of generative verbal responding. *The Psychological Record, 54,* 437–460.

O'HORA, D., PELAEZ, M., BARNES-HOLMES, D., & AMESTY, L. (2005). Derived relational responding and human language: Evidence from the WAIS-III. *The Psychological Record, 55*, 155–174.

O'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.

POWER, P. M., 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.

SIDMAN, S. (1994). *Equivalence relations and behavior: A research story.* Boston, MA: Authors Cooperative.

STEWART, I., BARNES-HOLMES, D., & ROCHE, B. (2002). Developing an ecologically valid model of analogy using the relational evaluation procedure. *Experimental Analysis of Human Behavior Bulletin, 20*, 12–16.

STEWART, I., BARNES-HOLMES, D., & ROCHE, B. (2004). A functional analytic model of analogy using the relational evaluation procedure. *The Psychological Record, 54*, 531–552.

VAHEY, N. A., BARNES-HOLMES, D., BARNES-HOLMES, Y., & STEWART, I. (2009). A first test of the Implicit Relational Assessment Procedure as a measure of self-esteem: Irish prisoner groups and university students. *The Psychological Record, 59*, 371–388.

WATT, A. W., KEENAN, M., BARNES, D., & CAIRNS, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41*, 33–50.

Study Questions

1. Why is it important to distinguish between automatic and deliberative responses in the study of attitudes and beliefs?
2. What does the acronym IRAP stand for?
3. What is the name of the behavior-analytic theory that provided the general conceptual basis for the IRAP?
4. Studies that involved pitting natural verbal relations against laboratory-induced derived relations provided a specific conceptual foundation for the IRAP. Identify and describe the first study to adopt this strategy.
5. The REP and the IAT were each important cornerstones for the development of the IRAP. What do these acronyms stand for?
6. The IRAP was initially called the IREP. What does the latter acronym stand for and why was the name changed?
7. Examine the diagram in Figure 1 and draw a similar figure that represents an IRAP designed to assess implicit attitudes in a domain of your choice. Below the four trial-type boxes, list all of the target words or pictures that you would employ in your IRAP.
8. Briefly describe the general sequence of practice and test blocks used in a typical IRAP and the performance criteria that are often used to determine if a participant is allowed to progress from practice to test.
9. Download and run the IRAP software from the IRAP website (http://psychology.nuim.ie/IRAP/IRAPSoftware.shtml) and briefly describe the purpose behind each of the settings presented down the left- and right-hand sides of the stimuli-and-settings input screen.
10. Identify the study that explored the impact of reducing the practice latency criterion from 3,000 to 2,000 ms and briefly describe the nature of that impact.
11. Why was the warning message “Too Slow” introduced into the 2008 version of the IRAP? Describe the process for discarding data when this warning message is used in a study.
12. Why is the D_{IRAP} algorithm used to transform latency scores from the IRAP? List the steps involved in calculating the four trial-type D_{IRAP} scores.
13. Examine the graph in Figure 2 and draw a similar graph that represents hypothetical IRAP data from the study you designed under point 7 above. Write a brief interpretation of the findings presented in your graph.
14. How does a “known-groups” analysis test the validity of an IRAP?
15. Briefly describe what may be considered the most interesting feature of the IRAP specifically and implicit measures more generally.
16. What does the acronym REC stand for?
17. Provide a brief description of the REC model’s explanation for the divergence between implicit and explicit measures of psychologically sensitive attitudes.
18. Describe three findings from IRAP research that the REC model appears to explain.
19. List three assumptions that the REC and APE models share.
20. How does the REC model differ from the APE model?
21. List four areas in which additional IRAP research is needed.