The Psychological Record, 2008, 58, 497–516

THE IMPLICIT RELATIONAL ASSESSMENT PROCEDURE (IRAP) AS A RESPONSE-TIME AND EVENT-RELATED-POTENTIALS METHODOLOGY FOR TESTING NATURAL VERBAL RELATIONS: A PRELIMINARY STUDY

Dermot Barnes-Holmes, Eilish Hayden, Yvonne Barnes-Holmes
National University of Ireland, Maynooth

Ian Stewart
National University of Ireland, Galway

The current article reports the first attempt to test the Implicit Relational Assessment Procedure (IRAP), as a group-based measure of natural verbal relations, using both response-latency and event-related potentials as dependent variables. On each trial of the IRAP, participants were presented with 1 of 2 attribute stimuli ("Pleasant" or "Unpleasant"), a positive (e.g., "Love") or negative (e.g., "Murder") target stimulus, and 2 relational terms, "Similar" and "Opposite," as response options. Participants were required to respond as quickly and accurately as possible across blocks of trials, with half of the blocks requiring responses that were deemed consistent (e.g., Pleasant–Love–Similar), and the other half inconsistent (e.g., Pleasant–Love–Opposite), with natural verbal relations. Shorter mean latencies were predicted for consistent than for inconsistent blocks. Two separate experiments supported this prediction. Event-related potentials, gathered during the second experiment, also proved to be sensitive to the IRAP, yielding more negative waveforms for inconsistent relative to consistent blocks of trials. A theoretical interpretation of the IRAP effect is offered, and important directions for future research are highlighted.

The study of human language and cognition has attracted increasing attention among behavior-analytic researchers, with a particular focus on stimulus equivalence and derived stimulus relations (e.g., Hayes, Barnes-Holmes, & Roche, 2001; Sidman, 1994). 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

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

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.

Much of the interest in stimulus equivalence arises from the argument that it may provide a functional-analytic model of semantic relations in natural language (e.g., Barnes & Holmes, 1991). Although the debate surrounding this claim is far from resolved (e.g., Hayes, Barnes-Holmes, & Roche, 2003), a number of researchers have attempted to use the equivalence procedure as a means of testing natural verbal relations or categories. The basic approach involves training and testing for laboratory-induced equivalence relations that are likely to conflict with the verbal or semantic relations that have been established previously by the wider verbal community. Critically, researchers predict that the emergence of laboratory-induced equivalence relations will be hindered when they compete with natural verbal relations.

The first study to adopt the foregoing strategy employed a sample of adult participants who resided in Northern Ireland and a group of English participants who 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 MTS 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 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 (e.g., Barnes, Lawlor, Smeets, & Roche, 1996; Dixon, Rehfeldt, Zlomke, & Robinson, 2006; Leslie, Tierney, Robinson, Keenan, Watt, & Barnes, 1993; Merwin & Wilson, 2005).

While behavior-analytic researchers were exploring stimulus equivalence procedures as a method for revealing natural verbal relations, a growing number of social psychologists were working on a procedure that was designed to evaluate what have been called implicit attitudes. Greenwald and Banaji (1995) defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (p. 8). Although this definition is open to debate, the basic argument is that individuals are often not aware of their implicit beliefs or attitudes or how they can manifest as judgments or actions. It follows, therefore, that traditional self-report methods, such as questionnaires and interviews, which require respondents to reflect on their "conscious" thoughts and feelings, are not best suited to measuring implicit cognitions (de Jong, Pasman, Kindt, & van den Hout, 2001; Dovidio & Fazio, 1992; Gemar, Segal, Sagrati, & Kennedy, 2001). Alternative methodologies that aim to evaluate implicit attitudes have thus been developed, and the best established of these procedures is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998).

The IAT is based on the idea that it should be easier to map two concepts

onto a single response when those concepts are somehow similar or associated in memory than when the concepts are unrelated or dissimilar (De Houwer, 2002). Greenwald et al. (1998) tested this idea by presenting participants with names of flowers (e.g., tulip), names of insects (e.g., beetle), positive words (e.g., love), and negative words (e.g., ugly). It was presumed that the concepts "flower" and "positive" are associated in memory, as are the concepts "insect" and "negative." Based on this assumption, responses should be faster when one key is assigned to both "flower" and "positive," and a second key to "insect" and "negative" (Consistent Task), than when one key is assigned to "flower" and "negative," and a second key to "insect" and "positive" (Inconsistent Task). The results of the experiment supported this prediction: group-based average reaction times were faster for the Consistent than for the Inconsistent tasks (Greenwald et al., 1998, Experiment 1).

Greenwald and colleagues have published a wide range of studies demonstrating that the IAT may reveal implicit attitudes, such as racial stereotypes, which participants typically deny when questionnaires and open-ended interviews are used. Research has shown, for example, that many participants who claimed *not* to hold racist attitudes nevertheless responded more quickly and more accurately on an IAT when asked to categorize names typical of white persons with positive words and names typical of black persons with negative words, than when asked to categorise white with negative and black with positive (see Greenwald, Banaji, Rudman, Farnham, Nosek, & Mellott, 2002, for a review). This basic IAT effect has now been replicated numerous times (e.g. de Jong, 2002; de Jong et al., 2001; Gemar et al., 2001; Teachman, Gregg, & Woody, 2001), and the methodology has become an increasingly popular measure of group-based implicit attitudes and dysfunctional beliefs.

Although IAT research emerged from mainstream social psychology and is often discussed in cognitive and mentalistic terms, the IAT effect itself may be interpreted behaviorally (see Barnes-Holmes, et al., 2004). The basic argument is that the IAT effect occurs because participants are required to respond to functionally similar equivalence classes as functionally equivalent during the consistent task (e.g., by pressing the same key for *positive* or *flower* words) but are required to respond to functionally nonequivalent classes as functionally equivalent during the inconsistent task (e.g., by pressing the same key for *negative* and *flower* words).[1] In effect, responses are slower for the latter task because they involve responding against previously established derived or verbal relations (see O'Toole, Barnes-Holmes, & Smyth, 2007, for empirical support). In very broad terms, this is the same behavioral explanation that was provided for the disruption of equivalence class formation when the stimuli involved participated in mutually exclusive verbal categories, such as *Protestant* and *Catholic.*

The current research drew on the foregoing behavioral explanation of

1 From this perspective, the IAT involves four separate equivalence classes, with stimuli in two of the classes possessing appetitive behavioral functions (e.g., Positive and Flower) and stimuli in the other two classes possessing aversive functions (e.g., Negative and Insect). During consistent tasks, the same response function (e.g., press left) is established for the two appetitive classes and another function (e.g., press right) is established for the two aversive classes. For inconsistent tasks, however, the same response function is established for both appetitive and aversive classes (e.g., left = Positive or Insect and right = Negative or Flower).

both the IAT effect and the equivalence-based studies as the basis for creating a new group-based procedure for assessing previously established verbal relations (Barnes-Holmes et al., 2006). We predicted that if a procedure requires participants, under time pressure, to switch between response patterns that are consistent and inconsistent with previously established natural verbal relations, an IAT-like effect should be observed. That is, average latencies for a group of participants should be slower for response patterns that are inconsistent rather than consistent with existing verbal relations.

In developing a relevant procedure, we drew on earlier work with what is called the Relational Evaluation Procedure (REP). 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 "Similar." In recent years, a number of studies have employed the REP in the analysis of relational responding in adult humans (O'Hora, Barnes-Holmes, Roche, & Smeets, 2004; O'Hora, Pelaez, Barnes-Holmes, & Amesty, 2005; Stewart, Barnes-Holmes, & Roche, 2002, 2004).

The procedure that is the focus of the current study, which we developed from the REP, is called the Implicit Relational Assessment Procedure (IRAP). In fact, initially the IRAP was called the IREP, but the former 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. In essence, the IRAP is a combination of the REP and the IAT. Like the REP, the IRAP involves presenting specific relational terms (e.g. SIMILAR, OPPOSITE, BETTER, WORSE) so that the properties of the relations among the relevant stimuli can be assessed. And similar to the IAT, 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 should be shorter across blocks of consistent relative to inconsistent trials. Or in other words, participants should respond more rapidly to relational tasks that reflect their current beliefs than to tasks that do not.

The current article reports the first IRAP study and as such is purely exploratory in nature. If the predicted IRAP effect is observed, then naturally many questions and issues will arise, and addressing these will require further systematic research across a large number of separate studies. In Experiment 1 of the current study, however, our primary concern is to determine simply whether the IRAP, using natural verbal relations, does indeed produce the predicted effect described above.

On each trial of the IRAP, participants were presented with one of two attribute stimuli ("Pleasant" or "Unpleasant"), a positive (e.g., "Love") or negative (e.g., "Murder") target stimulus, and two relational terms, "Similar" and "Opposite," as response options. The IRAP involved presenting alternating blocks of consistent and inconsistent trials (based on Greenwald et al.'s [1998] categorization of pleasant and unpleasant terms). Shorter mean latencies were predicted for consistent blocks of trials (e.g., Pleasant – Love – Similar) relative to inconsistent blocks (e.g., Pleasant – Love – Opposite). Experiment 1 tested this prediction, and Experiment 2 attempted to replicate the initial findings while also recording electroencephalograms (EEGs) as another dependent measure of the IRAP effect.

Experiment 1

Method

Participants. Twenty-eight undergraduate participants, 11 male and 17 female, agreed to participate. Sixteen participants were assigned to an experimental group, and twelve to a control group (described below). Ages ranged from 18 to 30 years. No financial payment or other inducements were offered for participation in the study.

Apparatus and materials. The experimental tasks were presented to participants in a small quiet room, on a standard Pentium 4 personal computer, programmed in Microsoft Visual Basic 6.0 (IRAP software may be downloaded from http://psychology.nuim.ie/IRAP/IRAP_1.shtml).

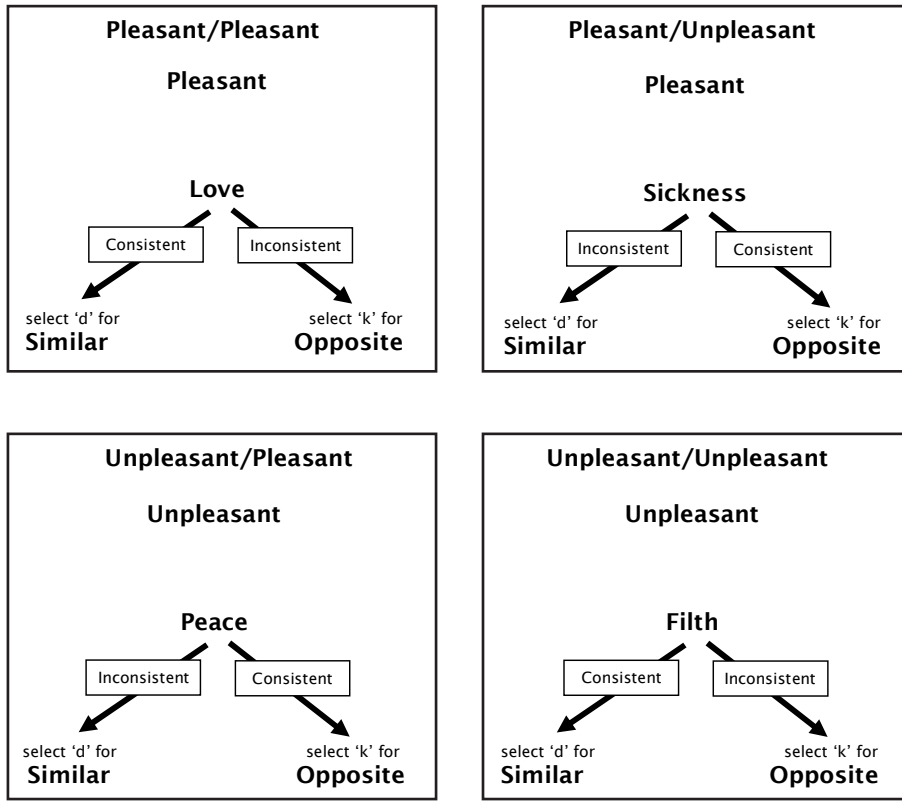

Figure 1. Examples of the four IRAP trial types. The attribute (Pleasant or Unpleasant), target word (love, sickness, peace, and filth, etc.), and response options (Similar and Opposite) appeared simultaneously on each trial. Arrows with superimposed text boxes indicate which responses were deemed consistent or inconsistent (boxes and arrows did not appear on screen). Selecting the consistent response option during a consistent block, or the inconsistent option during an inconsistent block, cleared the screen for 400 ms before the next trial was presented; if the inconsistent option was chosen during a consistent block, or the consistent option during an inconsistent block, a red X appeared on screen until the participant emitted the alternative response.

Procedure. Each participant sat in front of the computer, which presented the instructions and the stimuli and recorded all responses. Equal numbers of participants were assigned randomly to one of two conditions: consistent-relations-first or inconsistent-relations-first. On each trial of the IRAP, four words appeared simultaneously on the computer screen. An attribute stimulus, either "Pleasant" or "Unpleasant," appeared at the top, with a single target word presented in the center, and two relational terms, "Opposite" and "Similar," at the bottom left- and right-hand corners of the screen (see Figure 1). All of the stimuli remained visible until the participant pressed one of the response keys.

The task involved choosing one of the two relational terms ("Similar" or "Opposite"). To choose the term on the left, participants pressed the "d" key, and to choose the term on the right, participants pressed the "k" key. Choosing the relational term that was deemed correct for that block of trials removed all four stimuli from the screen for 400 ms before the next trial was presented. Choosing the relational term that was deemed incorrect for that block of trials produced a red X in the middle of the screen (immediately below the target word). The participant was not allowed to continue to the next trial until he or she chose the correct relational term (for that block of trials).

Before starting the first block of IRAP trials participants were presented with the following instructions:

> For each of several relating tasks you will be shown words one at a time in the middle of the computer screen.
>
> Your task is to use the feedback from the computer to learn to relate each item as fast as you can by pressing EITHER the "d" key or the "k" key.
>
> IMPORTANT: Press the "d" key with your left index finger, or "k" key using your right index finger.
>
> The response options associated with the "d" and "k" keys will be shown at the bottom of the screen. Please pay close attention to these options—they will change position unpredictably for each relating task!
>
> For the relating task one of the following two words will appear at the top of the screen
>
> "Pleasant" or "Unpleasant"
>
> And the following two response options will appear at the bottom of the screen.
>
> "Similar" or "Opposite"
>
> For each task, you must look at the word at the top, then look at the word in the middle, and finally choose one of the two words at the bottom by pressing the "d" or "k" key. The computer will tell you if you have made the correct or wrong choice.

> In some parts of the experiment the feedback from the computer may make sense to you, but in other parts it may not. This is part of the experiment.
>
> The most important thing for you to do is to respond quickly and to make as few errors as you can.
>
> If you make an error you will see a RED "X" below the stimulus—when this happens, you have to make the correct response to proceed.
>
> Press space bar to continue.

All participants were exposed to eight blocks of 24 trials; two practice blocks followed by six test blocks. Within each block, 12 target words were presented in a random order with the constraint that each word was presented twice, once in the presence of "Pleasant" and once in the presence of "Unpleasant." The 12 target words were Caress, Freedom, Health, Love, Peace, Cheer (all defined as positive), Abuse, Crash, Filth, Murder, Sickness, and Accident (all defined as negative; see Greenwald et al., 1998). The left-right position of the relational terms alternated randomly across all trials within each of the eight blocks.

The first block of trials in the consistent-relations-first condition reinforced responses that were deemed relationally consistent based on Greenwald et al.'s (1998) consistent and inconsistent categorization of pleasant and unpleasant terms. Given the attribute "Pleasant," for example, and any of the six positive target words, choosing the relational term "Similar" immediately progressed the computer program to the next trial (after 400 ms). If, however, "Opposite" was chosen, a red "X" appeared below the stimulus and the participant had to make the correct response by choosing "Similar" for the computer to progress to the next trial.

The second block of trials in the consistent-relations-first condition reinforced responses that were deemed relationally inconsistent based on Greenwald et al.'s (1998) categorization of pleasant and unpleasant terms. Given the attribute "Pleasant," for example, and any of the negative target words, choosing the relational term "Similar" immediately progressed the computer program to the next trial (after 400 ms); if "Opposite" was chosen, the red "X" appeared and a correction response was required to progress (see Figure 1 and caption for a description of all four IRAP trial types, including the definition of consistent versus inconsistent).

For the remaining six blocks of trials, blocks 3, 5, and 7 reinforced consistent relational responses and blocks 4, 6, and 8 reinforced inconsistent responses. This pattern of reinforcement contingencies was reversed for participants assigned to the inconsistent-relations-first condition; blocks 1, 3, 5, and 7 were inconsistent and blocks 2, 4, 6, and 8 were consistent.

The procedure for the control group differed from the experimental group (outlined above) in only one way. The relational terms "Similar" and "Opposite" were replaced with the nonwords "Cug" and "Zid," respectively. Thus participants chose between Cug and Zid on each IRAP trial by pressing either the d or k keys. Given that the response options were nonwords, the IRAP blocks cannot be described meaningfully as consistent and inconsistent. Nevertheless, the response functions for Cug and Zid were counterbalanced equally across participants to parallel the procedure applied to the experimental group. The

rationale behind running a control group was as follows. If an IRAP effect is dependent on the previously established relational functions for the words "Similar" and "Opposite," as suggested in the introduction, then the control condition, in which no such relational terms are used, should fail to produce the effect.

Before each of the two practice blocks, the following message appeared on the screen: "This is a practice - errors are expected." Before each of the six test blocks, the message read, "This is a test - go fast, making a few errors is OK." At the end of each practice and test block, two messages appeared. The first reported the participant's percentage of correct responses and median response latency for that prior block. The second message informed the participant that the previously correct and wrong answers would be reversed in the next block of trials (i.e., participants were made aware of the change in feedback contingencies before each block). After completion of all eight blocks, a final message appeared indicating that the experiment was over.

Results and Discussion

All 28 participants (16 experimental and 12 control) completed the experiment. The primary datum was response latency, defined as the time in milliseconds (ms) that elapsed between the onset of the trial and a correct response emitted by the participant. For the purposes of the current study, the IRAP response latency data were transformed using the basic C4 algorithm, which has been suggested for use with the IAT (Greenwald, Nosek, & Banaji, 2003). A number of the other algorithms also suggested by Greenwald, et al. were used, but they each yielded similar conclusions, for both this and the next experiment, and thus only the C4 analyses are reported here.

For the C4 algorithm, responses of more than 3,000 ms were recorded as 3,000 ms, and responses of less than 300 ms were recorded as 300 ms. Mean response latencies were calculated for each participant for each of the six test blocks, providing three mean latencies for consistent blocks and three mean latencies for inconsistent blocks. For the purposes of analysis, the data were also divided between the two test sequences: consistent- versus inconsistent-relations-first.

The overall mean latencies calculated across participants are presented in Figure 2. Within each pair of test blocks, and for both test sequences, mean latencies were shorter for the consistent test block relative to its corresponding inconsistent block. In addition, longer latencies were recorded for the consistent- relative to the inconsistent-relations-first sequence within each block. The latency data were subjected to a $2 \times 3 \times 2$ mixed repeated measures analysis of variance (ANOVA) with IRAP condition (consistent versus inconsistent) and IRAP blocks (first, second, and third pair) as repeated measures and test sequence (consistent- versus inconsistent-relations-first) as a between-participant variable. The main effect for IRAP condition proved to be significant, $F(2, 14) = 23.88$, $p < .001$, $\eta_p^2 = 0.6$, as did the effect for test sequence, $F(2, 14) = 8.214$, $p = .01$, $\eta_p^2 = 0.4$. The main effect for IRAP blocks and all four interaction effects were nonsignificant (all ps $> .3$). The predicted IRAP effect was thus observed and was not moderated significantly by either blocks or test sequence.

The latency data for the control group (not presented) were subjected to the same type of repeated measures ANOVA that was used for the experimental condition. For the purposes of analysis the two response-option stimuli

"Cug" and "Zid" were defined arbitrarily as "Similar" and "Opposite," respectively. The ANOVA yielded no significant main or interaction effects (all *p*s > .1). Furthermore, the overall mean latencies for what were defined as consistent and inconsistent trials differed by only 8 ms, and the difference was in the opposite direction to that of the experimental group (consistent trials, $M = 1{,}786$, $SE = 80$; and inconsistent trials, $M = 1{,}778$, $SE = 73$). As predicted, therefore, the experimental condition yielded an IRAP effect but the control condition did not. It is perhaps worth noting that subsequent control conditions conducted in our laboratory have also yielded absent IRAP effects when nonrelational but "real" words were used as response options (e.g., "Black" versus "White").

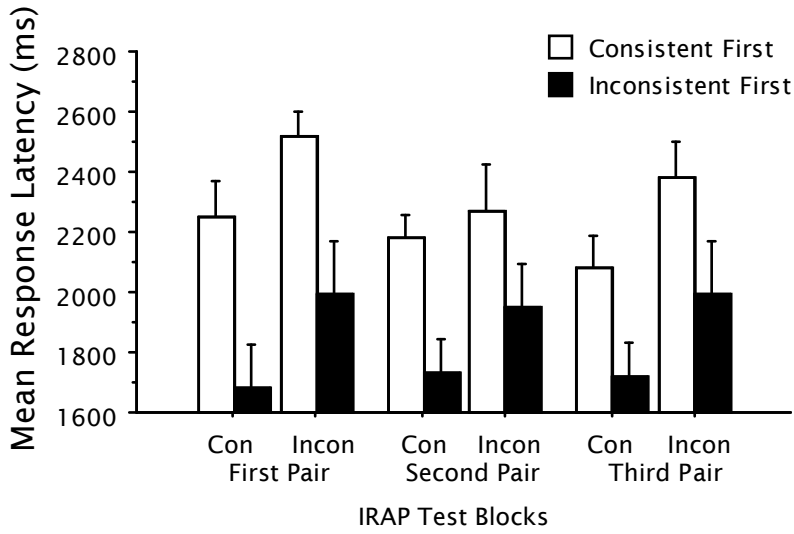

Figure 2. Overall adjusted mean response latencies (in milliseconds), including standard error bars, for each block of trials, for consistent- and inconsistent-relations-first test sequences, for the experimental condition of Experiment 1. "Con" indicates consistent test blocks and "Incon" indicates inconsistent test blocks.

Although this first experiment yielded the predicted IRAP effect, it seemed important, given the novelty of the procedure, to attempt to replicate the results with an experimentally naïve sample. Conducting a second experiment also afforded us the opportunity to assess the impact of the IRAP on participants' neural activity using multiple EEGs.

Experiment 2

Exploring the use of EEG measures was deemed important because the IRAP differs from the IAT in a possibly critical manner. With the IAT the locations of some of the stimuli change across consistent and inconsistent trials (e.g., the left-right positioning of the labels *positive* and *negative* switch). Consequently, any critical differences in EEG patterns observed with the IAT may be contaminated by these stimulus changes, thus obscuring the measurement of differential response strengths between consistent and inconsistent trials. With the IRAP, however, the same stimulus configurations are presented on consistent

versus inconsistent trials (only the feedback differs). Consequently, any differences in EEG activity across these trials, which occur *before* the feedback, would likely reflect the different private behavioral processes that occur just prior to emitting history-consistent versus history-inconsistent relational responses.

In Experiment 2 recordings were taken from multiple EEG signals, while participants completed the IRAP, and these signals were then transformed into event-related potentials (ERPs; e.g., Kutas, 1993; Kutas & Hilliard, 1984). This method of recording neural activity is relatively noninvasive and inexpensive, and allows researchers to investigate the neurophysiological processes underlying functions such as perception, semantic relations, and reasoning (see Barnes-Holmes, Staunton, et al. 2005; Barnes-Holmes, Regan, et al., 2005, for examples of ERP research within the behavior-analytic tradition).

Generating ERP data involves time-locking the EEG signals to a particular series of events and then averaging the signals across trials. The process of averaging allows the researcher to distinguish the brain's normal background activity from the activity produced by the stimuli presented in the experiment. In effect, each EEG signal for a particular set of stimuli is collated and averaged to produce a single waveform for each site, and then these waveforms are averaged across participants to provide "grand average" waveforms that provide group-based measures of the effect of the targeted stimulus or stimuli.

There is a range of waveforms associated with ERP measures. Some ERPs, for example, are thought to be correlated with specific cognitive processes, such as understanding words or discriminating one type of auditory stimulus from another. Such ERPs tend to occur at around 300 or 400 ms after the onset of the stimulus. The use of ERP measures with the IRAP in the current study was entirely exploratory, and thus we refrained from making specific predictions pertaining to the ERP waveforms that might emerge. Nevertheless, one ERP measure that seemed particularly relevant to the IRAP is the late negative waveform, known as the N400 (see Holcomb & Anderson, 1993; Kounios & Holcomb, 1992). This waveform is usually produced when participants are required to respond to stimuli that are deemed to be unrelated, unexpected, or incorrectly paired in some sense (referred to as low *cloze-probability).* For example, presenting pairs of words that are semantically unrelated tends to produce an N400, but words from the same semantic categories produce a much reduced or completely absent effect. Insofar as inconsistent trials on the IRAP require "incorrect" or "unexpected" responses, perhaps a more negative waveform will emerge for inconsistent relative to consistent trials. In Experiment 2, therefore, separate ERP waveforms for consistent and inconsistent IRAP trials were collected across a range of sites.

Method

Participants. Twelve participants, 4 male and 8 female, agreed to participate. Ages ranged from 18 to 28 years. No financial payment or other inducements were offered for participation in the study.

Apparatus and materials. The entire experiment was conducted in an electrically shielded room in the human neuroscience laboratory in the Department of Psychology at NUI, Maynooth. The stimuli and materials used were identical to those of Experiment 1. To record EEG signals during the IRAP task, a Brain Amp magnetic resonance (MR) compatible (Class IIa, Type BF) with approved control software (Brain Vision Recorder 1.0), and electrode cap (BrainCap/

BrainCap MR) were used. Two Dell personal computers (Pentium 4) were employed for the experiment. One computer controlled the Brain Amp, and a second the IRAP. The ERPs data were analyzed using approved analysis software (Brain Vision Analyser 1.0). Hardware and software were manufactured and supplied by Brain Products GmbH, Munich, Germany.

Procedure. The IRAP was identical to that of Experiment 1. Six participants were assigned to the consistent-relations-first group, and six to the inconsistent-relations-first group. No control group was employed in Experiment 2. Participants were first attached to the Brain Amp and were then exposed to the entire IRAP procedure. Each session, consisting of electrode placement and then the IRAP task, lasted on average 1 hr and 15 min. Only the ERPs data from the six test blocks were analyzed.

Evoked potentials were recorded and analyzed from 15 sintered AG/AG-CI scalp electrodes positioned according to the international 10-20 system. The 15 sites chosen for recording were Fp1, Fp2, F3, Fz, F4, F7, F8, C3, Cz, C4, P3, Pz, P4, O1, and O2. The central vertex electrode was used as reference and the FPz as ground. Amplifier resolution was 0.1 μV (range, ± 3.2768 mV) and the bandwidth set between 0.5 and 62.5 Hz, with a sampling rate of 250 Hz. The notch filter was set at 50 Hz. All electrode impedances were at or below 5 kΩ. The EEG was collected continuously and edited off-line.

Results and Discussion

Response latencies. The response latency data, calculated in the same manner as for Experiment 1, are presented in Figure 3. Once again shorter overall

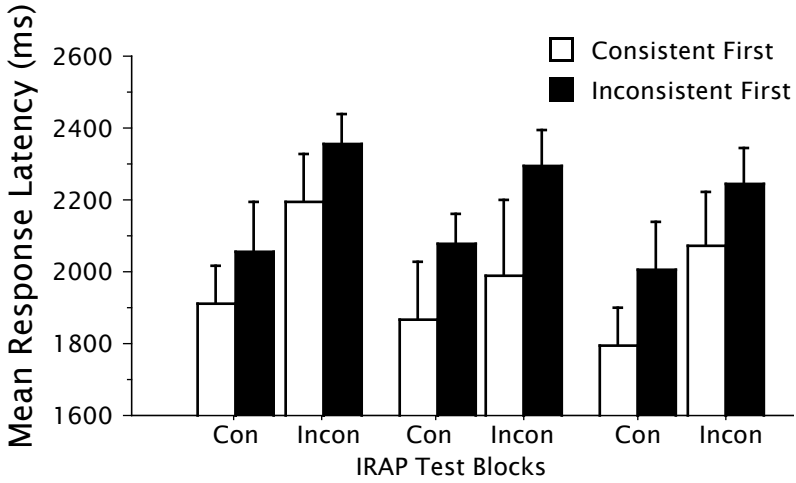

Figure 3. Overall adjusted mean response latencies (in milliseconds), including standard error bars, for each block of trials, for consistent- and inconsistent-relations-first test sequences, for Experiment 2. "Con" indicates consistent test blocks and "Incon" indicates inconsistent test blocks.

mean latencies were observed during the first, second, and third consistent test blocks relative to their corresponding inconsistent test blocks. Similar to Experiment 1, the latency data were subjected to a $2 \times 3 \times 2$ mixed ANOVA, with IRAP condition (consistent versus inconsistent) and IRAP blocks (first, second, and third pair) as repeated measures and test sequence (consistent-

versus inconsistent-relations-first) as a between-participant variable. Only the main effect for IRAP condition proved to be significant, $F(2, 14) = 18.9$, $p = .0014$, $\eta_p^2 = 0.6$ (all remaining ps > .2). Experiment 2 thus produced the predicted IRAP effect, which again was not moderated significantly by either blocks or test sequence. These results replicate the data from Experiment 1, except that the difference between consistent- versus inconsistent-relations-first was nonsignificant, and the effect was in the opposite direction to the previous experiment (we shall return to this issue in the general discussion, below).

ERPs data. The continuous EEG signals for each of 12 participants were filtered (0.53 Hz, time constant = 0.3 s, 24 dB/octave roll-off) and then segmented for consistent and inconsistent trial types. To reduce noise for the ERPs analyses, the data for all three consistent test blocks were collapsed, as were the data for inconsistent test blocks. To further reduce noise, the data for consistent- and inconsistent-relations-first participants were also combined. The segments were divided into 1,300 ms epochs commencing 100 ms before onset of the stimuli on each trial (overlapping segments were removed). Vertical and horizontal ocular artifacts were then corrected, and any segments on which EEG or electro-ocular activity exceeded ±75 μV were rejected (the data from participant 10 were removed from subsequent analyses because no segments were artifact free). The remaining segments were then baseline corrected (using the 100 ms prestimulus interval) and finally averaged for consistent and inconsistent IRAP trial types. The grand average waveforms for each of 10 electrode sites (F7, F8, C3, C4, Cz, P3, P4, Pz, O1, and O2) for consistent (light lines) versus inconsistent (dark lines) trials are presented in Figure 4. The five frontal sites (Fp1, Fp2, Fz, F3, F4) also were used, but differences between consistent and inconsistent trials were not reliably detectable; in accordance with common practice (e.g., Weisbrod et al., 1999), these data are not reported. Visual inspection of the waveforms from the 10 sites indicated little evidence of differential activity for the two trial types until approximately 600 ms after stimulus onset. Specifically, the inconsistent, relative to the consistent, trials produced greater negativity on all panels.

The area dimensions (μV × ms) for each ERP waveform (in the temporal region 600–1,000 ms) for each participant were calculated, yielding either positive or negative values with respect to the 0 μV level. The two central sites were subjected to a 2 × 4 ANOVA with location (central sites, Cz and Pz) and IRAP condition (Consistent versus Inconsistent) as repeated measures. The ANOVA revealed significant main effects for position, $F(1, 10) = 11.46$, $p = .0069$, $\eta_p^2 = 0.5$, and IRAP condition, $F(2, 9) = 8.113$, $p = .0173$, $\eta_p^2 = 0.4$, and an interaction effect was also found, $F(2, 9) = 9.2$, $p = .0126$, $\eta_p^2 = 0.5$. Two one-way ANOVAs were then performed for each site, and they revealed significant differences between consistent versus inconsistent waveforms for Cz, $F(1, 10) = 6.59$, $p = .02$, $\eta_p^2 = 0.4$, and Pz, $F(1, 10) = 8.894$, $p = .0138$, $\eta_p^2 = 0.5$.

A 2 × 4 × 2 ANOVA was then conducted with laterality (left and right), position (F7-F8, C3-C4, P3-P4, O1-O2), and IRAP condition as repeated measures factors. The ANOVA revealed significant main effects for position, $F(3, 30) = 24.189$, $p < .0001$, $\eta_p^2 = 0.7$, and IRAP condition, $F(1, 10) = 5.79$, $p = .0368$, $\eta_p^2 = 0.4$, but not for laterality, $p = .24$. All interaction effects were nonsignificant, all ps > .8, except for position by IRAP condition, $F(3, 30) = 4.29$, $p = .0124$, $\eta_p^2 = 0.3$. A series of Scheffé post hoc tests indicated that each of the four positions differed from each other significantly, all ps ≤ .01, except for the C3-C4 versus P3-P4 comparison, $p > .9$.

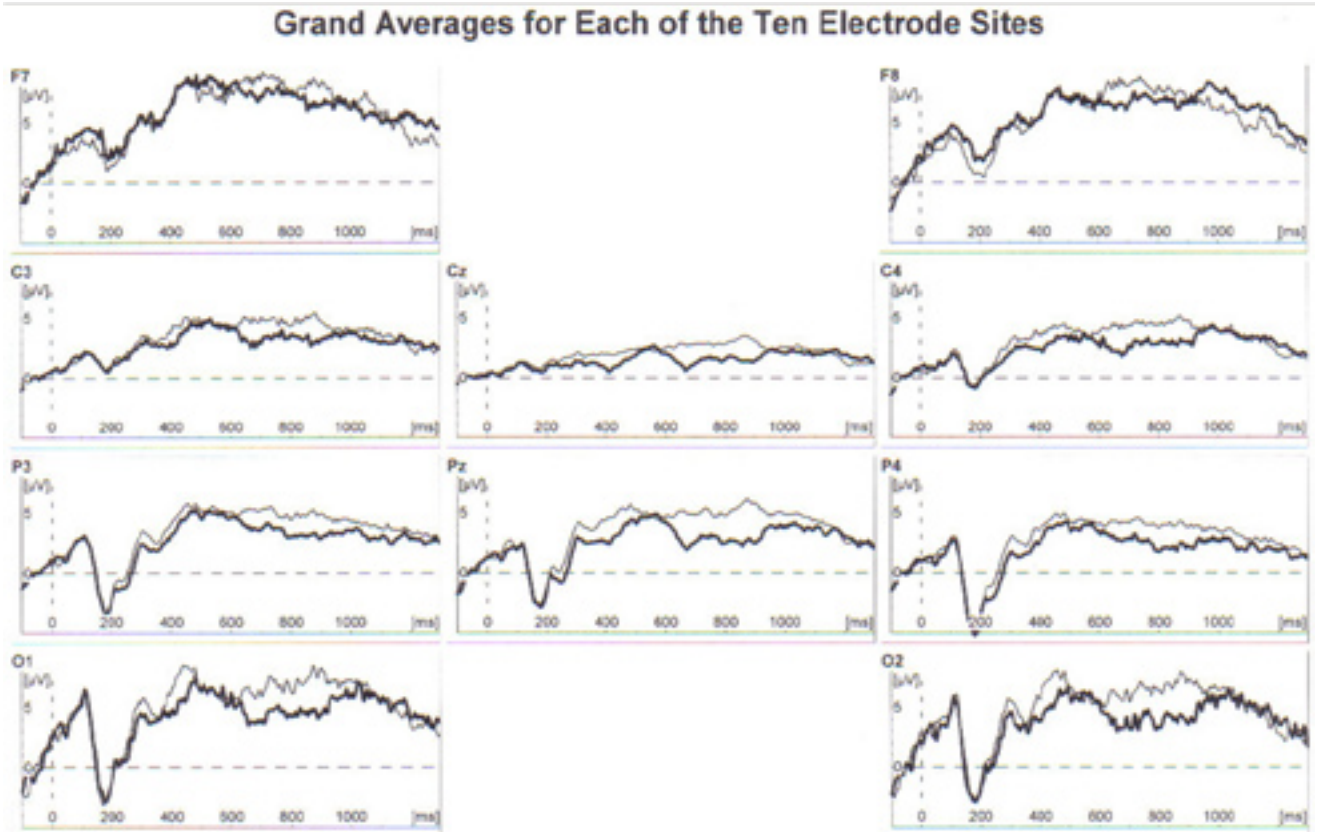

Figure 4. Grand average waveforms from 11 participants for consistent (light lines) and inconsistent (dark lines) trial types at electrode sites F7, C3, P3, O1 (left column), Cz, Pz (center column), F8, C4, P4 and O2 (right column). Stimuli were presented at 0 ms.

Given the significant interaction for position by IRAP condition, four separate 2 × 2 repeated measures ANOVAs were required to analyze each position separately. The ANOVA for position F7-F8 yielded no significant effects, *p*s > .2. Position C3-C4 yielded a significant main effect for IRAP condition, $F(1, 10) = 6.069$, $p = .0335$, $\eta_p^2 = 0.4$, but laterality and the interaction were nonsignificant, *p*s > .2. Position P3-P4 yielded main effects for both laterality, $F(1, 10) = 10.858$, $p = .0081$, $\eta_p^2 = 0.5$, and IRAP condition, $F(1, 10) = 7.595$, $p = .0203$, $\eta_p^2 = 0.4$, but no interaction effect, $p = .56$. Finally, position O1-O2 yielded significance for IRAP condition, $F(1, 10) = 7.138$, $p = .0234$, $\eta_p^2 = 0.4$, but not for laterality or for the interaction, *p*s > .09.

Summary. Experiment 2 replicated the basic IRAP effect obtained in Experiment 1 by producing longer mean response latencies for inconsistent relative to consistent trials. The ERPs analyses indicated significantly greater positivity in the two waveforms for the Pz relative to the Cz site; both sites also yielded significantly more negative waveforms for inconsistent relative to consistent IRAP trials. Overall, significant differences were also found between each position, F7-F8, C3-C4, P3-P4, and O1-O2, except for the central-parietal comparison (generally, the two waveforms were more positive in the frontal and occipital regions). Further analyses revealed significant differences between consistent and inconsistent waveforms for areas C3-C4, P3-P4, and O1-O2 but not for areas F7-F8. A significant difference was found for laterality in area P3-P4. In general, inconsistent waveforms became increasingly more negative than consistent waveforms, in the 600–1,000 ms interval, for both hemispheres, in all areas except for F7-F8.

General Discussion

Both Experiments 1 and 2 produced the predicted IRAP effect, with significantly shorter mean response latencies for consistent relative to inconsistent trials. Critically, neither experiment indicated that the IRAP effect was moderated significantly across successive test blocks or by the test sequence, suggesting that the effect was relatively stable and was not determined simply by whatever stimulus relations were practiced or tested first. These findings are broadly consistent with the IAT effect, which has also been found to be relatively reliable across repeated exposures and across the two different practice/test sequences that are typically employed in IAT research (Greenwald et al., 2003).

In Experiment 1, significantly longer latencies were observed for the consistent- relative to the inconsistent-relations-first test sequence. It remains unclear why this pattern emerged, but it failed to recur in Experiment 2, and in fact the effect was in the opposite direction (although nonsignificant). Given that subsequent IRAP research from our group has also repeatedly failed to reproduce the sequence effect (e.g., McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007), it seems wise to view it as a possible statistical artifact, the likes of which may emerge from time to time in the context of small *n* research. In other words, it seems likely that the 8 participants assigned to the inconsistent-relations-first condition tended, simply by chance, to be relatively fast responders on the IRAP. In any case, as noted above, the critical IRAP effect was observed across both experiments and it was not moderated by test sequence.

In Experiment 2, the grand average waveforms for areas Cz-Pz, C3-C4, P3-P4, and O1-O2 showed significantly greater negative deflections for

inconsistent relative to consistent trials, between 600 and 1,000 ms. There was little suggestion that the differences were greater for one hemisphere than the other, with only a significant difference between P3 and P4. As noted previously, research in the neurocognitive literature has reported that negative ERPs components are modulated by what has been called *cloze-probability,* which can be described as degree of expectedness. Some studies have investigated the effect in terms of the final word in a sentence. For example, the stimulus "it is hard to admit when one is *asleep*" elicits a more negative waveform than the stimulus "it is hard to admit when one is *wrong*" (Kutas, 1993; Kutas & Hilliard, 1984). Although somewhat speculative, it is possible that the relatively negative waveforms elicited by the inconsistent IRAP trials indicates that the required response on those trials is low probability relative to that of the consistent trials. In other words, the contradictory relational responses required on inconsistent trials (relative to the participant's prior learning history) may overlap functionally with a low-probability sentence completion, as described in the example above.

On balance, the waveforms observed for the inconsistent trials were only negative relative to the consistent trials, and they occurred during the 600–1,000 ms interval (i.e., the waveform did not occur around 400 ms [N400], which is typically associated with cloze-probability). Given the relative complexity of the IRAP task, however, the increased interval might well be expected, and the absence of a standard N400 effect is perhaps not surprising. At the present time, therefore, it should simply be noted that the waveforms for the inconsistent IRAP trials may be suggestive of a low cloze-probability effect, and this issue could be pursued in future research.

Subsequent studies might also explore the effect of repeated IRAP exposures on the ERP waveforms. The size of the difference in response latencies between consistent and inconsistent trials has been found to decrease when participants are exposed to a second complete IRAP (McKenna et al., 2007). Perhaps, therefore, the size of the waveforms for inconsistent trials might also attenuate during a second exposure. On balance, preliminary work using the IAT suggests that differential ERPs waveforms may be maintained across successive exposures, even when differences in response latencies are not (Barnes-Holmes et al., 2004). It thus remains to be seen if a similar persistence in waveforms is observed using the IRAP.

The ERPs data from Experiment 2 failed to produce a significant difference between the consistent and inconsistent waveforms for areas F7 and F8, although the difference was in the same direction as for the other sites—perhaps a larger *n*, more typical of ERPs research, might have reached significance. It is also worth noting that the two frontal sites produced the greatest positive activations (for both waveforms), relative to the other sites, which indicates that the IRAP produces a "heavy load" on an area that is typically associated with higher cognitive processes, and more recently with relational reasoning (Waltz et al., 1999). In short, the frontal site activity is consistent with both previous cognitive neuroscience research and the argument that the IRAP is an intensely relational task.

In the introduction, we predicted that average response latencies, emitted under time pressure, should be slower for response patterns that are inconsistent rather than consistent with existing verbal relations. The current data supported this prediction, and thus it seems important to suggest a tentative process-based explanation for the IRAP effect (see Barnes-Holmes et al., 2006).

At present, our explanation is as follows. Each trial of the IRAP presents a target stimulus with contextual cues that specify particular relational (e.g., *same* versus *opposite*) and functional dimensions (e.g. *pleasant* versus *unpleasant*), which produces an immediate and private relational response before the participant actually presses a response key (the participant may or may not be "consciously aware" of this private response). As operant theorists, we assume that the probability of the initial response will 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 during a consistent IRAP trial that response will tend to possess the correct key-pressing function; during an inconsistent trial, however, the response will tend to possess the wrong function. Thus, across multiple trials the average latency for inconsistent blocks will be longer than for consistent blocks. In short, the IRAP effect is based on immediate, private, and perhaps "unconscious" relational responding, which is made apparent when the behavioral system is put under pressure to respond quickly and accurately.[2]

Of course, the adequacy of the foregoing explanation of the IRAP effect will be determined only through systematic empirical inquiry, and indeed numerous other issues will require detailed analysis. For example, it remains to be determined to what extent the IRAP is an *implicit* measure. De Houwer (2006) argued recently that a measure is implicit if one or more of the following criteria apply: participants (a) are not aware that the relevant attitude is being measured; (b) do not have conscious access to the attitude; or (c) have no control over the outcome of the measure. In the context of the current study, participants were almost certainly aware that their "attitudes" to the pleasant and unpleasant target words were being assessed in some manner. Furthermore, participants were likely aware of their attitudes to the target words. Finally, the present research was not designed to determine the extent to which participants could control the measurement outcome. On balance, recent findings from our research group indicate that participants possess little control over the IRAP effect (McKenna et al., 2007), and in the context of "socially sensitive" stimuli the IRAP may produce effects that diverge from consciously reported attitudes (Barnes-Holmes et al., 2006). At the current time, therefore, there is evidence, albeit limited, that the IRAP meets the second two criteria for an implicit measure.

In closing, it must be acknowledged that considerable empirical work will be required to assess the reliability and validity of the IRAP as a measure of implicit attitudes and beliefs. Furthermore, it should also be recognized that the IRAP is being developed, at least at this stage, as a group-based measure. Perhaps when its potential in this regard has been determined, it may be useful to explore its value as an assessment tool for individual cases. In summary, therefore, the current response latency data and EEG recordings clearly indicate that the IRAP is sensitive to natural verbal relations, at least at the group level, and thus further systematic study is required to explore the IRAP's full potential.

2 In appealing to private responses and measuring brain activity, it is important to recognize that we are treating these events as behavioral in nature, and that they need to be explained by appealing to past and current contextual variables. In other words, it is the environmental contingencies that establish and maintain private responding, and its relationship to overt responding, that will provide a relatively complete behavior-analytic explanation (see Barnes-Holmes, 2003, for a relevant discussion of this issue).

References

BARNES, D., & HOLMES, Y. (1991). Radical behaviorism, stimulus equivalence, and human cognition. *The Psychological Record, 41*, 19–31.

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. (2003). For the radical behaviorist biological events are not biological and public events are not public. *Behavior and Philosophy, 31,* 145–150.

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., REGAN, D., BARNES-HOLMES, Y., COMMINS, S., WALSH, D., STEWART, I., ET AL. (2005). Relating derived relations as a model of analogical reasoning: Reaction times and event related potentials. *Journal of the Experimental Analysis of Behavior. 84,* 435–452.

BARNES-HOLMES, D., STAUNTON, C., BARNES-HOLMES, Y., WHELAN, R., STEWART, I., COMMINS, S., ET AL. (2004). Interfacing relational frame theory with cognitive neuroscience: Semantic priming, the implicit association test, and event related potentials. *International Journal of Psychology and Psychological Therapy, 4,* 215–240.

BARNES-HOLMES, D., STAUNTON, C., WHELAN, R., BARNES-HOLMES, Y., COMMINS, S., WALSH, D., ET AL. (2005). Derived stimulus relations, semantic priming, and event-related potentials: Testing a behavioral theory of semantic networks. *Journal of the Experimental Analysis of Behavior. 84,* 417–434.

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

DE HOUWER, J. (2002). The Implicit Association Test as a tool for studying dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behaviour Therapy and Experimental Psychiatry, 33,* 115–133.

DE HOUWER, J. (2006). What are implicit attitudes and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11–28). Thousand Oaks, CA: Sage Publishers.

DE JONG, P. (2002). Implicit self-esteem and social anxiety: Differential self-positivity effects in high and low anxious individuals. *Behaviour Research and Therapy, 40,* 501–508.

DE JONG, P., PASMAN, W., KINDT, M., & VAN DEN HOUT, M. A. (2001). A reaction time paradigm to assess (implicit) complaint-specific dysfunctional beliefs. *Behaviour Research and Therapy, 39,* 101–113.

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.

DOVIDIO, J. F., & FAZIO, R. H. (1992). New technologies for the direct and indirect assessment of attitudes. In J. Tanur (Ed.), *Questions about questions: Meaning, memory, expression and social interaction in surveys* (pp. 204–237) New York: Sage.

GEMAR, M. C., SEGAL, Z. V., SAGRATI, S., & KENNEDY, S. J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. *Journal of Abnormal Psychology, 110,* 282–289.

GREENWALD, A. G., & BANAJI, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and social stereotypes. *Journal of Personality and Social Psychology, 102,* 4–27.

GREENWALD, A. G., BANAJI, M. R., RUDMAN, L. A., FARNHAM, S. D., NOSEK, B. A., & MELLOTT, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109,* 3–25.

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., AND ROCHE, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition.* New York: Plenum.

HAYES, S. C., BARNES-HOLMES, D., & ROCHE, B. (2003). Behavior analysis, relational frame theory, and the challenge of human language and cognition: A reply to the commentaries on relational frame theory—A post-Skinnerian account of human language and cognition. *The Analysis of Verbal Behavior, 19,* 39–54.

HOLCOMB, P. J., & ANDERSON, J. E. (1993). Cross-modal semantic priming: A time-course analysis using event-related potentials. *Language and Cognitive Processes, 8,* 327–411.

KOUNIOS, S. A., & HOLCOMB, P. J. (1992). Structure and process in semantic memory: Evidence from event-related potentials and reaction times. *Journal of Experimental Psychology: General, 121,* 460–480.

KUTAS, M. (1993). In the company of other words: Electrophysiological evidence for simple-word and sentence-context effects. *Language and Cognitive Processes, 8,* 533–578.

KUTAS, M., & HILLIARD, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 1161–1163.

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., BARNES-HOLMES, D., BARNES-HOLMES, Y., & STEWART, I. (2007). Testing the fake-ability 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.

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., & SMYTH, S. (2007). A derived transfer of functions and the Implicit Association Test. *Journal of the Experimental Analysis of Behavior, 88,* 263–283.

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.

TEACHMAN, B. A., GREGG, A. P., & WOODY, S. R. (2001). Implicit associations of fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110,* 226–235.

WALTZ, J. A., KNOWLTON, B. J., HOLYOAY, K. J., BOONE, K. B., MISHKIN, F. S., SANTOS, M. D., ET AL. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10,* 119–125.

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

WEISBROD, M., KEIFER, M., WINKLER, S., MAIER, S., HILL, R., ROESCH-ELY, D., ET AL. (1999). Electrophysiological correlates of direct versus indirect semantic priming in normal volunteers. *Cognitive Brain Research, 8,* 289–298.