# Combining the Implicit Relational Assessment Procedure and the Recording of Event Related Potentials in the Analysis of Racial Bias: A Preliminary Study

Patricia M. Power[1], Colin Harte[2], Dermot Barnes-Holmes[2], and Yvonne Barnes-Holmes[2],

[1] Department of Psychology, National University of Ireland Maynooth, Ireland

[2] Department of Experimental, Clinical and Health Psychology, Ghent University, Belgium

Corresponding Author:

Colin Harte

Department of Experimental, Clinical, and Health Psychology

Ghent University

Henri Dunantlaan, 2

9000 Ghent

Belgium

Email: Colin.Harte@UGent.be

**Authors' Note** This article was prepared with the support of an Odysseus Group 1 grant awarded to the third author by the Flanders Science Foundation (FWO).

**Abstract**

The current study examined racial bias among white individuals residing in Ireland using the Implicit Relational Assessment Procedure (IRAP). In addition, neural activity, measured with electroencephalograms (EEG), was recorded while participants completed the IRAP. On some blocks of trials participants were required to respond quickly and accurately in a pro-white and anti-black manner, whereas on other blocks they were required to respond in the opposite manner (anti-white/pro-black). The difference in response latencies between these two types of trials provided an index of racial bias, whilst event-related potentials (ERPs), derived from the EEG signals, provided a simultaneous measure of brain activity during these responses. Results revealed anti-black and pro-white biased responding on the IRAP in terms of differential response latencies. In addition, greater positivity in the ERPs signals located in the frontal sites was recorded when participants responded in a pro-black/anti-white pattern, relative to a pro-white/anti-black pattern. These results are broadly consistent with previous literature in the area and suggest the IRAP as a potentially useful methodology for research in the field of affective neuroscience.

KEYWORDS: IRAP, EEG, Evoked potentials, Racial bias

The study of derived stimulus relations has been used widely in behavior analysis as a paradigm for analyzing human language and cognition, with an early example involving the study of social categorization and prejudice in Northern Ireland (Watt, Keenan, Barnes, & Cairns, 1991). The study involved training and testing participants for derived equivalence relations between Catholic names and Protestant symbols, which would be inconsistent with the verbal/social histories of Northern Irish residents. The results showed that some Northern Irish participants failed to form these equivalence relations, whereas non-Northern Irish participants readily formed the laboratory-induced relations. Broadly similar findings have since been reported across a range of domains (e.g., Barnes, Lawlor, Smeets, & Roche, 1996; Dixon, Rehfeldt, Zlomke, & Robinson, 2006; Leslie, et al., 1993; Merwin & Wilson, 2005).

The general strategy of comparing patterns of responding that are consistent versus inconsistent with the pre-experimental histories of participants has also been adopted in more recent efforts to develop behavior-analytic procedures for assessing verbal relations occurring in the natural environment (Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008). The currently most popular method in this regard is the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010), which was based on Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001), an account of human language and cognition that draws heavily on the concept of derived stimulus relations.

A typical IRAP presents word and/or image pairs, and participants are required to confirm or disconfirm the relation between them. Corrective feedback is provided and is delivered in a manner that is assumed to be consistent with participants' pre-existing verbal histories on some blocks, and inconsistent with that history on other blocks. Thus,

for example, responding "True" to "Flowers-Positive" on consistent blocks, but responding "False" on inconsistent blocks, would be required. The basic assumption of the IRAP effect is that, all things being equal, participants should respond more quickly on blocks that are consistent with their histories than on blocks that are inconsistent. Typically, participants are required to respond within a relatively narrow temporal window across all trials, such as 2000 ms. Thus, any differences between average response latencies that are history-consistent versus history-inconsistent are likely due to subtle response biases, rather than self-directed rules to respond more slowly or quickly on certain trials. A more detailed treatment of the RFT-based conceptual analysis of the IRAP effect has been articulated in terms of the Relational Elaboration and Coherence (REC) model (Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010).

The difference in latencies between the two different patterns of responding on the IRAP is often referred to as a positive or negative response bias, depending on whether it is above or below zero. Given the conceptual basis of the IRAP, an IRAP effect should not be interpreted as a proxy for a mental construct or implicit attitude in a cognitive or social psychological sense. Instead, the term simply denotes a tendency to respond in one particular direction over another on the IRAP. Use of the IRAP is steadily increasing in a wide range of domains, and a recent meta-analysis of its use in various clinical domains has reported relatively high predictive validity (Vahey, Nicholson, & Barnes-Holmes, 2015).

One of the earliest IRAP studies examined the response patterns of white participants toward pictures of black and white individuals (Barnes-Holmes, Murphy, et al., 2010). Specifically, participants were presented with one of two label stimuli ("Safe" and

"Dangerous") on each trial with a picture of a white or black man holding a gun as a target stimulus. The IRAP required responding in a pro-white and anti-black pattern on some blocks of trials (e.g., pressing a key for "True" rather than "False" when "Safe" appeared with a picture of a white man). On other blocks of trials, responding in a pro-black and anti-white pattern was required (e.g., pressing a key for "True" rather than "False" when "Safe" appeared with a picture of a black man). The IRAP revealed pro-white and anti-black biases, although the anti-black effect was restricted to one trial-type. That is, participants responded "True" more quickly than "False" when presented with "Dangerous" and pictures of black men holding guns; when the pictures were of white men holding guns, participants responded "False" more quickly than "True". Three other studies have also examined racial bias using the IRAP and similar positive in-group biases were found for white participants (Drake et al., 2010, 2015; Power, Harte, Barnes-Holmes, & Barnes-Holmes, in press).

In virtually all of the studies that have employed the IRAP, including those that have examined racial response biases, the standard latency-based measure has been the sole metric by which the IRAP effect has been assessed. The one exception is one of the earliest published IRAP studies, in which electroencephalograms (EEG signals) were recorded while participants completed an IRAP (Barnes-Holmes et al., 2008). The stimuli consisted of words that were either positively or negatively valenced. That is, the label stimuli were the words "pleasant" and "unpleasant", and were presented with target words, such as "love", "peace", "hate", and "war". The EEG signals, recorded while participants completed the IRAP, were transformed into event related potentials (ERPs; these are explained below) and indeed the results showed different patterns of EEG activity across blocks of consistent and inconsistent trials. Since this early study, no other

published research, of which we are aware, has reported the use of EEG measures with the IRAP. Since this early study, the IRAP itself has been developed and refined considerably, and has been used across a wide range of psychological domains, with meta-analytic evidence that it has impressive predictive validity with clinical phenomena. There is clear potential, therefore, that the IRAP could be used fruitfully as a method that may be combined with neurophysiological measures in a range of domains. The current study constitutes the first step in this regard, focusing in this case on racial bias.[1]

In the present study, recordings were taken from multiple EEG signals while participants completed a race-IRAP and these signals were then transformed into ERPs (e.g., Kutas, 1993; Kutas & Hillyard, 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, et al., 2005; Barnes-Holmes, Staunton, et al. 2005; for examples of ERP research within the behavior-analytic tradition).

Generating ERPs data involves time-locking the EEG signals to a particular series of events and subsequently 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 (Sur & Sinha, 2009). 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

[1] In deciding to employ EEG during exposure to an IRAP, we are not suggesting that neural activity reveals a causal variable in a behavior-analytic sense. Rather, EEG provides another property of the relational responding, in addition to response latency, that occurs on an IRAP. For a recent and detailed discussion on measures of neural activity in behavior analysis, particularly with respect to clinical phenomena, see Vahey, Bennett, & Whelan (in press).

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 differentiating different auditory stimuli from one another or understanding words. These ERPs commonly occur at around 300 or 400 ms after stimulus onset (e.g., Kutas & Hillyard, 1980, Sur & Sinha, 2009). The use of ERP measures with the race IRAP in the current study was entirely exploratory, and thus no specific predictions were made pertaining to the ERPs waveforms that might emerge. One ERPs measure, however, that seemed particularly pertinent to the IRAP is the N400, a late negative waveform (see Holcomb & Anderson, 1993; Kounios & Holcomb, 1992). The N400 is usually produced when participants are required to respond to stimuli that are unexpected, unrelated, or wrongly paired in some sense (known as low *cloze-probability)*. Presenting pairs of words that are semantically unrelated, for example, tends to produce an N400, while words from the same semantic categories do not. Insofar as pro-black/anti-white trials on the race-IRAP require "incorrect" or "wrongly paired" responses, a more negative waveform may emerge for these trials relative to pro-white/ anti-black trials. Indeed, this is the general pattern of results obtained in the only study that has measured EEG signals while participants completed an IRAP (Barnes-Holmes, et al., 2008). On balance, this previous IRAP study was conducted using verbal relations that would not be deemed socially sensitive (e.g., "Pleasant-Holiday-Similar") and a latency criterion of 3000 ms was applied. Given that the current study will employ socially sensitive verbal relations (e.g., "Black-Stupid-True") and a 2000 ms response latency criterion, it is possible that

different EEG results will emerge (although, as noted above, the study is exploratory because no other research has used EEG with the IRAP and socially loaded stimuli).

In the current research, separate ERPs waveforms, recorded across a range of sites, for blocks of pro-white/anti-black IRAP trials were collected. Similarly, waveforms were also collected for blocks of anti-white/pro-black trials. A comparison could thus be made between the ERPs waveforms associated with these two types of IRAP trials.

## Method

### Participants

Sixteen adults, 8 male and 8 female, participated in the study. Their ages ranged from 18 to 33 years ($M$ = 25). All participants were white, had been born and lived in Ireland for most of their lives, and were recruited via convenience sampling from the Dublin area. All data from 7 participants were excluded due to excessive noise in the EEG data (explained below). Participants were given a local record store voucher worth 10 euros upon completing the study.

### Setting

The entire experiment was conducted in an electrically shielded room in the human neuroscience laboratory in the Department of Psychology at NUI, Maynooth. All participants completed the experiment individually and in a single session. Each session lasted on average 1 hour, 15 minutes.

### Materials and Apparatus

**Implicit Relational Assessment Procedure (IRAP).** All participants completed the IRAP on a standard personal computer. The IRAP software presented the stimuli and recorded participant responses. Each trial presented the label statement; “I think BLACK

people are" or "I think WHITE people are". One of 12 target stimuli was also presented, 6 stereotypically positive words ("Friendly", "Honest", "Hardworking", "Peaceful", "Good", "Clever") and 6 negative ("Hostile", "Deceitful", "Lazy", "Violent", "Bad", "Stupid") along with two response options, "True" and "False". Based on the various label-target combinations, the IRAP comprised 4 trial-types; *White People-Positive, Black People-Negative, Black People-Positive,* and *White People-Negative* (see Figure 1).

**INSERT FIGURE 1 HERE**

**Electroencephalogram (EEG).** 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 standard 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**

Participants were first attached to the Brain Amp. Evoked potentials were recorded and analyzed from 32 sintered AG/AG-CI scalp electrodes positioned according to the international 10-20 system. The 32 sites chosen for recording were Fp1, Fp2, F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3,Cz, C4, T8, TP9, TP7, CP3, CPz, CP4, TP8, TP10, P7 P3, Pz, P4, P8, O1, Oz, 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 was 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 EEGs were collected continuously and edited off-line.

The IRAP program began with a set of instructions, which described the task by illustrating the layout of the screen and explaining the response options. The instructions informed participants that on each trial one of two statements, "I think BLACK people are" or "I think WHITE people are", would appear at the top of the screen along with a target word in the center. Participants were also told that the response options "True" and "False" would appear at the bottom, and that they were required to choose one of these options on each trial; they were told that the left-right positions of these response options would switch randomly from trial-to-trial. The instructions also informed participants that correct responses would allow them to progress to the next trial, but incorrect responses would produce a red 'X' in the middle of the screen, which could only be removed by pressing the correct key. In addition, participants were informed that if they took longer than 2000 ms on any IRAP trial, the phrase "Too Slow!" would be presented on the screen.

The IRAP task consisted of a minimum of two practice blocks and a fixed set of six test blocks. Only the ERPs data from the six test blocks were analyzed. Each block presented 24 trials, consisting of the four different trial-types. The first block of trials was consistent with pro-white/anti-black stereotyping (e.g., I think WHITE people are-Positive-True; I think BLACK people are-Positive-False; I think WHITE people are-Negative-False; I think BLACK people are-Negative-True). The feedback contingencies alternated from block to block. Thus, in the second block of trials, correct responses were consistent with anti-white/pro-black stereotyping (e.g., I think WHITE people are-Positive-False; I think BLACK people are-Positive-True; I think WHITE people are-Negative-True; I think BLACK people

are-Negative-False). Before each new block, participants were informed that the previously correct and wrong answers would be reversed.

For the first two practice blocks, participants were informed that it was a practice phase and errors were expected. Participants were required to reach a standard of >/=80% correct responses, and a median response time of </=2000ms. Participants were allowed three attempts (a total of six practice blocks) to achieve the practice criteria; all participants achieved these criteria and proceeded to the six test blocks. No performance criteria were applied during the test blocks in order to proceed, but performance feedback was provided at the end of each block to encourage participants to maintain the practice criteria. Those participants who provided data for the EEG analyses maintained these criteria throughout the test blocks.

## Results

### IRAP

**Data preparation**. The primary datum was response latency (i.e., time in ms between trial onset and a correct response). In accordance with previous IRAP studies, response latency data were transformed into *D*-IRAP scores (see Nicholson and Barnes-Holmes, 2012). The data transformation yielded positive *D*-IRAP scores for positive bias, and negative scores for negative bias. A separate overall *D*-IRAP score was calculated, with positive scores indicating a pro-white/anti-black bias and negative scores indicating an anti-white/pro-black bias.

**Analysis.** The mean *D*-IRAP scores for the four trial-types are presented in Figure 2. The results showed a positive bias for the two white trial-types and a negative bias for the two black trial-types. A one-way repeated measures ANOVA revealed a significant

main effect for trial-type, $F(3, 8) = 88.906$, $p < .001$, $\eta_p^2 = .92$. Fisher's PLSD post-hoc analyses indicated that the two white trial-types were significantly different from the two black trial-types (*p*s < .001). However, the two white trial-types did not differ significantly from each other (*p*s > .9), neither did the two black trial-types (*ps* > .4). One-sample *t*-tests indicated that all four trial-type effects differed significantly from zero (all *p*s < .001).

**INSERT FIGURE 2 HERE**

**ERPs Data**

The continuous EEG signals for all 16 participants were individually filtered (0.53 Hz, time constant = 0.3 s, 24 dB/octave roll-off) and then segmented. The segments were divided into 900 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 7 participants were removed from subsequent analyses because no segments were artifact-free). The remaining segments were then baseline corrected (using the 100 ms pre-stimulus interval). Finally, to reduce noise for the ERPs analyses, the data for the three pro-white/anti-black test blocks were collapsed, as were the data for the three pro-black/anti-white test blocks (for ease of communication, these two types of test block will be referred to as pro-white and pro-black, respectively).

The grand average waveforms for each of the 6 frontal electrode sites (Fp1, Fp2, F7, F3, F4, and F8) for pro-white (light lines) versus pro-black (dark lines) blocks are presented in Figure 3. No differences in evoked potentials between pro-white and pro-black trials were detectable at any of the other sites and thus, in accordance with common

practice (e.g., Weisbrod et al., 1999), these data are not reported. Visual inspection of the waveforms from the six sites indicated little evidence of differential activity between the pro-white and pro-black blocks until approximately 200 ms after stimulus onset. Thereafter, the two waveforms separated with the pro-black blocks producing greater positivity than the pro-white blocks. The waveforms for sites F3 and F4 tended to converge again around 500ms, whereas the waveforms for the remaining sites did not.

**INSERT FIGURE 3**

The area dimensions (μV x ms) for each ERPs waveform (in the temporal interval 300-800 ms) for each participant were calculated, yielding either positive or negative values with respect to the 0 μV level. For the purposes of statistical analysis, average area dimensions were calculated across the three left sites (Fp1, F7, F3) and across the three right sites (Fp2, F8, F4) for pro-white and pro-black waveforms. The data were entered into a 2x2 repeated measures ANOVA with laterality (left versus right) and IRAP (pro-white versus pro-black) as variables. The main effect for laterality proved to be significant, $F(1, 8) = 7.37$, p = .03, $\eta_p^2 = .48$, as did the effect for the IRAP, $F(1, 8) = 7.48$, p = .02, $\eta_p^2 = 48$; the interaction, however, was non-significant ($p > .6$). Follow-up paired *t*-tests for each of the six sites revealed significant differences between pro-white and pro-black waveforms at Fp1, Fp2, F7, and F8 (all $ps < .03$).

## Discussion

The results of the current study were broadly consistent with previous research that has used the IRAP as a measure of racial bias (Barnes-Holmes, et al., 2010; Drake et al., 2010; 2015; Power et al., in press). That is, participants generally showed pro-white and anti-black biases. One notable difference, however, between the current study and the

data reported by Barnes-Holmes et al. is that the anti-black effect was shown on both black trial-types here, but in the earlier study this effect was observed only on the Black-Negative trial-type (a non-significant pro-black effect was shown on the Black-Positive trial-type). At the present time, it remains unclear why this difference emerged, particularly given that the studies were both conducted by the same research group. On balance, it should be noted that in the earlier study, the labels were the words "safe" and "dangerous" presented with pictures of white and black men holding guns as targets, whereas in the current study the labels were "I think white people are" and "I think black people are", with positive and negative target words (but see Power et al., in press). Furthermore, participants completed the current study while having their EEGs recorded, the potential impact of which upon IRAP performances remains unknown. Finally, it is worth noting that the *N* here was relatively low due to the removal of data for entire participants, arising from noise in the EEG data. Although these factors suggest the need for caution in interpreting the current findings, they are still useful given that we have no other such data available to us at the present time.

The EEG recordings revealed that the ERPs grand average waveforms for the pro-black trials were more positive than for the pro-white trials across six of the frontal sites between 300-800 ms. Insofar as pro-black responding for white participants is considered history-inconsistent and pro-white responding history-consistent, the current experiment produced completely opposite effects to those reported in the only other IRAP study that employed EEG as a measure (Barnes-Holmes, et al., 2008). Specifically, waveforms associated with relational responding that was deemed inconsistent with the participants' prior history were more positive than those waveforms associated with history-consistent

responding. In addition, the previous study also reported significant differences between the waveforms for sites in the central and parietal areas; these were not observed in the current experiment.

At the present time, it remains unclear why these differences emerged in the EEG measures across the two studies. As noted earlier, however, the previous Barnes-Holmes et al. (2008) study employed stimuli that were not deemed socially sensitive, and used a response-latency criterion of 3000 ms (rather than 2000 ms.). Furthermore, participants in the earlier study were not required to remain within the latency criterion during the test blocks (this was required in the current study). Further research will be required to determine the variables responsible for the different ERPs patterns observed across the two studies. Nevertheless, the current findings do indicate that EEG signals may be used to differentiate between two different types of IRAP trial, even when socially-sensitive stimuli are employed.

Indeed, it is interesting that the differential ERPs patterns observed in the current study were restricted to the frontal sites and that greater positivity was recorded for the IRAP performances that required responding in a manner that was inconsistent with a white in-group racial bias. More informally, greater activation was observed in the frontal areas of the cortex when participants were asked to respond in a way that perhaps involved suppressing a socially-conditioned pro-white/anti-black response. Increasing evidence in the affective neuroscience literature suggests that the dorso-lateral prefrontal cortex is heavily involved in suppressing the activation of other areas of the cortex, such as the amygdala, responsible for the processing of emotional reactions (e.g., Siegle, Thompson, Carter, Steinhauer, & Thase, 2007). In this sense, therefore, the current data

may be seen as broadly consistent with this literature. That is, the frontal sites in the current study yielded differential levels of activity across consistent and inconsistent blocks of the IRAP; this differential activity may indicate that the frontal areas of the cortex were more engaged in suppressing the activity of other areas of the brain during those inconsistent blocks. Admittedly, this interpretation remains highly speculative because EEG signals do not readily reveal brain activity associated with the emotional centers of the brain, such as the amygdala. In any case, these findings suggest that the IRAP could provide a useful methodology for researchers working in the area of affective neuroscience.

In closing, one of the key weaknesses of the current study was the limited sample size and the implications this has for the statistical analyses that were conducted, particularly on the EEG data. Indeed, due to the small sample, it was not possible to analyze the ERPs data at the level of the individual trial-type (while reducing noise inherent in EEG data to a reasonable level for analysis). Ideally, future studies that attempt to use the IRAP, with EEG as a concomitant measure, should increase the number of trials per block and the number of participants who successfully complete the experiment without excessive EEG artifacts. This approach would allow us to determine if the differential activity observed here in the frontal sites would be replicated for both white and black stimuli. In other words, such research would tell us if differential activity in the pre-frontal cortex was correlated with negative reactions to black stimuli or the requirement to respond negatively to white stimuli. In summary, although the current findings were obtained with a limited dataset, and should be interpreted with caution,

they are broadly consistent with existing literatures, and they are instructive in terms of where future researchers may direct their efforts.

## Compliance with Ethical Standards

**Declaration of Interest:** This article was prepared with the support of an Odysseus Group 1 grant awarded to the third author by the Flanders Science Foundation (FWO). The data collected for this research was attained with the support of funding awarded to the first author by the Irish Research Council (IRC). The authors declare no other conflicts of interest.

**Ethical Approval:** All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

**Informed Consent:** Informed consent was obtained from all individual participants included in the study.

## References

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

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

Barnes, 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., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The Implicit Relational Assessment Procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record, 60*, 57-66.

Barnes-Holmes, D., Regan, D., Barnes-Holmes, Y., Commins, S., Walsh, D., Stewart, I., . . . Dymond, S. (2005). Relating derived relations as a model of analogical reasoning: Reaction times and event related potentials. *Journal of Experimental Analysis of Behavior, 84*, 435-452. doi: 10.1901/jeab.2005.79-04

Barnes-Holmes, D., Staunton, C., Barnes-Holmes, Y., Whelan, R., Stewart, I., Commins, S., . . . Dymond, S. (2005). 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.

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

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

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

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

Holcomb, P.J. & Anderson, J.E. (1993). Cross-model 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. doi: 10.1037/0096-3445.121.4.459

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. doi: 10.1080/01690969308407587

Kutas, M. & Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 1161-1163. doi: 10.1038/307161a0

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205. doi: 10.1126/science.7350657

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.

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

Nicholson, E. & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62,* 263-278.

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

Siegle, G.J., Thompson, W., Carter, C.S., Steinhauer, S.R., & Thase, M.E. (2007). Increased amygdala and decreased dorsolateral prefrontal BOLD responses in unipolar depression: related and independent features. *Biological Psychiatry, 61*(2), 198-209. doi: 10.1016/j.biopsych.2006.05.048

Sur, S. & Sinha, V.K. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal, 18*(1), 70-73. doi: 10.4103/0972-6748.57865

Vahey, N.A., Bennett, M., & Whelan, R. (in press). Conceptual advances in the cognitive neuroscience of learning: Implications for relational frame theory. *Journal of Contextual Behavioral Science*. doi: 10.1016/j.jcbs.2017.04.001

Vahey, N.A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48*, 59-65. doi: 10.1016/j.jbtep.2015.01.004

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

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. doi: 10.1016/S0926-6410(99)00032-4

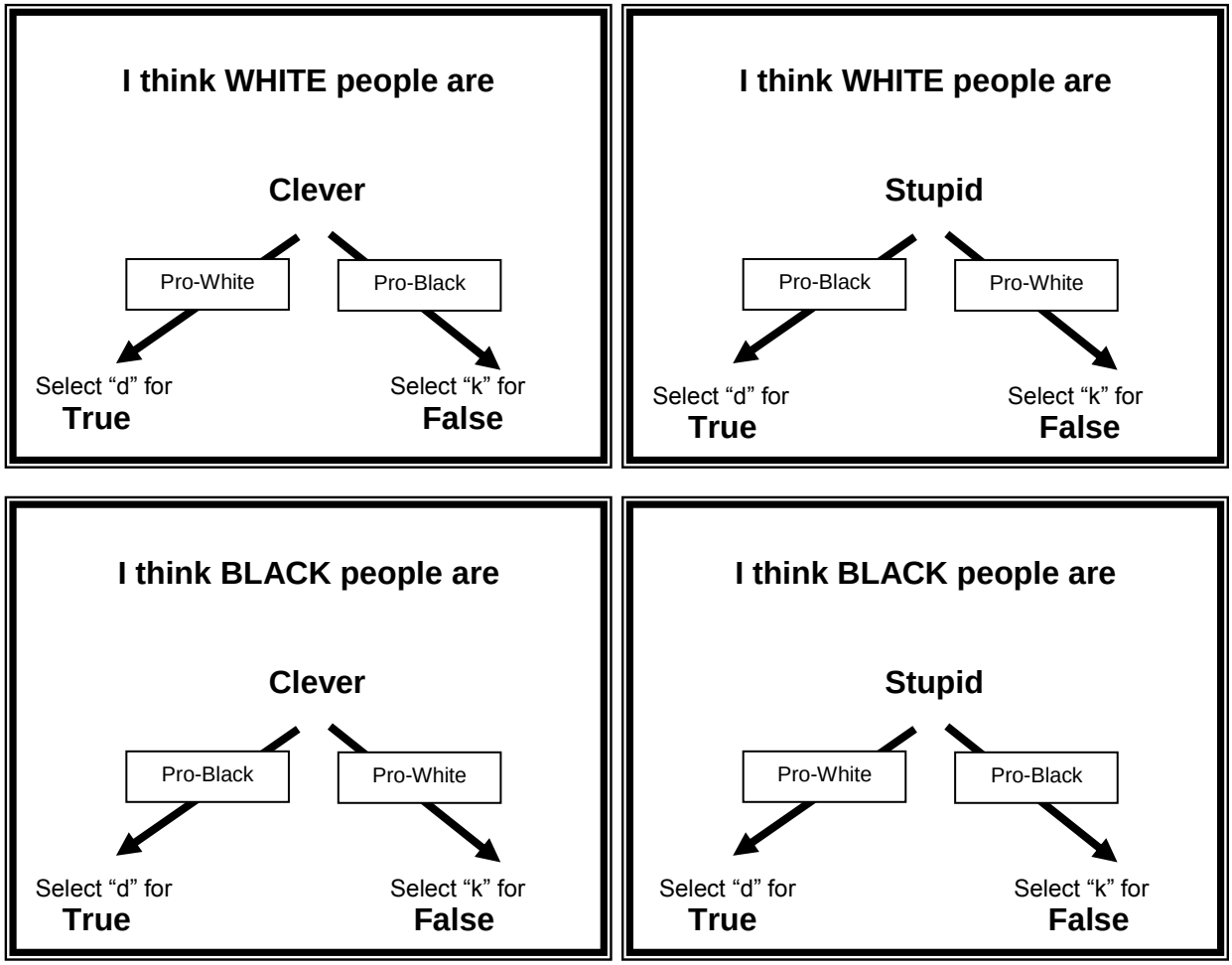


*Figure 1.* Diagrammatic representation of the four IRAP trial-types. Arrows and boxes containing the words *Pro-White* and *Pro-Black* did not appear on-screen.

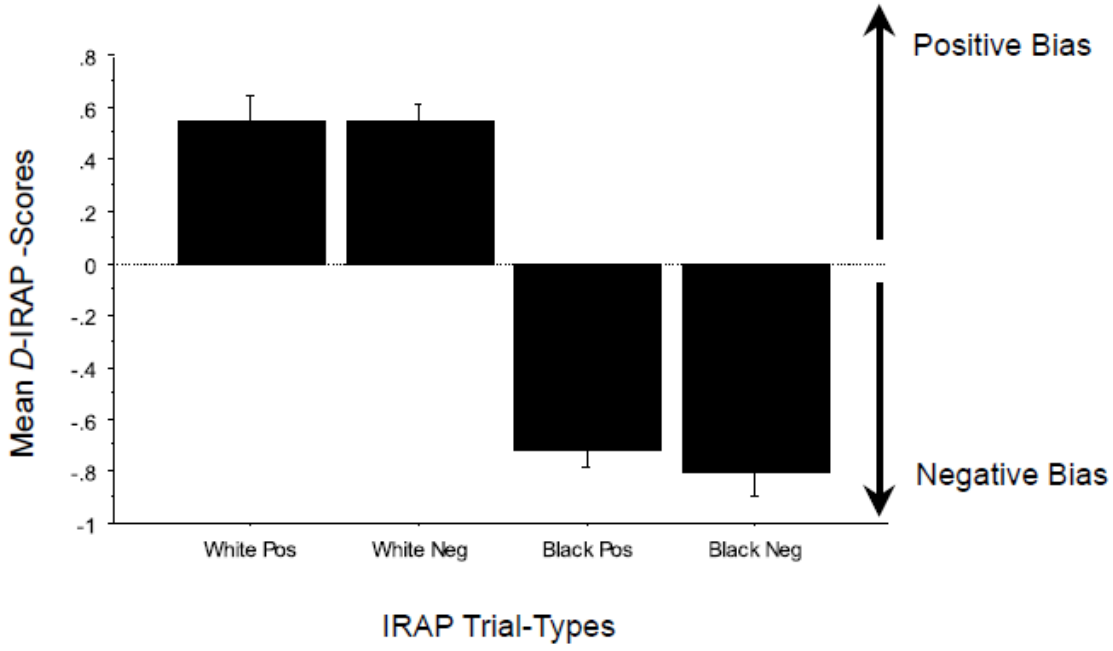


*Figure 2*. The mean *D*-IRAP scores, with standard error bars, for the four IRAP trial-types.

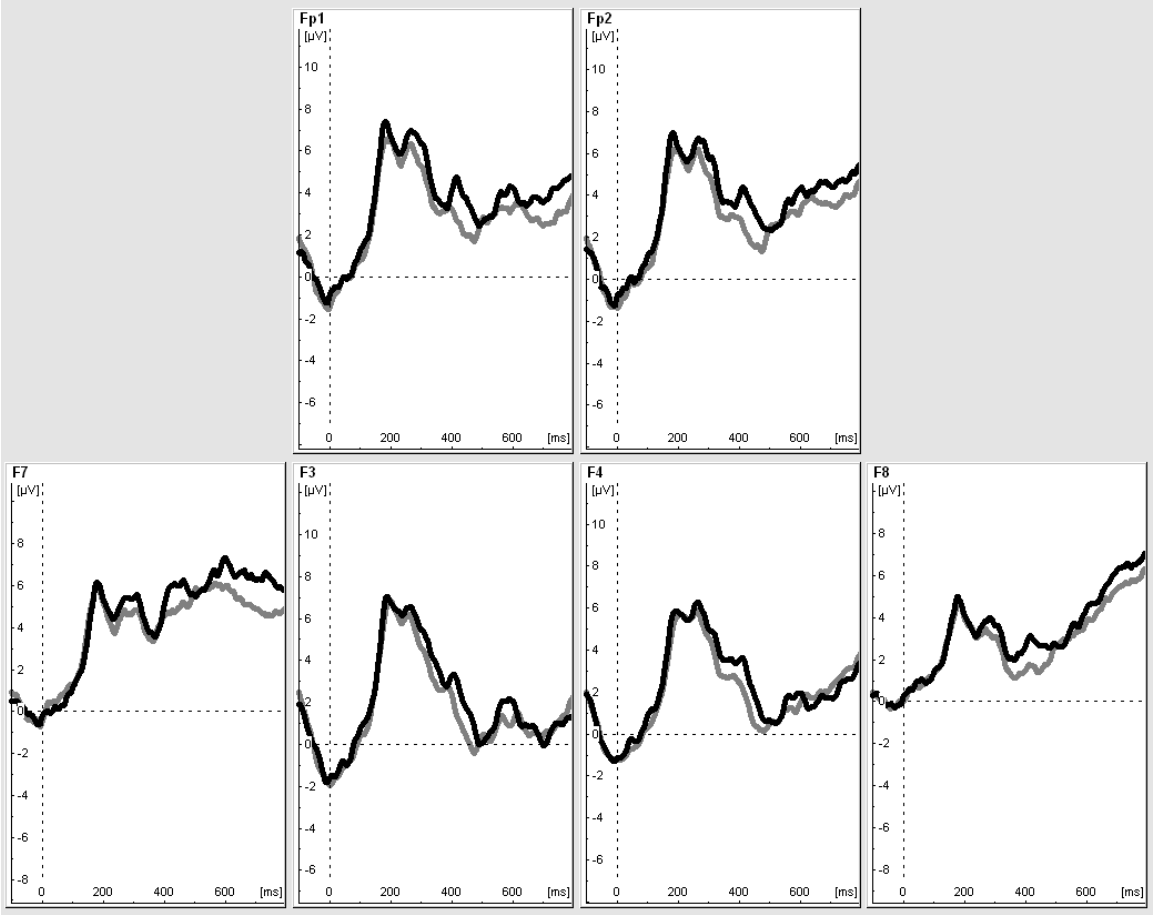


*Figure 3*. The grand average waveforms for each of the 6 frontal electrode sites (Fp1, Fp2, F7, F3, F4, and F8) for pro-white (light lines) versus pro-black (dark lines) blocks.