EXPLORING THE RELIABILITY AND CONVERGENT VALIDITY OF IMPLICIT RACIAL EVALUATIONS

Chad E. Drake, Sam Kramer, Travis Sain, Rachel Swiatek, Kristin Kohn, & Meagan Murphy
Southern Illinois University

ABSTRACT: Racial evaluations have received considerable attention by researchers of implicit cognition, especially with the Implicit Association Test (IAT). The IAT measures associative biases in a relativistic manner, whereby attitudes toward a given racial category are compared to attitudes toward another. The Implicit Relational Assessment Procedure (IRAP) is a new behavior analytic measure of cognition that may provide a less relativistic and more specific measure of cognitive repertoires. The current study utilized a race IRAP to assess evaluative biases among a balanced sample of Black and White undergraduates. The race IRAP was administered twice in a row in conjunction with a collection of self-report measures of racial attitudes. Results for reliability and convergent validity were generally supportive. Furthermore, observed biases appeared to reflect positive in-group biases rather than derogatory attitudes toward the out-group, an effect that would not be apparent with a similarly configured race IAT. Future research may benefit from consideration of the evaluative content of the IRAP as well as the racial demographics for both the participants and the experimenters.

KEYWORDS: Implicit Relational Assessment Procedure, implicit cognition, social cognition, attitudes, racism

Implicit cognition, a type of thinking characterized with descriptors such as “automatic” and “unconscious”, has been a topic of psychoanalytic theorizing for decades (for review, see Westen, 1998). In recent years a variety of behavioral measures of implicit cognition have proliferated rapidly among cognitive and social perspectives in psychology (Fazio & Olson, 2003; Nosek, Hawkins, & Frazier, 2011). Modern computers and computer software have made it possible to create programs that can sensitively and efficiently measure volumes of reaction latencies to complex arrays of verbal stimuli. The Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) is arguably the premiere measure among the growing list of implicit measures currently available (for review, see Gawronski & De Houwer, 2011). In less than two decades, more than a thousand empirical publications have resulted from IAT researchers addressing a variety of topics in cognitive, social, and clinical psychology, among others.
The IAT and Racial Attitudes

The value, and some might say notoriety, of the IAT is evidenced perhaps most notably on the topic of social cognition, particularly in respect to attitudes regarded as reflective of racism, prejudice, and discrimination. In the United States, implicit measures commonly reveal pro-White, anti-Black racial biases among White respondents that are not detected by traditional self-report measures (e.g., Green et al., 2007). A common explanation for this divergence refers to the prevailing social demands that may discourage self-reports of stigmatizing attitudes revealed by the IAT, a measure designed to circumvent social demand effects (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002). Nevertheless, individual performances vary substantially with the IAT, and environmental variables have been shown to influence IAT performance. For example, Frantz, Cuddy, Burnett, Ray, and Hart (2004) demonstrated that a race IAT was influenced by a variety of factors, including White participants’ desire to appear egalitarian, their perception of what the IAT was measuring, and pre-IAT affirmations of egalitarian values. However, in spite of the variability accounted for by these variables, in all three studies the samples overall exhibited racial biases. Even among studies containing Black participants (e.g., Nosek, Banaji, & Greenwald, 2002; Stewart, von Hippel, & Radvansky, 2009), average scores on a race IAT typically reveal relatively neutral biases rather than the pro-Black, anti-White biases that would mirror the pro-White, anti-Black biases witnessed with White participants. These and related findings have generated much discussion and debate among implicit cognition researchers (e.g., Arkes & Tetlock, 2004; Banaji, Nosek, & Greenwald, 2004).

Much of the debate resolves on the basic issue of defining what is measured by the IAT (e.g., Gawronski, LeBel, & Peters, 2007; van Ravenzwaaij, van der Maas, & Wagenmakers, 2011). To measure perceptions of race, an IAT administrator might populate the measure with stimuli for two classes for race (e.g., images of Black and White faces) and two classes for evaluation (e.g., positive and negative words). An individual trial presents a stimulus from one of these four classes in the center of the screen along with response options at each upper corner of the screen. Participants “associate” the center stimulus with the response options by pressing either the “e” key (for the options on the upper left) or the “i” key (for the options on the upper right). An example trial for a race IAT might contain an image of a black face appearing at the center of the screen along with “black or positive” at the upper left corner and “white or negative” at the upper right corner. Other trials might present at the center of the screen an image of a white face, a positive evaluative word, or a negative evaluative word. In each case, the participant must press the key that appropriately categorizes the stimulus at the center (e.g., a black face would go with “black or positive”, and the word “bad” would go with “white or negative”). The coupling of race category and evaluative word for the response options is counterbalanced over the duration of the procedure. One block of trials would present “black or positive” at the upper left and “white or negative” at the upper right, while another block would present “black or negative” at the upper left and “white or positive” at the upper right. Throughout the task, the computer records the latency required to correctly select the “e” or “i” key for each trial. A substantial difference in the average latencies between these two blocks is deemed to be a metric of attitudinal bias. The difference between latencies may be around zero, in which case no relative bias is inferred, or it may be a positive or negative value, in which case a pro-White, anti-Black or a pro-Black, anti-White bias is inferred.

Many IAT studies on racial attitudes utilize similar stimulus selection approaches. One popular approach resembles the example in the previous paragraph – displaying images of black
and white faces in conjunction with evaluatively valenced words (e.g., Cunningham, Preacher, & Banaji, 2001). Another approach (e.g., Dasgupta, McGhee, Greenwald, & Banaji, 2000) involves populating the IAT with two groups of common names, each group representative of members of a particular race (e.g., “Lamar”, “Malik”, “Andrew”, “Justin”). These names are presented in conjunction with relatively generic but evaluatively valenced words (e.g., “gentle”, “paradise”, “death”, “poison”). Thus, trials in this presentation involve text only, rather than a mixture of images and text as is done when presenting faces. Yet another approach may involve a modification of either of the preceding examples, except that instead of generic evaluative words, words bearing racial stereotypes are presented in conjunction with names or images (e.g., “lazy”, “unemployed”, “dangerous”). Although each configuration has generated promising results, some data indicates that the configuration of an IAT impacts psychometrics in important ways. For example, Rudman and Ashmore (2007) have demonstrated that an “attitude” race IAT (containing generic evaluative words) more strongly correlated with self-reports of racist attitudes while a “stereotype” race IAT (containing evaluative words indicative of racial stereotypes) more strongly correlated with a measure of discriminatory behavior.

The IAT has generated promising psychometrics across a variety of content domains. Most of the existing literature reports on the internal reliability of the measure, although much of this data is garnered from IATs measuring content other than racial attitudes. For example, a review by Nosek, Greenwald, and Banaji (2007) reported relatively strong internal consistencies for the IAT, particularly among a collection of studies on anxiety and self-esteem. A more recent study of the test-retest reliability of an ethnicity IAT (comparing Asian Americans and European Americans) revealed somewhat weak reliability (Rezaei, 2011). It appears that reliability data is variable overall and sparse in respect to Black/White racial attitudes. Many studies have explored relationships between a race IAT and other measures. As with the existing reliability data, the data on convergent validity is mixed, with some studies failing to obtain significant correlations (e.g., Dasgupta, McGhee, Greenwald, & Banaji, 2000), and others finding strong relationships with self-report (e.g., Cunningham, Preacher, & Banaji, 2001). Some studies have demonstrated contextual influences over convergent validity, such as the participant’s perception of the accuracy of the IAT (Frantz et al., 2004; Nier, 2005) and the data analysis strategy chosen by the experimenter (Greenwald, Nosek, & Banaji, 2003). Overall, it appears that the IAT offers a valid means of measuring race attitudes, although the quality of the psychometrics varies from one study to another.

IAT Limitations

Despite the substantive volume of empirical work with the IAT and related measures, the procedure is not without its critics (e.g., Arkes & Tetlock, 2004; Fiedler, Messner, and Bluemke, 2006). A common concern has referenced the relativistic nature of the measure, which limits the inferences one can make from the results. Although a race IAT involves responding in respect to two racial groups, the measure is configured in a way that makes it difficult to consider attitudes that are unique to either group (Blanton, Jaccard, Gonzalez, & Christie, 2006). Performance on the task reflects attitudes toward one race in respect to the other race, not each race separately and independent of the other. Thus, a pro-White score on a race IAT implies that the respondent exhibits positive associations in respect to White people and negative associations in respect to Black people. Other possibilities exist, yet are impossible to detect with this measure. For example, a respondent may regard each group positively but unequally, or negatively but
IMPLICIT RACIAL EVALUATIONS

unequally, or may view one group in a biased manner and the other group in an unbiased manner. It seems conceivable that this relativistic feature may account for some of the poor psychometrics exhibited in some race IAT studies.

The relativistic nature of the IAT is arguably a byproduct of the philosophical and theoretical assumptions that grounded its development (Hughes, Barnes-Holmes, & De Houwer, 2011). The IAT is founded on associative theorizing (Greenwald, et al., 2002), whereby “association strength” is regarded as the behavior of interest. Association strength is an inference that relies on hypothesized constructs presumably operating in conjunction with neural substrates. This mentalistic approach represents a well-known and discussed departure from behavior analytic theorizing (Barnes-Holmes, 2003; Hayes & Brownstein, 1986; Skinner, 1977). Although behavior analytic labs have used and interpreted data from the IAT (Gavin, Roche, & Ruiz, 2008; Gavin, Roche, Ruiz, Hogan, & O’Reilly, 2012; Weinstein, Wilson, Drake, & Kellum, 2008), an alternative procedure more comfortably nested within the functional contextual assumptions of behavior analysis may provide a means for a progressive body of work on verbal behavior and complex human behavior, including the topic of racial bias.

The Implicit Relational Assessment Procedure

The Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, et al., 2006) is a task methodologically similar to the IAT but developed from a body of behavior analytic research. It originated from a program of work involving matching-to-sample paradigms and a progression of related procedures configured to measure derived relational responding (Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008). The repertoire of interest focuses on bidirectional stimulus relations, a conceptual divergence from the purported unidirectional associations measured by the IAT. More specifically, the IRAP was designed as a convenient means of assessing arbitrarily applicable derived relational repertoires, the subject of Relational Frame Theory (Hayes, Barnes-Holmes, & Roche, 2001; Dymond & Roche, 2013). Known more colloquially as relational frames, the ability to derive stimulus relations has been examined extensively by RFT researchers and found to demonstrate the expected characteristics of an operant, including a developmental trajectory (Berens & Hayes, 2007), flexibility (Hughes & Barnes-Holmes, 2011), antecedent stimulus control (Barnes-Holmes, Barnes-Holmes, Smeets, & Luciano, 2004), and consequential control (Barnes-Holmes, Barnes-Holmes, Smeets, Strand, & Friman, 2004). The IRAP offers a flexible and relatively brief measure of this specific class of operant behavior.

Recently, a functional account of IRAP performance called the Relational Elaboration and Coherence (REC) model has appeared (Hughes, Barnes-Holmes, & Vahey, 2012). According to the REC model, the IRAP measures Brief and Immediate Relational Responses (BIRRs) and Extended and Elaborated Relational Responses (EERRs). BIRRs are regarded as relatively quick reactions that cohere with the respondent’s prevailing learning history in respect to the trial content. BIRRs are conceptually analogous to what a psychoanalyst or a cognitive psychologist might call “unconscious” or “implicit”, respectively. For example, the word “true” may be regarded as the more prominently socially reinforced response to the statement “Eating vegetables is healthy”. In contrast, EERRs are regarded as more complex relational responses requiring more time due to their relative incompatibility with the respondent’s learning history. Continuing with the previous example, responding “false” to the statement “Eating vegetables is healthy” may require additional time in comparison to responding “true” because there likely are
fewer contingencies supporting this derivation provided by the culture at large. When EERRs are substantially longer than BIRRs, the observed difference is known as an “IRAP effect” and suggests a difference in repertoires regarding the content of the procedure. In a similar manner, an IRAP populated with racial categories and evaluative words may provide a means of assessing for racial evaluations promoted explicitly or implicitly by one’s social environment.

The comparison of BIRRs and EERRs reflects a subtle but important divergence from the IAT paradigm. While a race IAT requires respondents to “associate” stimuli among combinations of racial categories and evaluative words, the IRAP requires respondents to derive stimulus relations among race categories and evaluations in respect to (1) responses that are consistent with pro-White, anti-Black attitudes during one half of the procedure, and (2) responses that are inconsistent with those attitudes (pro-Black, anti-White) during the other half. Instead of examining responses that compare perceptions of two races as with the IAT (e.g., responses to Black and an evaluation are compared with responses to White and an evaluation), the IRAP examines responses that compare conditioning histories (e.g., responses to “true” are compared to responses to “false” in respect to each possible combination of race and evaluation separately). In addition to being less relativistic, the strategy inherent to the IRAP provides access to a more specific set of repertoires than the IAT. Whereas the IAT produces a single score of racial bias (a monolithic metric of bias in respect to both races), the IRAP may produce scores for each combination of race category and evaluation known as trial-types. A single race IRAP administration reveals biases in respect to trial-types for (1) positive evaluations of Black people, (2) positive evaluations of White people, (3) negative evaluations of Black people, and (4) negative evaluations of White people. Thus, the methodology of the IRAP provides a behavior analytic basis for examining complex verbal repertoires that are more specific and less relativistic than the IAT.

The IRAP and Race Attitudes

The IRAP has been the subject of a few dozen studies (for review, see Hughes & Barnes-Holmes, 2013), including a handful on perceptions of race. Power, Barnes-Holmes, Barnes-Holmes, and Stewart (2009) demonstrated an in-group bias with an IRAP configured to assess for comparisons of likeability among Irish, Scottish, American, and African ethnicities. When participants were Irish, the results reflected a relatively pro-Irish, anti-African bias with the IRAP, while Americans exhibited a relatively pro-American, anti-African bias. Additionally, as with many race IAT studies, Power and colleagues demonstrated that IRAP effects may diverge from self-reports of racial attitudes, showing that the lesser regard for Africans at the implicit level was not corroborated by self-reported comparisons of the ethnicities.

Barnes-Holmes, Murphy, Barnes-Holmes, and Stewart (2010) assessed implicit attitudes among White participants regarding perception of dangerousness in respect to images of a Black or White male holding a gun. They demonstrated that lowering a latency requirement for responding to IRAP trials from three seconds to two increased pro-White, anti-Black biases. This reduction also increased internal reliability and to some extent enhanced concurrent validity with a self-report measure of discriminatory attitudes toward Black people. Interestingly, some of these results seemed to be driven by a particular trial-type (White + safe) rather than uniformly across the trial-types. Drake and colleagues (2010) administered a race IRAP contrasting the words “black” and “white” with positive and negative evaluations. Among White participants, pro-White biases were exhibited with trials containing the word white, and neutral biases with
trials containing the word black. As with the trial-type-specific findings for Barnes-Holmes and colleagues (2010), these results reveal a degree of specificity impossible with the IAT. Additionally, post-hoc analyses tentatively suggested that a small subsample of Black participants exhibited scores that diverged from those for the White participants.

Although sparse compared to the IAT database, initial studies suggest that racial attitudes may be fruitfully explored with the IRAP. Preliminary data has shown that the IRAP may assess racially biased evaluations, that it may distinguish respondents on the basis of race, and that it may correlate with other measures of racial bias. However, the existing studies have contained predominantly White samples. Given the positive nature of these preliminary results, a program of research that fully explores the applicability and utility of the IRAP as a measure of racial attitudes seems justifiable. The current manuscript details a preliminary study in that program – an exploration of the reliability and convergent validity of a race IRAP with a diverse and balanced sample.

Method

Participants

Fifty-seven participants were recruited from a population of college undergraduates enrolled in general psychology classes. Participants were randomly assigned to a variety of ongoing studies. At times, Black participants were non-randomly assigned to this study in an effort to keep subsamples relatively equal. The average age of the sample was 19.1 years (SD = 1.1). The majority of the sample consisted of males (68.4%; N = 39), reported an annual income of $25,000 or less (43.9%; N = 25), identified as freshman (78.9%; N = 45), and declared a major other than psychology (68.4%; N = 39). Regarding race, 22 participants (38.6%) identified as Black, 22 (38.6%) as White, 4 (7.0%) as both Black and White, and 9 (15.8%) as other races (Asian, Hispanic/Latino, and “Other”) or combinations of races.

Self-Report Measures

Demographics. A 10-item questionnaire assessed a variety of demographic categories, including for age, race, religion, and sex. An item for race/ethnicity asked the participant to “select as many as are appropriate for you” among American Indian or Alaska Native, Asian, Black or African-American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, White or Caucasian, and Other.

Modern Racism Scale. The MRS (McConahay, 1986) is a 6-item measure of subtle racism. High scores represent higher levels of racist perceptions of African-Americans. Items offer five options on a Likert-type scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). This study obtained a Cronbach’s alpha of .759 with this measure.

Semantic Differential Scale. The SDS contained each target word included in the race IRAPs and allowed participants to rank the evaluative valence for each. For each item, participants responded to an 11-point scale, ranging from -5 (Extremely Negative) to 5 (Extremely Positive). A mean ranking was calculated for each class of words. The Cronbach’s alphas obtained in the current study were .947 for negative words and .731 for positive words.

Social Dominance Orientation. The SDO scale (Pratto, Sidanius, Stallworth, & Malle, 1994) is a 14-item measure of social cognition regarding inequality among groups. Items on the scale are not specific to any race. Response options were presented in a 7-point Likert-type scale,
ranging from 1 (Very Negative) to 7 (Very Positive). Due to an administrator’s error, the 11th item of the measure (“If people were treated more equally we would have fewer problems in this country.”) was omitted and therefore did not factor into scoring or analyses. The current study obtained a Cronbach’s alpha of .868.

**IRAPs**

The IRAP is a computerized measure of relational repertoires (Barnes-Holmes et al., 2008). A protocol for IRAP administration was developed from the “Experimenter’s script” downloaded from irapresearch.org/downloads-and-training/. Over three successive IRAP administrations, participants responded to a collection of randomly ordered trials presented repeatedly within a series of blocks. All trials presented a sample at the top of the screen, a target below the sample, and a response option at the bottom corners of the screen. Participants selected a response option by pressing the “d” key for the left corner option and the “k” key for the right corner option. Before each block of trials, participants were presented with a rule that illustrated how they were to respond to the upcoming trials within that block, as well as images depicting examples for each trial-type. During a block, correct selections resulted in a subsequent trial after a 400ms delay, and incorrect selections added a red “X” to the trial under the target and remained until a correct selection was made. The pattern of “correct” responses for odd-numbered blocks (e.g., pro-Black and anti-White responses) was in direct opposition to the pattern for even-numbered blocks (e.g., pro-White and anti-Black responses).

The first IRAP was administered to familiarize participants with the demands of the procedure. It contained the samples “flower” and “sewage” and four evaluative targets (the words “good”, “pleasant”, “bad”, and “repulsive”). This practice IRAP was limited to 2 practice blocks-pairs and 1 test block-pair. The second and third IRAPs contained the samples “black people” and “white people”, six positive evaluative targets (“deserving”, “good”, “motivated”, “smart”, “superior”, and “worthy”), and six negative evaluative targets (“bad”, “deficient”, “inadequate”, “inferior”, “lazy”, and “stupid”). All target words were the same as those used by Drake and colleagues (2010), while “people” was appended to “black” and “white” to clarify the social nature of the categories. Thus, each trial of the race IRAPs contained four trial-types: (1) black people + positive evaluation, (2) black people + negative evaluation, (3) white people + positive evaluation, and (4) white people + negative evaluation (see Figure 1). Trial-types were ordered randomly within each block, with 2 trials set as the maximum run for the same trial-type. Response options for all IRAPs were “true” and “false”, and their locations were set to alternate randomly across trials with 2 trials set as the maximum run for the same positions.

To encourage quick responding during practice blocks, the message “Too slow!” appeared near the center of the screen after 2000ms elapsed during a trial. The first race IRAP provided up to 2 training block-pairs and 3 test block-pairs. Participants who failed to provide greater than 80% correct answers with a median latency less than 2000ms for each block of a training block-pair proceeded to administration of the second race IRAP. The second race IRAP was identical to the first, except that no training blocks were presented. The experimenter provided guidance and feedback during the training blocks of the practice IRAP and the first race IRAP. For the second race IRAP, participants were informed that they would repeat the previous IRAP and were allowed to proceed without detailed instructions. Experimenters encouraged maintaining high accuracy and speed and observed performance from a distance during test blocks.
**General Procedure**

An experimenter conducted data collection in 1-hour sessions with one participant. All experimenters were White. After consenting to the study, participants were assigned a participant number and then engaged two tasks on an IBM-compatible computer running a recent version of Microsoft Windows. The tasks were a collection of self-report measures and a sequence of three IRAPs. The self-reports were presented in conjunction with each other, with their order randomized by the online survey software. The first IRAP was the experimentally unrelated practice IRAP, and the second and third IRAPs were evaluative race IRAPs. The first task (surveys or IRAPs) and the first block-type of the IRAPs (pro-black or pro-White) were randomly counterbalanced across participants. Participants were credited and debriefed about the study when finished with all measures or when 1 hour had elapsed.

---

**Figure 1: IRAP trial-types.** Dashed arrows depict correct responses during pro-White, anti-Black blocks, while solid arrows depict correct responses during pro-Black, anti-White blocks (arrows did not appear on the screen during the procedure).
Results

Attrition and Demographic Subgroups

During administration of the first race IRAP, 12 participants (21.1%; 9 Black, 1 White, and 2 others) failed the practice block criteria and did not produce test block data. Seven additional participants (12.3%; 3 Black, 2 White, 1 Black and White, and 1 other) failed to maintain at least 70% correct on all of the test blocks. These 19 participants were excluded from subsequent analyses involving data from the first race IRAP, leaving an analyzed sample of 38 participants (10 Black, 19 White, 3 Black and White, and 6 “Others” who identified as races other than Black or White).

For the second race IRAP, 1 White participant did not provide any IRAP data due to an experimenter’s error. Four participants (10.0%; 2 Black, 1 White, and 1 Black and White) failed to maintain at least 70% correct on all of the test blocks. These 5 participants were excluded from subsequent analyses involving data from the second race IRAP, as well as the 12 participants who failed the practice block criteria of the first race IRAP (because it was ambiguous whether to categorize their data as a first or second IRAP, given that they did not encounter test blocks from the first IRAP). This left an analyzed sample of 40 participants for the second race IRAP (11 Black, 19 White, 3 Black and White, and 7 Others).

Forty-two participants produced valid data for at least one race IRAP (12 Black, 20 White, 3 Black and White, and 7 Others). Thirty-six participants produced valid data for both race IRAPs (9 Black, 18 White, 3 Black and White, and 6 Others). Although the study was designed as a comparison of data provided by Black and White racial groups, the acquired data resulted in a substantial number of participants who were either (1) affiliated with both groups or (2) affiliated with neither group. Given that neither of these subgroups reported an affiliation exclusive to Black or White, it was decided to combine these two subsamples as a single third “Other” group to allow for comparisons to each of the exclusively affiliated racial groups. All subsequent investigations involved examination of these three self-identified racial groups (see Table 1 for descriptive statistics for all measures among each group as well as for the full sample). Furthermore, given the amount of attrition among Black participants, the unanticipated inclusion of a third racial subgroup, and the sheer number of comparisons available in the dataset, a descriptive approach was embraced for all analyses.

IRAP D-Scores and Reliability

Raw IRAP latency data was processed via the D-score algorithm detailed in previous studies (e.g., Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart, 2010). The race IRAPs were configured such that positive values indicated stereotypical racial attitudes toward Blacks and Whites in the United States (i.e., pro-White and anti-Black evaluations). For example, positive D-scores for the Black-positive and Black-negative trial-types would suggest anti-Black biases, while negative D-scores would suggest pro-Black biases. Alternatively, positive D-scores for the White-positive and White-negative trial-types would suggest pro-White biases, while negative D-scores would suggest anti-White biases. Given that the overall D-score reflects an average of the four trial-type D-scores, a positive value would suggest some combination of pro-White and/or anti-Black biases, while a negative value would suggest pro-Black and/or anti-White biases.
Table 1. Means (M) and Standard Deviations (SD) of Measures for the Full Sample and Racial Subsamples

<table>
<thead>
<tr>
<th>Measure</th>
<th>Full Sample</th>
<th>Black Participants</th>
<th>White Participants</th>
<th>Other Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>M (SD)</td>
<td>M (SD)</td>
<td>M (SD)</td>
</tr>
<tr>
<td>Self-Reports</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modern Racism Scale</td>
<td>12.62 (4.23)</td>
<td>8.92 (2.02)</td>
<td>15.45 (4.14)</td>
<td>11.40 (1.65)</td>
</tr>
<tr>
<td>Social Dominance Orientation</td>
<td>36.05 (12.74)</td>
<td>30.92 (12.26)</td>
<td>39.20 (14.01)</td>
<td>35.90 (9.16)</td>
</tr>
<tr>
<td>Positive Evaluative Words</td>
<td>3.02 (1.17)</td>
<td>3.14 (1.61)</td>
<td>3.07 (0.95)</td>
<td>2.77 (1.02)</td>
</tr>
<tr>
<td>Negative Evaluative Words</td>
<td>-1.96 (2.09)</td>
<td>-0.90 (3.16)</td>
<td>-2.66 (1.08)</td>
<td>-1.82 (1.61)</td>
</tr>
</tbody>
</table>

1st IRAP

| Overall D-Score             | 0.07 (0.24) | -0.14 (0.23)       | 0.19 (0.18)        | 0.05 (0.22)       |
| Black-positive              | -0.17 (0.36)| -0.40 (0.32)       | -0.02 (0.30)       | -0.23 (0.41)      |
| Black-negative              | -0.05 (0.35)| -0.15 (0.26)       | 0.06 (0.36)        | -0.17 (0.39)      |
| White-positive              | 0.34 (0.47) | 0.03 (0.48)        | 0.48 (0.44)        | 0.36 (0.38)       |
| White-negative              | 0.17 (0.34) | -0.04 (0.36)       | 0.24 (0.33)        | 0.24 (0.26)       |

2nd IRAP

| Overall D-Score             | 0.02 (0.24) | -0.06 (0.20)       | 0.10 (0.20)        | -0.06 (0.31)      |
| Black-positive              | -0.23 (0.33)| -0.29 (0.38)       | -0.18 (0.30)       | -0.27 (0.36)      |
| Black-negative              | -0.03 (0.41)| -0.10 (0.29)       | 0.07 (0.43)        | -0.15 (0.48)      |
| White-positive              | 0.31 (0.34) | 0.23 (0.31)        | 0.45 (0.23)        | 0.14 (0.44)       |
| White-negative              | 0.01 (0.37) | -0.08 (0.24)       | 0.04 (0.41)        | 0.04 (0.42)       |
**Split-half reliability.** The D_{IRAP} algorithm was modified such that two overall D-scores were generated: one for odd-numbered trials within each trial-type and one for even-numbered trials within each trial-type. For each race IRAP, these two D-scores were subjected to a Pearson correlation. A large correlation was obtained for the first race IRAP ($r = .464, N = 38$). A Spearman-Brown correction resulted in an estimation of the split-half reliability of the IRAP data at $r = .634$. A medium correlation was obtained for the second race IRAP ($r = .329, N = 40$), with a Spearman-Brown correction resulting in $r = .495$. For the first race IRAP, internal consistency was comparable to an average $r = .653$ for the existing IRAP data reported in a review by Golijani-Moghaddam, Hart, and Dawson (2013); consistency for the second race IRAP was somewhat lower.

**Test-retest reliability.** The overall D-score as well as each trial-type D-score from the first race IRAP was subjected to a Pearson correlation with each complimentary D-score from the second race IRAP. The overall D exhibited a medium correlation ($r = .366, N = 36$), the Black negative trial-type exhibited a medium to large correlation ($r = .413, N = 36$), and the White positive trial-type exhibited a large correlation ($r = .480, N = 36$). Correlations for the Black positive ($r = .209, N = 36$) and White negative trial-types ($r = .203, N = 36$) were small to medium.

**IRAP Effects within Race Affiliations**

Figures 2 and 3 display average D-scores, Cohen’s D effect sizes, and 95% confidence intervals within each race affiliation for each race IRAP administration. With the first race IRAP, the average overall D-scores revealed a medium to large effect size for Black participants and a very large effect size for White participants, both in a pro-in-group direction. Other participants exhibited a small pro-White/anti-Black effect size. For the second race IRAP, effect sizes for the overall D-scores for Black and White participants were still pro-in-group but substantially smaller. Other participants still exhibited a small effect size but in the opposite (pro-Black/anti-White) direction.

For the trial-type D-scores, effect sizes were variable within each racial subgroup. With the first IRAP, Black participants exhibited large and medium pro-Black effects for the Black positive and Black negative trial-types, respectively, and negligible effects for both White trial-types. With the second IRAP, pro-Black effects with the Black trial-types were smaller, a large pro-White effect was exhibited for White positive, and a small anti-White effect was exhibited for White negative.

White participants exhibited large pro-White effects for both White trial-types and negligible effects for both Black trial-types with the first IRAP. With the second IRAP, the pro-White effect remained large for the White positive trial-type, and a medium to large pro-Black effect was exhibited for Black positive. Trial-types involving a negative evaluation were negligible.

Other participants exhibited medium pro-Black effects for both Black trial-types and large pro-White effects for both White trial-types on the first IRAP. On the second IRAP, pro-Black effects were also present for the Black trial-types, but was large for Black positive and small for Black negative. A pro-White effect was small for White positive and negligible for White negative.
Figure 2: Average IRAP $D$-scores and error bars depicting 95% confidence intervals for each race category of the first race IRAP. Cohen’s $D$ effect sizes are displayed along the right margin.

Figure 3: Average IRAP $D$-scores and error bars depicting 95% confidence intervals for each race category of the second race IRAP. Cohen’s $D$ effect sizes are displayed along the right margin.
Table 2. *Cohen’s d Effect Sizes for Each Comparison of Racial Subgroups*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Subgroup Comparison</th>
<th>1st IRAP</th>
<th>2nd IRAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Black/White</td>
<td>Black/Other</td>
</tr>
<tr>
<td>Self-Reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modern Racism Scale</td>
<td>1.86</td>
<td>1.33</td>
<td>1.15</td>
</tr>
<tr>
<td>Social Dominance Orientation</td>
<td>0.62</td>
<td>0.45</td>
<td>0.26</td>
</tr>
<tr>
<td>Positive Evaluative Words</td>
<td>0.06</td>
<td>0.27</td>
<td>0.31</td>
</tr>
<tr>
<td>Negative Evaluative Words</td>
<td>0.84</td>
<td>0.36</td>
<td>0.66</td>
</tr>
<tr>
<td>1st IRAP</td>
<td>Overall D-Score</td>
<td>1.67</td>
<td>0.84</td>
</tr>
<tr>
<td>White-positive</td>
<td>0.99</td>
<td>0.76</td>
<td>0.28</td>
</tr>
<tr>
<td>White-negative</td>
<td>0.82</td>
<td>0.88</td>
<td>0.00</td>
</tr>
<tr>
<td>2nd IRAP</td>
<td>Overall D-Score</td>
<td>0.80</td>
<td>0.00</td>
</tr>
<tr>
<td>Black-positive</td>
<td>0.33</td>
<td>0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>Black-negative</td>
<td>0.44</td>
<td>0.13</td>
<td>0.49</td>
</tr>
<tr>
<td>White-positive</td>
<td>0.84</td>
<td>0.24</td>
<td>0.98</td>
</tr>
<tr>
<td>White-negative</td>
<td>0.34</td>
<td>0.36</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**IRAP and Self-Report Effects between Race Affiliations**

Table 2 displays Cohen’s D effect sizes for comparisons of each possible pairing of the racial subgroups (Black vs. White, Black vs. Other, and White vs. Other) for all measures. The MRS obtained very large effect sizes for all comparisons, while the SDO obtained a large, medium, and small effect size among the comparisons. Furthermore, the magnitude of effect sizes from one comparison to another was identical for the MRS and SDO; Black vs. White obtained the largest while White vs. Other obtained the smallest effect sizes. Relative to each other, Black participants reported the lowest and White participants reported the highest subtle racism and socially dominant attitudes, with Other participants reporting moderate scores on average.

For the full sample, ratings of the evaluative words on the SDS were consonant with expectations, revealing a very large effect size for positive evaluations ($d = 2.581$), as calculated by their deviations from a neutral evaluation value of zero, and a large effect size for negative evaluations ($d = -0.937$). Overall, positive words were viewed as very positive and negative words were viewed as very negative, at least on the average. Between each pairing of the racial subgroups, the positive evaluative words obtained a negligible effect size for the Black vs. White comparison, and for the remaining two comparisons the effect sizes were small. The racial subgroups were fairly comparable in their views of the positive words. For ratings of the negative evaluative words on the SDS, effect sizes for Black vs. White and White vs. Other were large, while Black vs. Other revealed a small effect size. White participants viewed the negative words as substantially more negative than both Black and Other participants.
Among the comparisons of IRAP D-scores for Black and White participants, the effect size was large to very large for all D-Scores for the first race IRAP as well as for the overall D-score and the White positive trial-type of the second race IRAP. The effect sizes were small to medium for Black negative and small for Black positive and White negative of the second IRAP. In all cases, White participants exhibited more pro-White and/or less pro-Black D-scores.

Among comparisons of D-scores for Black and Other participants, effect sizes ranged from medium to large for the first race IRAP, except for a negligible effect size for Black negative. For the second race IRAP, a small effect size was obtained for White negative, and all other D-scores were negligible. Except for the Black negative trial-type for each IRAP, as well as the overall D-score and the White positive trial-type of the second IRAP, average D-scores for Others was always more pro-White and/or less pro-Black than for Black participants.

Among comparisons of D-scores for White and Other participants, for both race IRAPs effect sizes for the White negative trial-type were negligible, while effect sizes for the other D-scores ranged from small to medium to large. Except for the White negative trial-type for each race IRAP, average D-scores for Others were always less pro-White and/or more pro-Black.

**IRAP Convergent Validity**

The overall D-score as well as each trial-type D-score was subjected to a Pearson correlation with scores for the MRS and the SDO. All correlations and their respective 95% confidence intervals are displayed in Table 3. The overall D-scores from the first and the second race IRAP exhibited large correlations with the MRS and medium to large correlations with the SDO. Among the trial-types, the MRS exhibited small to large correlations with each race IRAP. Correlations between the trial-types and the SDO were small to medium.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Modern Racism Scale</th>
<th>Social Dominance Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>95% CI</td>
</tr>
<tr>
<td>Self-Reports (N = 42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Dominance</td>
<td>.57</td>
<td>[0.32, 0.74]</td>
</tr>
<tr>
<td>Orientation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st IRAP (N = 38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall D-Score</td>
<td>.51</td>
<td>[0.23, 0.71]</td>
</tr>
<tr>
<td>Black-positive</td>
<td>.32</td>
<td>[0.00, 0.58]</td>
</tr>
<tr>
<td>Black-negative</td>
<td>.29</td>
<td>[-0.03, 0.56]</td>
</tr>
<tr>
<td>White-positive</td>
<td>.45</td>
<td>[0.15, 0.67]</td>
</tr>
<tr>
<td>White-negative</td>
<td>.18</td>
<td>[-0.15, 0.47]</td>
</tr>
<tr>
<td>2nd IRAP (N = 40)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall D-Score</td>
<td>.50</td>
<td>[0.22, 0.70]</td>
</tr>
<tr>
<td>Black-positive</td>
<td>.30</td>
<td>[-0.01, 0.56]</td>
</tr>
<tr>
<td>Black-negative</td>
<td>.44</td>
<td>[0.15, 0.67]</td>
</tr>
<tr>
<td>White-positive</td>
<td>.32</td>
<td>[0.01, 0.57]</td>
</tr>
<tr>
<td>White-negative</td>
<td>.25</td>
<td>[-0.07, 0.52]</td>
</tr>
</tbody>
</table>
The current findings provide supportive evidence for the reliability and validity of the IRAP as a measure of relational repertoires regarding racial evaluations. Psychometrically, there were supportive results for internal reliability, test-retest reliability, and convergent validity. Furthermore, the trends in the data were more nuanced than would be apparent from more widely used measures of implicit cognition such as the IAT. In the first race IRAP, Black and White participants exhibited effect sizes suggestive of a pro-in-group bias and little to no evidence suggesting either a pro- or anti-out-group bias. The pro-Black effects obtained among Black participants in this study have been elusive for IAT researchers. For the second race IRAP, similar pro-Black effects with Black participants and pro-White effects with White participants were observed, although in-group trial-types containing a negative evaluation were smaller than observed for the first race IRAP. In contrast, Other participants exhibited a unique combination of pro-Black effects and pro-White effects for the first IRAP. For the second IRAP, these patterns were similar although substantially smaller for each of the White trial-types.

The additional specificity provided by the IRAP trial-types over the one-dimensional score provided by a comparable IAT appears to offer additional insights into racial attitudes that may not be apparent with IAT studies. For the first race IRAP, the expected pro-White bias among White participants was indeed pro-white but not anti-Black. Black participants exhibited a complimentary pro-Black and not anti-White bias. These data may conflict with typical views of discriminatory attitudes, as the biases exhibited were in favor of the in-group and neutral toward the out-group, rather than in opposition to the out-group. Although the end result may be the same in regard to discriminatory behavior, the foundation for the observed biases appears to be driven by favoritism for one group rather than opposition toward another, a finding that seems more reflective of a self-serving bias in respect to social categorization (e.g., Tajfel & Turner, 1979) than stigmatizing attitudes toward members of an out-group. Interestingly, the two trial-types most indicative of stereotypical racist attitudes were the most reliable in the test-retest analysis (Black negative and White positive), and the least reliable trial-type was in regard to negative evaluations of the racial majority (White negative).

Curiously, the trial-type D-scores among Black and White participants (but not Others) appeared to shift to a more egalitarian pattern with the second race IRAP, which may raise questions about the stability and validity of the measure over successive administrations. In contrast to the small to negligible effects observed with out-group trial-types in the first race IRAP, on the second race IRAP Black participants exhibited a large pro-White bias for White-positive, and White participants exhibited a medium to large pro-Black bias for Black positive. Perhaps exposure to the first race IRAP made it possible for the participants to respond to the second race IRAP more as they would on self-reports of racial attitudes. Also of note, Black participants exhibited a small anti-White effect for White negative for the second race IRAP, which presents a unique combination of pro- and anti-out-group biases in this sample. Again, a similar IAT could not detect these kinds of complex (and somewhat contradictory) patterns.

In addition to the apparent differences in effect sizes among the trial-types, the overall pattern of biases within each race subsample was not fully consistent across each IRAP. Effect sizes changed in various ways from the first to the second IRAP, including transitions from negligible to large (e.g., White positive for Black participants) and from large to negligible (e.g., White negative for White participants). Only one large effect was shown to increase with the second IRAP (White positive for White participants). Although split-half and test-retest
IMPLICIT RACIAL EVALUATIONS

correlations were medium to large, they were not large enough to provide strong support for the reliability of the IRAP (for discussion, see Golijani-Moghaddam, Hart, & Dawson, 2013). In any case, it appears that much of the variability accounted for by self-report was preserved from the first to the second race IRAP, as comparable correlations with self-reports were observed with both. For both the MRS and the SDO, r values with the overall D-scores of both IRAPs were almost identical. The generally larger correlations with MRS scores over SDO scores offers additional support for convergent validity, as the race-specific nature of the MRS items was more consistent with the race-specific content of the IRAP. Self-report correlations with some of the trial-type D-scores were also substantive. On the whole, it appears that the level of specificity provided by the IRAP trial-types can reveal noteworthy variations in the function of interlocking sets of attitudes.

Limitations and Future Directions

The attrition rate was substantial and skewed on the basis of participant race. This may justify some degree of reservation about the results among Black participants, given that about half of them did not produce IRAP data. Black participants had a relatively more difficult time achieving criterion responding. Perhaps the unbalanced attrition for the racial subsamples was due to all experimenters being White. Having an equal number of Black experimenters administering the IRAP to White participants may have balanced the attrition rates among the racial groups. Although a previous IRAP study did not detect any effects for experimenter ethnicity (Barnes-Holmes et al., 2010), Lowery, Hardin, and Sinclair (2001) found that the magnitude of implicit bias measured on the IAT among White participants varied depending on whether the experimenter was Black or White. The current data suggest that further investigation of this potential influence on IRAP performance may be merited. Future researchers might consider selecting IRAP administrators that represent the diversity of participants in respect to the social categories that are being examined, which may be a challenge for a lab examining racial bias when the experimenters are racially homogenous.

Alternatively, perhaps the selection of IRAP targets (evaluative words) inadvertently favored retention of White participants and attrition of Black participants, as some of the negative evaluative words may be regarded as negative stereotypes for African Americans (e.g., “lazy”), while some of the positive evaluative words may stereotype Caucasians (e.g., “superior”). Tables 1 and 2 display some data suggesting that Black participants viewed the negative evaluative words less negatively than White participants. It is not clear if the negative words were viewed differently because of their stereotyped racial significance or some other factor. Nevertheless, it appears that populating an IRAP with stereotypical evaluations bears some complexities, as there may be positive and negative stereotypes for each race, and they may not be the same for each race.

An experimenter could populate the target words of the IRAP in a variety of ways. For example, all the positive words could stereotypically reference one race sample, while all the negative words could reference the other. However, this strategy resembles that for the current study, and may impact attrition or effects in a manner similar to the current results because each race may relate to the evaluative words differently. An alternative strategy could involve populating the IRAP with positive and negative target words that stereotypically reference only one race and not the other. This also may result in unbalanced effects between the races because all of the stimuli are more relevant to one race than the other. A third option might be to mix
positive and negative stereotypes for both races, which would provide a methodologically balanced approach. However, it may also be a more challenging task, as the mixture of different stereotypes into single classes of target stimuli may complicate the repertoires required for successful engagement with the task. A fourth option might entail avoiding stereotyped evaluations entirely and relying exclusively on generic, non-racial evaluations. These four strategies may or may not differentially impact IRAP data and their relationship to other measures and seem worthy of additional research. The optimal approach likely depends on the goals and creativity of the researcher. Future studies should explore these concerns by systematically manipulating the stereotypy of the words, positive or negative, in respect to the racial affiliations being compared, as this will impact the design of subsequent IRAP studies across a variety of content domains.

Although additional research on reliability and concurrent validity is needed, the current data are generally positive, albeit incomplete. This study did not address the predictive validity of the IRAP, a crucial test of the value of any measure. Race IATs have been shown to predict laboratory examples of discriminatory behavior (e.g., Green, et al., 2007; Heider & Skowrons, 2007), although some recent meta-analyses have suggested that the predictive validity data as a whole is relatively unimpressive (Blanton, et al., 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). It seems important to subject the IRAP to an examination of predictive validity, given the divergence in methodology from the IAT and the more specificity of results it can provide. Considering the generally positive results for the current study, a program of predictive validity research might be a productive next step.

References
IMPLICIT RACIAL EVALUATIONS


Gawronski, B., & De Houwer, J. (2011). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology (2nd edition)*. New York, NY: Cambridge University Press. [http://dx.doi.org/10.1017/cbo9780511996481.016](http://dx.doi.org/10.1017/cbo9780511996481.016)


