The Psychological Record, 2009, 59, 371–388

A FIRST TEST OF THE IMPLICIT RELATIONAL ASSESSMENT PROCEDURE AS A MEASURE OF SELF-ESTEEM: IRISH PRISONER GROUPS AND UNIVERSITY STUDENTS

Nigel A. Vahey, Dermot Barnes-Holmes, and Yvonne Barnes-Holmes
National University of Ireland, Maynooth, Ireland,

Ian Stewart
National University of Ireland, Galway, Ireland

The study examined the Implicit Relational Assessment Procedure's (IRAP) validity as a computerized response-latency-based measure of implicit self-esteem. University undergraduates and 2 sets of convicted prisoners participated. One set of prisoners resided in the main block, and the other in a privileged lower security "open area" of a medium-security Irish prison. The IRAP required participants to maintain relational responses that were self-positive on half of the IRAP trials ("Consistent"), and self-negative on the other half ("Inconsistent"). As predicted, the students and the prisoners in the open area showed stronger IRAP effects (shorter latencies during consistent vs. inconsistent trials) than the main block prisoners. Additionally, the IRAP's convergent validity was supported by its moderate positive correlation with an explicit self-esteem measure. The findings provide preliminary support for the analytic utility of the IRAP and suggest future avenues of investigation afforded by the IRAP's design.

A number of researchers have argued that the study of implicit cognition could be important in the analysis and treatment of human psychopathology (e.g., Wiers, Teachman, & De Houwer, 2007). Numerous methodologies have been developed that aim to provide measures of such cognitions, including evaluative priming procedures, the *Implicit Association Test* (IAT), the *Go/No Go Association Test* (GNAT), and the *Extrinsic Affective Simon Test* (EAST; De Houwer, 2003, 2008, for reviews). Although these and other measures differ in their procedural details, they are all best considered *indirect* measures. That is, in each case the

Preparation of the current article was supported by a John and Pat Hume postgraduate scholarship awarded to Nigel Vahey, from the National University of Ireland, Maynooth.

The authors wish to express their gratitude to the Irish Prison Service Ethics Committee, Governor Daniel Scannell, Riona Finn, and Maria Lahiff for their assistance in completing the current research.

Address correspondence to either Nigel Vahey or Dermot Barnes-Holmes, Department of Psychology, National University of Ireland, Maynooth, Ireland. E-mail: NigelVahey@gmail.com, Dermot.Barnes-Holmes@nuim.ie.

procedure involves asking participants to engage in some form of task that does *not* involve confirming or denying the belief or attitude under study. Instead, participants are asked to categorize attitude-relevant stimuli with positive and negative emotional functions. Thus, on a self-esteem IAT, for example, participants are sometimes required to press a left key for *self* words (e.g., “me”) and *positive* words (e.g., “love”), and to press a right key for *not-self* (“other”) and *negative* (e.g., “hate”) words. If response times are faster on this task than on the other, when categorizing *self* with *negative* and *not-self* with *positive,* the difference in response times is taken to be an index of positive self-esteem (if the response time difference is reversed, an index of negative self-esteem is assumed).

The IAT and other associative measures are indirect in the sense that they are designed to tap into underlying associations without asking participants to respond directly to questions concerning specific beliefs or attitudes (see De Houwer, 2002). Although it seems reasonable to infer attitudes from associations, it would also seem prudent to attempt to develop additional methodologies that aim to provide relatively direct measures of implicit cognition. The Implicit Relational Assessment Procedure (IRAP) constitutes a first step in this direction, and it is the focus of the current study (see Barnes-Holmes et al., 2006).

The IRAP procedure is based on a relatively recent account of human language and cognition known as Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). According to RFT, the core units of human language and cognition are not associations per se, but derived stimulus relations. One of the main methodologies to emerge from the theory is the Relational Evaluation Procedure (REP). The REP allows participants to report on a 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.” The REP has now been used across a range of studies to examine reasoning and other forms of higher cognition (O’Hora, Barnes-Holmes, Roche, & Smeets, 2004; O’Hora, Peláez, & Barnes-Holmes, 2005; Stewart, Barnes-Holmes, & Roche, 2002, 2004). Critically, the REP provided the basis for the development of the IRAP, which is basically a combination of the IAT and the REP (see Barnes-Holmes et al., 2006; Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008).

The IRAP is a relatively direct latency-based measure, in that participants must confirm or deny a specific belief or attitude by responding to a relation between a sample stimulus and a target term (e.g., Similar – Good = Me or Not Me?). The computer-based task requires participants to respond quickly and accurately in ways that are either consistent or inconsistent with their prior learning histories. It is assumed that overt relational responses defined as consistent on the IRAP will be preceded by incipient or private responses that occur at a higher probability than those responses defined as inconsistent; the probability of such responses is assumed to be determined by historical and current contextual variables. The basic rationale behind the IRAP is that participants’ responding should be faster on consistent relative to inconsistent trials because incipient relational responding will coordinate more frequently with the consistent overt responding. In other words, during inconsistent trials, participants’ responding is expected to be slower, as they respond against their more probable incipient relational responses.[1] The extent of the observed difference between consistent and inconsistent trials is assumed to provide a relatively[2] direct index of the strength of the specific belief being assessed.

The basic IRAP effect has been replicated across a number of domains in previous and ongoing research at our laboratory (Barnes-Holmes et al., 2006; Barnes-Holmes, Hayden, et al., 2008; Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart, in press; Barnes-Holmes, Waldron, Barnes-Holmes, & Stewart, in press; McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007; O'Toole & Barnes-Holmes, 2009). At the current time, however, there is no published evidence to support the IRAP's validity as a potential tool for measuring clinically relevant variables. De Houwer (2002) suggested that one method for testing the validity of an implicit measure was to determine if it produced different results in accordance with known-group differences. This was achieved in the context of the current study by drawing on a population that the literature indicates typically presents with lower than average self-esteem (Irish male prisoners; see Oser, 2006). For comparison, the study also employed participants from a population with a tendency for higher levels of self-esteem than convicted prisoners: a random sample of Irish undergraduate students (for empirical evidence see Gullone, Jones, & Cummins, 2000; Oser, 2006). Indeed, the well-documented positive bias in measures of self-esteem, both implicit (Yamaguchi, Greenwald, Banaji, Murakami, Chen, Shiomura, et al., 2007) and explicit (Greenwald & Banaji, 1995), led us to predict that the undergraduates, a relatively normative group in the literature, would indicate positive levels of self-esteem on the IRAP.

On balance, it is important to recognize that self-esteem may fluctuate as a function of numerous variables. Social comparison theory (Festinger, 1954) and its successor theories predict that self-esteem may vary because of the types of social comparisons an individual makes between the self and *available* peers (Martinot & Redersdorff, 2006; Mussweiler, 2001, 2003; Stiles & Kaplan, 2004). Furthermore, both upward and downward social comparisons tend to be related to lower self-esteem in persons with a perceived lack of control over their circumstances, whereas the converse appears to hold with those having greater perceived control over their circumstances (Judge, Erez, Bono, & Thoresen, 2002; Michinov, 2001; Stiles & Kaplan, 2004). Thus, prisoners who have greater control over their daily routines and who perceive themselves to be better than many of their fellow prisoners are likely to experience higher self-esteem than mainstream prisoners, even though relative to the wider culture outside the prison they would compare rather negatively. Crucially, findings in the prison literature concur with these predictions: Prisoners reporting higher internal locus of control, those having achieved relatively higher status, and in particular those benefiting from the greater autonomy of trustee status tend to experience higher levels of self-

[1] We assume that responding against incipient relational responses may occur at an unconscious or nondeliberative level (Barnes-Holmes et al., 2006; Barnes-Holmes et al., 2008). This view is broadly consistent with recent evidence indicating that when participants are successfully motivated to reduce an IAT effect, it appears that they do so through unconscious processes (Boysen, Vogel, & Madon, 2006).

[2] The qualifier "relative" is used because although participants are asked to respond directly to stimulus relations that confirm or deny the relevant beliefs, response latency, rather than specific verbal reports, is used as the index of those beliefs (see Barnes-Holmes et al., 2006, for a detailed discussion).

esteem than those subject to higher security prison regimes (e.g., Blatier, 2000; Jacques & Chason, 1977; Oser, 2006; Regens & Hobson, 1978).

To test this basic prediction in the current study, a second group of male prisoners was sampled who had *earned* residency in a special "open area" within the prison through a protracted application process (open to all prisoners). The prisoners from the open area, as "trustees," were publicly accorded self-affirming responsibilities, freedoms, and privileges not granted to the other prisoners. Thus it was predicted that the open area prisoners, who had been identified as worthy of special treatment, would possess higher self-esteem relative to their "less worthy" counterparts resident in the highly controlled environment of the main block.

The IRAP and a self-report "feeling-thermometer" (adapted from Greenwald & Farnham, 2000) were used with all participants to measure implicit and explicit self-esteem, respectively. The feeling-thermometer is a very brief self-report measure that requires participants to rate how warmly they feel toward themselves on an illustrated thermometer. The IRAP procedure involved a computer-based task in which participants were asked to respond quickly and accurately to "questions" that asked them to confirm or deny positive and negative self-evaluations. On one type of trial, for example, the relational word *Similar* appeared at the top of the computer screen with the positive term *Honest* immediately below. For this trial, participants were required to choose *own name* for some blocks of test trials (referred to as Consistent) and to choose *not own name* on other blocks (referred to as Inconsistent). Thus, if the participant's name was Nigel, the trial was to be read as "Who is Similar to Honest: 'Nigel' or 'Not Nigel'?"

The full IRAP task consists of four trial-types; these are explained via exemplars in Figure 1. Overall, the IRAP requires participants to emit responses during consistent blocks that confirm self-positive and deny self-negative evaluations; during inconsistent blocks this requirement is reversed (i.e., confirm self-negative and deny self-positive evaluations). If the IRAP functions as a valid measure of implicit self-esteem, shorter response latencies on consistent relative to inconsistent blocks would indicate higher levels of self-esteem. Furthermore, based on previous IAT studies (e.g., Bosson, Swann, & Pennebaker, 2000; Franck, De Raedt, & De Houwer, 2007; Franck, De Raedt, Dereu, & Van den Abbeele, 2007; Glen & Banse, 2004; Greenwald & Farnham, 2000; Karpinski & Steinman, 2006; Nosek, Greenwald, & Banaji, 2007), it was anticipated that there would be a weak to moderate positive correlation between the IRAP and the explicit measure of self-esteem.[3]

Method

Participants

Undergraduates. Thirty participants (17 females and 13 males) were recruited from a convenience sample of Irish undergraduate students from a range of disciplines.

Prisoners. A total of 21 male convicted prisoners from a medium-security

[3] It is worth noting that research focusing on more socially sensitive attitudes has found a divergence between the IRAP and explicit measures similar to that reported in the IAT literature (e.g., Barnes-Holmes et al., 2006).

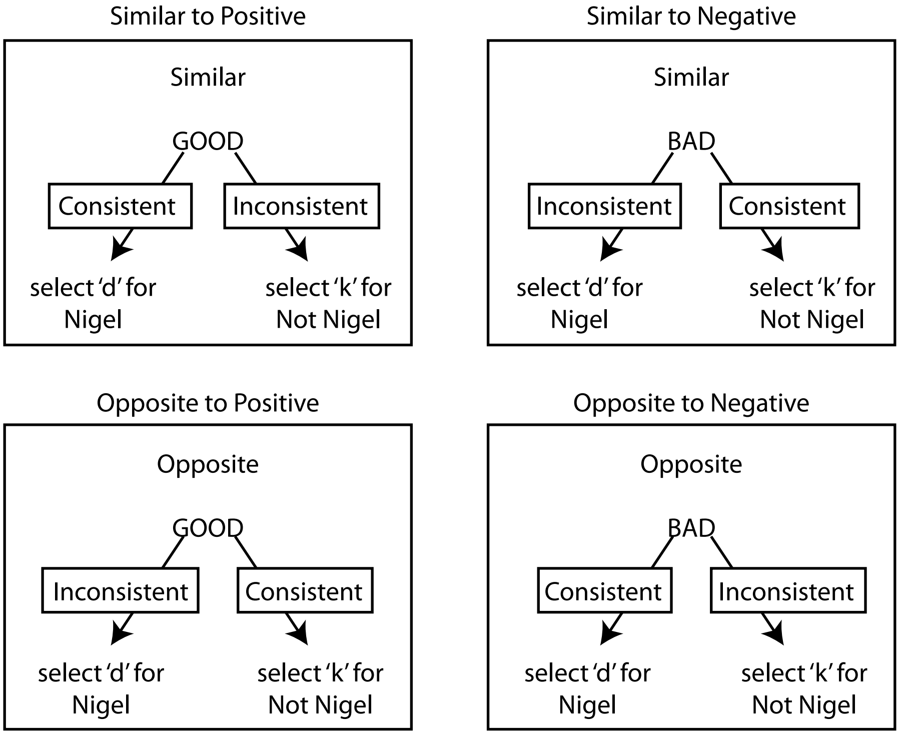

Figure 1. Examples of the four trial types employed in the self-esteem IRAP: one for each combination of the two sample stimuli (Similar or Opposite) with the two types of target stimuli (self-positive or self-negative evaluative words). A sample stimulus, a target word (e.g., *Good*, *Bad*, etc.), and both of the response options (Participant's First Name, and Not Participant's First Name) appeared simultaneously on screen at the onset of each trial. The left–right positions of the response options varied randomly across trials. The superimposed arrows with text boxes indicate the responses deemed consistent (i.e., self-positive) or inconsistent (i.e., self-negative); boxes and arrows did not appear on screen during the experiment. The critical comparison in calculating the IRAP effect is between consistent and inconsistent responses within each trial type.

Irish prison were recruited according to availability and the guidelines of the Irish Prison Service Ethics Committee. Fifteen of the prisoners were recruited in the school wing of the prison. These prisoners resided in the main prison block. The remaining 6 prisoners,[4] also students of the prison school, resided in an open area of the same prison—a compound with low-security status contained within a 20-foot-high wall surrounding the prison compound. The open area is approximately four acres in size and contains seven two-story houses, rather than cell blocks, so that the prisoners are allowed access to a wide range of privileges, including dedicated workshops

[4] The number of open area prisoners available to participate in the research was low because only a limited number reside therein, and of those, many had to be excluded on the grounds indicated herein. Furthermore, the open area provision, within a "mainstream" prison, is currently unique in the Irish Prison Service.

and classrooms, an agricultural garden and greenhouses, a prisoner-staffed canteen, and a gym. Also, in contrast to the weekly 30-minute visit allowed other prisoners, open area prisoners are allowed visitors on a daily basis from 10 a.m. to 6 p.m. Furthermore, a number of open area prisoners tend goats, rabbits, and a variety of fowl within their compound; notably, the integration of animal care into daily prison life is known to enhance prisoner self-esteem (Furst, 2006). In summary, prisoners in the open area benefit from a domestic-type environment and have a very high level of autonomy in comparison to prisoners in the main block, who spend the majority of their time locked in their shared cells (approximately 16 hours per 24-hour period).

Only those imprisoned for sexual offenses or paramilitary activities were excluded from the study. Sexual offenders were excluded because of the heterogeneity of self-esteem among sub-categorizations of such prisoners (e.g., Kalichman, 1991; Marshall, Marshall, Sachdev, & Kruger, 2003; Shine, McCloskey, & Newton, 2002). Paramilitary prisoners were excluded because they frequently report that they are not criminals, even to the extent that they do not associate with nonparamilitary prisoners and retain their rank structure within the prison (e.g., Jamieson & Grounds, 2002; Mitchell, 2003). The prisoners' (and undergraduate students') ages ranged from 18 to 40 years, as is typical of the Irish prisoner population.

Measures

IRAP Self-Esteem Measure. During pilot testing, IRAP stimuli were chosen so as to facilitate social evaluations meaningful to the undergraduates, the main block prisoners, and the open area prisoners in ways relevant to the known-groups predictions we made via Social Comparison Theory (see Table 1). For example, the target word *dishonest* would likely be interpreted as negative by both undergraduates and prisoners, whereas *dangerous* might be interpreted less negatively by prisoners than by students (because a "dangerous" prisoner may accrue higher status among other prisoners; Johnson, as cited in Oser, 2006).

Table 1
Stimulus-response combinations deemed consistent in the self-esteem IRAP

Sample 1	Positive targets	Sample 2	Negative targets	Sample 1	Negative targets	Sample 2	Positive targets
Similar	Good	Opposite	Bad	Similar	Bad	Opposite	Good
	Success		Failure		Failure		Success
	Honest		Dishonest		Dishonest		Honest
	Capable		Worthless		Worthless		Capable
	Pleasant		Nasty		Nasty		Pleasant
	Confident		Ashamed		Ashamed		Confident
Response option 1				Response option 2			
Participant's name				Not participant's name			

Note. By implication all other stimulus-response combinations are deemed inconsistent.

Subjective self-esteem measure. A paper-based feeling-thermometer adapted from Greenwald and Farnham (2000) was used as an explicit measure of self-esteem. Feeling-thermometers are commonly used in the IAT literature; they possess the advantage of being very brief yet nevertheless correlate to a high degree with questionnaires such as the *Rosenberg Self-esteem Scale* ($.68 < r < .74$; Greenwald & Farnham, 2000; Karpinski, 2004). The current feeling-thermometer was composed of an illustrated thermometer with a continuous vertical scale anchored below at 0 (cold), rising in intervals of 10 to its upper bound at 99 (warm); a caption above the thermometer read, "Please Indicate on the Thermometer Below How Warmly You Feel Towards Yourself." A further caption underneath the illustrated thermometer prompted the participant to provide the thermometer score to be used in subsequent analyses: "Once you have marked the position you consider appropriate, please then write the number from 0–99 that your mark indicates."

Procedure

Undergraduates. Participants were asked to complete the feeling-thermometer measure, giving their initial "gut reaction." Participants read a series of IRAP instructions, and the researcher then used laminated illustrations of four example trials to probe participants' understanding of the IRAP tasks (see the Appendix). The illustrations were similar to those presented in Figure 1 but without any indication of which responses were deemed consistent versus inconsistent. To check for understanding, participants were asked to explain what each response choice indicated about them personally in the context of each of the four IRAP trial-types. In other words, the meaning of each of the four types of stimulus combinations in Figure 1 was explained to participants; for example, they were told how responding with their own name to the stimulus combination "Similar – Good" meant that they were designating themselves as similar to good.

The IRAP computer program required participants to complete two practice blocks and then six test blocks, with each block containing 24 trials (software available from the authors upon request). Participants in the consistent-relations-first condition commenced with a block of consistent trials (confirm self-positive and deny self-negative relations) and thereafter alternated between blocks of inconsistent (confirm self-negative and deny self-positive relations) and consistent trials; participants in the inconsistent-relations-first condition were exposed to the blocks in the opposite sequence (i.e., inconsistent followed by consistent). Participants were assigned randomly to the consistent- and inconsistent-relations-first conditions.

In each block, the trials were presented in a quasi-random order with the constraint that each of the two sample stimuli ("Similar" and "Opposite") appeared once with each of the 12 target stimuli (see Table 1). Across trials, the left–right positions of the two response options varied randomly. On each trial, all stimuli appeared simultaneously on screen. During consistent blocks, a consistent response cleared the screen for 400 ms and then the next trial was presented. If an inconsistent response was emitted, a red X appeared immediately under the target stimulus. To remove the red X and continue to the 400-ms intertrial interval, the participant was required to emit the consistent response. In contrast, during inconsistent blocks, participants were required to make the inconsistent response in order to progress from one trial to the next. When the participant had completed all 24 IRAP trials, the screen cleared and two types of feedback were presented for that block: the percentage of correct responses and the median response latency. Participants were reminded via the laminates mentioned above that their primary task was to learn response rules that would allow them to avoid the red Xs. It was further explained, and then role-played via the laminates, that these response rules would alternate between blocks of tasks. Participants were then reminded of these instructions as needed throughout the practice phase of the IRAP.

Between each block of trials the following instructions were presented on screen: "Important: during the next phase the previously correct and wrong answers are reversed. This is part of the experiment. Please try to make as few errors as possible—in other words, avoid the red X." Before each test block, the following message also appeared: "This is a test. Go fast; making a few errors is okay." After the feedback for the sixth and final test block, a message appeared informing the participant that the experiment was complete.

Prisoners. Pilot work with the prisoner population indicated that some participants experienced difficulty in learning the IRAP task, relative to the students. Consequently, for the prisoners, the researcher monitored the performance feedback displayed at the end of each practice block. When these participants failed to attain at least 80% accuracy on the practice blocks, or did not respond quickly across trials, the researcher then reloaded the software and the practice blocks were presented again. When participants responded at 80% accuracy or above on both practice blocks, they were allowed to progress through the subsequent six test blocks of the IRAP.

Results

The IRAP Measure

The raw IRAP data comprise response latencies, defined as the time in milliseconds from the onset of a trial to the first emission of the required response for that trial. The data from six undergraduates and two main block prisoners were excluded from the analyses because they each failed to achieve at least 70% accuracy during the test blocks. Low accuracy scores that persisted across test blocks were taken to indicate unreliable control of participants' relational responses by the IRAP, and thus any differences in response latency between consistent and inconsistent blocks would be difficult to interpret. The overall mean response latencies averaged across the three test blocks for each of the three groups were shorter for the consistent (C) than the inconsistent (I) trials (undergraduates, C = 2,479 ms, σ_{Con} = 528, I = 3,081 ms, σ_{Incon} = 1009; main block prisoners, C = 2,944 ms, σ_{Con} = 526, I = 3,136 ms, σ_{Incon} = 606; open area prisoners, C = 2,929 ms, σ_{Con} = 464, I = 3,513 ms, σ_{Incon} = 742). Each group produced patterns of accuracy concordant with the response latencies for the consistent and inconsistent test blocks: Where latencies were lower, accuracy scores tended to be higher (undergraduates, C = 94%, I = 86%; main block prisoners, C = 89.4%, I = 88.5%; open area prisoners, C = 92%, I = 83%).[5]

D_{IRAP} algorithm. Statistical analyses first involved transforming the individual response latencies for each participant using the D_{IRAP} algorithm,

[5] Although the analyses described were conducted with the inclusion of female participants (in the undergraduate sample), the same statistical conclusions were yielded when female participants were excluded from the analyses.

derived from the D-algorithm developed by Greenwald, Nosek, and Banaji (2003) for the IAT (see also Back, Schmukle, Egloff, & Gutenberg, 2005; Cai, Sriram, Greenwald, & McFarland, 2004; Mierke & Klauer, 2003). Specifically, the D-algorithm is used to control for individual variations in speed of responding, caused by differences in cognitive ability, that may act as a possible confound when analyzing between-group differences. Indeed, the use of the D-algorithm was deemed particularly important in the current study because overall response latencies for the prisoners were longer than for the undergraduates, thus indicating possible between-group differences in cognitive ability.[6]

The D_{IRAP} algorithm transforms the raw latency data for each participant using the following steps:

1. Only response-latency data from test blocks are used;
2. Latencies above 10,000 ms are eliminated from the dataset;
3. The data are eliminated for a participant for whom more than 10% of test-block trials have latencies less than 300 ms;
4. Compute 12 standard deviations for the four trial types: 4 for the response latencies from across test blocks 1 and 2, 4 from across the latencies from test blocks 3 and 4, and a further 4 from across test blocks 5 and 6;
5. Compute the 24 mean latencies, one for each of the four trial types in each of the six test blocks;
6. For each pair of test blocks, use step 5 to compute difference scores for each of the four trial types, by subtracting the mean latency of each trial type's consistent test trials from the mean latency of their corresponding inconsistent test trials;
7. Divide each difference score by its corresponding standard deviation from step 4, yielding 12 D_{IRAP} scores—1 score for each trial type for each of the 3 pairs of test blocks;
8. Calculate 4 overall trial-type D_{IRAP} scores by averaging the 3 scores for each trial type across the three pairs of test blocks;
9. Two compound D_{IRAP} scores, one for positive target words ($D_{IRAP\text{-}POS}$) and one for negative target words ($D_{IRAP\text{-}NEG}$), were then calculated by averaging the two positive and then the two negative trial-type D_{IRAP} scores from step 8; and
10. We calculated a single overall D_{IRAP} score called $D_{IRAP\text{-}Total}$ by averaging the 4 trial-type D_{IRAP} scores from step 8.

Preliminary analyses. The design of the current study involved assigning half of the participants in each participant group to a consistent-relations-first condition and the remaining halves to an inconsistent-relations-first condition (order). To determine if order interacted with the critical IRAP effects, a 2 × 2 mixed repeated measures ANOVA was conducted with order as the between-groups variables and IRAP effect type ($D_{IRAP\text{-}POS}$ or $D_{IRAP\text{-}NEG}$) as

[6] O'Toole and Barnes-Holmes (2009) found that their raw IRAP effect — the response latency differences between consistent and inconsistent trials — correlated significantly with various measures of intelligence; yet when the D_{IRAP}-transformation was performed on the data (not reported in the article), no significant correlations with intelligence were observed.

the within-participants variable. The main effect for IRAP effect type was significant, $F(1, 41) = 15.18$, $p < .001$, $\eta^2_p = .19$, but neither the main effect for order nor the interaction were significant, Fs < .08, ps > .78, $\eta^2_p < .002$. Thus, order was not included in subsequent analyses.

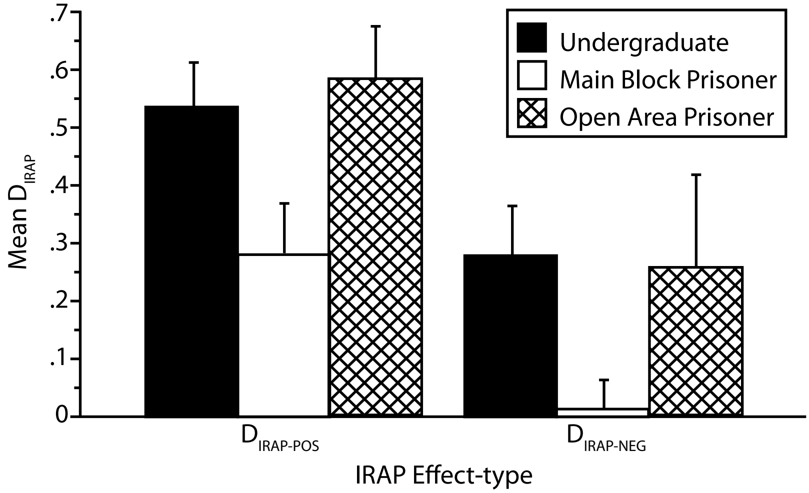

Figure 2. The mean $D_{\text{IRAP-POS}}$ and mean $D_{\text{IRAP-NEG}}$ scores with standard error bars for the undergraduate students, main block prisoners, and open area prisoners.

Participant-type analyses. The mean $D_{\text{IRAP-POS}}$ and $D_{\text{IRAP-NEG}}$ scores calculated for each of the three groups of participants are presented in Figure 2. The data show similar mean IRAP effects for the undergraduates ($D_{\text{IRAP-POS}} = .54$; $D_{\text{IRAP-NEG}} = .28$) and the open area prisoners ($D_{\text{IRAP-POS}} = .59$; $D_{\text{IRAP-NEG}} = .26$), but far smaller IRAP effects for the main block prisoners ($D_{\text{IRAP-POS}} = .29$; $D_{\text{IRAP-NEG}} = .01$). In other words, the former two groups, relative to the main block prisoners, responded more rapidly for trials that required confirmation of self-positive and denial of self-negative relations (i.e., consistent trials) over tasks requiring confirmation of self-negative and denial of self-positive relations (i.e. inconsistent trials). In short, the IRAP data indicated that the undergraduates and open area prisoners both possessed higher self-esteem relative to the main block prisoners. The D_{IRAP} scores for each participant were entered into a 2 × 3 mixed repeated measures ANOVA with participant type as the between-participants variable (undergraduate, main block, and open area) and IRAP effect-type as the within-participants variable ($D_{\text{IRAP-POS}}$ and $D_{\text{IRAP-NEG}}$). Participant type proved to be significant, $F(2, 40) = 4.55$, $p = .017$, $\eta^2_p = .19$, as did the main effect for IRAP effect type, $F(2, 40) = 12.262$, $p = .001$, $\eta^2_p = .23$; however the interaction did not reach significance[7], $F(2, 40) = .067$, $p = .94$, $\eta^2_p = .003$. Post hoc Fisher PLSD tests indicated significant differences between main block and open area prisoners ($p = .006$, $d = 1.58$) and between main block prisoners and undergraduates ($p = .04$, $d = 1.04$), but nonsignificance between open area prisoners and undergraduates ($p = .9$, $d = .08$).

One-group t tests were computed on each group's $D_{\text{IRAP-Total}}$, $D_{\text{IRAP-POS}}$, and

[7] Given the low n for the open area prisoner group, appropriate nonparametric analyses were also conducted, which yielded similar statistical conclusions.

$D_{IRAP\text{-}NEG}$ scores to identify which groups produced positive IRAP effects differing significantly from zero (indicating positive self-esteem). All three groups produced $D_{IRAP\text{-}Total}$ effects that were significantly different from zero: Undergraduates, $t(23) = 6.385$, $p < .0001$, $d = 1.3$; main block, $t(12) = 3.317$, $p = .006$, $d = .92$; open area, $t(5) = .249$, $p = .002$, $d = 2.3$, and this pattern was also observed for the $D_{IRAP\text{-}POS}$ scores, undergraduates, $t(23) = 6.761$, $p < .0001$, $d = 1.38$; main block, $t(12) = 3.541$, $p < .002$, $d = .98$; open area, $t(5) = 6.973$, $p < .0005$, $d = 2.9$. For the $D_{IRAP\text{-}NEG}$ scores, however, the effect was significant for the undergraduates, $t(23) = 3.415$, $p = .001$, $d = .699$; marginally significant for the open area prisoners, $t(5) = 1.635$, $p = .08$, $d = .68$, and nonsignificant for the main block prisoners, $t(12) = .249$, $p = .4$, $d = .07$. Applying a Bonferroni correction for familywise type-1 error among the nine t tests ($p < .006$), results in the same statistical conclusions with the exception of the comparison for the open area prisoner's $D_{IRAP\text{-}NEG}$ scores. Therefore, the undergraduates and open area prisoners produced IRAP effects that tended to confirm positive and deny negative evaluations of self; in contrast, the effects for the main block prisoners indicated confirmation of positive but not denial of negative self-evaluations.

The feeling-thermometer. The participants' thermometer scores ranged from 24 to 96 on the 0–99 scale. Open area prisoners produced marginally higher thermometer scores than the undergraduate participants, while the main block prisoners produced the lowest thermometer scores ($M = 69.8$, $\sigma = 10.5$; $M = 68.2$, $\sigma = 16.9$; $M = 53.7$, $\sigma = 10.4$, respectively). The thermometer scores were entered into a between-participants one-way ANOVA, and this yielded a significant effect for group (see Note 7), $F(2, 40) = 4.77$, $p = .01$, $\eta^2_p = .19$. Fisher's PLSD post hoc tests revealed a pattern of known-group differences in keeping with the D_{IRAP} scores: main block and open area prisoners, $p = .03$; main block and undergraduates, $p = .006$; open area and undergraduates, $p = .8$.

Correlational analyses of the IRAP versus the feeling-thermometer. Pearson's r was calculated to determine if the feeling-thermometer measure was indeed related to the self-esteem IRAP's global effect: the $D_{IRAP\text{-}Total}$. The correlation proved to be both positive and significant ($r = .34$; $n = 43$; $p = .024$).

Discussion

The current study employed the known-groups approach to assess the validity of the IRAP as a potential measure of self-evaluative verbal relations underlying the self-esteem construct. Preliminary statistical analyses indicated that the IRAP effects (the difference between consistent and inconsistent trials as indexed by $D_{IRAP\text{-}POS}$, $D_{IRAP\text{-}NEG}$, and $D_{IRAP\text{-}Total}$) were unaffected by order of testing (consistent-first vs. inconsistent-first). Critically, the IRAP effects were significantly different across the three participant groups, with the main block prisoners producing significantly lower D_{IRAP} scores than either the open area prisoners or the undergraduates. In fact, the open area prisoners produced marginally higher D_{IRAP} scores than the undergraduates. Interestingly, concordant known-group effects were observed for the feeling-thermometer scores.

The current results suggest that the freedoms, privileges, and artificially positive social comparisons *available* to the open area prisoners (relative to their main block counterparts) served to increase their levels of self-esteem (e.g., Blatier, 2000). On balance, of course, the direction of causality cannot be determined within the current cross-sectional design; perhaps prisoners with high self-esteem were more likely to be selected for the open area. Indeed, similar ambiguities are frequent in the self-esteem literature, because self-esteem is often observed to both predict and be predicted by the same clinically relevant outcome variables (cf. Marsh & Craven, 2006). In any case, our data tally with the descriptive findings in the explicit self-esteem literature: The undergraduates and lower security prisoners with trustee status tended to have higher self-esteem than the prisoners from mainstream prisoner populations (Blatier, 2000; Gullone et al., 2000). In particular, the difference in both the implicit and self-report measures of self-esteem between the main block and open area prisoners is remarkable, given that the study was conducted with relatively small samples from a single prison. Overall, therefore, the IRAP data serve to support the validity of the measure.

Interestingly, all three groups produced significantly larger IRAP effects on trials involving positive target words ($D_{\text{IRAP-POS}}$) relative to those involving negative target words ($D_{\text{IRAP-NEG}}$); that is, participants found it more difficult to respond negatively about themselves using positive rather than negative descriptors. A possible explanation for this effect is that individuals may find it easier to evaluate self in terms of how they *are* rather than in terms of how they *are not*; indeed, the overall IRAP effect, $D_{\text{IRAP-Total}}$, was significantly positive for all three groups, and thus each would be operating from a self-positive perspective. Certainly, the literature on reasoning and problem solving has provided strong evidence for what is called *confirmation bias* (cf. Nickerson, 1998), and thus participants may have found it easier to respond to positive than to negative self-descriptors. If this was the case, the reduced IRAP-effect observed for the negative target words likely resulted from the reduced “automaticity” involved in responding to more difficult trials (cf. Moors & De Houwer, 2006). But of course, this interpretation remains open to further empirical inquiry.

We refrained from making specific predictions concerning the absolute levels of the prisoners’ implicit self-esteem due to the limited research in this area (Yamaguchi et al., 2007). Nevertheless, the positive bias observed in each group’s $D_{\text{IRAP-Total}}$ score is clearly compatible with the general tendency of implicit and explicit self-esteem measures to be positively biased even in clinical samples such as depressed individuals (Franck, De Raedt, Dereu, & Van den Abbeele, 2007). On balance, it is interesting that the open area prisoners, similar to the undergraduates, demonstrated a positive bias in their $D_{\text{IRAP-POS}}$ and $D_{\text{IRAP-NEG}}$ scores, but the main block prisoners showed a clear divergence across these two measures (significantly positive for $D_{\text{IRAP-POS}}$ with virtually no effect for $D_{\text{IRAP-NEG}}$). On the one hand, this finding suggests that, unlike the other two groups, the main block prisoners responded with roughly equal speed when denying as well as affirming negative statements about self. On the other hand, all three groups showed a reduction in $D_{\text{IRAP-NEG}}$ relative to $D_{\text{IRAP-POS}}$, and thus the near-zero effect for the main block prisoners may simply reflect an overall reduction in IRAP effects resulting from the confirmation bias effect mentioned above. In other words, it remains to be determined if the near-zero $D_{\text{IRAP-NEG}}$ measure for the main block prisoners has important theoretical and perhaps applied implications beyond the reduction observed for the other two groups. In any case, the fact that this issue arises with the IRAP serves to illustrate that it may well provide a level of analytic precision that could be very useful in the study of implicit self-esteem.

The correlational analyses between the feeling-thermometer and the D_{IRAP} scores yielded correlations *within* the higher range of those obtained in self-esteem IAT research (e.g., Bosson et al., 2000, p. 638; Franck, De Raedt, Dereu, & Van den Abbeele, 2007, p. 78; Greenwald & Farnham, 2000, p. 1026). In this context, it is worth noting that the current study employed a more diverse sample in terms of predicted levels of self-esteem than is typical in the IAT literature, with its focus on undergraduate populations. One study, however, produced a relatively broad range of self-esteem via a mood-induction procedure designed to manipulate self-esteem, and higher-range IAT-explicit correlations were found (Glen & Banse, 2004, p. 145). In any case, the fact that the IRAP, like the IAT, correlates with the feeling-thermometer measure of self-esteem provides some support for the convergent validity of the IRAP.

One possible criticism of the IRAP is that the response latencies are relatively long and this undermines the claim that it is tapping into an implicit or automatic process. Our assumption, however, is that the critical process does not occur until the end of the trial, when the relevant stimulus relations have been understood. In effect, the difference in response latencies between consistent and inconsistent trials occurs in the act of making the choice between the two response alternatives and not during the initial processing of the information presented within the trial itself (see Barnes-Holmes et al., 2006; Barnes-Holmes, Hayden, et al., 2008). In any case, one of the defining features of an implicit measure is the participant's lack of control over the measurement outcome (e.g., when participants are asked to fake a particular attitude), and recent evidence indicates that the IRAP does indeed possess this feature (McKenna et al., 2007). Furthermore, the robustness of IRAP responses against participants' manipulations has been further supported in ongoing research addressing other attitudinal domains, such as attitudes toward smoking (Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2008). Of course, the controllability of the IRAP effect remains to be investigated in the specific domain of self-esteem, as do more general issues pertaining to the validity and reliability of the IRAP as a tool for the assessment of clinically relevant processes. The current findings do suggest, however, that such research may well be worthwhile.

References

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

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

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

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

BARNES-HOLMES, D., WALDRON, D., BARNES-HOLMES, Y., & STEWART, I. (in press). Testing the validity of the Implicit Relational Assessment Procedure (IRAP) and the Implicit Association Test (IAT): Measuring attitudes towards Dublin and country life in Ireland. *The Psychological Record.*

BLATIER, C. (2000). Locus of control, causal attributions, and self-esteem: A comparison between prisoners. *International Journal of Offender Therapy and Comparative Criminology, 44*, 97–110.

BOSSON, J. K., SWANN, W. B., & PENNEBAKER, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631–643.

BOYSEN, G. A., VOGEL, D. L., & MADON, S. (2006). A public versus private administration of the Implicit Association Test. *European Journal of Social Psychology, 36*, 845–856.

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

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. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Erlbaum.

DE HOUWER, J. (2008). Comparing measures of attitudes at the procedural and functional level. In R. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures*. Mahwah, NJ: Erlbaum.

FESTINGER, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

FRANCK, E., DE RAEDT, R., & DE HOUWER, J. (2007). Implicit but not explicit self-esteem predicts future depressive symptomatology. *Behavior Research and Therapy, 45*, 2448–2455.

FRANCK, E., DE RAEDT, R., DEREU, M., & VAN DEN ABBEELE, D. (2007). Implicit and explicit self-esteem in currently depressed individuals with and without suicidal ideation. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 75–85.

FURST, G. (2006). Prison-based animal programs: A national survey. *The Prison Journal, 86*, 407–430.

GLEN, I. S., & BANSE, R. (2004). Probing the malleability of implicit and explicit self-esteem: An interview approach. *Current Psychology of Cognition, 22*, 133–158.

GREENWALD, A. G., & BANAJI, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

GREENWALD, A. G., & FARNHAM, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79,* 1022–1038.

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.

GULLONE, E., JONES, T., & CUMMINS, R. (2000). Coping styles and prison experience as predictors of psychological well-being in male prisoners. *Psychiatry, Psychology, and Law, 7*, 170–181.

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

JACQUES, J. M., & CHASON, K. J. (1977). Self-esteem and low status groups: A changing scene? *The Sociological Quarterly, 18,* 399–412.

JAMIESON, R., & GROUNDS, A. (2002). *No sense of an ending: The effects of long-term imprisonment amongst republican prisoners and their families.* Northern Ireland: SEESYU Press Ltd.

JUDGE, T. A., EREZ, A., BONO, E. A., & THORESEN, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology, 83*, 693–710.

KALICHMAN, S. C. (1991). Psychopathology and personality characteristics of criminal sexual offenders as a function of victim age. *Archives of Sexual Behavior, 20*, 187–197.

KARPINSKI, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin, 30,* 22–34.

KARPINSKI, A., & STEINMAN, R. B. (2006). The single category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*, 16–32.

MARSH, H. W., & CRAVEN, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163.

MARSHALL, W. L., MARSHALL, L. E., SACHDEV, S., & KRUGER, R. (2003). Distorted attitudes and perceptions, and their relationship with self-esteem and coping in child molesters. *Sexual Abuse: A Journal of Research and Treatment, 15*, 171–181.

MARTINOT, D., & REDERSDORFF, S. (2006). The variable impact of upward and downward *social comparisons* on self-esteem: When the level of analysis matters. In S. Guimond (Ed.), *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 127–150). New York: Cambridge University Press.

MCKENNA, I. M., BARNES-HOLMES, D., BARNES-HOLMES, Y., & STEWART, I. (2007). Testing the fake-ability of the Implicit Relational Assessment Procedure (IRAP): The first study. *The International Journal of Psychology and Psychological Therapy, 7*, 253–268.

MICHINOV, N. (2001). When downward comparion produces negative affect: The sense of control as moderator. *Social Behavior and Personality, 29,* 427–444.

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

MITCHELL, B. (Ed.). (2003). *Conflict transformation papers volume 4: Ex-prisoners and conflict transformation.* Belfast: LINC Resource Centre.

MOORS, A., & DE HOUWER, J. (2006). Automaticity: A conceptual and theoretical analysis. *Psychological Bulletin, 132*, 297–326.

MUSSWEILER, T. (2001). 'Seek and ye shall find': Antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31*, 499–509.

MUSSWEILER, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110*, 472–489.

NICKERSON, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–180.

NOSEK, B. A., GREENWALD, A. G., & BANAJI, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Philadelphia, PA: Psychology Press.

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

OSER, C. B. (2006). The criminal offending–self-esteem nexus: Which version of the self-esteem theory is supported? *The Prison Journal, 86,* 344–363.

O'TOOLE, C., & BARNES-HOLMES, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligence test: The importance of relational flexibility. *The Psychological Record, 59,* 119–132.

REGENS, J. L., & HOBSON, W. G. (1978). Inmate self-government and attitude change. *Evaluation Review, 2*, 455–479.

SHINE, J., MCCLOSKEY, H., & NEWTON, M. (2002). Self-esteem and sex offending. *Journal of Sexual Aggression, 8*, 51–61.

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.

STILES, B. L., & KAPLAN, H. B. (2004). Adverse social comparison processes and negative self-feelings: A test of alternative models. *Social Behavior and Personality, 32*, 31–44.

VAHEY, N. A., BARNES-HOLMES, D., BARNES-HOLMES, Y., & STEWART, I. (2008, May). *Implicit Relational Assessment Procedure (IRAP): Examining the context-dependent nature of smoking-related implicit attitudes*. Paper presented at the 34th annual meeting of the Association for Behavior Analysis International, Chicago.

WIERS, R. W., TEACHMAN, B. A., & DE HOUWER, J. (2007). Implicit cognitive processes and psychopathology: An introduction. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 95–104.

YAMAGUCHI, S., GREENWALD, A. G., BANAJI, M. R., MURAKAMI, F., CHEN, D., SHIOMURA, K., ET AL. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science, 18, 498–500.*

Appendix

Abridged Version of the Written Instructions and Illustrated Laminates

During the experiment you will be asked to respond quickly *and* accurately on a computer task. To make the correct responses for the task, sometimes you will be required to respond in a way that *agrees* with what you believe and at other times you will be required to respond in a way that *disagrees* with what you believe. Crucially, this means you must learn how to provide the answers required by the computer program and NOT your own personal opinions. **This is part of the experiment.**

Note that when an incorrect response for a task is made, it is signalled by the appearance of a red "X" on the center of the screen. To remove the red "X" and continue, please make the correct response.

IMPORTANT: Please note that in the computer task that follows, YOUR first name will be on the screen in place of the name Nigel. Please bear this in mind while reading the following instructions. Also, from task to task, the positioning of the response options will vary randomly between left and right.

It is intended that you interpret the words displayed in these pictures in the following way:

Similar

NICE

Select 'd' for	Select 'k' for
Nigel	**Not Nigel**

If you make a response of "Nigel" by pressing the 'd' key, you are stating that "Nigel" is "Similar to NICE".

If you make a response of "Not Nigel" by pressing the 'k' key, you are answering "Not Nigel" is "Similar to NICE". In other words, you are saying that "Nigel is not nice".

Opposite

NICE

Select 'd' for	Select 'k' for
Nigel	**Not Nigel**

If you make a response of "Not Nigel" by pressing the 'k' key, you are answering "Not Nigel" is "Opposite to NICE". In other words, you are saying that "Nigel is nice".

If you make a response of "Nigel" by pressing the 'd' key, you are in other words stating that "Nigel" is "Nigel is not nice".

Similar

ROTTEN

Select 'd' for	Select 'k' for
Not Nigel	**Nigel**

If you make a response of "Not Nigel" by pressing the 'd' key, you are answering "Not Nigel" is "Similar to ROTTEN". In other words, you are saying that "Nigel is not rotten".

If you make a response of "Nigel" by pressing the 'k' key, you are stating that "Nigel" is "Similar to ROTTEN".

Opposite

ROTTEN

Select 'd' for	Select 'k' for
Not Nigel	**Nigel**

If you make a response of "Not Nigel" by pressing the 'd' key, you are answering "Not Nigel" is "Opposite to ROTTEN". In other words, you are saying that "Nigel is rotten".

If you make a response of "Nigel" by pressing the 'k' key, you are stating in other words that "Nigel" is "not rotten".