Testing the Fake-ability of the Implicit Relational Assessment Procedure (IRAP): The First Study

Ian M. McKenna*[1], Dermot Barnes-Holmes[1], Yvonne Barnes-Holmes[1], and Ian Stewart[2]

[1] *National University of Ireland, Maynooth, Ireland*

[2] *National University of Ireland, Galway, Ireland*

Abstract

This was the first study that aimed to examine the effects of instructing participants to "fake" their performance on the Implicit Relational Assessment Procedure (IRAP). Thirty-six participants were first exposed to an IRAP. Consistent blocks involved responding to pleasant target words as pleasant, and unpleasant words as unpleasant; inconsistent blocks involved the opposite response pattern. As predicted, latencies were significantly shorter on consistent relative to inconsistent blocks. Subsequently, all participants were informed about how the IRAP works; 12 participants were also asked to try to fake the next IRAP by thinking of pleasant things as unpleasant and unpleasant things as pleasant; and another 12 were also told how to fake the IRAP (slow down on consistent trials). Results showed no evidence of faking, which contrasts with previous research in which the Implicit Association Test (IAT) was successfully faked.

Keywords: Implicit relations, Assessment, Faking

Resumen

Este es el primer estudio en el que se examinan los efectos de instruir a los participantes para fingir o simular su actuación en el Procedimiento de Evaluación de Relaciones Implícitas (IRAP). Inicialmente, 36 participantes fueron expuestos a una tarea de IRAP. En los bloques de relaciones consistentes los participantes debían responder a las palabras-objetivo agradables como si fuesen agradables, y a las palabras-objetivo desagradables como si fuesen desagradables. En los bloques de relaciones inconsistentes, debían responder según el patrón opuesto. Las latencias fueron significativamente menores en los bloques de relaciones consistentes, de acuerdo con lo esperado. Posteriormente, todos los participantes fueron informados acerca del funcionamiento del IRAP. A doce de ellos se les pidió, además, que intentasen fingir su siguiente actuación. Se les instruyó para que pensasen que las palabras agradables eran desagradables y las desagradables agradables. Otros doce fueron instruidos cómo fingir el IRAP aumentando voluntariamente su latencia en los bloques de relaciones consistentes. Los resultados muestran que en ningún grupo fue posible para los participantes fingir su actuación en el IRAP, lo que contrasta con la investigación previa con otros procedimientos como el Test de Asociaciones Implícitas (IAT), para los que sí se ha logrado fingir la actuación en la tarea.

Palabras clave: relaciones implícitas, fingimiento.

*Preparation of the current article was supported by postgraduate scholarships awarded to Ian McKenna from the Irish Research Council for Science, Engineering and Technology, and from NUI, Maynooth (John and Pat Hume Postgraduate Scholarships Scheme). Address correspondence to either Ian McKenna, Department of Psychology, National University of Ireland, Maynooth, Ireland (Email: Ian.M.McKenna @nuim.ie), or Dermot Barnes-Holmes, Department of Psychology, National University of Ireland, Maynooth, Ireland (Email: Dermot.Barnes-Holmes@nuim.ie).

The Implicit Association Test (IAT) was designed to tap into so called implicit attitudes based on a rather simple associative assumption; responses should be faster when two closely associated concepts are assigned to the same key, than when those concepts are assigned to different keys. In a seminal study, Greenwald, McGhee, and Schwartz (1998, Experiment 1) demonstrated that participants responded faster when pleasant items and flowers were categorized together, and unpleasant items and insects were categorized together, than when these categorizations were reversed (pleasant-insects and unpleasant-flowers). This basic IAT effect has been replicated numerous times (e.g., de Jong, 2002; de Jong, Pasman, Kindt, & van den Hout, 2001; Gemar, Segal, Segratti, & Kennedy, 2001; Teachman, Gregg, & Woody, 2001), and as such the IAT has become a very popular assessment tool for measuring implicit attitudes and dysfunctional beliefs in social and clinical psychological research.

One of the possible strengths of the IAT, it has been argued, is that it may be less sensitive to self-presentational biases or deliberate attempts to conceal socially-sensitive attitudes than questionnaires and other explicit measures (see Nosek, Greenwald, & Banaji, in press). A number of studies have tested this assumption directly (e.g., Fiedler & Bluemke, 2005; Kim, 2003). The study by Kim (2003), for example, investigated the voluntary controllability of the IAT by examining participants' ability to fake their attitudes towards strongly valenced word categories. Across two experiments, three different faking conditions were employed; No Faking/Control, Faking/No Strategy, and Faking/Strategy. Participants in each of the groups completed two IAT exposures in short succession. After the first exposure, participants in the No Faking/Control group received instructions explaining how the IAT operated before completing the second exposure. Participants in the Faking/No Strategy condition were provided with the same instructions but also were instructed to try to conceal their attitudes towards positive and negative stimuli. The Faking/Strategy group received the same instructions as the No Strategy group but were also explicitly told how to fake their responses (by speeding up on some tasks and deliberately slowing down on others). The results from Kim's study indicated that neither the No Faking/Control nor the Faking/No Strategy conditions impacted significantly upon the participants' IAT performances; however, the Faking/Strategy condition clearly reversed the IAT effect. In other words, participants' could control their IAT performance when they were given explicit instructions on how to fake the procedure (but see Fiedler & Bluemke, 2005).

Very recently, another procedure for assessing implicit cognitions has been proposed, the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Power, Hayden, Milne, & Stewart, 2006). The IRAP was developed from a modern behavioral theory of human language and cognition, Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). According to this account, the core elements of human higher cognition are relational acts or processes, not mere associations. The IRAP is a latency-based response measure used to assess previously established stimulus or verbal relations between sample and target stimuli by presenting relational response options, such as *Similar* and *Opposite*, or *Better* and *Worse* on a computer-based task. Participants are asked to respond as quickly and accurately as possible across trials that

are deemed relationally consistent or inconsistent with currently held beliefs. For example, during a consistent IRAP block, participants may be required to respond to pleasant target words as pleasant and unpleasant target words as unpleasant, but during an inconsistent block the opposite response pattern is required (pleasant–unpleasant and unpleasant–pleasant). The core IRAP prediction is that shorter average response latencies should be observed for consistent relative to inconsistent blocks. Recent empirical evidence has supported this prediction (Barnes-Holmes et al, 2006).

The theoretical basis for the predicted IRAP effect is as follows. According to RFT, contextual cues are presented on each IRAP trial that specify particular relational (e.g. same, opposite, more/less) and functional dimensions (e.g. pleasant/unpleasant, likeability). Given these cues, a participant will likely emit an incipient or private relational response before he or she actually presses the appropriate key. As an operant theory of human language, RFT assumes that the probability or strength of the private response will be determined by the verbal and non-verbal history of the participant and current setting factors. By definition the most probable response will be emitted first most frequently, and thus during a consistent block of a properly designed IRAP the first private response will tend to possess the correct key-pressing function. In contrast, that response on the same trial during an inconsistent block will tend to possess the wrong function. In the second case, therefore, additional private responding will be required to match the correct function. In other words, during consistent trials participants can simply do the first thing that comes to mind, but during inconsistent trials they must "work against" this tendency before pressing a response key. The need to "self-correct" before pressing the correct key during inconsistent trials requires time, and thus across multiple trials the average latency for inconsistent blocks will be longer than for consistent blocks.

From an RFT perspective, therefore, the IRAP effect is based on immediate and often private relational responding, which is difficult to "hide" when the behavioural system is put under pressure to respond quickly and accurately. In making this argument, however, it is important to determine empirically if participants do indeed find it difficult to hide high-probability relational responding on the IRAP. Or, in other words, how difficult is it to fake an IRAP performance? The purpose of the current study was to explore the extent to which the IRAP may be faked. In summary, the research involved replicating Kim's (2003) three faking conditions, outlined previously, using the IRAP instead of the IAT.

Three groups of participants were presented with the same stimuli across two IRAP exposures. After the first IRAP, one group of participants received information about the IRAP and how it worked, and then completed a second IRAP. The second group received similar information but were instructed to try to fake the next IRAP, but were not provided with a strategy for doing so. The third group received similar information to the previous group, but were explicitly told how to fake the IRAP by deliberately slowing down on the consistent trials and going faster on the inconsistent trials. Would the IRAP prove to be as difficult to fake as the IAT, with a reversed effect on the second exposure emerging only in the third condition in which participants were explicitly told how to fake the IRAP?

METHOD

Participants

Thirty-six participants, 18 male and 18 female aged 18 to 30 years (M= 23 years) completed the study. All participants were undergraduates at the National University of Ireland, Maynooth. The participants were randomly assigned to one of three experimental groups, with 6 males and 6 females in each group. None of the participants had previous exposure to the IRAP.

Materials and Stimuli

The experiment was conducted in an experimental laboratory in the Department of Psychology at the National University of Ireland, Maynooth. Before performing the computer-administered IRAP task, participants responded to a questionnaire consisting of two self-report attitude measures: a feeling thermometer and a Word Pleasantness Rating Scale. Each measure was used to assess participants' evaluations of each of the 12 target words that were to be used in the IRAP. For the former measure, participants placed a mark on each of 12 thermometers, which were labeled at the bottom, middle, and top with "0 degrees (cold, or unfavorable)," "50 degrees (neutral)," "99 degrees (warm and favorable)," respectively (Robinson, 1974). The latter measure was a 12-item questionnaire. For each item, participants were asked to place a mark on a seven-point likert scale labeled on the left, middle and right with "–3 (Very Unpleasant)," "0 (Neutral)," and "3 (Very Pleasant)".

A *Faking Strategy Questionnaire* was employed with those participants who were asked to fake an IRAP performance. The first item on this questionnaire was designed to assess a participant's understanding of the experimental instructions ("Write down on the page below a description of what is being asked of you in the next part of the experiment"). The final two items on the Faking Strategy Questionnaire were completed by participants immediately after completing the second IRAP exposure. The second item asked participants to "Please report and describe on the page below the strategies you used in the second part of the experiment to respond as you think a person would who likes Unpleasant words more than Pleasant words?" The third item asked "Do you think that the strategies you applied were successful at allowing you to respond as you think a person would who likes Unpleasant words more than Pleasant words? (Yes, No, or Other?) Indicate and explain your answer on the page below".

The stimuli used in the IRAP procedure were 12 target words. Six were defined as 'pleasant' (caress, freedom, health, love, peace, and cheer) and six were defined as 'unpleasant' (abuse, crash, filth, murder, sickness, and accident) based on Greenwald et al's (1998) consistent and inconsistent categorisation of pleasant and unpleasant terms. Two sample words 'Pleasant' and 'Unpleasant', and two relational terms 'Similar' and 'Opposite' were also used in the IRAP.

Design

The experiment involved a combined assignment mixed 3x2x2 factorial design. The between-participant independent variable was the strategy that participants were instructed to use and was operationalised on three levels: No Faking/Control; Faking/No Strategy; and Faking/Strategy Instructions. The first within-participant independent variable was IRAP exposure (1st versus 2nd), and the second variable was IRAP condition (consistent versus inconsistent tests). The main dependent variable was response latency on each trial, defined as time in milliseconds from the point of stimulus onset to the first correct response.

Procedure

Experimental sequence and participant assignment. Twelve participants each were assigned randomly, balancing for gender, to one of the three groups; No Faking/Control, Faking/No Strategy, Faking/Strategy. Within each of the three groups, 6 participants were randomly assigned to either a consistent- or inconsistent-relations-first IRAP (explained below). This assignment was reversed for the second exposure to the IRAP.

Phase 1: Self-report questionnaire measures. Each participant was asked to complete the feeling thermometer and the word pleasantness rating scale.

Phase 2: IRAP, first exposure. After completing the questionnaires, each participant sat in front of the computer, which presented a consent form, a brief description of the procedures and instructions for completing the IRAP, and then the IRAP itself. The instructions for IRAP, exposure 1 simply described the procedure and did not identify trials as consistent or inconsistent nor did it articulate any specific predictions concerning the participants' performances (all *verbatim* instructions are available from the first author upon request).

For each trial of the IRAP, four words were presented on the computer screen simultaneously. The sample stimulus, either 'pleasant' or 'unpleasant', appeared at the top, the target word appeared in the center, and the two relational words, 'similar' and 'opposite' appeared at the bottom left and right corners (see Figure 1). All four stimuli remained on screen until the participant chose one of the two relational terms at the bottom ('similar' or 'opposite') by pressing one of the two response keys. Participants chose the term on the left by pressing the 'd' key with their left index finger or the term on the right by pressing the 'k' key with their right index finger. Participants were asked to rest their left and right index fingers on the 'd' and 'k' keys respectively, for the duration of each block of trials. The left-right position of the relational terms alternated randomly across trials.

If the participants emitted a correct response for a particular trial, all four stimuli were removed from the screen for a 400 ms inter-trial interval before the next trial was displayed. If a participant emitted an incorrect response (or pressed any other key apart from 'd' and 'k') a red 'X' was presented directly under the target word. The X remained on screen until the correct response was emitted. Only when a participant emitted the

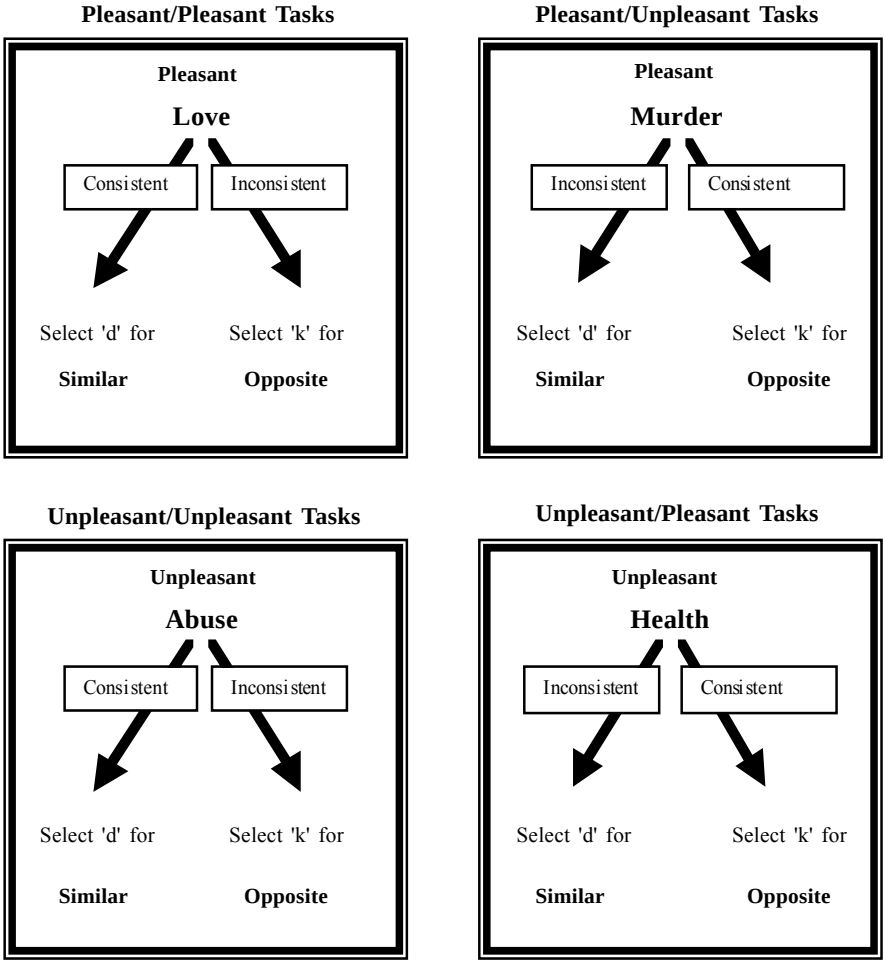

Figure 1. Examples of the four IRAP trial-types. The sample (Pleasant or Unpleasant), target words (Caress, Freedom, Accident, etc.), and response options were presented on screen at the same time. Note, the superimposed arrows and text boxes used to illustrate which responses were deemed consistent and inconsistent did not appear on screen during the IRAP.

correct response was the X and all other stimuli removed. After 400ms the next trial was presented.

The IRAP consisted of two practice blocks and six test blocks, each containing 24 trials. During each block, the 12 target words were displayed in a quasi-random sequence, with each word presented twice, once with each sample (see Table 1). The first block of 24 trials in the consistent-relations-first condition required participants to emit responses that were predicted to be relationally consistent with their previously learned verbal history. For example, if the sample word ‘pleasant’ and any of the

pleasant target words ('caress', etc.) appeared on screen a correct response was defined as choosing the relational term 'similar'. Choosing the other relational term, 'opposite,' on this trial was defined as an incorrect response. Alternatively, if the sample word was 'pleasant' and the target word unpleasant ('murder', etc.), 'opposite' was 'correct' and 'similar' incorrect. After participants completed the 24 trials they were presented with feedback indicating the percentage of correct responses and the median response time (calculated across the 24 trials). Subsequently, on-screen instructions informed the participant that during the next block of trials all of the previously correct and wrong answers will be reversed.

The second block of 24 trials in the consistent-relations-first condition required participants to emit responses that were relationally inconsistent with their previously learned verbal history. For example, given the sample 'pleasant' and a pleasant target word ('love', etc.), 'opposite' was correct and 'similar' incorrect; but if the target word was unpleasant ('abuse', etc.), 'similar' was correct and 'opposite' incorrect. After completing all 24 trials the feedback indicating the percentage of correct responses and the median response time was presented, followed by an instruction that the previously correct and wrong answers will be reversed in the next block. In addition, in preparation for the first block of test trials, the instructions informed participants that the next block would be a test and they were asked to respond quickly.

The third, fifth, and seventh blocks in the consistent-relations-first condition required participants to emit responses that were predicted to be relationally consistent with their previously learned verbal history; the fourth, sixth, and eighth blocks required the opposite response pattern. Between each of the test blocks participants were informed of the percentage of correct responses and the median response time for that test block; they were also informed before each test block that the previously correct and incorrect answers would be reversed in the next block. After completing the eighth and final block, the screen cleared and a message appeared indicating that part of the experiment was over and the participant should report to the experimenter.

The procedure for participants in the inconsistent-relations-first condition was identical to the consistent-relations-first sequence, except that the order of practice and test blocks was reversed (i.e., Block 1= Inconsistent-Practice; Block 2= Consistent-Practice; Block 3= Inconsistent-Test; Block 4= Consistent-Test, and so on).

Table 1. IRAP Stimuli

Sample 1: Pleasant Response option 1: Similar	Sample 2: Unpleasant Response option 2: Opposite
Target words rated as consistent with Sample 1:	Target words rated as consistent with Sample 2:
Caress	Abuse
Freedom	Crash
Health	Filth
Love	Murder
Peace	Sickness
Cheer	Accident

Phase 3: Faking instructions. Immediately after completing IRAP exposure 1, participants in each of the three groups were presented with a typed message on a sheet of paper, which read as follows:

> The computer task you have just completed shows that we find it easier to report a true relation rather than a false one. So, for example, we typically find it easier to choose "Similar" rather than "Opposite" when pleasant words are presented with the word "Pleasant", than when unpleasant words are shown. Similarly, we find it easier to choose "Opposite" rather than "Similar" when pleasant words are presented with the word "Unpleasant", than when pleasant words are shown.

The 12 participants in the No Faking/Control group proceeded immediately after reading this message to Phase 4. The 24 participants in the two faking groups (Faking/No Strategy and Faking/Strategy) were instructed to read another message. The message was as follows:

> Your task in the next part of the experiment is to respond as if you find all the pleasant words unpleasant, and all the unpleasant words pleasant. For example, please try to imagine that you find the word "murder" pleasant and the word "love" unpleasant. It is still important for you to respond rapidly on each trial, but to avoid making errors as in the previous computer task.

Only the 12 participants in the Faking/Strategy group received the additional instructions on how to fake the IRAP, which were as follows:

> To fake your responses to pleasant and unpleasant words, you need to do two things.
> First, try to respond slowly on those tasks that ask you to choose the true response. For example, respond slowly when you have to choose "Similar" given a pleasant meaning word and "Pleasant", and respond slowly when you have to choose "Opposite" given an unpleasant meaning word and "Pleasant".
> Second, try to respond fast on those tasks that ask you to choose the false response. For example, respond fast when you have to choose "Similar" given a pleasant meaning word and "Unpleasant", and respond fast when you have to choose "Opposite" given an unpleasant meaning word and "Unpleasant".

After reading the instructions, the experimenter handed the participants in both the Faking/No Strategy and Faking/Strategy groups a sheet of paper with Faking Strategy Questionnaire item 1 on it, which read "Write down on the page below a description of what is being asked of you in the next part of the experiment?" If participants were unable to write a complete and accurate description, they were provided with additional time to re-read the instructions, and were only permitted to continue with the experiment after writing the appropriate description.

Phase 4: IRAP, second exposure. All three groups of participants completed the second IRAP exposure. The second IRAP was the same as the first IRAP except that the order in which the IRAP blocks were presented was reversed within participants (i.e., if a participant received a consistent-relations-first IRAP in Exposure 1, an inconsistent-relations-first IRAP was presented in Exposure 2, and vice versa).

Phase 5: Completing the Faking Strategy Questionnaire. Following the second IRAP exposure participants in the Faking/No Strategy and Faking/Strategy groups completed items two and three of the *Faking Strategy Questionnaire* (participants in the No Faking/Control group were not asked to complete these items and were simply thanked and debriefed).

The two items read as follows:

> 2) Please report and describe on the page below the strategies you used in the second part of the experiment to respond as you think a person would who likes Unpleasant words more than Pleasant words.
> 3) Do you think that the strategies you applied were successful at allowing you to respond as you think a person would who likes Unpleasant words more than Pleasant words? (Yes, No, or Other?)
> Indicate and explain your answer on the page below.

Having responded to the above questions, the remaining participants were then thanked and debriefed.

RESULTS

Explicit Measures

Each participant completed 12 feeling thermometers (scored from 0 to 99) and 12 word pleasantness scales (scored from -3 to 3) for each of the 12 target words used in the IRAP. Four mean scores, two for each measure, were obtained for each participant, two calculated across the six pleasant words and another two calculated across the six unpleasant words (Thermometer, Pleasant M= 84.592, SE= 1.448; Thermometer, Unpleasant M= 11.193, SE= 1.865; Words Scale, Pleasant M= 2.015, SE= .095; Words Scale, Unpleasant M= -2.163, SE= .104). Each of the three groups rated the pleasant target words as "warmer" and "more pleasant" than the unpleasant target words. The data were subjected to two mixed 2x3 repeated measures analyses of variance (ANOVA), one for each measure, with faking strategy (No Faking/Control, Faking/No Strategy, and Faking/Strategy) as the between-participant variable and word-type (Pleasant Rating/ Unpleasant Rating) as the repeated measure. A significant main effect for word-type was observed for both the feeling thermometers [F (1, 33)= 603.585, p< .0001, η_p^2= .9481], and for the word pleasantness scales [F (1, 33)= 561.547, p= .0001, η_p^2= .9437]. Neither ANOVA yielded a significant effect for faking strategy or for the interaction (all ps> .38). In summary, the two explicit measures produced the same results for all three groups. Any differences that emerge among the groups for the IRAP measure cannot be readily attributed, therefore, to different explicit ratings of the target words.

IRAP Preliminary Analyses

The primary raw datum was response latency on each trial of the IRAP, defined as time in milliseconds (ms) from the point of stimulus onset to the first correct response

emitted by the participant. The latency data were normalized using Greenwald, Nosek, and Banaji's (2003) C4 algorithm. Specifically, all latencies longer than 3000ms were recorded as 3000ms and those shorter than 300ms were recorded as 300ms. Six mean adjusted response latencies for each of the six test blocks for each participant were calculated for each exposure to the IRAP. Table 2 presents the overall mean response latencies, for each test-block, divided in terms of test-order and exposure, for each group. Within each pair of consistent versus inconsistent test blocks, mean response latencies were shorter for the former than for the latter blocks, and this was the case for each test-order, faking group, and IRAP exposure.

Six 2x2x3 mixed repeated measures ANOVAs were conducted for each IRAP exposure for each of the three groups. The primary purpose of these preliminary analyses was to determine if the consistent versus inconsistent IRAP effect interacted with either the test-order (consistent-first versus inconsistent-first) or test-sequence (test blocks, 1, 2, and 3). If no significant interactions were obtained, the test-order and test-sequence variables would be removed from subsequent analyses. Only one of the six ANOVAs yielded a significant interaction effect with IRAP condition (No Faking/Control, Exposure

Table 2. The Mean Adjusted Latencies and Standard Errors for Each of the Consistent and Inconsistent Test Blocks for Each Test Order and Faking Strategy Group for IRAP Exposures 1 and 2.

IRAP Mean Adjusted Latencies and Standard Errors per Test Block												
IRAP Condition	Consistent First						Inconsistent First					
	Test 1		Test 2		Test 3		Test 1		Test 2		Test 3	
	M	SE	M	SE	M	SE	M	SE	M	SE	M	SE
No Faking/Control Exposure 1												
Consistent	2175	179	2141	175	2174	164	2237	116	2066	137	2205	126
Inconsistent	2387	156	2377	135	2380	141	2721	83	2535	95	2581	95
Faking/No Strategy Exposure 1												
Consistent	2069	117	2014	76	2098	89	1878	107	2000	136	1879	142
Inconsistent	2304	119	2163	88	2168	102	2239	104	2188	201	2164	154
Faking/Strategy Exposure 1												
Consistent	2211	194	2179	155	2120	172	1891	190	1926	201	1805	180
Inconsistent	2385	158	2404	173	2333	136	2100	231	2238	183	2254	204
No Faking/Control Exposure 2												
Consistent	2109	175	2005	201	2082	178	2031	87	2040	116	1957	91
Inconsistent	2245	179	2247	171	2149	194	2271	175	2276	137	2250	157
Faking/No Strategy Exposure 2												
Consistent	1867	71	1939	76	1864	52	1734	98	1724	100	1753	109
Inconsistent	2042	108	2129	109	2100	97	1898	89	1807	95	1870	82
Faking/Strategy Exposure 2												
Consistent	2079	161	2078	191	1956	139	1906	128	1886	89	1738	121
Inconsistent	2273	138	2221	115	2266	140	2121	126	2067	151	2020	174

1), and this was with test-order [F (1, 10)= 9.649, p= .0111, η_p^2= .4911]. The difference between consistent and inconsistent IRAP conditions was in the same direction for both test orders (i.e., faster response latencies for consistent versus inconsistent trial-types), and two additional post-hoc ANOVAs indicated that these differences were both significant; consistent-first [F (1, 5)= 17.929, p= .0082, η_p^2= .7819] and inconsistent-first [F (1, 5) = 74.644, p= .0003, η_p^2= .9377]. At this point, it was decided to remove test-order and test-sequence variables from subsequent analyses because; (i) only one of the six ANOVAs yielded a significant interaction effect with IRAP condition (a Bonferroni correction for six ANOVAs would require p< .0083), (ii) the difference between IRAP conditions was in the same direction and both were significant, and (iii) this interaction effect had not been observed in any other IRAP experiment conducted in our laboratory.

The IRAP Effect and Faking Instructions

The overall mean adjusted response latencies calculated for consistent and inconsistent trials for each group and for each exposure to the IRAP are presented in Figure 2, confirming the predicted IRAP effect in each case. The figure also indicates that response latencies for both consistent and inconsistent trials were shorter for the second relative to the first IRAP exposure (i.e., a standard practice effect). The six 2x2x3 ANOVAs conducted for each group and for each exposure indicated a significant main effect for consistent versus inconsistent IRAP trials (see Table 3). The fact that a significant IRAP effect was observed across all six ANOVAs suggests that the faking instructions did not impact substantively upon the IRAP performances.

To explore further any possible effects of faking instructions on the IRAP, a difference score was calculated for each participant. The mean latency for consistent trials was subtracted from inconsistent trials, and thus a positive score was indicative of an IRAP effect. A 2x3 mixed repeated measures ANOVA was applied to the difference scores with faking instructions as the independent variable and exposure as the repeated measure. The ANOVA yielded a significant main effect for exposure [F (1, 33)= 4.684, p= .0378, η_p^2= .1229], but the effect for faking instruction was non-significant [F (2, 33)= 1.126, p= .3364], as was the interaction [F (2, 33)= .597, p= .5564]. Once again, the statistical analyses indicated that faking instructions had no significant impact on IRAP performance, although the IRAP effect did reduce from the first to second exposure.

Table 3. Results From the Six ANOVAs Comparing Consistent with Inconsistent IRAP Trial-Types for Each Group for Each IRAP Exposure.

Source	*df*	*F*	*p*	h_p^2
No Faking/Control Exposure 1	1,10	82.814	.0001	.8923
No Faking/Control Exposure 2	1,10	29.128	.0003	.7444
Faking/No Strategy Exposure 1	1,10	53.021	.0001	.8413
Faking/No Strategy Exposure 2	1,10	7.437	.0213	.4265
Faking/Strategy Exposure 1	1,10	24.606	.0006	.7110
Faking/Strategy Exposure 2	1,10	15.844	.0026	.6131

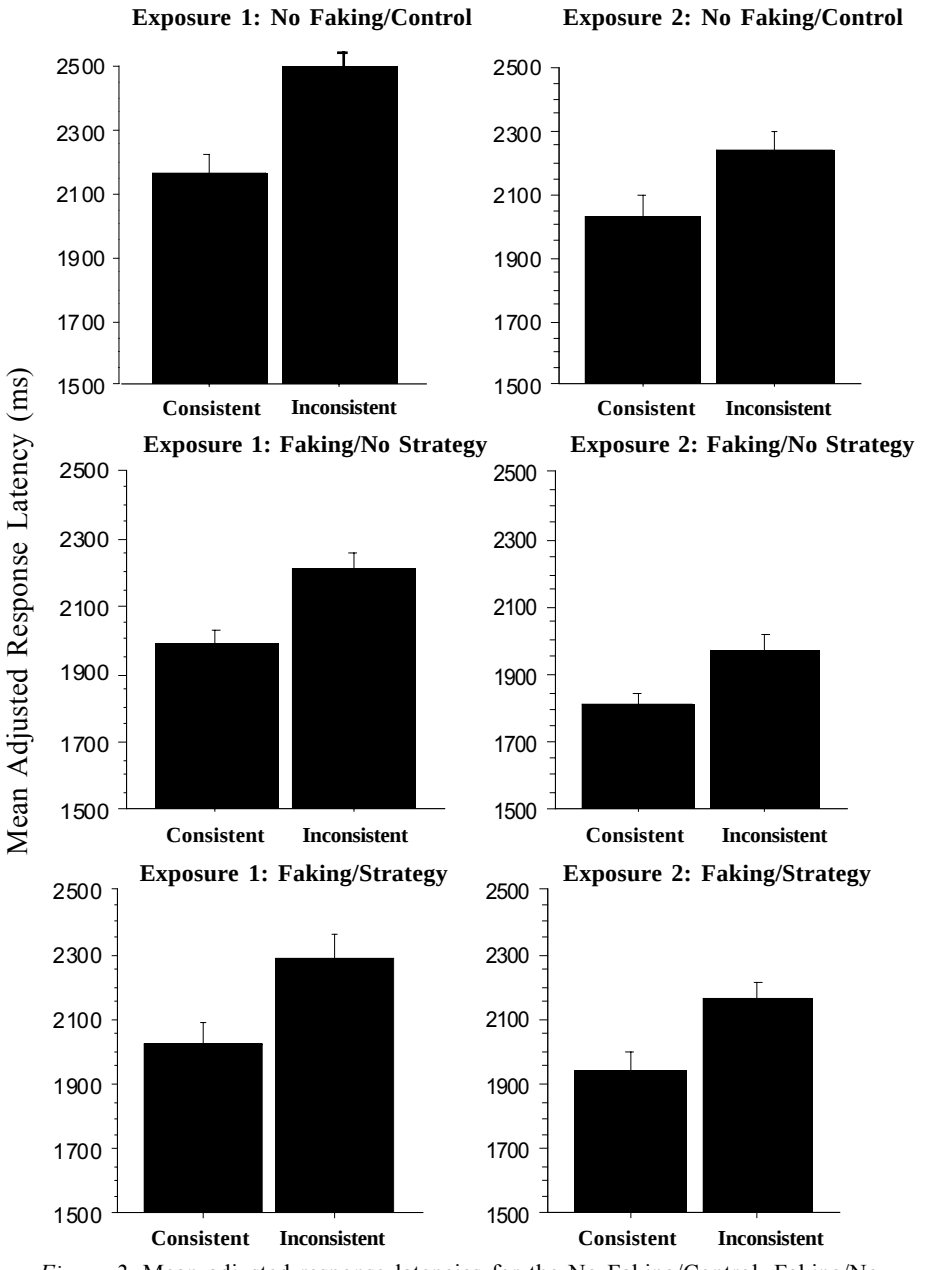

Figure 2. Mean adjusted response latencies for the No Faking/Control, Faking/No Strategy and Faking/Strategy groups for consistent and inconsistent trials for IRAP exposures one and two.

Faking Strategy Questionnaire

The results of the *Faking Strategy Questionnaire* are presented in Table 4. All 12 participants in the Faking/No Strategy group reported using the faking strategy, and 7 of the 12 participants reported that they believed they were successful at faking their responses. Inspection of the data for individual participants (not presented) indicated

Table 4. Faking Strategy Questionnaire Results for the Faking/No Strategy and Faking Strategy Groups.

Participant Number	Understood what was requested of them	Did they use the strategy provided	Believed they were successful	Faked the IRAP
		Faking/No Strategy		
13	Yes	Yes	No	No
14	Yes	Yes	No	No
15	Yes	Yes	Yes	No
16	Yes	Yes	Yes	No
17	Yes	Yes	Yes	No
18	Yes	Yes	Yes	No
19	Yes	Yes	Yes	No
20	Yes	Yes	Yes	Yes
21	Yes	Yes	Yes	No
22	Yes	Yes	No	No
23	Yes	Yes	No	No
24	Yes	Yes	No	No
		Faking/Strategy		
25	Yes	Other (thought of pleasant words as unpleasant and vice versa)	Yes	No
26	Yes	Other (read carefully then made decision)	Not Sure	No
27	Yes	Other (associating emotions with responses)	Not Sure	Yes
28	Yes	Yes	Yes	No
29	Yes	Other (2 x pleasant words or 2 x unpleasant = similar all other = opposite)	Yes	No
30	Yes	Other (thought of pleasant words as unpleasant and vice versa)	No	No
31	Yes	Other (reversed correct to incorrect)	Yes	No
32	Yes	Other (said words aloud)	Yes	No
33	Yes	None used (too difficult)	Not Sure	No
34	Yes	Other (thought of pleasant words as unpleasant and vice versa)	Yes	No
35	Yes	Other?	Yes	No
36	Yes	Yes	Don't know	Yes

that only one participant reversed the IRAP effect from exposure 1 to 2.

Only two of the 12 participants in the Faking/Strategy group reported using the strategy provided. Three of the remaining 10 participants used a strategy that involved "thinking of pleasant words as unpleasant and vice versa." The other six reported using various different strategies, the description of which was difficult to interpret. One participant reported that it was too difficult to use the strategy. Seven of the 12 participants reported that they believed they were successful at faking their responses, and of these none were successful. Four of the participants indicated that they did not know if they were successful, and of these two participants did reverse the IRAP effect in exposure 2. Finally, one participant reported that he definitely did not fake the IRAP performance. In sum, for the two participants that faked their responses, one used the strategy provided and the other used a different strategy (i.e., "associating emotions with responses") but neither was sure if the attempts at faking were successful.

Summary

The current data indicate that the two sets of faking instructions did not impact significantly upon the IRAP performances of the two faking strategy groups. The Faking Strategy Questionnaire indicated that all participants in the Faking/No Strategy group, but only 2 in the Faking/Strategy group, followed the specific instruction to fake their responses. Overall, it appears that participants found it difficult to fake an IRAP performance, even when they were provided with direct instructions on how to do so.

DISCUSSION

The current research was based on a previous study (Kim, 2003), which found that participants could fake the IAT when provided with explicit instructions on how to do so. In contrast to the IAT data, providing participants with explicit instructions on how to fake the IRAP did *not* reverse, or even remove, the significant IRAP effect. On balance, it should be noted that 90% of the participants in the Faking/Strategy group in Kim's (2003) study reported using the faking strategy compared to 17% of the Faking/Strategy group in the current study. At the present time, it remains unclear why adherence to the faking instructions was relatively low. One possible reason is that the increased complexity of the IRAP, relative to the IAT, rendered the faking instructions very difficult to follow. Indeed, during debriefing a number of participants reported finding the task very difficult.

Although the relative difficulty involved in faking the IRAP will have to await further empirical inquiry, it is worth noting that three participants in the current study apparently did manage to fake their responses (i.e., reversed the consistent-inconsistent difference across the two exposures). Thus, faking the IRAP may be very difficult, relative to the IAT, but faking is not impossible. On balance, however, IRAP research that was conducted concurrently with the present study has shown that a small number of participants on occasion spontaneously reverse their IRAP performances across repeated exposures *without* faking instructions (but consistent with the current study the IRAP

effect remains significant at the group level). The fact that other studies have found occasional spontaneous reversals for individual participants renders the reversals obtained in the current study difficult to interpret; perhaps these reversals would have occurred in the absence of the faking instructions, and thus do not reflect an attempt on behalf of the participant to fake the IRAP. Future studies will need to explore this issue very carefully in order to separate out spontaneous from faking-induced reversals.

After the current study was conducted, a second successful IAT-faking study was published (Fiedler & Bluemke, 2005). Overall, the results of this recent study indicated that when participants had been exposed to an IAT they could spontaneously derive a strategy that allowed them to fake a subsequent IAT without being instructed explicitly on how to do so (i.e., by slowing down on the consistent tasks). The current data are in stark contrast with the Fiedler and Bluemke research, in that neither of the current faking groups managed to fake an IRAP performance even when provided with an explicit instruction on how to do so. This provides additional support for the suggestion that the IRAP may be more difficult to fake than the IAT. On balance, this is the first attempt to study the fake-ability of the IRAP and thus any conclusions arising from the current work must be viewed with caution. Nevertheless, the present research provides a solid foundation upon which to base subsequent empirical inquiry into the fake-ability of the IRAP.

In closing, it is worth noting that the current findings provide some support for the IRAP as an implicit measure, but only in one sense. In reviewing the relevant literature, DeHouwer (in press) suggested that a measure may be defined as implicit if participants; (a) are not aware of the fact that the relevant attitude or cognition is being measured, (b) do not have conscious access to the attitude or cognition, or (c) have no control over the measurement outcome. In the current study, it seems likely that participants were aware that their "attitudes" to the target stimuli as pleasant or unpleasant were being assessed in some way. Furthermore, it is almost certain that the participants were reasonably aware of their attitudes to the target stimuli. Given the results of the two faking conditions, however, it appears that there was limited control over the measurement outcome. In short, the current IRAP may be considered an implicit measure based on definition c, but not definitions a or b. On balance, ongoing research conducted by our group suggests that if an IRAP presents "socially sensitive" stimuli it may tap into attitudes or beliefs of which participants claim they were unaware (definition b). In any case, considerable empirical work will be required to investigate this area fully and to assess the reliability and validity of the IRAP as a measure of implicit beliefs and attitudes.

References

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.

De Houwer J (in press). What are implicit attitudes and why are we using them? In RW Wiers & AW

Stacy (Eds.), *The handbook of implicit cognition and addiction.* Thousand Oaks, CA: Sage Publishers.

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

de Jong P, Pasman W, Kindt M, & van den Hout MA (2001). A reaction time paradigm to assess (implicit) complaint-specific dysfunctional beliefs. *Behaviour Research and Therapy, 39,* 101-113.

Fiedler K, & Bluemke M (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27,* 307-316.

Gemar MC, Segal ZV, Segratti S, & Kennedy SJ (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. *Journal of Abnormal Psychology, 110,* 282-289.

Greenwald AG, McGhee DE, & Schwartz JLK (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald AG, Nosek BA, & Banaji MR (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197-216.

Hayes SC, Barnes-Holmes D, & Roche B (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition.* New York: Plenum.

Kim D-Y (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66,* 83-96.

Nosek BA, Greenwald AG, & Banaji MR (in press) The Implicit Association Test at age 7: A methodological and conceptual review. In JA Bargh (ed.), *Automatic processes in social thinking and behaviour.* New York: Psychology Press.

Robinson JP (1974). Public opinion during the Watergate crisis. *Communication Research, 1,* 391-405.

Teachman BA, Gregg AP, & Woody SR (2001). Implicit associations of fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110*, 226-235.

Received, 30 November 2006
Accepted, 21 May 2007