*Social Psychology Quarterly*
2003, Vol. 66, No. 1, 83–96

# Voluntary Controllability of the Implicit Association Test (IAT)*

DO-YEONG KIM
*University of Washington*

*Greenwald, McGhee, and Schwartz showed that white participants indicated a more positive evaluative association with whites than with blacks in the Implicit Association Test (IAT), and were being neutral on explicit measures. Their results suggested that the IAT might resist self-presentational forces which can mask personally or socially undesirable racial attitudes. In the current study, two experiments tested whether participants could voluntarily suppress the tendency to appear (1) more favorable to flowers than to insects on the IAT of those attitudes, or (2) pro-white on the racial IAT of whites and blacks. Both experiments found that participants could not fake the IAT effectively when merely asked to do so; they could produce a faked implicit attitude only when they were instructed to respond slowly to a subset of the stimuli. Overall, participants did not spontaneously discover the apparently controllable strategy for faking the IAT; they had to be taught how to implement it.*

Although national surveys report a reduction in racism over the past 50 years (Schuman et al. 1997), many researchers believe that subtle and implicit forms of stereotypes and prejudice against minority populations still exist (Crosby, Bromley, and Saxe 1980; Fazio et al. 1995; Fiske 1998; Wittenbrink, Judd, and Park 1997). Two explanations for these discrepant results are (1) that self-report measures of attitudes may be susceptible to self-presentation bias and/or (2) that subtle form of stereotypes and prejudice are not captured by the explicit self-report measures (Dasgupta et al. 2000).

For the past several decades, researchers who use self-report measures frequently have expressed concerns about the susceptibility of such measures to socially desirable self-presentations (Helmes and Holden 1986; Sigall and Page 1971; Weber and Cook 1972). For this reason, developing an assessment technique that is less vulnerable (if not invulnerable) to voluntary control or conscious distortion of responses on a measure would be an important step toward more valid measurement.

In response to these problems, substantial efforts have been invested in developing new methods for psychological research. In the past decade, significant advances of this sort have been achieved in the study of implicit social cognition (Bargh 1994; Bornstein and Pittman 1992; Greenwald and Banaji 1995; Kihlstrom 1990; Uleman and Bargh 1989). Because responses on implicit measures are assumed to be uncontrollable, attempts to use indirect measures for research on socially sensitive topics such as prejudice and stereotypes have been particularly noteworthy (Devine 1989; Dovidio and Fazio 1992; Gaertner and McLaughlin 1983).

### *Introduction to the Implicit Association Test (the IAT)*

Greenwald, McGhee, and Schwartz (1998) recently described a new implicit method, the Implicit Association Test (IAT),

* This research was supported by grants from the National Science Foundation (Grants SBR-9422242 and SBR-9710172), and from the National Institute of Mental Health (Grants MH 41328, MH 01533, and MH 57672) awarded to Tony Greenwald. Preparation of this report was supported, in part, by the National Institute of Mental Health Grant MH 60849. I wish to thank Tony Greenwald for letting me use his laboratory to complete this research and for providing valuable feedback on this manuscript. I would also like to thank Melissa Caldwell, Ed Diener, Karen Rudolph, Barbara Sarason, and Ronald Smith, and two anonymous reviewers for wonderful comments on a draft of this article. Address correspondence regarding this manuscript to Do-Yeong Kim, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820; kimd@s.psych.uiuc.edu.

which uses a latency-based indirect measure to assess automatic operation of attitudes. The IAT illustrated in Figure 1 uses the four concepts *insect*, *flower*, *pleasant*, and *unpleasant* to provide a measure of attitude toward flowers versus insects (Greenwald et al. 1998, Exp. 1).

The IAT begins by introducing participants to the four categories used in the task. In this procedure, participants are asked to sort stimuli representing four concepts into just two categories, each including two of the four concepts. The usefulness of the IAT as a measure of association strength depends on an empirically tested assumption: when the two concepts that share a response are associated strongly, the sorting task is considerably easier than when the two response-sharing concepts are associated weakly. If the participant responds more rapidly when *flower* and *pleasant* share a response than when *insect* and *pleasant* do so, this indicates that the *flower-pleasant* association is stronger than the *insect-pleasant* association and that the participant has a more positive attitude toward flowers than toward insects.

Greenwald et al. (1998, Exp. 3) also used the IAT to examine white college students' implicit racial attitudes toward whites and blacks. Their study showed that white participants performed the task more easily and more quickly when *white* was associated with *pleasant* than when *black* was associated with *pleasant*, indicating a more positive evaluation associated with white than with black. The same pattern of in-group positive association was replicated in a study of Korean Americans' and Japanese Americans' implicit attitudes toward Korean and Japanese ethnic groups (Greenwald et al. 1998, Exp. 2).

### *Voluntary Controllability of the IAT*

Greenwald et al. (1998) suggested that one useful quality of the IAT method may be its resistance to self-presentation strategies: IAT attitude measures may reveal attitudes even in those who seek to suppress the expression of an attitude when providing responses to the measure. This claim was partly supported in their study of white college students' implicit attitudes toward whites and blacks (Greenwald et al. 1998, Exp. 3), which showed discrepancies between explicit and implicit attitude measures of whites toward blacks. That is, on self-report measures white participants, on average, were impartial or nonprejudiced. On an IAT measure, however, all but one of 26 participants demonstrated more positive automatic evaluation of whites than of blacks.

### *Purpose of the Present Study*

Research on implicit attitudes has included the assumption that participants cannot control their responses on implicit measures. Yet previous research using implicit measures did not test explicitly whether participants were able to misrepresent their attitudes on the implicit measure. The purpose of the two experiments described here was to test this assumption, with a focus on the voluntary controllability of the IAT, by investigating the participants' ability to misrepresent their attitudes using three different IAT measures: flower versus insect, musical instrument versus weapon, and (racial) white versus black.

## EXPERIMENT 1

In Experiment 1 I examined participants' ability to control their implicit attitudes in IAT measures involving two pairs of attitude objects: (1) flowers versus insects and (2) musical instruments versus weapons. In a previous study, Greenwald et al. (1998, Exp. 1) showed that participants performed a classification task better for evaluatively compatible combinations than for noncompatible combinations: that is, they responded faster and made fewer errors when target concepts were combined with closely associated attributes (flower + pleasant and insect + unpleasant) than with less closely associated attributes (flower + unpleasant and insect + pleasant). These results were consistent with the expectation that participants had more positive automatic evaluations of flowers than of insects, and more positive attitudes toward musical instruments than toward weapons.

Experiment 1 involved two treatment groups: (1) a faking treatment group (faking group), which was instructed to respond as if

weapons or insects were more pleasant than musical instruments or flowers, and (2) a (nonfaking) control group, which received the usual IAT instructions.

## METHOD

### *Sample*

A total of 73 students from introductory psychology courses at the University of Washington provided data in exchange for an optional credit. Nine of those participants were dropped from the analysis: five in the faking group, for not understanding experimental instructions, and four additional participants (two in each condition) because they lacked fluency in English. The study was left with 64 participants (32 for the faking group and 32 for the nonfaking group) for whom the data were analyzed.

### *Materials and Measures*

The two IAT measures employed in Experiment 1 used 15 flower names, 15 insect names, 14 names of musical instruments, 14 weapon names, 15 words with pleasant meanings, and 15 words with unpleasant meanings. I selected these items from the category lists used by Greenwald et al. (1998); they are listed in Appendix A1.

Before performing the computer-administered IAT task, participants responded to a questionnaire containing two self-report attitude measures: a feeling thermometer and a semantic differential. On the feeling thermometer, participants were asked to place a mark on each of four pictures of a thermometer, which were labeled at bottom, middle, and top with "0 degrees (cold, or unfavorable)," "50 degrees (neutral)," and "99 degrees (warm, or favorable)." The marks were to indicate the warmth (i.e., positiveness) of the respondents' feelings toward insects, flowers, musical instruments, and weapons (Robinson 1974). The resulting attitude measure was computed as the rating of flowers (or musical instruments) minus the rating of insects (or weapons). This measure had a potential range of –99 to 99.

Next, the participants completed a set of five semantic differential items for each of the four object categories. They used five seven-point bipolar adjective scales (beautiful/ugly, good/bad, pleasant/unpleasant, honest/dishonest, nice/awful) to indicate their evaluations of each of the four objects (flowers, insects, musical instruments, and weapons). The semantic differential items were scored –3 to 3; greater numbers indicated greater liking. The difference between participants' average ratings of insects and flowers (or weapons and instruments) had a potential range of –6 to 6. (Positive numbers indicate a preference for flowers over insects, or for instruments over weapons.)

### *Procedure*

Each participant first responded to the feeling thermometer and semantic differential measures. Participants completed these questionnaires in their cubicles; at the outset they were instructed that they were to place their completed questionnaires in an envelope, which in turn would be placed in a covered box.

After completing the questionnaire, participants performed a series of two IAT tasks (the preliminary IAT and the test IAT). Participants were assigned randomly to one of two treatment conditions: faking (experimental) and nonfaking (control). After they completed the computer tasks, we administered a questionnaire requesting reports of the strategies and methods they used to comply with the experimental (faking) instruction.

*The preliminary IAT.* Participants initially were given the opportunity to remove unfamiliar words from the list of names used for the IAT, leaving a minimum of 10 items in each category. Then they were instructed: "[R]espond rapidly in categorizing each stimulus, but don't respond so fast that you make many errors. (Occasional errors are okay.)"

Figure 1 shows the sequence of tasks constituting each IAT measure and illustrates this sequence with materials from Experiment 1. In the preliminary IAT, half the participants were assigned to the flowers-versus-insects measure, and the remainder to the musical instruments-versus-weapons measure. For each of these groups, half the participants performed the compatible combined task first (flower + pleasant versus

| Block Sequence | Response Key on the Keyboard: Left | Right |
|---|---|---|
| 1. Initial Target-Concept Discrimination | **INSECT** | **FLOWER** |
| | AZALEA _ | |
| | _ ANT | |
| 2. Associated Attribute Discrimination | **PLEASANT** | **UNPLEASANT** |
| | Abuse _ | |
| | _ Joy | |
| 3. Initial Combined Task | **INSECT** | **FLOWER** |
| | **PLEASANT** | **UNPLEASANT** |
| | Failure _ | |
| | CARNATION _ | |
| | _ Joy | |
| | _ DRAGONFLY | |
| 4. Reversed Target-Concept Discrimination | **FLOWER** | **INSECT** |
| | _ AZALEA | |
| | ANT _ | |
| 5. Reversed Combined Task | **FLOWER** | **INSECT** |
| | **PLEASANT** | **UNPLEASANT** |
| | Failure _ | |
| | _ CARNATION | |
| | _ Joy | |
| | ANT _ | |

The IAT procedure for the present experiment involved a series of five discrimination tasks (numbered rows). A pair of target concepts and an attribute dimension are introduced in the first two steps. Categories for each of these discriminations are assigned to a left or a right response, indicated by the black dot. These are combined in the third step and then recombined in the fifth step, after response assignments are reversed (in the fourth step) for the target-concept discrimination. The illustration uses stimuli for the specific tasks for one of the task-order conditions of Experiment 1; correct responses are indicated as black dots (see Greenwald, McGhee, and Schwartz 1998).

Figure 1. Description and Illustration of the Implicit Association Test (IAT)

insect + unpleasant), and the other half began with the noncompatible combined task (flower + unpleasant versus insect + pleasant).

For the combined tasks (Steps 3 and 5 in Figure 1), the stimuli came alternately from one category pair (e.g., pleasant versus unpleasant) and from the other (e.g., flower versus insect). On each trial, the stimulus

item was visible until the correct response was made; then the next item appeared after a 150ms. delay (intertrial interval). If the response was incorrect, the stimulus was replaced by the word ERROR for 300 ms.

*The test IAT.* Immediately after the preliminary IAT, both the faking and the nonfaking group of participants were informed, "The experiment you just participated in typically produces data showing that participants associate flowers or musical instruments with pleasant meaning-words more easily than they associate insects or weapons with pleasant meaning-words. In other words, you probably have noticed that it was relatively easy to respond to flower names and pleasant meaning-words with the same key, but more difficult to respond to insect names and pleasant meaning-words using the same key."

Then both groups of participants were informed that they would perform a different task for the next part of the experiment: flowers-insects if they had worked previously on weapons-instruments, and vice versa.

Next, participants in the faking group (illustrated here if they were completing the weapons-instruments IAT second) were instructed, "What we want you to do in the next task is to try to respond as if you have a more positive attitude toward weapons than you do toward musical instruments. In other words, respond like you think a person would who likes weapons more than musical instruments."

Participants were not instructed explicitly on how to accomplish this task, but were asked to indicate whether they understood the faking instruction. Participants in the nonfaking (control) group simply were reminded of the previous instructions before completing the test IAT.

## RESULTS

In keeping with the methods introduced by Greenwald et al. (1998), all trials with latencies greater than 3,000ms were recoded to 3,000ms; all trials with latencies less than 300ms were recoded to 300ms. To reduce the skew associated with response-latency data, I log-transformed participants' response latencies and dropped the first two trials of each block because their latencies typically were lengthened.

### *A Summary Measure of IAT Effect*

I calculated the IAT effect as the mean performance for the noncompatible combined task (insect + pleasant or weapon + pleasant) minus that for the compatible combined task (flower + pleasant or instrument + pleasant). A positive IAT effect indicated preference for flower (or musical instrument) over insect (or weapon).

### *The Effect of the Faking Instruction*

Table 1 illustrates mean IAT effects of both the preliminary and the test IAT condi-

Table 1. Summary Statistics, Preliminary IAT and Test IAT, Experiment 1, for Nonfaking and Faking Groups

| | Mean | SD | Effect Size ($d$)[a] |
|---|---|---|---|
| *Preliminary IAT* | | | |
| Nonfaking Group | | | |
| IAT effect (latency) | 232.21 | 135.09 | 1.72 |
| IAT effect (log latency) | .15 | .11 | 1.36 |
| Faking group | | | |
| IAT effect (latency) | 203.56 | 147.78 | 1.38 |
| IAT effect (log latency) | .13 | .13 | 1.00 |
| Test IAT | | | |
| Non-faking Group | | | |
| IAT effect (latency) | 158.09 | 116.50 | 1.36 |
| IAT effect (log latency) | .10 | .10 | 1.00 |
| Faking group | | | |
| IAT effect (latency) | 173.18 | 162.33 | 1.10 |
| IAT effect (log latency) | .10 | .10 | 1.00 |

[a] The effect size measure $d$ = (mean/SD). Conventional small, medium, and large values of $d$ are .2, .5, and .8 respectively.

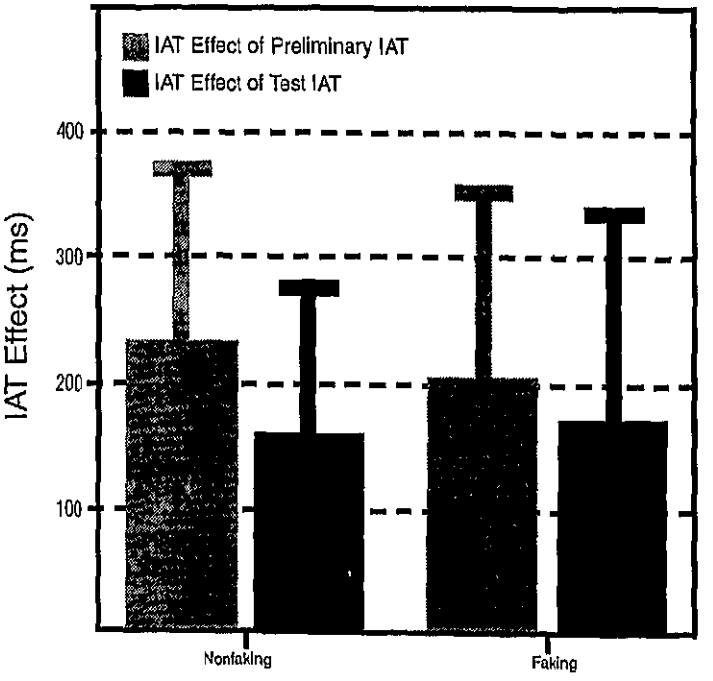
tions for the faking and the nonfaking groups. In the preliminary IAT, both of these groups responded more rapidly for the compatible grouping (flower + pleasant and insect + unpleasant) than for the noncompatible grouping (flower + unpleasant and insect + pleasant): $t$ (31) = –7.28, $p < .0001$ (two-tailed) for the nonfaking group; $t$ (31) = –5.84, $p < .0001$ (two-tailed) for the faking group.

A mixed-design analysis of variance showed no statistically significant interaction between group and test, $F(1,60) = .41, p > .50$ (two-tailed). The difference in the IAT effects of the preliminary IAT before the experimental manipulation between the two groups was not statistically significant, $t$ (62)= .43, $p > .60$ (two-tailed). For the test IAT conducted after the manipulation, the participants who received faking instructions did not show a significant difference in IAT effects from the control group, $t$ (62)= –.36, $p > .70$ (two-tailed). The difference between the mean IAT effects of each group corresponded to a standardized effect size, Cohen's (1988) $d$ = .3. (Conventional small, medium, and large values of $d$ are .2, .5, and .8 respectively.) (See Figure 2.)

*Effects of Individual Differences in Strategy*

On the basis of participants' responses to whether they used any strategy in following the faking instruction, I further classified participants in the faking group into two subgroups: strategy and no strategy. The data suggested that participants who reported using a strategy to fake the test could not do so effectively. I found no statistically significant differences in the IAT effect between those who reported using a strategy and those who reported using none: $F$ (1, 28) = .01, $p > .90$ (two-tailed); standardized effect size, Cohen's $d$ = .02.

Experiment 1 revealed that participants who tried to suppress their attitudes favoring flowers or musical instruments (relative to insects or weapons) could not do so. Within the faking group, strategies that participants




Thirty-two subjects in the nonfaking control group and 32 in the faking group received instruction in faking. Error bars indicate within-cell standard deviations.

*Figure 2. Mean IAT Effects: Results of Experiment 1*

reported using did not enable them to fake effectively.

## EXPERIMENT 2

Experiment 2 was designed to extend the findings from Experiment 1 to a more socially interesting domain, namely implicit racial attitudes. In a previous study, Greenwald et al. (1998, Exp. 3) reported that white participants responded more rapidly for the groupings of (white + pleasant and black + unpleasant) than for the alternative groupings (white + unpleasant and black + pleasant) on the IAT task, even though they showed no racial preference on explicit self-report measures of these attitudes. That is, the participants showed a more positive automatic evaluation toward whites than toward blacks in the IAT, but were neutral on explicit measures. This finding supports previous demonstrations of automatic expressions of race-related stereotypes and attitudes, which are disavowed by the participants who display them (Crosby et al. 1980; Devine 1989; Fazio et al. 1995; Gaertner and McLaughlin 1983; Greenwald and Banaji 1995; Wittenbrink et al. 1997).

In Experiment 2, I tested the effects of faking instructions on the race (black-white) IAT, and also observed the effects of providing participants with specific instructions about how to fake the IAT.

### METHOD

#### *Sample*

The participants were students from introductory psychology courses at the University of Washington who participated in exchange for course credit. A total of 73 white and Asian participants (49 white Americans and 24 Asian Americans) were recruited in the study and were classified on the basis of a demographic questionnaire completed at the beginning of the experiment. Among participants who identified themselves racially as Caucasian, only participants who identified themselves ethnically as American were included in white groups for the analysis. Data from one white American subject were excluded in the analysis because that subject described herself as Russian. Of the remaining 72 participants, 19 participants were males and 53 were females.

#### *Materials and Procedure*

Stimuli for the IAT were 15 white male names (e.g., Frank and Paul), 15 black male names (e.g., Deion and Lamar), 15 white female names (e.g., Emily and Nancy), and 15 black female names (e.g., Lashandra and Tanisha), all borrowed from Greenwald et al. (1998), along with the same 15 pleasant- and 15 unpleasant-meaning words used in Experiment 1 (see Appendix A2).

As in Experiment 1, participants first responded to two self-report racial attitude measures, the feeling thermometer and the semantic differential (white versus black), while alone in a cubicle. At the outset, they were instructed that they were to place the completed questionnaires in an envelope, which in turn would be placed in a covered box.

Next, participants completed a series of two IAT tasks: the preliminary IAT and the test IAT. They were assigned randomly to one of three conditions: the nonfaking control group ($n = 24$), the faking-no-strategy group ($n = 24$, the same as the faking group in Experiment 1), and the faking-strategy group ($n = 24$). The last condition, the faking-strategy group, was intended to show the effects of providing participants with specific strategies for faking the IAT.

Half the participants performed the white + pleasant and black + unpleasant combined task first; the other half completed the white + unpleasant and black + pleasant combined task first. Among each half of the participants, half performed the preliminary IAT with male names; the other half performed this test with female names. After the participants completed the test IAT, a questionnaire requesting a report of strategies used in the IAT was administered to all participants.

*The preliminary IAT.* Except for the replacement of flower and insect (or weapon and instrument) names with white and black male and female names, instructions for the preliminary IAT of Experiment 2 were iden-

tical to those for the preliminary IAT in Experiment 1.

*The Test IAT.* Before taking the test IAT, participants in the two faking groups (faking-no-strategy and faking-strategy) were instructed: "Regardless of your performance in the first computer task, please treat the second computer task as if it may indicate that you possess prejudice, but you prefer not to give that indication. It is still important for you to respond rapidly in categorizing each stimulus, but not to make many errors as in the first computer task." As in Experiment 1, participants in the nonfaking (control) group received no added instructions for the test IAT.

For the faking-strategy group, I provided additional instructions on how to fake the test, as follows: "Try to respond slowly for the condition in which white and pleasant (and black and unpleasant) are assigned to the same response and try to respond rapidly for the condition in which black and pleasant (and white and unpleasant) are assigned to the same response." As in Experiment 1, faking instructions were not given to the nonfaking participants. The rest of the procedure for the test IAT was the same as for the preliminary IAT.

## RESULTS

### *Tests of Voluntary Controllability of the IAT*

As in Experiment 1, the IAT effects difference score was used for the analysis. Pretreatment of the data (recoding of extreme scores, log transformation, and so on) was the same as for Experiment 1.

Table 2 shows the mean IAT effects for all groups on both the preliminary and the test IAT. I first performed a mixed-design analysis of variance to test an interaction between group and test; this interaction was statistically significant, $F(2, 69) = 3.93, p < .05$ (two-tailed).

In the preliminary IAT conducted before experimental manipulation, all three groups responded more rapidly for the grouping (white + pleasant and black + unpleasant) than for the alternative grouping (black + pleasant and white + unpleasant): $t(23) = -8.69, p < .0001$ (two-tailed) for the nonfaking group; $t(23) = -7.95, p < .0001$ (two-tailed) for the faking-no-strategy group; $t(23) = -6.16, p < .0001$ (two-tailed) for the faking-strategy group. In further analysis using a one-way analysis of variance for group difference in the IAT effects, the three

Table 2. Summary Statistics, Preliminary IAT and Test IAT, Experiment 2, for Nonfaking, Fake-No-Strategy, and Fake-Strategy Groups

| | Mean | SD | Effect size ($d$)[a] |
|---|---|---|---|
| Preliminary IAT | | | |
| Nonfaking group | | | |
| IAT effect (latency) | 207.98 | 125.62 | 1.66 |
| IAT effect (log latency) | .20 | .12 | 1.67 |
| Faking-no-strategy group | | | |
| IAT effect (latency) | 185.18 | 113.89 | 1.63 |
| IAT effect (log latency) | .20 | .12 | 1.67 |
| Faking-strategy group | | | |
| IAT effect (latency) | 181.49 | 171.47 | 1.06 |
| IAT effect (log latency) | .18 | .14 | 1.29 |
| Test IAT | | | |
| Nonfaking group | | | |
| IAT effect (latency) | 164.01 | 97.63 | 1.68 |
| IAT effect (log latency) | .17 | .10 | 1.36 |
| Faking-no-strategy group | | | |
| IAT effect (latency) | 151.53 | 172.30 | 1.38 |
| IAT effect (log latency) | .18 | .21 | 1.00 |
| Faking-strategy Group | | | |
| IAT effect (latency) | −32.21 | 283.06 | 1.38 |
| IAT effect (log latency) | −.003 | .25 | .01 |

[a] The effect size measure $d$ = (mean/SD). Conventional small, medium, and large values of $d$ are .2, .5, and .8 respectively.

groups of participants (nonfaking, faking-no-strategy, and faking-strategy) did not differ in their performance, $F$ (2, 69) = .30, $p$ > .60 (two-tailed), and thus uniformly showed participants' strong automatic positivity toward whites.

Table 3 provides correlations between explicit and implicit racial attitude measures of the preliminary IAT. The "feeling thermometer" explicit measure was correlated more highly with the "semantic differential" explicit measure than with the IAT measure.

*Effects of Faking Instructions*

In the test IAT conducted after the experimental manipulation, the three groups showed significant difference in the IAT effects, $F$ (2, 69) = 2.96, $p$ < .01 (two-tailed). Further analyses using planned comparisons suggested that the faking-no-strategy participants who were instructed to fake the test (but without specific strategy instruction) did not show a significant difference from the nonfaking group on the IAT effect, $t$ (69) = −.16, $p$ > .80 (two-tailed) (see Figure 3). In contrast, the faking-strategy group, whose members received specific instruction on how to fake the IAT, showed a significant difference from both the nonfaking and the faking-no-strategy group on the IAT effect: $t$ (69) = −3.15, $p$ < .01 (two-tailed); $t$ (69) = 3.31, $p$ < .01 (two-tailed)(see Figure 3).

Further analyses suggested that faking-strategy participants partially followed the instructions. They were able to slow down in the easy condition (white + pleasant), showing a statistically significant difference in latencies between the preliminary and the test IAT of that condition, $t$ (23) = −3.26, $p$ < .01 (two-tailed). They were unable, however, to speed up their responses in the difficult condition (white + unpleasant), $t$ (23)= 1.65, $p$ > .10 (two-tailed).

*Tests of Racial Difference*

In a mixed-design analysis of variance test for the three-way interaction on the performance of white and Asian participants, (group x race x test) showed no significant effect, $F$ (2, 63) = .09, $p$ > .90 (two-tailed).

*Awareness of Success in Faking the IAT*

For the strategy questionnaire that requested reports of strategies and methods used to comply with the experimental faking instructions, the data suggested that only three of the 24 participants in the faking-strategy group believed they were successful in faking the test. The remainder said they were not successful (11 participants) or not aware of their success (eight participants), or gave no response (two participants). Moreover, in regard to the question about the strategies employed (other than the two strategies provided), about 90 percent of the faking-strategy participants responded that they simply used the two strategies provided. The remaining 10 percent attempted to use other strategies: for example, thinking of famous positive black figures such as (then) General Colin Powell.

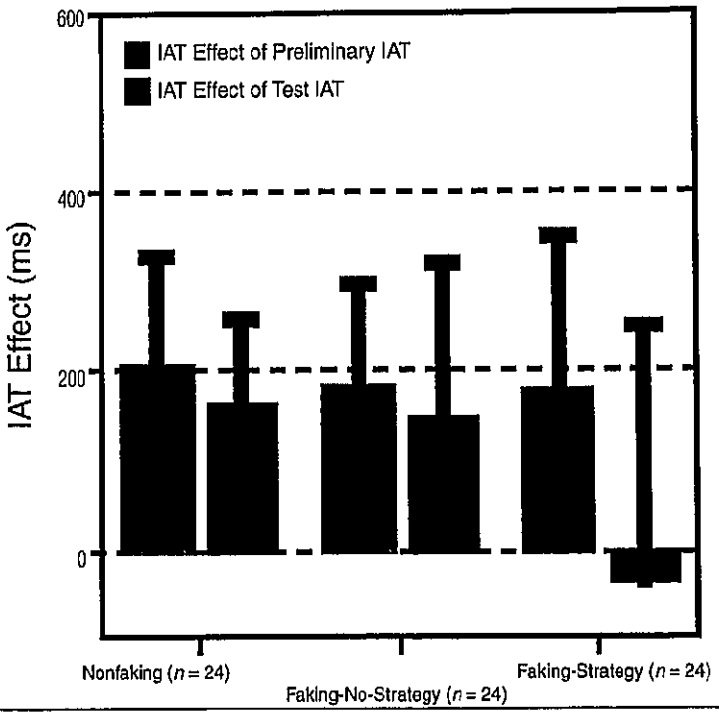
In Experiment 2 I sought to replicate the findings of Experiment 1 in the more socially relevant domain of racial attitudes. As indicated by the results from the preliminary IAT of all groups, a previous finding (Greenwald et al. 1998, Exp. 3) was replicated. This result demonstrated that the IAT revealed an implicitly stronger positive association with whites than with blacks among white and Asian participants. As in Experiment 1, par-

Table 3. Correlations Among Explicit and Implicit Measures in the Preliminary IAT, Experiment 2

| | Explicit Attitude | | Implicit Attitude |
|---|---|---|---|
| Measure | 1 | 2 | 3 |
| Feeling Thermometer | — | | |
| Semantic Differential | .62** | — | |
| IAT (log latency) | .10 | .20 | — |

*Note:* All measures were scored so that positive scores indicate preference for whites over blacks. Latency measures were transformed to natural logarithms for this analysis. IAT in this table is the data from the preliminary IAT.

** $p$ < .01 (two-tailed)




The graph indicates (1) no evidence of successful faking in the faking-no-strategy condition (replication of Experiment 1) and (2) the effectiveness of strategy instruction in the faking-strategy condition. Error bars indicate within-cell standard deviations.

Figure 3. Mean IAT Effects: Results of Experiment 2

ticipants who were asked to fake, but who received no specific instruction in strategy, could not do so reliably.

Participants who were given explicit strategies were partly able to fake the IAT, but only by slowing their performance in the white + pleasant condition. This result indicates that it is possible to control performance by slowing responses in the ordinarily easy white + pleasant condition, but not by attempting to speed up responses in the black + pleasant condition.

## DISCUSSION

In two experiments, I examined participants' ability to voluntarily suppress their attitudinal associations toward strongly valenced semantic categories (flower, insect, musical instrument, and weapon) and racial categories (black and white). I found that participants, when instructed to indicate a favorable attitude towards insect, weapon, or black, were unable to do so. Only those who were given specific instructions to go slowly in the typically easier (white + pleasant) condition displayed a faked implicit attitude in Experiment 2.

### *The Dissociation Between Implicit and Explicit Measures*

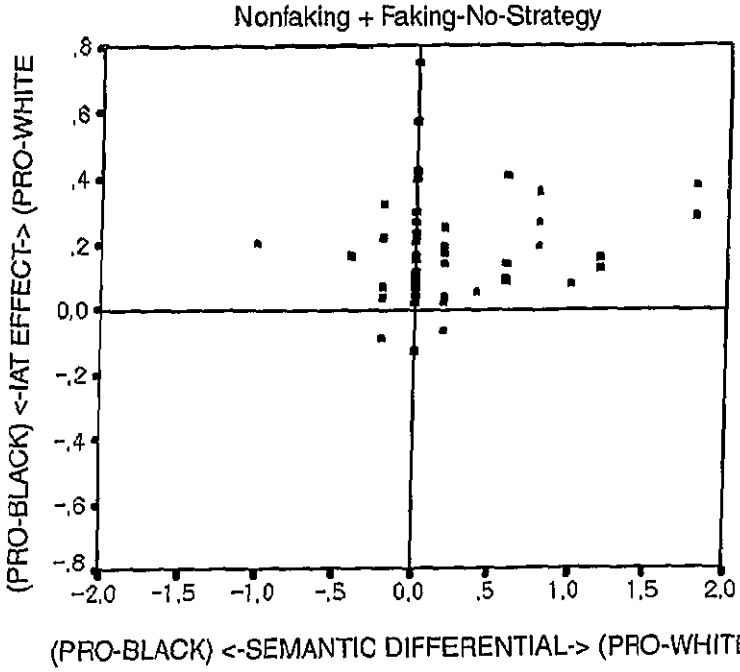
The current study confirmed previous findings of implicit racial preference among whites, favoring whites over blacks (Devine 1989; Dovidio and Gaertner 1993, 1998; Fazio et al. 1995; Judd et al. 1995; Lepore and Brown 1997). This result also appeared among Asians.

In a previous study, Greenwald et al. (1998, Exp. 3) reported that white participants (19 of 26) explicitly endorsed either black-white indifference or black preference on the same semantic differential measure as I used in Experiment 2. On the implicit IAT measure in that earlier experiment, however, all but one of the 26 white participants demonstrated a more positive association with whites than with blacks on the IAT.

As shown in Figure 4, the data from both the nonfaking and the faking-no-strategy control groups in Experiment 2 replicated the pattern observed previously: participants who expressed neutral or positive attitudes toward blacks on the semantic differential self-report measure were almost uniformly pro-white on the IAT measure, $r$ (46) = .12, $p > .40$ (two-tailed). This finding suggests dissociation between the explicit and the implicit attitudes. In contrast, participants in the treatment condition (faking-strategy) who produced a faked implicit attitude, expressing positive attitudes toward blacks on the semantic differential, also showed evidence of positive attitudes on the IAT measure, $r(22) = .46, p < .05$ (two-tailed).

In summary, the results revealed that participants did not spontaneously discover the apparently controllable strategy that they could use to fake an IAT: they had to be instructed to implement it. Further, the somewhat successful strategy (deliberate slowing in the easier condition) ultimately may not be satisfactory because deliberately slowed responses are likely to be identifiable as an attempt to manipulate the high error rate (20% or more) in a careful examination of IAT data (e.g., Greenwald and Farnham 2000). If participants could have speeded their responses in the more difficult condition, they would have produced a more effective faked IAT pattern, but it is apparent that they could not.

The results from two experiments suggested that the Implicit Association Test could be a useful tool which resists participants' attempts to mask their automatic expression of attitudes in typical conditions of administration. I strongly recommend that future research follow suit by testing other types of indirect tools and IAT measures as well.




Data are taken from the nonfaking and the faking-no-strategy groups of experiment 2 ($N$ = 48 white and Asian American subjects). Both measures have meaningful zero points that indicate absence of preference. The major feature of the data is the replication of a previous study by Greenwald et al. (1998, Exp. 3), showing absence of preference on the explicit measure and substantial preference for whites on the IAT measure. (Data taken from Experiment 2.)

Figure 4. Relationship of Semantic Differential and Implicit Association Test (IAT) Measures of Black-White Evaluative Preference

Appendix A1. IAT Stimuli, Experiment 1

| *Flowers* | *Insects* | *Musical Instruments* | *Weapons* |
|---|---|---|---|
| azalea | ant | banjo | arrow |
| bluebell | bee | clarinet | cannon |
| buttercup | beetle | drum | dagger |
| carnation | blackfly | fiddle | *firearm* |
| daffodil | centipede | flute | grenade |
| daisy | cockroach | guitar | gun |
| geranium | dragonfly | mandolin | hatchet |
| iris | flea | piano | *knife* |
| lilac | gnat | saxophone | missile |
| lily | maggot | trombone | *pistol* |
| marigold | mosquito | trumpet | rifle |
| petunia | spider | tuba | spear |
| rose | mite | violin | sword |
| tulip | wasp | harp | whip |
| violet | locust | | |

| *Pleasant Words* | *Unpleasant Words* |
|---|---|
| caress | abuse |
| cuddle | agony |
| glory | assault |
| gold | brutal |
| health | corpse |
| joy | death |
| kindness | failure |
| lucky | filth |
| peace | killer |
| snuggle | poison |
| success | slime |
| sunrise | slum |
| talent | stink |
| triumph | torture |
| warmth | vomit |

Appendix A2. IAT Stimuli, Experiment 2

| WHITE NAMES | | BLACK NAMES | |
|---|---|---|---|
| *Male* | *Female* | *Male* | *Female* |
| Adam | Amber | Alphonse | Aiesha |
| Allen | Betsy | Deion | Ebony |
| Andrew | Colleen | Everol | Lakisha |
| Brad | Donna | Jamel | Lashandra |
| Frank | Ellen | Kenyon | Latisha |
| Fred | Emily | Lamar | Latonya |
| Greg | Katie | Lavon | Malika |
| Harry | Kristin | Leroy | Shanise |
| Jack | Lauren | Malik | Sharise |
| Jed | Megan | Marcellus | Tamesha |
| Jonathan | Nancy | Rasaan | Tanisha |
| Justin | Sara | Theo | Tawanda |
| Paul | Shannon | Torrance | Temeka |
| Peter | Stephanie | Tyree | Tia |
| Roger | Wendy | Wardell | Yolanda |

| *Pleasant Words* | *Unpleasant Words* |
|---|---|
| caress | abuse |
| cuddle | agony |
| glory | assault |
| gold | brutal |
| health | corpse |
| joy | death |
| kindness | failure |
| lucky | filth |
| peace | killer |
| snuggle | poison |
| success | slime |
| sunrise | slum |
| talent | stink |
| triumph | torture |
| warmth | vomit |

## REFERENCES


Bargh, John A. 1994. "The Four Horsemen of Automacity." Pp. 1–40 in *Handbook of Social Cognition,* edited by Robert S. Wyer, and Thomas K. Srull. Hillsdale. NJ: Erlbaum.

Bornstein, Robert F. and Thane S. Pittmann. 1992. *Perception Without Awareness: Cognitive, Clinical, and Social Perspectives.* New York: Guilford Press.

Crosby, Faye, Stephanie Bromley, and Leonard Saxe. 1980. "Recent Unobtrusive Studies of Black and White Discrimination and Prejudice: A Literature Review." *Psychological Bulletin* 87:546–63.

Dasgupta, Nilanjana, Debbie E. McGhee, Anthony G. Greenwald, and Mahzarin R. Banaji. 2000. "Automatic Preference for White Americans: Eliminating the Familiarity Explanation." *Journal of Experimental Social Psychology* 36:316–28.

Devine, Patricia. G. 1989. "Stereotypes and Prejudice: The Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56:5–18.

Dovidio, John F. and Russell H. Fazio. 1992. "New Technologies for the Direct and Indirect Assessment of Attitudes." Pp. 204–37 in *Questions About Survey Questions: Meaning, Memory, Attitudes, and Social Interaction,* edited by Judith Tanur. New York: Russell Sage Foundation.

Dovidio, John F. and Samuel L. Gaertner. 1993. "Stereotypes and Evaluative Intergroup Bias." Pp. 167–93 in *Affect, Cognition, and Stereotyping,* edited by Diane M. Mackie and David L. Hamilton. San Diego: Academic Press.

———. 1998. "On the Nature of Contemporary Prejudice: The Causes, Consequences, and Challenges of Aversive Racism." Pp. 289–304 in *Racism: The Problem and the Response,* edited by Jennifer L. Eberhardt and Susan T. Fiske. Newbury Park, CA: Sage.

Fazio, Russell H., Joni R. Jackson, Bridget C.

Dunton, and Carol J. Williams. 1995. "Variability in Automatic Activation As an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Social and Personality Psychology* 50:229–38.

Fiske, Susan T. 1998. "Stereotyping, Prejudice, and Discrimination." Pp. 357–411 in *The Handbook of Social Psychology*, edited by Daniel T. Gilbert and Susan T. Fiske. Boston: McGraw-Hill.

Gaertner, Samuel L. and John P. McLaughlin. 1983. "Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics." *Social Psychology Quarterly* 46: 23–30.

Greenwald, Anthony G. and Mahzarin R. Banaji. 1995. "Implicit Social Cognition: Attitudes, Self- Esteem, and Stereotypes." *Psychological Review* 102:4–27.

Greenwald, Anthony G. and Shelly D. Farnham. 2000. "Using the Implicit Association Test to Measure Self-Esteem and Self-Concept." *Journal of Personality and Social Psychology* 79:1022–38.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L.K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74:1464–80.

Helmes, Edward and Ronald R. Holden. 1986. "Response Styles and Faking on the Basic Personality Inventory." *Journal of Consulting and Clinical Psychology* 54:853–59.

Judd, Charles M., Bernadette Park, Carey S. Ryan, and Markus Brauer. 1995. "Stereotypes and Ethnocentrism: Diverging Interethnic Perceptions of African American and White American Youth." *Journal of Personality and Social Psychology* 69:460–81.

Kihlstrom, John F. 1990. "The Psychological Unconscious." Pp. 445–64 in *Handbook of Personality: Theory and Research*, edited by Lawrence A. Pervin. New York: Guilford Press.

Lepore, Lorella and Rupert Brown. 1997. "Category and Stereotype Activation: Is Prejudice Inevitable?" *Journal of Personality and Social Psychology* 72: 275–87.

Robinson, John. P. 1974. "Public Opinion During the Watergate Crisis." *Communication Research* 1:391–405.

Schuman, Howard, Charlotte Steeh, Lawrence Bobo, and Maria Krysan. 1997. *Racial Attitudes in America: Trends and Interpretations*. Cambridge, MA: Harvard University Press.

Sigall, Harold and Richard Page. 1971. "Current Stereotypes: A Little Fading, A Little Faking." *Journal of Personality and Social Psychology* 18: 247–55.

Uleman, James S. and John A. Bargh. 1989. *Unintended Thought*. New York: Guilford Press.

Weber, Stephen J. and Thomas D. Cook. 1972. "Subject Effects in Laboratory Research: An Examination of Subject Roles, Demand Characteristics, and Valid Inference." *Psychological Bulletin* 77:273–95.

Wittenbrink, Bernd, Charles M. Judd, and Bernadette Park. 1997. "Evidence for Racial Prejudice at the Implicit Level and Its Relationship With Questionnaire Measures." *Journal of Personality and Social Psychology* 72: 262–74.

***Do-Yeong Kim** is a postdoctoral fellow in the Department of Psychology at the University of Illinois at Urbana-Champaign. His research interests include implicit social cognition in prejudicial attitudes, culture/acculturation, and subjective well-being. Much of his work has investigated various aspects of implicit psychological phenomena, developing implicit measures of the psychological constructs, and applying those measures to test empirical research questions.*