Faking the IAT:

Aided and Unaided Response Control on the Implicit Association Tests

Klaus Fiedler & Matthias Bluemke

University of Heidelberg, Germany

Basic and Applied Social Psychology

(in press)

Abstract

One pragmatic goal of implicit tools like the Implicit Association Tests (IAT) is to rule out self-presentation and controlled responding. Three experiments examined whether the IAT meets this goal, using Turkish and German groups along with positive and negative traits. Experiment 1 was an Internet study. After completing a naïve IAT pretest, participants were instructed to fake on a posttest in three graded conditions that differed in the explicitness of faking instructions. Experiment 2 replicated and extended the approach in the laboratory, including no-pretest condition. Results demonstrate that participants who intended to fake were successful, provided the experience of a pretest. Experiment 3 ruled out an alternative account of faking in terms of pretest experience. Faking was mostly due to slow-down on compatible trials, but a notable speed-up on incompatible trials also occurred. Faking remained inconspicuous, especially with non-blatant instructions; experts failed to identify faked data sets.

Faking the IAT:

Aided and Unaided Response Control on the Implicit Association Tests

A major purpose of implicit measurement tools like the Implicit Association Tests (IAT; Greenwald, McGhee, & Schwartz, 1998) is to enable unadulterated assessment of attitudes that may not be frankly expressed in explicit self-reports. To rule out social-desirability and self-presentation tendencies, an IAT uses a Stroop-like speeded classification task (Banaji, 2001). For attitude assessment, the stimulus series to be classified under speed instructions includes two subsets of items, denoting attitude targets and evaluative concepts. For instance, to measure attitudes toward Blacks, trials might consist of names or pictures referring to Black versus White people along with unambiguously positive versus negative adjectives or nouns. As each stimulus is presented, participants have to respond, as fast as possible, with one of two response keys. In one trial block, they have to use one response key for White and positive stimuli and the other key for Black and negative stimuli. In another trial block, they are to map White and negative onto one key and Black and positive onto the other. Shorter latencies in the former than in the latter condition is taken as evidence for a more positive attitude towards White than Black people. The underlying assumption is that the relative speed with which the same motor response can be assigned to an attitude target and a specific evaluative meaning provides a measure of implicit attitude.

Whoever participated in an IAT, swearing not to be prejudiced at all against Blacks, will have found it nevertheless much easier to use the same response for White and positive and for Black and negative than vice versa. It is this apparent lack of control or impossibility to counteract the IAT effect that has nourished the claim that an unobtrusive instrument has been found, which does not lend itself to controlled responding.

In this article, we will not tackle the tricky question of whether a high IAT score provides diagnostic evidence for an attitude (cf. Fiedler, Messner, & Bluemke, 2005;

Rothermund & Wentura, 2004). The focus of the present research is exclusively on the question of whether the IAT indeed evades voluntary control. In three experiments, we test the ability to "fake" on an IAT devised to measure attitudes towards Turks as compared to Germans – a major inter-ethnic topic in Germany that was shown to produce regular IAT effects independent of varying stimulus selections (Florack, Scarabis, & Bless, 2001; Messner & Freytag, 2003; Neumann & Seibt, 2001).

As in faking studies in other diagnostic domains, such as polygraph lie detection tests (cf. Honts, Hodes, & Raskin, 1985), we based our empirical approach on the assumption that for a test to be unsusceptible to strategic manipulation it must resist two influences. First, the outcome must be invariant to deliberate variations in the test persons' performance intentions. Second, the test's usefulness must be maintained when respondents acquire knowledge about the test. Just as an intelligence test or polygraph test must remain valid when people inform themselves about how the test works, an IAT must still be useful when testees find out that the test score is computed as the difference in reaction speed measured in two trial blocks. Assuming that such freely available information can be concealed would be unrealistic. Thus, we pose the following empirical criterion of insusceptibility: IAT results must be unaffected by test *outcome intentions* as well as test *knowledge*. Accordingly, we encouraged participants to fake an intended IAT outcome, using all available test knowledge.

Given the flood of IAT studies conducted in recent years and the emphasis on its implicit and automatic nature, systematic research on faking is surprisingly scarce. A few exceptional studies (Banse, Seise, & Zerbes, 2001; Kim, 2003; Egloff & Schmukle, 2002) have not directly encouraged participants to fake and have not solicited instrumental strategies to shift test scores in a specific direction. For instance, Banse et al. (2001) found little evidence for control over test outcomes in participants who were exposed to an IAT for the very first time, being fully naïve about the test. Egloff and Schmukle (2002), using an anxiety IAT in the context of job application, based their faking manipulation merely on the

instruction to "try to make a very good impression … without exaggerating too much" (p.1446) – an instruction that might be just too weak to elicit successful faking. Kim (2003) instructed participants to "… 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 ...". With such indirect and unclear instructions, little evidence was found for successful manipulation of IAT scores. Only when Kim advised the subjects to "try to respond slowly" for the compatible condition and to "try to respond quickly" for the incompatible condition, subjects were partly able to fake.

Given the unspecific and partially inhibitory instructions used in most of these studies, the null findings provide no ultimate proof for the IAT's resistance to faking. Indeed, other studies have shown systematic influences of voluntary behaviour on IAT outcomes: through watching a movie clip showing Blacks in positive context (Wittenbrink, Judd, & Park, 2001), through counter-stereotypical imagery (Blair, Ma, & Lenton, 2001), or in the presence of a Black experimenter (Lowery, Hardin, & Sinclair, 2001). If such controllable behaviors can influence IAT responding, then it is in principle possible to modify test outcomes voluntarily. The bottom line of a recent review by Blair (2002) was indeed that many so-called automatic procedures are not at all unsusceptible to controlled influences. However, in spite of increasing evidence for voluntary and strategic influences influence (cf. Fazio & Olson, 2003), the IAT's relative immunity from faking continues to be emphasized. For instance, Gray, MacCulloch, Smith, Morris, & Snowden (2003) hold that "the IAT has successfully quantified beliefs that people may wish to disguise" (p. 497). Greenwald and Nosek (2001) interpret the aforementioned faking studies (Banse et al., 2001; Kim, 2003) as showing that faking failed. Gray et al. (2003) felt that even psychopaths high in manipulation skills may be identified using the IAT. And Banaji (2001) emphasized "a lack of control over one's response" (p. 20).

Overview of present research. Logically, null findings in a few studies cannot prove the *impossibility* of manipulating IAT scores substantially, whereas the mere existence of

successful faking under some conditions is sufficient to prove its *possibility*. However, our goal was not confined to demonstrating faking under more auspicious conditions, but we tried to provide a more refined answer to several aspects of the basic research question. We used graded manipulations varying in the explicitness of strategies offered for successful faking. In a pretest-posttest design, we assessed naïve and faked performance in the same respondents. We analyzed the impact of faking instructions on compatible and incompatible trial blocks. By including an Internet-based study as well as an experimentally controlled laboratory study, we also varied the task setting and the motivational boundary conditions. Finally, we pursued the intriguing question of whether experienced IAT testers can discriminate faked from authentic data.

As the Internet has afforded a prominent medium for distributing the IAT all over the world, we started with a net-based experiment (Reips, 2000). Experiment 1 was made available for several search engines (e.g., Google) and announced in various newsgroups (e.g., de.alt.umfragen) as a study on "How to beat a psychological test". Through this recruitment procedure, we presumably addressed participants who were motivated to fake test results, considering the task like a sportive challenge. Also, we may have solicited participants with some prior experience with the IAT. However, Internet participants could as well be less motivated, less accountable, and invest less effort in the experimental task than participants who get into personal contact with researchers in the lab.

Experiment 1 included three conditions, graded by the amount of instruction provided to facilitate faking. Participants in all conditions were first administered a normal German/Turk-IAT (i.e., without the instruction to fake). Then, the same IAT was administered once more with the explicit instruction to avoid a result that would indicate a negative implicit attitude against Turks. Participants in hierarchically ordered conditions received (a) only faking instructions without further advice; (b) an additional implicit hint that the IAT score reflects a comparison of two critical trial blocks; (c) plus the explicit strategy to slow down on

compatible trials rather than trying to speed up on incompatible trials. All three conditions were within the range of measures that real test persons employ to fake successfully.

Experiment 2 was a deliberate attempt to replicate and extend the results obtained in the first experiment under controlled laboratory conditions, warranting genuine randomization, exclusion of possible Internet artifacts (e.g., repeated participation), and enhanced chronometric accuracy. In addition to the subtlest and the most explicit instruction condition of the first experiment, a new condition received neither strategic aids nor even pretest experience but they were asked to fake on the very first IAT encounter. The latency data from this controlled experiment were then presented to experts asked to discriminate between faked and authentic data sets. Finally, Experiment 3 served to disentangle pretest experience from strategic aids, which were combined in Experiment 2.

Experiment 1

Method

Participants. Sixty-five German participants (46% male) were recruited through newsgroups postings and Internet-Labs (e.g., http://wwwhomes.uni-bielefeld.de/psylab). As Internet participants can easily drop out or break up the session, the web page layout was set up to recruit participants who are interested in their own true test performance (cf. Reips, 2000). Thus, instead of a financial incentive, the only reward consisted in the revelation of insights about one's own test performance. Mean age was 25.5 years (SD = 5.75); about one half of the participants were non-students.

Design, Materials and Procedure.

Upon entering the web page, participants were asked to indicate their German language ability (native, fluent, or insufficient) and their explicit attitudes toward both ethnic groups on 9-point rating scales (endpoints labelled "very much", "very little appreciation") in response to the question "What view do you hold of Turks / Germans in general?". Presentation order was counterbalanced. Next, they received general information about the purpose of the IAT and

concrete instructions of how to use the response keys. The specific relevance of the two crucial trial blocks for the resulting IAT score was not revealed in these initial instructions.

Then participants performed the first IAT, involving Turkish and German target items along with evaluatively positive and negative adjectives (see Table 1). The procedure was the same as in Greenwald et al. (1998), except for a reduced number of training trials (16) and combined trials (64 compatible, 64 incompatible), to keep the total duration of the Web-experiment quite short (10-15 min.). The presentation order of compatible and incompatible blocks was counterbalanced across participants.

To avoid cross-categorical confounds (i.e., German and Turkish stimuli not only differing in ethnicity but also in valence; Steffens & Plewe, 2001; Mitchell, Nosek, & Banaji, 2003; Steffens, Lichau, Still, Jelenec, Anheuser, Goergens, & Hülsebusch, 2004; Fiedler et al., 2005; Bluemke & Friese, in press), a balanced stimulus set was used (Table 1). An equal number of positive and negative items were included among the eight German target stimuli (e.g., poet; skinhead) and the Turkish target stimuli (e.g., bazar; torture). The evaluative adjectives were unrelated to ethnicity and were chosen in the style of Rudman, Greenwald, Mellott, and Schwartz (1999). Each stimulus item was presented only once in a training block and twice in a combined block, adding up to 16 training trials and 64 trials in the compatible and incompatible blocks. To keep valence items and target stimuli separable, they not only differed in linguistic class (nouns vs. adjectives) but also in letter format (upper vs. lower case). These stimuli had been shown in prior studies to produce the expected Turk IAT effect in German participants (Messner & Freytag, 2003).

After the first IAT, participants received an online feedback about their relative speed in the two crucial trial blocks. Participants whose mean latency in the Turks+Negative (and Germans+Positive) combination was shorter than in the Germans+Negative (and Turks+Positive) combination were told this outcome is interpreted as a negative attitude toward Turks. An opposite interpretation was given for an inverse latency difference. To

increase faking motivation, participants were reminded of their explicit attitude ratings, which were generally quite favorable.

The crucial manipulation was administered immediately before the second IAT. In the *uninformed* condition, participants were told, "Please reflect ahead of each block on how to influence your reaction times to avoid the inference that you are prejudiced against Turks". The *implicitly informed* condition were additionally informed, "The shorter reaction times are in the compatible block and the longer reaction times are in the incompatible block, the more you could be judged as being prejudiced against Turks". In the *explicitly informed* condition, the strategic instruction was added: "It is most important trying to be slower in the compatible block. It doesn't pay off trying to be faster in the incompatible block".

After completing the second IAT, participants received detailed feedback about their latencies on both tests, the attitudinal implications, and whether they succeeded in faking. They were questioned for seriousness of participation and hyperlinks to related Internet sites were supplied.

Results and Discussion

After eliminating 15 participants with error rates higher than 30%, 19 *uninformed*, 17 *implicitly informed*, and 14 *explicitly informed* participants remained in the analysis. Outliers were handled by trimming the latency distribution to values between 300 ms and 3000 ms (Greenwald et al., 1998). Erroneous responses were treated as missing data.

Error Rates. Notably, faking attempts did not increase error rates. Mean error rates on compatible trials were .09 on the pretest without faking and .07 on the posttest with faking instructions. On incompatible trials, the mean error rate was .11 on the pretest and .09 in the posttest. This pattern – enhanced error rates on incompatible trials but no impairment resulting from faking – remained similar across conditions. In the uninformed and the implicitly informed condition, separate test repetitions (pretest vs. posttest) x trial blocks (compatible vs. incompatible) analyses of variance (ANOVA), yielded only a trial-block main effect, $F(1,18)$

= 7.99, $p < .05$, and $F(1,16) = 4.79$, $p < .05$, respectively (all other Fs ≤ 1). In the explicitly informed condition, the trial blocks main effect, $F(1,13) = 13.55$, $p < .01$, came along with a non-significant main effect tendency for test repetitions, $F(1,13) = 3.44$, and a trial blocks x test repetitions interaction, $F(1,13) = 4.75$, $p < .05$, reflecting a decrease in error rates from pretest to posttest, which was even more pronounced for incompatible trials (.13 to .06) than for compatible trials (.10 to .06). Altogether, these error rates do not raise any suspicion that faking attempts came at the expense of accuracy.

Latencies and IAT Scores. Let us now turn to the latency data proper, which provide the basis of IAT scores. Means and standard deviations of the resulting average latencies within the compatible and incompatible trial as well as the resulting average IAT scores are given in Table 2, as a function of experimental conditions. The average pretest scores tended to be clearly positive in all three conditions, confirming the premise of a typical IAT effect among Germans vis-à-vis Turks. This basic effect amounts on average to latency differences of 180, 159, and 209 ms, in the *uninformed*, *implicitly informed*, and *explicitly informed* conditions, respectively. The consensus among participants was quite high, rather moderate standard deviations (Table 2). The IAT effect size amounts to approximately half a standard deviation. A positive IAT score was obtained by 12 of 19 participants in the *uninformed* condition, 11 of 17 in the *implicitly informed* condition, and 11 of 14 in the *explicitly informed* condition. Altogether, 34 of 50 participants received positive IAT scores.

Turning to the data from the second, manipulated test, Table 2 shows that all conditions succeeded in producing negative IAT scores. The average score on the posttest was at least as negative as the pretest score had been positive. However, the variance across respondents is markedly enhanced. Nevertheless, the proportions of participants obtained a negative IAT score on the posttest is as high as 9 of 19 *uninformed* participants, 15 out of 17 given *implicit information*, and 10 of 14 in the *explicitly informed* group. Overall, two thirds or 34 of all 50 participants were able to fake. Among those 34 who started with a positive IAT score on the

pretest, 21 succeeded in changing the sign of their test score.

These apparent differences were corroborated in a 2 (pretest vs. posttest) x 3 (uninformed vs. implicitly informed vs. explicitly informed) analysis of variance (ANOVA) with repeated measures on the first factor. A main effect for test repetitions, $F(1,47) = 15.16$, $p < .001$, reflected a marked influence of faking instructions. Neither the instructions main effect, $F(2,47) = 0.22$, nor the interaction term, $F(2,47) = 0.85$, were significant. Together the latter two results indicate that although faked IAT scores became somewhat more negative with increasing explicitness of the instruction, faking success generalized across all conditions. Posttest scores were negative (i.e., significantly below zero) in the *explicitly informed* condition, $t(13) = -3.18$, $p < .01$, the *implicitly informed* condition, $t(16) = -3.94$, $p < .01$, and in the *uninformed* condition, $t(18) = -2.39$, $p < .05$. Thus, faking was even possible in the *uninformed* condition that had to find their own spontaneous strategy.

Consistent with our intuition concerning the most effective strategy, the reversal of IAT scores was mainly brought about by increased latencies in compatible trial block (see Table 2). However, interestingly, there was also a notable (though less pronounced) speed-up on incompatible trials. This speed-up hardly reflects a practice effect because no practice gains whatsoever were observed within tests from the first to the second half of trials.

The naïve pretest IAT score correlated slightly with the difference in explicit attitude rating toward Turk and slightly positively with the attitude rating toward German minus Turks, $r = .26$, $p < .07$. The corresponding correlation with the manipulated posttest IAT scores was negligible $r = -.06$ (Germans – Turks). There were no gender differences in either explicit or implicit measures. Posttest latencies on incompatible trials correlated substantially with pretest latencies on incompatible trials ($r = .67$) and, to a lesser extent but with the same positive sign, with prestest latencies on compatible trials ($r = .39$), reflecting that regular individual differences in processing speed contribute to IAT performance.

In summary, the first experiment clearly demonstrated that explicit instructions to

respond on a German-Turk intergroup IAT in a contra-prejudiced way led to strong and highly significant reaction time reversals, at all levels of graded instruction. We believe that the present instructions in all conditions were representative of faking in reality. Individuals who intend to influence their test score in a desired direction would inform themselves about test scoring or ask an expert what test strategy is most likely successful. Given such information, we found clear evidence that IAT respondents *can* in principle exert considerable control. That other studies, using weak faking instructions, were less successful is subject to the well-known problems of null findings.

Experiment 2

The aim of the second experiment was to replicate successful faking under controlled experimental conditions and to rule out self-selection of good cheaters in an Internet study announced as "How to beat a psychological test". In Experiment 2 we also pursued the practical question of whether faking can be detected. In Experiment 1, a strong reversal of IAT scores could be obtained especially in the explicitly informed condition, but mainly at the price of conspicuous long latencies on compatible trials. Faking of this kind might be easily detected. The almost equal success at faking in the two less blatant conditions may thus be the stronger finding. When concealing faking is the goal, weaker instruction may in fact be more effective than too blatant instructions.

Accordingly, we tried to replicate the *explicitly informed* and the *uninformed* conditions but we also included a completely naïve condition, which was not even exposed to an IAT pretest (*exploratory* condition). Thus, faking in this condition required participants to invent a successful strategy spontaneously, on the first encounter. Instructions in all conditions were supplemented by the reminder to fake in an inconspicuous manner. For a check on this additional task, we provided two IAT experts with the latency data from 24 participants and asked them to distinguish faked from authentic latencies. If it turned out that participants can not only produce intended outcomes, but can also conceal their manipulative intent, this

would be cogent evidence for successful faking.

Method

Participants and Design. Seventy-four students (68% male) from Mannheim University (Germany) were recruited. Their average age was 23.4 years (*SD* = 2.5). Participants were invited to participate in a study on "How to beat a psychological test"; they received 3 Euro payment. The experiment took place in small groups in a computer lab.

Materials and Procedure. Materials were almost identical to Experiment 1; the same server delivered the web pages and the JAVA-applet providing the IAT. After reading about the study purpose and a general introduction, participants received instructions for the strategic knowledge treatment. These included, in all three conditions, the following additional statement: "In former experiments, participants were very well capable of faking the IAT. In this study, a researcher will examine your data for indications of manipulation. Your task, therefore, is to influence the test outcome in a way that no prejudice against Turks can be inferred from your results, and that others are unable to recognize the manipulation and convict you as a cheater." This treatment was meant to reduce any blatant faking attempts. While the *explicitly informed* and the *uninformed* condition received faking instructions between the pretest- and the posttest-IAT, participants in the *exploratory* condition read the same instructions as the uninformed condition, but before their first and only IAT.

Results and Discussion

Two participants had to be eliminated because they were not in full command of the German language. Another 13 participants were excluded because they produced more than 30% erroneous responses on at least one IAT. Of the remaining 59 participants, 16 belonged to the *explicitly informed* condition, 19 to the *uninformed* condition, and 24 to the completely-naïve *exploratory* condition. The same trimming criteria (minimum = 300 ms; maximum = 3000 ms) were applied as in the first study. Latencies of erroneous responses were treated as missing data. Influences of participant gender were negligible.

Error Rates. Accuracy was not reduced as a result of faking attempts. In the *uninformed* condition, the mean error rate for compatible trials dropped from .08 on the (authentic) pretest to .04 on the (faked) posttest, and for incompatible trials from .10 to .07. Significant main effects for trial blocks, $F(1,15) = 4.66$, $p < .05$, and for test repetitions, $F(1,15) = 8.48$, $p < .05$, reflect the disadvantage of incompatible trials but actually an accuracy gain on the posttest involving faking. In the *explicitly informed* condition, the test repetitions main effect, $F(1,18) = 7.41$, $p < .05$, came along with a similar interaction, $F(1,18) = 13.23$, $p < .01$, as for the same condition in Experiment 1. Error rates slightly decreased for incompatible trials (from .11 to .08) but increased for compatible trials (.06 to .08). In the *exploratory* condition, error rates for compatible (.13) and incompatible trials (.16) were generally elevated.

Latencies and IAT Scores. Table 3 provides an overview of means and standard deviations of latencies and IAT scores, as a function of experimental conditions. Again, a positive sign of an IAT score indicates slower responding on incompatible trials. A negative score on the second test indicates successful manipulation.

As in the Internet study, a majority of participants attained a positive pretest score so that the premise for a faking experiment was met. Of all 35 participants who conducted a pretest (in the *explicitly informed* and *uninformed* conditions), 26 received a positive IAT score. On the posttest, in contrast, 28 participants from the same set of 35 were able to produce a negative score. Twenty of the 26 participants with a positive pretest score changed the sign of their score. However, participants in the completely-naïve *exploratory* condition were obviously not successful. Their average IAT score was not significantly different from the pretest scores in the other two conditions, $t(57) = 0.66$, but significantly higher than the negative posttest scores in the other two conditions, $t(57) = 2.25$, $p < .05$. This failure of the completely naïve participants suggests, not surprisingly, that prior experience with at least one IAT is necessary to plan and execute test manipulation. However, the results in the other two conditions reflect the relative ease of faking, even when no strategy is offered.

A test repetitions (pretest vs. posttest) x instruction conditions (explicitly informed vs. uninformed) ANOVA yielded a significant main effect for test repetitions, $F(1,33) = 9.16$, $p < .005$, but no other effect (both Fs < 1). Again this reversal was mainly due to a slow-down on compatible trials, although there was also a notable speed-up on incompatible trials (see Table 3). Table 4 shows that the change in response speed is quite equally distributed across stimulus subsets (German items with positive and negative connotations, Turkish items with positive and negative connotations, positive valence terms, and negative valence terms), indicating a generally simple, non-sophisticated strategy. Rather than responding differentially to trials that have different implications, it appears that participants simply changed their overall speed of responding on the two combined trial blocks.

Detectability of faking. Can faking be detected when experienced IAT testers inspect the latency data representing authentic and manipulated IAT responses? If the variance between naïve responding on the pretest and controlled responding on the posttest were conspicuously high, relative to normal interpersonal variance in IAT latencies, then testers should have a good chance to identify fakers.

To examine whether faking can remain undetected, we used a quasi-random selection of 24 data vectors of participants stemming from Experiment 2, including all latency data from 12 authentic (pretest) and 12 manipulated (posttest) IATs. Data vectors were randomly chosen such that half of the 12 *manipulated sets* were stereotype-consistent (i.e., positive IAT score, pointing to anti-Turk prejudice), whereas the other half was stereotype-inconsistent (i.e., negative score, pointing to anti-German attitude). Within each half, all three conditions (explicitly informed, uninformed, exploratory) were equally represented. Of the 12 *authentic sets*, one third was stereotype-consistent, one third stereotype-inconsistent, and one third almost neutral (exhibiting a minimal IAT effect in the range of –18 to +79 ms). Within each of the resulting subsets, the ordering of compatible and incompatible blocks was counterbalanced. The rules of the detection task were intended to simulate the case that the

expert detector neither knows which data sets are manipulated nor whether the person's implicit attitude were pro or anti Turkish. Accordingly, faker detectors were not told which trial block in the data array reflected compatible trials and which block reflected incompatible trials, thus assuming that the initial attitude of a participant is unknown and a block cannot be declared "compatible" or "incompatible" beforehand. Otherwise, faking could have been trivially inferred or guessed whenever latencies on compatible trials were relatively high. Two experts experienced on several IAT studies served as expert detectors.

Ignorant about trial block condition as well as respondents' attitudes, one expert classified three data vectors as faked. For the remaining respondents, he did not make any judgment, because the individual mean response latencies and standard deviations seemed to converge with former IAT studies. Two of the three respondents classified as faking in fact came from the authentic subset. The other expert applied a multiplicative algorithm based on variance, response latencies, and error rates. Only 14 out of 24 respondents (58%) were classified correctly.

Pretest IAT scores correlated at a modest but significant level with differences in the explicit attitude ratings for Germans minus Turks ($r = .27$, $p < .05$). Posttest scores in the *explicitly informed* and *uninformed* conditions did not ($r = .16$). IAT scores in the *exploratory* condition correlated *negatively* with the explicit attitude difference ($r = -.33$, $p < .05$). We refrain from interpreting this unpredicted finding.

Experiment 3

The notable finding of Experiment 2 that faking failed in the exploratory condition, without any pretest experience, raises a critical question: Could the impact of faking be confounded with the impact of a pretest? Maybe, the reduction or even reversal of IAT effects on the posttest was not caused by the intention to fake deliberately but by the experience of two IATs in close succession. Although this possibility appears to be unlikely, given that IAT effects have not been erased in retest reliability studies (Bosson, Swann, & Pennebaker, 2000;

Cunningham, Preacher, & Banaji, 2001; Dasgupta & Greenwald, 2001; Gemar, Segal, Sagrati, & Kennedy, 2001; Greenwald & Farnham, 2000; Rudman, Ashmore, & Gary, 2001; Steffens & Buchner, 2003), it needs to be ruled out empirically, with regard to the present study context. We thus conducted a third experiment including three conditions: a *control* condition including pretest and posttest but without any instruction to fake; a replication of the *explicitly informed* condition of Experiments 1 and 2; and a slightly modified version of the *exploratory* condition of Experiment 2, again without any hint as to how to fake and how the IAT works, but this time including a pretest. If the experience of a pretest alone were sufficient for successful faking, then all three conditions of Experiment 3 should produce negative IAT scores on the posttest. However, if intention to fake is a necessary condition, faking should be successful in the *exploratory* condition (instructed to fake) but not in the *control* condition (instructed not to fake).

Method

Participants and Design. In order not to base our conclusions on specifically skilled test cracker samples, the recruitment strategy (via emails and hyperlinks) highlighted the need for participation in the development of accurate tests. The topic of prejudice was only briefly mentioned on the starting page, but not the target group (Turks). Faking instructions were not administered before the pretest was finished. Participants were randomly allocated to one of three experimental groups: *control* condition, *explicitly informed*, and *exploratory*. Approximately half of the participants in each group received the compatible trial block before or after the incompatible trial block. After discarding one participant because of too many errors (as in Experiment 1 and 2), 32 male and 27 female Internet participants remained in the sample.

Materials and Procedure. The same Internet-experimentation software was used as Experiment 1. Pretest instructions of all conditions simply introduced the IAT as a measurement device for assessing prejudice via response latencies. Posttest instructions in the

control condition stated: "You are to repeat the test another time. Please concentrate as much as you did in the first round. Do not try to fake or influence the results."[1] *Explicitly informed* participants received instructions identical to those of Experiment 1 and 2. The instructions of the *exploratory* condition, as far as the posttest was concerned, were identical to that of Experiment 2. To increase the generality of results, male and female Christian (matched for first letters) names were used as target stimuli, replacing the stereotypical nouns (cf. Table 1).

Results and Discussion

The mean latencies and IAT effects (as summarized in Table 5) provide a straightforward answer to the research question. Practice gains from two successive IATs cannot account for successful faking. Although pretests were run in all three conditions, posttest IAT scores were only reversed when participants were instructed to fake intentionally (*explicitly informed*, mean IAT score = –171 ms, and *exploratory* condition, –177 ms) but not when the posttest was a mere replication task, without any instruction to fake (*control* condition, +73 ms). IAT scores decreased significantly from pretest to posttest in the *explicitly informed* condition, $t(18) = 3.28$, and in the *exploratory* condition, $t(18) = 3.52$, but somewhat weaker in the *control* condition, $t(20) = 2.81$, all ps $< .01$. A planned contrast confirmed that the decrease was significantly stronger in the former two conditions than in the latter condition, $t(56) = 2.37$, $p = .022$, corrected for unequal variances this contrast is even more significant, $t(52) = 2.91$, $p = .005$.

General Discussion

The three experiments reported here provide a clear-cut answer to the question of whether controlled responding or faking on the IAT is possible. The answer is an unqualified Yes. Moreover, the degree of voluntary manipulation that was possible was quite impressive and was not dependent on explicit instructions of how to accomplish successful faking. A vast majority of participants found out themselves that one only has to slow down on the compatible trial block if one intends to produce a negative IAT difference score. All

instruction conditions, even the most blatant one, were not too different from the advantage that real test persons take from freely available information about tests and their scoring rules.

Only one condition failed and the contrast between this condition and the others was so obvious that the crucial boundary between success and failure can be assumed to lie somewhere between the *uninformed*, but IAT-experienced condition and the completely naïve, *exploratory* condition. The most plausible interpretation of the crucial distinctive feature between these two conditions is that participants must have experience with at least one prior IAT in order to apply a useful strategy. However, Experiment 3 clearly showed that pretest experience per se is not sufficient. Even with a pretest, posttest IAT scores were only reversed when respondents were intended to fake on instruction.

The strategy leading to successful faking was simple and straightforward: Don't try to speed up the limited performance on incompatible trials, but you will find it easy to slow down on compatible trials. In quantitative terms, faking took participants about 300 ms longer on compatible trials than naïve responding. However, this slowdown was not too obvious against the background of normal performance variation. The standard deviation in mean response latencies on compatible trials across non-faking participants was in the range of 150 ms to 300 to ms, affording efficient camouflage for manipulation through retarded responding. For two experienced experts, it was virtually impossible to identify IAT fakers among 24 respondents above chance. Thus, fakers not only managed to produce desired test outcomes but also to conceal their controlled responding effectively.

Although the dominant strategy was to slow down on compatible trials, we also observed a surprising speed-up on incompatible trials. Other than the slow-down on compatible trials, however, this speed-up also occurred in the control condition of Experiment 3, when respondents did not intend to and did not succeed in faking but merely practiced the same IAT twice, in close temporal succession. It can be concluded, then, that a successful faking strategy – which did not require any instruction but could be applied spontaneously –

was to retard responding on congruent trials. At the same time, an additional, unintended attenuation of IAT effects resulted from practice alone. Eventually, then, it would appear that volitional processes – such as the motive to fake or practice on the test – can not only bring about reversed IAT scores but even a genuine enhancement of seemingly resource-limited performance of the same kind as it was recently demonstrated for other Stroop-like tasks (Kuhl & Kazén, 1999).

Author Note

Klaus Fiedler and Matthias Bluemke, Psychological Institute, University of Heidelberg.

The present research was supported by a personal grant awarded to the first author by the Deutsche Forschungsgemeinschaft. We gratefully appreciate the help of Henning Plessner, Claude Messner, Christoph Schmitz, and Christian Unkelbach.

Correspondence concerning this article should be addressed to Klaus Fiedler, Psychological Institute, University of Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Germany; e-mail: kf@psychologie.uni-heidelberg.de.

Footnotes

[1] If they wished, participants in the control condition were allowed to try to fake their results in another round (but were then excluded from the data pool).

References

Banaji, M.R. (2001). Implicit attitudes can be measured. In H.L. Roediger, J.S. Nairne, I. Neath, A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48*, 145-160.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242-261.

Blair, I., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81*, 828-841.

Bluemke, M. & Friese, F. (in press). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*.

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.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*, 163-170.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800-814.

Egloff, B. & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83*, 1441-1455.

Fiedler, K., Messner, C. & Bluemke, M. (2005). *Unresolved problems with the "I", the "A" and the "T": Logical and psychometric critique of the Implicit Association Test (IAT).*

Manuscript submitted for publication.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition. *Annual Review of Psychology, 54*, 297-327.

Florack, A., Scarabis, M., & Bless, H. (2001). When do associations matter? The use of automatic associations toward ethnic groups in person judgments. *Journal of Experimental Social Psychology, 37*, 518-524.

Gemar, M. C., Segal, Z. V., Sagrati, S., & Kennedy, S. J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. *Journal of Abnormal Psychology, 110*, 282-289.

Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Violence viewed by psychopathic murderers. Adapting a revealing test may expose those psychopaths who are most likely to kill. *Nature, 423*, 497-498.

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., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Social Psychology, 70*, 177-187.

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

Kuhl, J., & Kazén, M. (1999). Volitional facilitation of volitional intentions: Joint activation of intention memory and positive affect removes Stroop interference. *Journal of Experimental Psychology: General, 128*, 382-399.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology, 81*, 842-855.

Messner, C. & Freytag, P. (2003). *The German-Turk IAT and the problem of dimensional overlap*. Unpublished research, University of Heidelberg.

Mitchell, J. P., Nosek, B. A., & Banaji, M. A. (2003). Contextual Variations in Implicit Evaluation. *Journal of Experimental Psychology: General, 132*, 455-469.

Neumann, R., & Seibt, B. (2001). The structure of prejudice: Associative strength as a determinant of stereotype endorsement. *European Journal of Social Psychology, 31*, 609-620.

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.). *Psychological Experiments on the Internet* (pp. 89-117). San Diego: Academic Press.

Rothermund, K. & Wentura, D. (2003). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*, 139-165.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: The *malleability* of *implicit* prejudice and stereotypes. *Journal of Personality and Social Psychology, 81*, 856-868.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition, 17*, 437-465.

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology, 49*, 1-16.

Steffens, M., Lichau, J., Still, Y., Jelenec, P., Anheuser, J., Goergens, N. K., & Hülsebusch, T. (2004). Individuum oder Gruppe, Exemplar oder Kategorie? Ein Zweifaktorenmodell

zur Erklärung der Reaktionszeitunterschiede im Implicit Association Test (IAT) [Individual or group, exemplar or category? A two-factor model for the explanation of reaction time differences in the Implicit Association Test (IAT)]. *Zeitschrift für Psychologie, 212*, 57-65.

Steffens, M. & Plewe, I. (2001). Items' cross-category associations as a confounding factor in the Implicit Association Test. *Zeitschrift für Experimentelle Psychologie, 48*, 123-134.

Wittenbrink, B., Judd, C.M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815-827.

Table 1

Stimulus materials used for the German-Turk IAT: Stereotypical nouns (Experiment 1 and 2) and Christian names (Experiment 3) as target stimuli plus evaluative stimuli (Experiment 1-3)

German Stimuli	Turkish Stimuli	Positive Valence	Negative Valence
Environmentalism +	Basar +	Loyal	Ugly
Poets +	Belly dancing +	Happy	Ill
Christmas +	Vapor bath +	Honest	Sad
Democracy +	Hospitality +	Affectionate	Cruel
Xenophobia –	Dirt –	Cheerful	Brutal
Heart attack –	Macho –	Talented	Stinking
Skinheads –	Death penalty –	Tender	Poisonous
Nazis –	Torture –	Glorious	Slimy
Yvonne	Yilmaz		
Saskia	Selma		
Almut	Aysel		
Nina	Nuray		
Hanno	Hakan		
Markus	Mehmet		
Erich	Emre		
Moritz	Murat		

Table 2

Means and standard deviations of latencies and latency differences (in ms) as a function of experimental conditions in Experiment 1

	Condition					
	Uninformed		Implicitly Informed		Explicitly Informed	
	Mean	SD	Mean	SD	Mean	SD
Pretest						
Compatible Block	1064	278	999	168	1063	185
Incompatible Block	1244	290	1158	188	1272	250
IAT Difference Score	+180	270	+159	174	+209	162
Posttest						
Compatible Block	1375	342	1392	384	1584	447
Incompatible Block	1137	178	981	237	1148	224
IAT Difference Score	-238	423	-411	418	-436	494

Table 3

Means and standard deviations of latencies and latency differences (in ms) as a function of experimental conditions in Experiment 2

	Condition					
	Explicitly Informed		Uninformed		Exploratory	
	Mean	SD	Mean	SD	Mean	SD
Pretest						
Compatible Block	1160	263	1158	278	----	----
Incompatible Block	1292	282	1361	264	----	----
IAT Difference Score	+132	228	+203	171	----	----
Posttest						
Compatible Block	1481	415	1402	396	1282	261
Incompatible Block	1132	298	1144	305	1382	301
IAT Difference Score	-348	303	-258	333	+100	189

Table 4

Mean latencies in the IAT pretest and subset as a function of stimulus subsets, instruction conditions, and trial blocks in Experiment 2

	Condition					
	Explicitly Informed			Uninformed		
	Pretest	Posttest	Difference	Pretest	Posttest	Difference
Compatible Trials						
German +	1000	1303	-303	1042	1325	-282
German –	1250	1389	-140	1196	1354	-157
Turkish +	1065	1382	-317	1013	1351	-339
Turkish –	1207	1508	-301	1290	1490	-200
Positive	1212	1496	-238	1142	1431	-290
Negative	1183	1624	-441	1225	1416	-191
Incompatible Trials						
German +	1211	1148	+63	1250	1232	+19
German –	1253	1125	+128	1194	1068	+125
Turkish +	1153	1098	+55	1194	1057	+137
Turkish –	1405	1299	+106	1358	1381	-23
Positive	1333	1124	+209	1503	1097	+407
Negative	1336	1086	+249	1447	1139	+309

Table 5

Means and standard deviations of latencies and latency differences (in ms) as a function of experimental conditions in Experiment 3

	Condition					
	Explicitly Informed		Exploratory		Control Condition	
	Mean	SD	Mean	SD	Mean	SD
Pretest						
Compatible Block	859	184	896	161	959	267
Incompatible Block	1027	237	1073	217	1141	314
IAT Difference Score	+168	149	+177	205	+182	176
Posttest						
Compatible Block	1090	385	1139	431	862	215
Incompatible Block	919	195	963	233	936	228
IAT Difference Score	-171	405	-177	387	+73	85