Is the Implicit Association Test Immune to Faking?

Melanie C. Steffens

University of Trier, Germany

Abstract. One of the main advantages of measures of automatic cognition is supposed to be that they are less susceptible to faking than explicit tests. It is an empirical question, however, to what degree these measures can be faked, and the response might well differ for different measures. We tested whether the Implicit Association Test (IAT, Greenwald, McGhee, & Schwartz, 1998) cannot be faked as easily as explicit measures of the same constructs. We chose the Big-Five dimensions conscientiousness and extraversion as the constructs of interest. The results show, indeed, that the IAT is much less susceptible to faking than questionnaire measures are, even if no selective faking of single dimensions of the questionnaire occurred. However, given limited experience, scores on the IAT, too, are susceptible to faking.

Key words: implicit cognition, automatic cognitive processes, faking, Implicit Association Test, Big Five

"Physicians would not ask their patients to estimate their own white blood cell count" (McCrae & Costa, 1999, p. 141). However, psychologists mostly rely on self-reports for finding out how dependable or sociable people are – even though we know that, in addition to systematic errors like self-deception and self-enhancement, demand characteristics, evaluation apprehension, and impression management factors influence self-reports (e.g., Furnham, 1986; Nederhof, 1985; Nisbett & Wilson, 1977; Strack, 1994). The employer who wants to know, say, how conscientious a job interviewee is cannot easily disentangle a person's conscientiousness from that person's self-presentation as conscientious. What makes matters worse is: Interviewees know that employers are interested in hiring people high in conscientiousness. Such personality assessment as a preemployment screening procedure seems to be quite popular (see Barrick & Mount, 1991), even if it "provides an almost ideal setting for dissimulation: Job applicants are motivated to present themselves in the best possible light; transparency of items makes it possible to endorse items that will make them look good, and there is little apparent chance of being caught in a lie" (Rosse, Stecher, Miller, & Levin, 1998, p. 635). Consequently, there is ample evidence that faking not only occurs in groups of participants instructed accordingly (e.g., Dalen, Stanton, & Roberts, 2001; Furnham, 1997; Jackson & Francis, 1999), but also, systematic differences between the test scores of job applicants and those of volunteers suggest that faking routinely happens (see Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001, for a review).

If faking introduced a constant additive factor to applicant scores such that their rank order remained, the effect would not be of much concern because it is the rank order of applicants that is crucial. However, Rosse et al. (1998) showed that faking also affects hiring decisions (but see, e.g., Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). According to Rosse et al., correcting for response distortion does improve criterion-related validity, and, for methodological reasons, suppression effects of faking on validity are "extremely hard" (p. 636) to detect because these validities commonly are of only moderate magnitude. These authors were able to show in a large study that among the small percentage of applicants "hired" the majority had extreme scores on a response-distortion measure, and taking those scores into account had a practically significant effect on

Requests for reprints should be sent to Melanie Steffens. The writing of this article was supported by a grant from the Deutsche Forschungsgemeinschaft (German Science Foundation). I thank Steve Arendt, Alex Besemer, Julia Breuer, Pascale David, Susanne Engelke, Bettina Erdmann, Katrin Modabber, and Christine Wurster for their great help with data collection and the preparation of materials, and Petra Jelenec and Stefanie Schulze König for valuable comments on earlier drafts of this article.

DOI: 10.1027/1618-3169.51.3.165

who was hired. Given findings such as these, it has been concluded that one "should begin to consider alternative models or methods of test construction that could be employed to develop personality measures that are less susceptible to faking than the currently used inventories" (Stark et al., 2001). This, of course, seconds a request already made by Cattell (Cattell, 1955; Cattell & Warburton, 1967).

One group of candidates for tests not susceptible to faking can be derived from the literature on indirect or implicit testing in cognitive psychology and social cognition (e.g., Fazio & Olson, 2003). Such tests would measure to what degree people automatically endorse certain traits, thus delivering somewhat different information than questionnaires which assess more controlled, reflected aspects of the self-concept that are open to introspection. To the degree that the behavior measured by an implicit test is automatic, it should occur without intention (e.g., Hasher & Zacks, 1979). Thus, implicit tests have generally been assumed to be beyond individuals' control. However, empirical evidence concerning this is scarce. If a test becomes transparent to the test-taker, faking may occur. The goal of the present research was to compare faking a test of automatic cognition, the Implicit Association Test (IAT, Greenwald et al., 1998), to faking explicit scales measuring the same construct: Are participants unable to fake the IAT? Are they, in contrast, able to selectively fake explicit scales?

In order to investigate faking on tests that may be of interest for practical applications, we focused on two of the *Big Five* (see, e.g., McCrae & Costa, 1999) personality traits: conscientiousness and extraversion. We developed the respective IATs. As an explicit test, the NEO-FFI was administered, which is apparently used routinely, and is evidently susceptible to faking (see Furnham, 1997; Scandell & Wlazelek, 1999).

The Implicit Association Test

In IATs, instances of four concepts are presented in a random order and require only two different reactions. The IAT's rationale is that people are able to react fast if a pair of closely associated concepts requires one reaction and another pair, another reaction. In contrast, if closely associated concepts require different reactions, reactions should be relatively slow. The difference in reaction times between these two tasks, the IAT effect, is taken as an indicator of the degree of association between concepts. In the present case, consider a person's reaction time in a task in which (a) instances of *self* (e.g., mine) and instances of *conscientious* (e.g., well-organized) require the same reaction, and (b) instances of *others* (e.g., your) and *not conscientious* (e.g., lazy) require a different reaction (henceforth, the self+conscientious task). The average reaction time in this task is compared to that in a task in which instances of self and not conscientious require the one response, and instances of others and conscientious, the other (henceforth, the self+not conscientious task). People who react faster in the self+conscientious than in the self+not conscientious task seem to more closely associate self and conscientious than people showing a smaller, no, or a reversed reaction time difference between tasks (Banaji, 2001). In other words, faster reactions in the self+conscientious task show the implicit endorsement of conscientiousness.

There is an ever-growing body of evidence as to IATs' validity for measuring automatic cognition, stemming, for instance, from known-groups approaches (Banse, Seise, & Zerbes, 2001; Greenwald et al., 1998; Kühnen et al., 2001; Rudman, Greenwald, Mellott, & Schwartz, 1999; Steffens, in press; Teachman, Gregg, & Woody, 2001). Moreover, some studies have found that IATs predict behavior (McConnell & Leibold, 2001; Rudman & Glick, 2001; Steffens, Günster, Hartmann, & Mehl, 2004). Finally, the IAT was the only one of seven implicit self-esteem measures that correlated significantly with several criterion variables (Bosson, Swann, & Pennebaker, 2000). Moderate correlations between IATs and related explicit measures are also found often (e.g., Steffens & Buchner, 2003; Steffens & Plewe, 2001).

In addition to the body of literature in which IATs were applied and indicators of their validity were found, a considerable number of studies have tested confounds in the IAT effect (e.g., Brendl, Markman, & Messner, 2001; Mierke & Klauer, in press; Ottaway, Hayden, & Oakes, 2001; Steffens & Plewe, 2001). What is most important in the present case, if all the instances of one pole of the attribute dimension (e.g., conscientious) were positive and all those of the other pole (not conscientious) were negative, then a participant may react fast in the IAT by attending only to stimulus valence and ignoring the attributes (e.g., they might use the rule that "all the self-related and positive stimuli require the left reaction", instead of the rule that "all the self-related and conscientiousness-related stimuli require the left reaction"). For participants using that rule, the IAT would be turned into a more general self-esteem IAT (see Steffens et al., 2003, for an extended discussion). It is possible to avoid this confound by selecting some more extreme instances of the presumably more positive concepts and less extreme ones for the more negative ones (e.g., "pedantic" for conscientious).

In the first study in which the IAT was adapted to the measurement of personality facets (in that case, shyness, Asendorpf, Banse, & Mücke, 2002), it was

argued that implicit measures should be more robust against deception than explicit measures. Indeed, participants instructed to try to appear non-shy were unable to fake their IAT scores. This finding is in line with others showing that participants did not distort scores on an anxiety IAT when instructed to make a good impression (Egloff & Schmukle, 2002), and showing that people could not deliberately fake positive attitudes towards homosexuals (Banse et al., 2001), insects, or black people (Kim, 2003). It seems that Kim (2003) found absolutely no faking effects, with a nonsignificant effect in the direction opposite to that expected for faking positive attitudes towards insects or weapons, and an effect of less than $d = .09$ for racial attitudes. In contrast, Asendorpf et al. (2002) reported a nonsignificant faking effect of $d = .39$ and Egloff and Schmukle (2002), $d = .16$. Similarly, the size of the nonsignificant faking effect in the study of Banse et al. (2001) was $d = .13$ and $d = .25$ in the IAT with nonalternating and alternating target-attribute trials, respectively[1], and an effect of $d = .23$ was observed in an as-yet-unpublished IAT faking study (Asendorpf, Banse, & Schnabel, 2003). All of these are nonsignificant effects in the expected direction that are smaller than medium effects (see Cohen, 1977).

In addition to the conclusion that participants cannot control the IAT, several other explanations for these null findings are conceivable. Most importantly, participants in those studies had no experience at all with the IAT, and faking becomes more likely with test experience (see Dalen et al., 2001). Therefore, we were interested in testing whether faking a personality IAT is possible after participants have taken it once. The rationale behind this is that we wanted to know how easy it would be to coach people to show certain IAT effects. A particularly easy means of coaching would be to let people take the test once and then ask them to fake it in a second trial. Also, once implicit tests become established diagnostic instruments, it is quite likely that people will take a given test more than once.

Experiment 1: Faking Conscientiousness

We selected implicit and explicit *conscientiousness* as the first dimension to be tested because conscientiousness is one of the most important trait motivation variables in personnel psychology (Mount, Barrick, & Strauss, 1994). After taking the conscientiousness IAT once (base rate trial), we asked participants to fake being either conscientious or not conscientious (faking trial). An instruction to fake the IAT might lure participants into committing more errors during the IAT, instead of manipulating their reaction times; therefore, we supplemented reaction-time analyses with analyses of error differences between the two IAT tasks. Error differences between IAT tasks usually show similar effects as reaction time differences, but the associated effect sizes are typically smaller. In addition, we tested whether participants would be able to fake selectively the explicit test dimension they were asked to. In order to separate more controlled from more automatic aspects of conscientious behavior, we conceptualized a behavioral indicator of conscientiousness. We handed participants a concentration test *after* they had received credit for their participation in the experiment and asked them to fill it in and drop it in a box within the next week. The test used, the d2 (Brickenkamp, 1981), asks participants to cross out each d accompanied by two dots among similar-looking letters with various numbers of dots. We reasoned that returning the test should imply a controlled (explicit) aspect of an individual's conscientiousness. In contrast, the number of errors committed when filling in such a monotonous test should reflect a more automatic, spontaneous aspect of conscientiousness: People high in spontaneous conscientiousness are free to fill in the test slowly and conscientiously, thus committing few errors, whereas people low in spontaneous conscientiousness who work on assigned chores in a less conscientious way may go through the test fast and thus commit more errors. More generally speaking, with regard to speed-accuracy tradeoffs, spontaneous conscientiousness should imply reducing speed in order to increase accuracy.

Method

Materials

For the IAT, the concepts used were *self*, *others*, *conscientious*, and *not conscientious*. Five items were selected as instances of each of the concepts. All stimuli presented in the IAT are listed in the Appendix. Instances for the concept self were pronouns clearly related to one's person or group, and for the concept others, pronouns clearly related to another person or group. Instances for the concepts conscientious and not conscientious were selected from the German version of the NEO-Five Factor Inventory (NEO-FFI, Borkenau & Ostendorf, 1993), which consists of 60 items measuring the five dimensions of the FFM with 12 items each, or they were generated by the experimenters. We made sure that there was no perfect confound between the to-be-measured

[1] I thank Rainer Banse for making available the standard deviations so I could compute effect sizes.

dimension and stimulus valence. Still, the stimuli for conscientious were probably more positive on average than those for not conscientious. Whereas this should boost the self+conscientious association found in the IAT across all participants (see Steffens et al., 2003 for details; Steffens, Lichau et al., 2004), this does not pose a problem if we refrain from interpreting the absolute size and direction of the IAT effect.

A computerized version of the NEO-FFI with the items presented in their original order was administered. In addition, the d2 was used, which is normally a timed test of selective attention (Brickenkamp, 1981).

Procedure

The procedure largely followed that of Steffens and Plewe (2001). Participants were tested individually in experimental cubicles equipped with iMacs. The Mac-IAT computer program (Steffens, 1999) was used. In order not to confound procedural features with person or group effects, that is, in order to keep the procedure as identical per participant as possible (see Banse et al., 2001, for a discussion), each participant received IAT items in the same random order. Words were sampled without replacement from the list of instances. The IAT consisted of 3 practice tasks (10 trials each) and 2 critical discrimination tasks (2 blocks of 60 trials each). The first critical task was the self+not conscientious task in which the left response key was to be pressed for instances of self and of not conscientious, whereas instances of others and conscientious required the right key. Similarly, in the self+conscientious task, the left response key was to be pressed for instances of self and of conscientious, whereas instances of others and not conscientious required the right key. This IAT task order works against finding the expected self+conscientious association (see Greenwald et al., 1998). After the first IAT, the NEO-FFI items were presented with a five-point scale.

For the subsequent faking trial, participants were randomly assigned to the conditions conscientious or not conscientious. Ostensibly, their impression management capabilities were under scrutiny. The not conscientious instruction asked them to imagine they were interested in renting a room in an apartment-sharing community. People in the community were totally cool and would not rent to bourgeois people. The least conscientious person would be selected on the basis of two tests, the same ones as they had just completed. Participants were further asked to imagine that they urgently needed a room, and they should thus fake the tests as best as could in order to appear not conscientious. A reward was offered to the participant appearing the least conscientious. In the conscientious condition, the instruction was modified in so far as participants should imagine they were applying for a job where the employer was most interested in conscientious work. All critical passages were analogously modified ("The most conscientious person…"). The instructions did not mention other personality dimensions and thus there were no directions as to replying to statements irrelevant to conscientiousness. The German translation of conscientiousness ("Gewissenhaftigkeit") is a commonly used trait, so there was no need to explain its meaning.

The IAT was then administered again. Before the second NEO-FFI, participants were reminded of the pretence. Finally, they were given a questionnaire including scales for subjective ratings of successful faking and they were asked to describe how the IAT worked. The d2 test was handed to them in the end.

Design and Participants

The main dependent variables were the IAT effect (the difference between the self+not conscientious and the self+conscientious IAT task) and scores on the five NEO-FFI dimensions. Independent variables were instruction (faking being conscientious, or not conscientious) and trial (base rate vs. faking). We expected the instruction effect to be considerably larger than a conventional "large effect" (Cohen, 1977). Given error probabilities of $\alpha = \beta = .05$, 46 participants were needed to detect an effect of $f = .55$. Data were collected from 48 participants, 25 of them were placed in the conscientious condition; 40 were female. Participants were psychology students at the University of Trier, who received credit for participating. Their mean age was 23 years ($SD = 4$).

Results

All significance tests in the present studies were conducted with $\alpha < .05$. Therefore, individual p values are omitted for statistically significant effects. Instead, in order to evaluate the effect sizes of statistically significant effects, R^2_p is reported (see Cohen, 1977), which is the proportion of variance explained by a given factor in relation to the variance not explained by any other factor. Where applicable, the Pillai-Bartlett V is reported as a multivariate measure of effect size (see Bredenkamp & Erdfelder, 1985).

IAT Analyses

The IAT D effect was computed (Greenwald, Nosek, & Banaji, 2003). Specifically, no reaction times

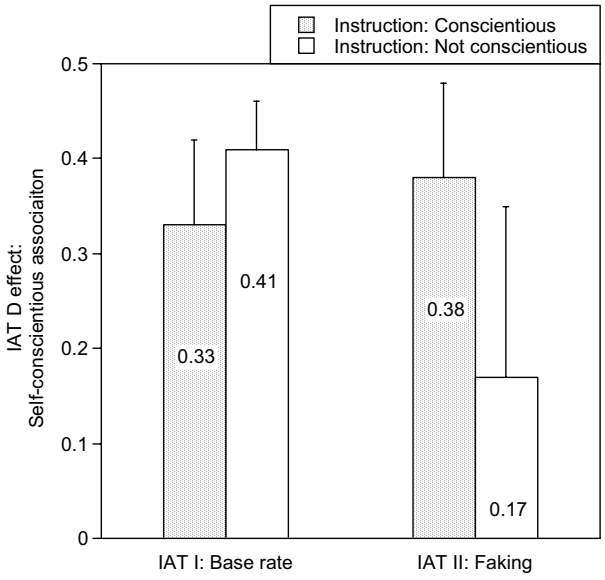

Figure 1. Experiment 1: Mean IAT D effects at base rate and at faking, separately for the instruction conditions asked, at faking, to appear conscientious or not conscientious. Error bars show standard errors of means.

were excluded from analyses nor recoded to other values; an error penalty of 500 ms was added to the reaction time of each trial in which an error was committed; and for each participant the difference between the mean reaction times in the congruent and the incongruent IAT task was divided by their overall standard deviation. The mean reaction times for each of the critical IAT tasks can be obtained from Table 2.

Internal consistencies of the IAT effects were calculated by treating the average IAT D effect for each stimulus as one scale item (see Steffens & Buchner, 2003). The internal consistencies for the resulting 20-item scale were $\alpha = .88$ at base rate and $\alpha = .95$ at faking. As Figure 1 shows, IAT effects were positive, that is, our participants reacted faster in the self+conscientious task than in the self+not conscientious task, automatically endorsing conscientiousness. Reassuringly, there is hardly a difference between instruction conditions at base rate. At faking, participants instructed to appear conscientious again showed a strong self+conscientious association. That IAT effect was smaller, but not reversed, for participants instructed to appear not conscientious.

A 2 × 2 ANOVA with the instruction condition as the between-subjects variable and trial (base rate vs. faking) as the repeated-measures variable on the IAT effect showed neither a main effect of trial ($F < 1$) nor of instruction ($F < 1$), nor a statistically significant interaction of both factors, $F(1, 46) = 2.26$, $p = .14$. The effect size of the nonsignificant difference between instruction conditions at Faking was $d = .32$. If, instead, difference scores between base rate and faking are computed, the between-groups effect size of the faking effect was $d = .44$ (n. s.).

Turning to our supplementary analyses, as an inspection of the errors made in each IAT tasks shows (see Table 2), at faking, participants instructed to appear conscientious made more errors in the self+not conscientious task, whereas participants instructed to appear not conscientious made somewhat more errors in the self+conscientious task. At base rate, participants in both instruction conditions made more errors in the self+not conscientious task than in the self+conscientious task. Accordingly, a 2 × 2 × 2 ANOVA of the errors in the congruent and incongruent IAT task at base rate and faking, with instruction condition as the between-subjects factor, showed interactions of trial and instruction, $F(1, 46) = 9.10$, $R^2_p = .17$, and of IAT task and instruction, $F(1, 46) = 5.53$, $R^2_p = .11$ (all other Fs < 3.79). It thus seems that participants tried to fake the IAT by making more errors. The effect size associated with the error difference between the instruction conditions was $d = .63$.

Explicit Ratings

Before scale scores were computed, internal consistencies were computed. The base rate values were all satisfactory and are shown in the diagonal of the upper half of Table 1. At faking, all internal consistencies were again high enough (all α values above .68) to summarize the variables in scales for each instruction condition.

The left half of Table 2 shows scores on each dimension of the NEO-FFI split by trial and instruction. As expected, at base rate, participants in the instruction conditions did not differ much on any of the NEO-FFI dimensions. After the different instructions were given, the picture is more complicated. The hypothesis that participants are able to *selectively* fake conscientiousness does not seem to be supported. However, reported conscientiousness changes as expected, and more dramatically than the other dimensions. A 2 (Trial) × 2 (Instruction) × 5 (NEO-FFI Dimension) MANOVA confirmed an interaction of trial and instruction, $F(1, 46) = 50.72$, $V = .52$. All other main effects and interactions were also statistically significant (all Fs > 16.33). A separate MANOVA showed that there was no effect of instruction at base rate, $F < 1$. However, there was an overall instruction effect at faking, $F(5, 42) = 82.64$, $V = .91$. Univariate tests for each dimension showed large instruction effects for conscientiousness, $F(1, 46) = 373.56$, $R^2_p = .89$, as well as for

Table 1. Scale Reliabilities and Correlations Between the Measures at Base Rate in Experiments 1 and 2

Scale	IAT C	N	E	O	A	C
		Experiment 1				
Implicit conscientiousness (IAT C)	.88	.08	−.14	−.11	.09	.23
Neuroticism (N)		.86	*−.31*	−.01	.09	−.13
Extraversion (E)			.71	.12	.13	−.04
Openness (O)				.76	.15	−.22
Agreeableness (A)					.75	.23
Conscientiousness (C)						.77
		Experiment 2				
Scale	IAT E	N	E	O	A	C
Implicit extraversion (IAT E)	.85	−.07	*.18*	.02	−.02	*−.19*
Neuroticism (N)		.88	*−.54*	−.09	*−.21*	*−.28*
Extraversion (E)			.81	*.12*	*.36*	*.15*
Openness (O)				.74	.06	−.06
Agreeableness (A)					.74	.08
Conscientiousness (C)						.85

Notes. Correlations in italics are statistically significant ($p < .05$).

agreeableness, $F(1, 46) = 42.57$, $R^2_p = .48$ (all other Fs < 3.24). The faking effect on the conscientiousness scale reached a Cohen's $d = 5.7$.

Correlations

As the upper panel of Table 1 shows, implicitly and explicitly measured base rate conscientiousness correlated in a medium order of magnitude, but missed statistical significance ($p = .06$). The only significant correlation between NEO-FFI dimensions, given the rather small sample size for detecting correlations, was a negative one between neuroticism and extraversion.

What cannot be inferred from the table, is that the IAT effects at base rate and at faking correlated significantly, $r = .27$. Please note that this correlation was much higher if the traditional scoring algorithm of the IAT was used (Greenwald et al., 1998), $r = .50$, which is in the order of magnitude of test-retest correlations typically observed in IATs (see Steffens & Buchner, 2003). Other than that, all results reported in the present article changed only slightly with traditional IAT scoring.

Behavioral measures of conscientiousness were first, returning the d2, and second, the number of errors committed in the d2. The d2 was returned by 58% of the participants. There was only a tendency for returning the d2 to correlate with explicit conscientiousness at base rate, $r = .19$, $p = .10$, and none with the IAT. For correlations with d2 errors, one participant was excluded whose number of errors was more than 3 SD above the mean number of errors committed. The number of d2 errors correlated significantly with the base rate IAT D effect, $r = -.53$, but not with explicit conscientiousness.

For testing with maximal statistical power (i.e., for the whole sample at once) whether there were relations between the ability to fake and the respective self-reports, the differences between scores at base rate and at faking were z transformed per condition, and the scores in the not conscientious condition were multiplied by (−1). Thus, higher scores show more faking in the required direction. There were statistically significant correlations between faking explicit conscientiousness and the reply to the questions how well one succeeded in faking the explicit test, $r = .24$, as well as between faking errors in the IAT and the reply to the question how well one succeeded in faking the implicit test, $r = .38$. There was no such correlation for the IAT D effect ($r = .13$, $p = .20$). Apparently, participants thought they were faking well if they manipulated errors in the IAT. Faking the IAT D effect correlated with faking conscientiousness in the NEO-FFI, $r = .27$. A content analysis of the replies to the question how the IAT works (coded as 0: no understanding, 1: some understanding, 2: accurate description) showed that replies did not correlate significantly with the degree of manipulation of the IAT D effect, Spearman's $\varrho = .17$ ($p = .12$), even though 18 participants had at least some understanding of the gist of the IAT (along the lines: "if one is conscientious it should be easier to react to the task in which self and conscientious require the same reaction").

Table 2. Mean NEO-FFI Scores and IAT Reaction Times at Base Rate and at Faking in Experiments 1 and 2 (SEMs Are Shown in Brackets), Separately for the Instruction Conditions

	Experiment 1		Experiment 2		
	Conscientious	Not conscientious	Extraverted	Reliability	Introverted
Neuroticism					
Base rate	2.0 (.2)	1.8 (.2)	1.8 (.1)	1.9 (.1)	1.8 (.1)
Faking	.9 (.1)	1.1 (.1)	.9 (.1)	1.9 (.1)	2.9 (.1)
Extraversion					
Base rate	2.4 (.1)	2.5 (.1)	*2.6 (.1)*	*2.5 (.1)*	*2.5 (.1)*
Faking	2.7 (.1)	2.9 (.1)	*3.7 (.1)*	*2.4 (.1)*	*.7 (.1)*
Openness					
Base rate	3.0 (.1)	3.0 (.1)	3.0 (.1)	2.8 (.1)	3.2 (.1)
Faking	2.4 (.2)	2.0 (.2)	2.4 (.1)	2.8 (.1)	2.8 (.1)
Agreeableness					
Base rate	2.4 (.1)	2.7 (.1)	2.7 (.1)	2.7 (.1)	2.6 (.1)
Faking	2.9 (.1)	1.6 (.2)	2.3 (.1)	2.7 (.1)	2.4 (.1)
Conscientiousness					
Base rate	*2.3 (.1)*	*2.4 (.1)*	2.5 (.1)	2.3 (.1)	2.5 (.1)
Faking	*3.7 (.1)*	*.5 (.1)*	2.0 (.1)	2.2 (.1)	3.0 (.1)
IAT					
Base rate RTs					
Congruent	956 (47)	912 (62)	1096 (35)	1010 (40)	995 (34)
Incongruent	1125 (61)	1098 (72)	1033 (27)	920 (35)	959 (30)
Faking RTs					
Congruent	820 (28)	1014 (62)	941 (33)	890 (29)	1081 (40)
Incongruent	988 (51)	1159 (123)	1016 (36)	820 (36)	894 (27)
Base rate errors					
Congruent	5.48 (.56)	5.60 (.56)	7.30 (.71)	8.45 (1.42)	6.49 (.69)
Incongruent	8.76 (1.31)	7.26 (.84)	9.31 (.78)	9.82 (1.29)	7.94 (.79)
Faking errors					
Congruent	3.32 (.56)	10.61 (2.06)	5.69 (.61)	6.64 (1.05)	10.10 (1.4)
Incongruent	7.40 (1.78)	8.26 (1.09)	11.30 (1.72)	6.23 (.89)	6.00 (.67)

Note. Numbers in italics show the critical NEO-FFI dimension. For the IAT scores, the upper numbers are the IAT reaction times (including error penalties) in the self+conscientious (Experiment 1) and self+extraverted (Experiment 2) task, the lower numbers, those in the self+not conscientious and self+introverted task, respectively.

Discussion

Participants did not succeed in faking the conscientiousness IAT to a statistically significant degree. At the same time, the effect size associated with the difference between the IAT effect in the conscientious condition and the not conscientious condition was in between a small and a medium effect according to Cohen's (1977) conventions. Participants were very well able to manipulate the NEO-FFI in order to appear conscientious or not conscientious: Conscientiousness as measured by the NEO-FFI was close to ceiling or bottom at faking. However, additional items of the NEO-FFI varied by instruction, most of all, those related to agreeableness. Thus, while our participants were able to manipulate conscientiousness in the questionnaire, "selectively" is not the most accurate description of that ability, at least given our rather general instructions.

The findings of Experiment 1 are not as clear as one might have hoped – one should neither conclude that participants were able to fake their IAT scores, nor that they were unable to do so. Reaction times at base rate and at faking still correlated, implying that participants could not completely hide their "true" implicit associations when trying to fake the test. The faking effect was more pronounced on errors than on reaction times. Along with the effect size, these findings suggest a rather limited ability to fake the IAT. There are two possible explanations for this.

On the one hand, it could be that in general, participants are bound to assume that the number of errors is the crucial dependent variable in the IAT. On the other hand, their hypothesis with regard to errors could be specific to the concept conscientiousness because facets of that trait are related to completing one's tasks with few errors. If this were so, the relatively small effect of instruction on reaction times could be confined to a conscientiousness IAT, and participants would be better able to fake implicit tests of other traits. Alternatively, the faking effect could, in general, be smaller than we expected, and a larger sample size is needed to detect it. For these reasons, we tested the faking hypothesis with a larger sample and a different FFI dimension in Experiment 2.

Our data allow some conclusions about the quality of the conscientiousness IAT as a measurement instrument. Internal consistency of the IAT at base rate was excellent, as has been found for other IATs (e.g., Steffens & Buchner, 2003; Steffens & Plewe, 2001). In addition, the IAT's tendency to correlate with NEO-FFI conscientiousness, but not with any of the other four NEO-FFI scales hints at convergent and discriminant validity. While the correlation with explicit conscientiousness is not high, it is what we expect if implicit and explicit constructs are related, but different. Finally, scoring higher on implicit conscientiousness correlated with a measure of spontaneous behavior, working on a boring chore more conscientiously.

IAT effects at base rate and at faking correlated substantially and much higher when traditionally scored than given IAT D scoring. With the traditional scoring algorithm, when people were hardly able to manipulate reaction times in the IAT, the test-retest correlation was almost as high as those found without faking instructions (see Steffens & Buchner, 2003). This finding differed for the IAT D effect. Across various analyses, we found that neither the error penalty nor whether the data are log-transformed or not, nor whether outliers are excluded or not much affects the correlation between the IAT effect at base rate vs. faking. Instead, ipsatizing the reaction time difference between IAT tasks had a major influence, making the correlation drop by $r = .14$. In Experiment 2, a control group was included that allowed us to test whether the test-retest correlation in the IAT is lower for the IAT D effect than for an IAT effect computed the traditional way.

Self-report data as to knowledge of the IAT or the ability to fake either the explicit or the implicit test were of limited value, at least given the rather small sample size we had for detecting correlations. Experiment 2 will shed more light on this question.

Experiment 2: Faking Extraversion

The main aim of Experiment 2 was to investigate whether the IAT can be faked with a larger sample and a different dimension of the NEO-FFI. We again hypothesized that participants would be able to fake the implicit and the explicit personality test, with smaller effects on the implicit test, and with additional effects on the other dimensions of the explicit test. In addition, we tested whether knowing that reaction times were the critical element in the implicit test would increase participants' ability to exert control over their test results. The reliability of the implicit and the explicit test was compared via repeated testing in a control group that was not asked to fake. An advantage of the extraversion IAT is that it is rather neutral, that is, on average, there is no self-extraverted or self-introverted association (see Mierke & Klauer, 2004; Steffens & Schulze König, 2003). In addition, with the exception of conscientiousness, extraversion is the most important personality dimension for predicting job success (Barrick & Mount, 1991).

Method

Materials

The target dimension was *extraverted* vs. *introverted*. All stimuli are listed in the Appendix. Other than that, the materials were identical to those used in Experiment 1.

Procedure

The procedure was identical to the one of Experiment 1, with the following exceptions. All participants started with the self+introverted task. At Faking, participants were told to try to appear as extraverted or introverted as they could in both tests. They learned that we were interested in knowing whether they could fake the tests. Participants in the control group were asked to take both tests again because we were interested in the tests' measurement quality.

Design and Participants

The main dependent variables were IAT effects on reaction times and errors (differences between the self+extraverted and in the self+introverted IAT task), and scores on the five NEO-FFI dimensions. The main independent variables were instruction (extraversion vs. introversion vs. control: reliability) and

trial (base rate vs. faking). Half of the participants in the extraversion and in the introversion condition had been informed that reaction times were crucial for faking the first-presented test; the other half received no such hint. Preliminary analyses yielded that this hint did not show a main effect on the IAT D effect or the error IAT effect at faking or enter into any interactions, all Fs < 1.58. For clarity of presentation, the data were therefore collapsed across this factor. Our total sample comprised 125 participants, 54 were in the extraversion, 49 in the introversion, and 22 in the control condition; 100 were female. A somewhat larger than medium effect ($f = .30$) of instruction with $\alpha = .05$ and $1 - \beta = .85$ (see Cohen, 1977) could be detected between the extraversion and the introversion condition. Participants were first-year psychology students at the University of Trier, who received credit for participating. Their mean age was 22 years ($SD = 4$).

Results

IAT Analyses

The 20 different IAT items reliably measured the same construct, $\alpha = .85$ and $\alpha = .93$ at base rate and at faking, respectively. As Figure 2 shows, at Base rate, participants in all three groups reacted somewhat faster in the self+introverted than in the self+extraverted task, which resulted in IAT D effects below zero. At Faking, participants instructed to appear extraverted reacted faster in the self+extraverted task than in the self+introverted task. Participants in the reliability condition showed the same effects as at base rate. Participants instructed to appear introverted showed a large negative IAT effect. A 2×2 ANOVA with instruction condition and trial as independent variables and the IAT effect as the dependent variable confirmed an interaction of both factors, $F(2, 122) = 18.70$, $V = .24$, which qualified the main effect of instruction condition, $F(2, 122) = 11.84$, $R^2_p = .16$ (the other $F < 1.03$).

As tests of simple main effects showed, there was no effect of instruction condition at base rate, $F < 1$, but at faking, $F(2, 122) = 22.40$, $R^2_p = .27$. Pairwise comparisons using a Bonferroni adjustment showed that the faked IAT effect in the extraversion condition was significantly larger than both the effect in the reliability condition and in the introversion condition, and, in turn, the effect in the reliability condition was significantly larger than in the introversion condition. The within-subjects analysis of simple main effects showed that the IAT effect at faking was significantly increased as compared to base rate in the extraversion condition, $F(1, 122) = 10.50$, $V = .08$, and significantly decreased in the introversion condition, $F(1, 122) = 28.22$, $V = .19$, whereas it was not changed in the reliability condition, $F < 1$. The effect size associated with the mean difference between the two faking conditions in the faking trial was $d = 1.24$, and using difference scores (see Experiment 1), it was $d = 1.13$.

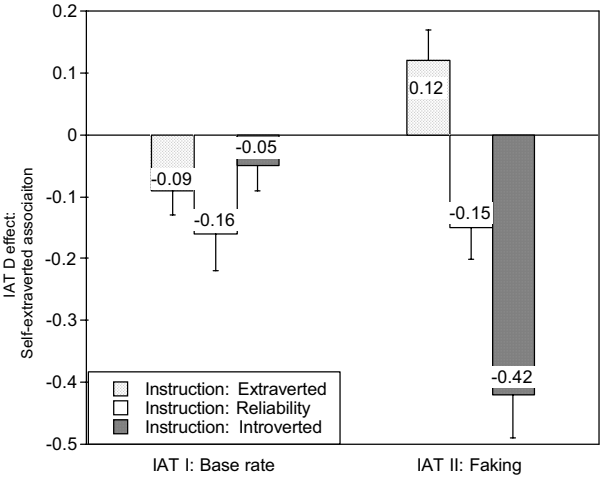

Figure 2. Experiment 2: Mean IAT D effects at base rate and at faking, separately for the extraversion, the reliability, and the introversion condition. Error bars show standard errors of means.

Supplementary analyses concerned the error rates that can be found along with reaction times in Table 2. The self+extraverted trial is referred to as "congruent" in the table. At base rate, participants, on average, made somewhat more errors in the self+introverted than in the self+extraverted task, which is unusual given that the reaction time differences are the other way round. At faking, participants instructed to appear extraverted again made more errors in the self+introverted task. In contrast, participants instructed to appear introverted made more errors in the self+extraverted task. Errors did not vary by task for participants in the reliability condition. The same ANOVA as above with error differences as the dependent variable replicated the interaction observed for the IAT D effect, $F(2, 122) = 8.91$, $V = .13$, again qualifying the main effect of instruction, $F(2, 122) = 12.93$, $R^2_p = .18$ (the other $F < 1.36$). The simple main effects found for the IAT D effect were all replicated, except that the pairwise comparisons showed that the faked error difference in the extraversion condition was different from that in the other two conditions, but those conditions did not differ significantly from each other.

Explicit Ratings

Base rate reliabilities of all measures are shown in the diagonal of the lower panel of Table 1. We sum-

marized the variables in scales for each instruction condition at faking after checking reliabilities (neuroticism: α = .71/.87/.78 in the extraversion, reliability, and introversion conditions, respectively; extraversion: α = .55/.77/.82; openness: α = .81/.62/.68; agreeableness: α = .76/.66/.77; and conscientiousness: α = .90/.73/.76). The reason for the low reliability of the extraversion scale in the extraversion condition is that some of the respective items had very limited or zero variance.

As can be seen in the right half of Table 2, there were only minor differences between the three instruction conditions at base rate for a given NEO-FFI dimension. In contrast, in addition to extraversion, faking scores on all other dimensions changed. A 2 (Trial) × 3 (Instruction) × 5 (NEO-FFI Dimension) MANOVA with repeated measures on the last two factors showed no main effect of instruction, $F < 1.41$, but an interaction of instruction and trial, $F(2, 122) = 5.68$, $V = .09$, along with all other main effects and interactions (all Fs > 3.13). A subsequent 3 (Instruction) × 5 (NEO-FFI Dimension) MANOVA confirmed that there was no chance effect of instruction at base rate, $F < 1.71$. The same MANOVA for the faking trial showed an overall effect of instruction, $F(10, 238) = 23.01$, $V = .98$. Subsequent univariate tests showed significant effects of instruction, in order of size, on extraversion, neuroticism, conscientiousness, openness, and agreeableness, Fs(2, 122) = 584.44, 133.57, 23.50, 5.59, and 3.52, respectively, and R^2_ps = .91, .69, .28, .08, and .06, respectively. The faking effect on the extraversion scale approached a Cohen's $d = 6$.

Correlations

The lower panel of Table 1 shows the correlations between measures at base rate. Implicit extraversion correlated with explicit extraversion, as expected, and there was a surprising negative correlation with explicit conscientiousness. Given the larger power than in Experiment 1, this time explicit neuroticism correlated (negatively) with extraversion, agreeableness, and conscientiousness, and extraversion correlated with agreeableness and conscientiousness.

In the introversion and the extraversion conditions taken together, the reaction time difference between self+extraverted and self+introverted at base rate did not correlate with the same difference at faking, r = .02. In each of the faking conditions regarded separately, the test-retest correlation of the IAT D effect was only r = .06. In contrast, in the reliability condition, the test-retest correlation was r = .60. This correlation was lower if the data were not ipsatized (r = .50). The test-retest correlation of NEO-FFI extraversion in the reliability condition was r = .96 (for the other NEO-FFI dimensions, all r values $\geq$ .89).

It could be that our participants tried to fake either reaction times or errors in the IAT, depending on their hypothesis of how the IAT works. This does not seem to be the case. Faking the IAT D effect correlated with faking errors in the IAT with r = .58 and r = .66 in the extraversion and introversion conditions, respectively. Note that these high correlations are not determined by the error penalty included in the IAT D effect (which might inflate correlations) because of a low r = .09 in the reliability condition. For a traditionally scored IAT effect, the same correlations are r = .40, .48 and −.03, respectively.

In the extraversion and the introversion conditions taken together (after recoding scores as in Experiment 1), there were statistically significant correlations between replies to the question how well one succeeded in faking the explicit test and faking explicit extraversion, r = .21, and between replies to the question how well one succeeded in faking the IAT and faking the IAT D effect, r = .21, as well as faking errors in the IAT, r = .28. Being able to explain how the IAT works also correlated positively with faking the IAT D effect, Spearman's ϱ = .32, and faking errors in the IAT, Spearman's ϱ = .29. The correlation between faking the IAT D effect and faking extraversion on the NEO-FFI was r = .27.

As mentioned above, several dimensions of the NEO-FFI correlated significantly with each other, a finding that is not unprecedented (e.g., Becker, 1999, 2000; Egan, Deary, & Austin, 2000; Schmitz, Hartkamp, Baldini, Rollnik, & Tress, 2001). In fact, on the basis of such correlations, Becker suggested a higher-order two-factor structure of the basic dimensions of personality. With the Experiment 2 data, we tested the factor structure of the NEO-FFI. In a confirmatory approach, extracting five factors and forcing an orthogonal rotation, the emerging pattern is largely in line with the FFM, with 51 of the 60 items loading only on the factor on which they are supposed to load and with factor loadings over .30 (9 of them under .40, however). The rest of the items do not load on any of the factors, or they load on several (cf. also Egan et al., 2000). Similarly, in other studies, confirmatory factor analyses using Structural Equation Modeling could not confirm the five-factor structure of the NEO-FFI (Egan et al., 2000; Schmitz et al., 2001). In a review of the German NEO-FFI, Schwenkmezger (1995) concluded that the instrument is suited for research purposes, but not yet approved enough for clinical use, similar to what has been stated with regard to using the NEO-FFI in Britain (Egan et al., 2000). We second those conclusions.

Discussion

The results of Experiment 2 were clear and in accordance with the hypotheses. Whereas at base rate, reaction times of participants in all conditions showed a weak association of self+introverted, at faking, the reaction time and error difference depended on the instruction. IAT effects of participants faking extraversion showed a self+extraverted association. Participants faking introversion showed a clear association of self+introverted. Reaction times of participants in the reliability condition showed a weak self+introverted association, as at base rate. IAT error differences replicated that pattern. Participants were very well able to manipulate the NEO-FFI in order to appear extraverted or introverted. However, they inadvertently manipulated other NEO-FFI dimensions as well.

The base rate correlation we found between implicit and explicit extraversion attests to the validity of the IAT. Replicating Experiment 1, reliability as assessed by inter-item correlations was satisfactory for all instruments at base rate. In addition, the test-retest-correlations were very high for the NEO-FFI, as one would expect. For the IAT effect, that correlation was $r_{tt} = .60$ in the reliability condition, which is, although rather low by established standards, very high in comparison to what was found in most other IAT studies (see Steffens & Buchner, 2003, for a review). This correlation was lower if the IAT D effect data were not ipsatized, in line with the recommendation of Greenwald et al. (2003) that ipsatizing improves IAT data (it should be mentioned, however, that traditional IAT scoring also resulted in a good test-retest correlation of $r = .61$).

General Discussion

Tests measuring automatic aspects of behavior are, supposedly, not susceptible to faking. We inspected one test measuring automatic cognition, the IAT, and found that it is susceptible to faking, but to a limited degree, certainly less so than an explicit test, and faked scores on the IAT still correlate with those at base rate. In turn, the results of the explicit test were not faked selectively without affecting scores on the other traits. Self-report data concerning one's ability to fake the tests and explaining the mechanism behind the implicit test bore some relation with the ability to fake.

Scores on the relevant dimensions of the NEO-FFI were close to ceiling or bottom at faking, showing that participants had no difficulties at all faking their responses to the respective items. However, they inadvertently faked scores on the other dimensions of the NEO-FFI, too. This is not surprising, given the correlations we found between those dimensions. Our faking instruction may have resulted in extreme faking effects because participants were not concerned with plausibility, but only with appearing as conscientious or as nonconscientious as possible. Therefore, caution is mandatory when generalizing from the present situation to others. Probably, there will be less faking on the NEO-FFI in practical applications, even if participants try to create a positive impression.

Correlations between explicit and implicit measures of the same construct were in a medium order of magnitude. We think we found higher correlations than are found in most studies because our IATs comprise many trials, and longer IATs lead to more consistent and more valid IAT effects (Steffens & Buchner, 2003). In the present studies, medium correlations were found despite the fact that in the NEO-FFI, the items that measure the five dimensions of the FFM are not presented in a blocked fashion, and participants are not told which traits are measured. Thus, extraversion, for example, is inferred from responses to statements that are instances of extraversion, and the self-ascription of extraversion is not measured directly, but so-to-speak, 'implicitly.' Participants are not asked directly how extraverted they are. Instead, statements are simply presented from which extraversion is inferred. It is not made transparent how many traits are measured, and which items form a group. Other tests use direct and transparent self-ascriptions of traits. Evidently, there is a moderately strong overlap between scores obtained on the NEO-FFI and such self-ascriptions (Scandell & Wlazelek, 1999). The IAT measures the automatic self-ascription of a given trait, so-to-speak, "explicitly" because that trait is presented in an unveiled fashion and throughout the IAT on the computer screen. Indeed, the endorsement of the trait is the main factor determining the IAT effect (see Steffens et al., 2003; Steffens, Lichau et al., 2004). In other words, it should become clear to participants in an IAT that the test has something to do with endorsing a given trait. Therefore, we would expect a stronger relationship between the implicit and the explicit test results if participants were directly asked about their self-ascriptions, or if an implicit test was administered in which the different facets of traits exert a larger influence on reaction times, that is, those facets presented in the statements of the explicit test.

Not only were our participants well able to fake the explicit test, but on a group level, they also showed the ability to fake the IAT given some experience with it. This finding is in line with other recent findings (Fiedler & Bluemke, 2003) and needs to be evaluated in combination with studies in which the

IAT could not be faked significantly by naïve participants. In Fiedler and Bluemke's study, participants were asked to fake an IAT after they had once reacted to it. They showed large faking effects of d = 1.14 (Experiment 1) and d = 1.05 (Experiment 2). Those participants had been informed in advance that reaction times were the crucial feature of the test. Taken together, findings on faking the IAT suggest that faking effects are often present descriptively, even if rather small. With test experience, faking becomes much more likely, or faking effects become much larger. It is an important finding that IATs can be faked, all the more so because given the studies that found no statistically significant faking effect, it is often presented as a fact that IATs cannot be faked. Importantly, the ability to fake the IAT was not restricted to a few individuals in our study. Rather, the whole distribution of IAT effects in the faking trial was moved in the respective direction. However, whereas many participants faked their IAT scores in the required direction, they could not completely hide their base rate performance, as shown in significant correlations of IAT effects at base rate and faking. Moreover, they could not selectively fake reaction times to the exclusion of faking errors. Combining those findings with the rather small relations between faking on the one hand and on the other the self-reported ability to fake the IAT and knowledge how the IAT works, we conclude that faking the IAT does not seem to be a perfectly controlled, deliberate process. Rather, participants seem to be able to fake the IAT somewhat without knowing exactly what they are doing or what they should be doing.

The correlations we found between faking reaction times and faking errors show that naïve participants tried faking reaction times and errors to a similar degree. They probably assumed that reaction time *and* error differences are crucial in the IAT. In addition, these correlations demonstrate that there was no speed-accuracy trade-off during faking. One might have thought that faking in one IAT task results in faster responses at the expense of making more errors. In contrast, it seems that faking was achieved by slowing down somewhat in the other IAT task, and committing more errors in that task, too. This interpretation is in line with the reaction times and errors reported in Table 2. Reaction times and errors were increased in the task that was "incongruent," according to the faking instruction, but not decreased in the other task. Given this moderate increase and the large variability in response speed across participants, faking cannot easily be detected. In line with this reasoning, it was recently demonstrated that IAT experts could not readily detect faked data patterns (Fiedler & Bluemke, 2003).

We think our findings show that caution is mandatory when regarding the IAT as a test that is not controllable by the individual performing it. Whereas it may be true that the IAT usually measures automatic behavior, test scores can be contaminated by intentional, controlled behavior. If one is striving for an uncontrollable test for individual diagnosis, the relevant question is not whether there are individuals who cannot control it, but whether there are test-takers who can. If such individuals exist, the test can be faked. In our Experiment 2, there were many individuals who were able to fake the IAT. Thus, we conclude that, while being much harder to fake than the NEO-FFI and, presumably, other personality inventories, the IAT is not immune to faking.

Socially desirable responding seems to consist of two factors, one of them being the deliberate tailoring of answers in order to create a positive impression. The use of implicit measures like the IAT may provide limited guard against such impression management. The other factor is unconscious ego-enhancement in the form of overly positive beliefs about the self (see Furnham, 2001; Rosse et al., 1998). We believe that the currently existing implicit tests cannot shield against that factor. For instance, a self-esteem IAT should measure the spontaneous subjective association of "self" and "positive," and that may be nothing but an instance of a spontaneous personal belief about the self that may or may not be overly positive. Therefore, implicit tests may well be distorted by unconscious ego-enhancement.

It could be that our participants complied with our faking instruction by selectively activating situations in which they behaved especially (non)conscientious, introverted, or extraverted (see Asendorpf et al., 2003, for related evidence). Much research has shown that IAT effects are sensitive to context manipulations (e.g., Wittenbrink, Judd, & Park, 2001). In other words, implicit cognition is malleable, including implicit self-related cognition (Steffens, Kirschbaum, & Müller, 2004). On the one hand, one might argue that the selective activation of instances of instruction-consistent behavior should not be considered faking because these instances are actual instances of a person's behavior. On the other hand, if sloppy people can score higher on a test of conscientiousness simply by bringing to mind the few occasions on which they were not so sloppy, is this not an important piece of information about the test? Put differently, we cannot yet exactly pin down the mechanisms by which faking instructions change IAT scores, nor do we know how and why experiments involving context manipulations change IAT scores. What we know for now is that this test score, among other things, depends on the circumstances and instructions present during test administration, and that we cannot draw from an IAT score firm conclusions about stable person-related factors such as the degree to which a given person possesses a given trait.

References

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83,* 380–393.

Asendorpf, J. B., Banse, R., & Schnabel, K. (2003). Fakability of an Implicit Association Test (IAT) and a new Implicit Association Procedure (IAP) for shyness. *Manuscript submitted for publication*.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, III & J. S. Nairne (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder. Science conference series* (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.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Becker, P. (1999). Beyond the Big Five. *Personality and Individual Differences, 26,* 511–530.

Becker, P. (2000). The "Big Two" Seelische Gesundheit und Verhaltenskontrolle: Zwei orthogonale Superfaktoren höherer Ordnung? [The "Big Two" psychological health and behavior test: Two orthogonal super factors of higher order?. *Zeitschrift für Differentielle und Diagnostische Psychologie, 21,* 113–124.

Borkenau, P., & Ostendorf, F. (1993). *NEO-FFI. Neo-Fünf-Faktoren Inventar nach Costa und McCrae – deutsche Fassung* [NEO-FFI. Neo-Five-Factor inventory according to Costa and McCrae – German version. Göttingen, Germany: Hogrefe.

Bosson, J. K., Swann, W. B., Jr., & 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.

Bredenkamp, J., & Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium (Multivariate variance analysis with the V criterium]. *Psychologische Beiträge, 27,* 127–154.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81,* 760–773.

Brickenkamp, R. (1981). *Aufmerksamkeits-Belastungs-Test d2* (Test of Attention d2]. Göttingen, Germany: Hogrefe.

Cattell, R. B. (1955). Psychiatric screening of flying personnel; personality structure in objective tests: A study of 1,000 air force students in basic pilot training. (Proj. No. 21-0202-0007. [Rep. No. 9]). *USAF School of Aviation Medicine, 50,* 50.

Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests: a theoretical introduction and practical compendium*. Champaign, IL: University of Illinois Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (revised ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.*

Dalen, L. H., Stanton, N. A., & Roberts, A. D. (2001). Faking personality questionnaires in personnel selection. *Journal of Management Development, 20,* 729–742.

Egan, V., Deary, I., & Austin, E. (2000). The NEO-FFI: Emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. *Personality and Individual Differences, 29,* 907–920.

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.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition: Their meaning and uses. *Annual Review of Psychology, 54,* 297–327.

Fiedler, K., & Bluemke, M. (2003). Faking the IAT: Aided and unaided response control on the Implicit Association Test. *Manuscript submitted for publication*.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality & Individual Differences, 7,* 385–400.

Furnham, A. (1997). Knowing and faking one's Five-Factor personality score. *Journal of Personality Assessment, 69,* 229–243.

Furnham, A. (2001). Test-taking style, personality traits, and psychometric validity. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 289–304). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85,* 197–216.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Jackson, C. J., & Francis, L. J. (1999). Interpreting the correlation between neuroticism and lie scale scores. *Personality and Individual Differences, 26,* 59–63.

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

Kühnen, U., Schiessl, M., Bauer, N., Paulig, N., Pöhlmann, C., & Schmidthals, K. (2001). How robust is the IAT? Measuring and manipulating implicit attitudes of East- and West-Germans. *Zeitschrift für Experimentelle Psychologie, 48,* 135–144.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37,* 435–442.

McCrae, R. R., & Costa, P. T., Jr (1999). A Five-Factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research (2nd ed.)* (pp. 139–153). New York: The Guilford Press.

Mierke, J., & Klauer, K. C. (2004). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85,* 1180–1192.

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79,* 272–280.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15,* 263–280.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition, 19,* 97–144.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634–644.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57,* 743–762.

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.

Scandell, D. J., & Wlazelek, B. (1999). The relationship between self-perceived personality and impression management on the NEO-FFI. *Personality and Individual Differences, 27,* 147–154.

Schmitz, N., Hartkamp, N., Baldini, C., Rollnik, J., & Tress, W. (2001). Psychometric properties of the German version of the NEO-FFI in psychosomatic outpatients. *Personality and Individual Differences, 31,* 713–722.

Schwenkmezger, P. (1995). NEO-Fünf-Faktoren Inventar [NEO-Big-Five Inevntory]. Borkenau, P. & Ostendorf, F. (Testrezension). *Zeitschrift für Differentielle und Diagnostische Psychologie, 15,* 237–238.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 943–953.

Steffens, M. C. (1999). Mac-IAT [Computer program]. Trier, Germany: University of Trier.

Steffens, M. C. (in press). Implicit and explicit attitudes towards lesbians and gay men. *Journal of Homosexuality.*

Steffens, M. C., & Buchner, A. (2003). Implicit Association Test: Separating transsituationally stable and variable components of attitudes toward gay men. *Experimental Psychology, 50,* 33–48.

Steffens, M. C., Günster, A., Hartmann, J., & Mehl, B. (2004). Veränderte Märkte, feminisiertes Management, neue Chancen? Der steinige Weg der Frauen in Führungspositionen [Changed markets, feminized management, new chances? Women's stoney path to management positions]. In C. Baltes-Löhr, K. Hölz (Eds.), *Gender-Perspektiven: interdisziplinär – transversal – aktuell* (pp. 113–127). Frankfurt a. M., Germany: Peter Lang Verlag.

Steffens, M. C., Jelenec, P., Anheuser, J., Goergens, N. K., Lichau, J., & Still, Y. (2003). A two-factor model of reaction time differences in the Implicit Association Test. *Manuscript submitted for publication*.

Steffens, M. C., Kirschbaum, M., & Müller, P. (2004). Avoiding stimulus effects in the Implicit Association Test: The Concept Association Task. *Manuscript submitted for publication*.

Steffens, M. C., 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 explaining the reaction-time differences in the IAT]. *Zeitschrift für Psychologie, 212,* 57–65.

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

Steffens, M. C., & Schulze König, S. (2003). Predicting spontaneous behavior with Implicit Association Tests. *Manuscript submitted for publication*.

Strack, F. (1994). Response processes in social judgment. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition (2nd ed.)* (pp. 287–322). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110,* 226–235.

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.

Received December 2, 2003
Revision received February 13, 2004
Accepted February 16, 2004

Melanie Steffens

FB I – Psychology
University of Trier
D-54286 Trier
Germany
Tel: +49 651 201 2017
Fax: +49 651 201 2017
E-mail: steffens@uni-trier.de

Appendix

Items used in Experiment 1

Gewissenhaft: ausdauernd, willensstark, diszipliniert, organisiert, zuverlässig (persevering, strong-willed, disciplined, organized, dependable)
nicht gewissenhaft: ziellos, faul, chaotisch, unordentlich, unpünktlich (aimless, lazy, chaotic, untidy, late)
selbst: ich, mein, mir, wir, uns (I, my, me, we, us)
andere: du, dein, dir, euer, euch (you, your, you, your, you)

Items used in Experiment 2

Extravertiert: selbstsicher, aktiv, gesprächig, energisch, optimistisch (self-assured, active, talkative, energetic, optimistic)
Introvertiert: zurückhaltend, unabhängig, ausgeglichen, zuhörend, ruhig (withdrawn, independent, balanced, attentive, quiet)
selbst: ich, mein, mir, mich, meins (I, my, me, me, mine)
Andere: du, dein, dir, dich, deins (you, your, you, you, yours)