BASIC AND APPLIED SOCIAL PSYCHOLOGY, 32:302–314, 2010

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2010.519236

Psychology Press
Taylor & Francis Group

Faking of the Implicit Association Test Is Statistically Detectable and Partly Correctable

Dario Cvencek and Anthony G. Greenwald
University of Washington

Anthony S. Brown
G4S Justice Services

Nicola S. Gray
Pastoral Cymru, Ltd

Robert J. Snowden
Cardiff University

Male and female participants were instructed to produce an altered response pattern on an Implicit Association Test measure of gender identity by slowing performance in trials requiring the same response to stimuli designating own gender and self. Participants' faking success was found to be predictable by a measure of slowing relative to unfaked performances. This combined task slowing (CTS) indicator was then applied in reanalyses of three experiments from other laboratories, two involving instructed faking and one involving possibly motivated faking. Across all studies involving instructed faking, CTS correctly classified 75% of intentionally faking participants. Using the CTS index to adjust faked Implicit Association Test scores increased the correlation of CTS-adjusted measures with known group membership, relative to unadjusted (i.e., faked) measures.

In a poker game, you might look for a "tell" in another player's behavior as an indicator of bluffing. In psychological assessments, data provided by respondents may likewise contain evidence of attempts to fake. In the Minnesota Multiphasic Personality Inventory–2, for example, faked profiles are identifiable, in part, by elevated scores on the Superlative Self-Assessment Scale (Butcher & Han, 1995). The present study looked for a "tell" that might reveal attempted faking in the behavior of respondents to the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998).

The IAT provides a measure of strengths of associations among socially significant categories. Previous research has revealed that participants asked to fake on IAT measures or make a good impression on them without being instructed how to do so are either unsuccessful (Asendorpf, Banse, & Mücke, 2002; Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim, 2003) or moderately successful (Fiedler & Bluemke, 2005; Schnabel, Banse, & Asendorpf, 2006; Steffens, 2004). Even when successful, faking of the IAT appears to be limited, comparatively less successful than on explicit tests (Steffens, 2004) and dependent upon prior IAT experience (Fiedler & Bluemke, 2005). In contrast, novel attitudes toward fictitious social groups appear to be relatively easy to fake on an IAT (De Houwer, Beckers, & Moors, 2007). The apparently best strategy for faking the IAT is to deliberately slow responses when given the task of responding with the same key to two well-associated categories, although few participants spontaneously discover this strategy without an IAT pretest experience (Fiedler & Bluemke, 2005; Kim, 2003).

Correspondence should be sent to Dario Cvencek, Institute for Learning & Brain Sciences, University of Washington, Box 357988, Seattle, WA 98195. E-mail: dario1@u.washington.edu

For example, in the gender identity IAT used in the present research, men can fake to appear as female-identified if they deliberately slow responding in a task that requires the same response to both male words and self-referring pronouns.

The present research sought to identify an indicator of deliberate slowing of responses that might mark faked IAT performances. Participants first took a baseline gender identity IAT and were later asked to fake one of their two subsequent IATs by using a slowing strategy.

Instructing participants how to fake in order to develop an index of faking may appear circular (or even trivial) at first. However, imagine a researcher who is trying to develop a measure of presence of HIV infection. To develop such a measure, the researcher needs to know who is infected and who is not. Similarly, a researcher who is trying to develop a measure for detecting faking needs to have knowledge of who is faking to estimate the accuracy of the new measure.

In the present research, participants first took a baseline gender identity IAT and were later asked to fake one of their two subsequent IATs by using a slowing strategy. Several possible indexes of slowing were then evaluated for their ability to predict amount of faking on a subsequent IAT. The best performing index, which distinguished fakers from nonfakers with 80% accuracy, was tested using data from two previous experiments in which participants had been instructed to fake their IAT scores and a third in which participants were possibly motivated to fake. Last, the use of this indicator to statistically adjust potentially faked IAT scores was tested.

STUDY 1: NEW EXPERIMENT: INSTRUCTED FAKING OF GENDER IDENTITY

Method

Participants. Participants were 47 introductory psychology students (23 male, 24 female; M age = 18.9, SD = .85). All participants were tested individually and received course credit for participation.

Materials. Each participant was seated in an individual cubicle equipped with a desktop computer. After completing consent and demographic forms, participants learned that they would be classifying words representing four concepts: self (represented by *self, me, I, mine, my*), other (*other, they, them, theirs, their*), male (*male, man, boy, him, he*), and female (*female, woman, girl, her, she*). Inquisit (Millisecond Software, Seattle, WA) was used to present stimuli as well as record the response times.

Procedure. Participants completed three gender identity IATs, each assessing association of *self* with *male* or *female* gender. The second or third of these three IATs was faked in response to instructions. In the first gender identity IAT, participants started with a block of 20 trials, in which they practiced sorting *self* and *other* items. They responded to *self* items by pressing a response button on the left side of the keyboard (i.e., D) and to *other* items by pressing a response button on the right side of the keyboard (i.e., K). After that, participants completed another block of 20 trials in which they practiced sorting *male* items and *female* items using the same two response buttons.

Following these two *single* discrimination tasks, participants completed two *combined* discrimination tasks in which all four categories were used. Each combined task consisted of two blocks of trials: The first block consisted of 30 trials, and the second block consisted of 40 trials. During the combined tasks, two of the four categories were mapped onto the same response key. In the *self*/*male* pairing, *self* items and *male* words shared a response key as did *other* and *female* items. In the *self*/*female* pairing, these were reversed—*self* was paired with *female*, and *other* with *male*. Before the second combined task, participants completed an additional 30-trial, single task block, which practiced the reversal of key assignments for the *self* and *other* words to create the second combined pairing (see Table 1 for the details of all three IATs used in the Gender Identity study). Initial assignment of the two pairings was counterbalanced across participants. After committing an error, participants were obliged to provide the correct response before presentation of the next stimulus. As is standard for IAT measures, trial latency was recorded to the correct response, thus creating a built-in error penalty (cf. Greenwald, Nosek, & Banaji, 2003). The intertrial interval was 400 ms.

The IAT D measure (Greenwald et al., 2003) was computed so that positive values indicated stronger association of *self* with *female* (with computational lower and upper D measure bounds of −2 and +2 corresponding to strongest implicit male and female gender identity, respectively).

Following the completion of the nonfaked baseline measure (IAT1) consisting of 3 single and 4 combined blocks of trials, participants completed 4 combined task blocks of trials for IAT2 and IAT3. This provided data for 12 combined task blocks (4 from each of 3 IATs) and 3 single task blocks from IAT1. Half of the participants received the following faking instructions prior to IAT2. The remainder received these instructions prior to IAT3.

If you are FEMALE:

1. Try to go deliberately slowly in the condition in which SELF and FEMALE get the left response (and OTHER and MALE get the right response).

TABLE 1
Example of the 3-IAT-Structure Used in the Gender Identity Study

IAT	*Task*	*Block*	*Trial No.*	*Items Assigned to Left Key Response*	*Items Assigned to Right Key Response*
IAT1	Single Task 1	1	20	self items	other items
IAT1	Single Task 2	2	20	male words	female words
IAT1	Combined Task 1	3	30	self items + male words	other items + female words
IAT1	Combined Task 1	4	40	self items + male words	other items + female words
IAT1	Single Task 3	5	30	other items	self items
IAT1	Combined Task 2	6	30	other items + male words	self items + female words
IAT1	Combined Task 2	7	40	other items + male words	self items + female words
IAT2	Combined Task 1	8	30	self items + male words	other items + female words
IAT2	Combined Task 1	9	40	self items + male words	other items + female words
IAT2	Single Task 3	10	30	other items	self items
IAT2	Combined Task 2	11	30	other items + male words	self items + female words
IAT2	Combined Task 2	12	40	other items + male words	self items + female words
IAT3	Combined Task 1	13	30	self items + male words	other items + female words
IAT3	Combined Task 1	14	40	self items + male words	other items + female words
IAT3	Single Task 3	15	30	other items	self items
IAT3	Combined Task 2	16	30	other items + male words	self items + female words
IAT3	Combined Task 2	17	40	other items + male words	self items + female words

Note. For half the participants, the positions of Combined Task 1 Blocks in each IAT were switched with those of Combined Task 2 Blocks, respectively. IAT = Implicit Association Test.

2. Also try to respond rapidly for the condition in which OTHER and FEMALE get the left response (and SELF and MALE get the right response).

You will get reminders about this just before each block.

The wording of instructions was suitably reversed for male participants (see the appendix). Participants who received these instructions prior to IAT2 were instructed to "stop trying to respond as a person of the opposite sex" prior to IAT3 and were instructed to "try to respond rapidly for all tasks, while making few errors." This provided four combined task blocks of faked data from one IAT (either IAT2 or IAT3) for each participant. Three IATs were used to avoid confounding faking with position in the experimental sequence while effectively doubling the amount of data for a statistical comparison with faked data.

Results

Effects of faking instructions on IAT performance. Within each IAT, the combined task with longer average response time (in seconds) was the *slower* combined task, and the one with shorter average response time was the *faster* combined task. Slower and faster combined tasks were expected to vary by participants' gender and faking status. During the nonfaked IAT performances, the *self/female* pairing (a *congruent* pairing for female participants) was expected to be the slower combined task for male participants, whereas the *self/male* pairing (a *congruent* pairing for male participants) was expected to be the slower combined task for female participants. Conversely, during the faked IAT performances, the *self/female* pairing (an *incongruent* pairing for male participants) was expected to be the faster combined task for male participants, whereas the *self/male* pairing (an *incongruent* pairing for female participants) was expected to be the faster combined task for female participants.

As expected, faking participants responded slower in congruent blocks and nonfaking participants responded slower in incongruent blocks. Figure 1 presents mean response times (RTs) for single, congruent and incongruent tasks in the three gender identity IATs.

For faking participants, average RTs in congruent blocks were slower than average RTs in incongruent blocks in both IAT2 and IAT3 (all $ps < .03$). For nonfaking participants, average RTs in incongruent blocks were slower than average RTs in congruent blocks in both IAT2 and IAT3 (all $ps < .02$).

Across all three IATs, the average error rates in incongruent blocks were higher than average error rates in congruent blocks. This difference was statistically significant for faking as well as for nonfaking participants in both IAT2 and IAT3 (all $ps < .05$). This pattern is consistent with Fiedler and Bluemke's (2005) findings, which have shown that attempts to fake need not result in an increase in error rates (but cf. Steffens, 2004). Taken together, these preliminary results suggest that participants followed the instructions to fake by *slowing* their performance down in what was expected to be a congruent combined task for them. While doing so, participants did not appear to try to accompany slowing

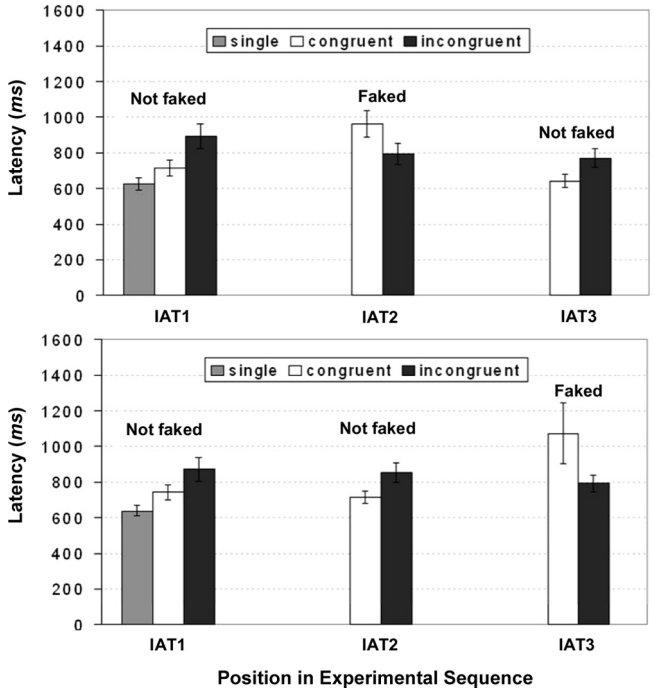

FIGURE 1 Mean latencies for participants who were instructed to fake Implicit Association Test 2 (IAT2; upper panel, $N = 24$) and participants who were instructed to fake IAT3 (lower panel, $N = 23$). *Note.* Error bars = 95% confidence intervals. The *self/male* and *self/female* pairings were congruent tasks for male and female participants, respectively. The *other/male* and *other/female* pairings were incongruent tasks for male and female participants, respectively.

down by increasing error rates. It should be noted that participants were instructed only how to manipulate their response speed. There was no consideration of complicating that by adding an instruction to increase errors.

Faking success: IAT D change. To quantify participants' faking, an index of faking success (D change) was computed as a difference between the faked IAT D score and the immediately preceding nonfaked IAT D score. For participants who faked IAT2, D change was calculated relative to IAT1, and for participants who faked IAT3, D change was calculated relative to IAT2 (i.e., a D score difference between one faked and the immediately preceding nonfaked IAT performance). The decline in response times across nonfaked IAT performances that is visible in Figure 1—nonfaked latencies faster in the second position (bottom panel) than in the first position (both panels) and in the third position (top panel) than in the second position (bottom panel)—was, in part, the basis for not making both calculations relative to IAT1. Following the baseline IAT1, each participant contributed an additional nonfaked IAT performance. For these nonfaked IAT performances, D change was calculated relative to the preceding nonfaked IAT, which for all nonfaked performances was IAT1 (i.e., a D score difference between two nonfaked IAT performances). D change scores were reversed for female participants so that, for all participants, positive values indicated successful faking in the opposite gender direction.

Faking success did not vary as a function of the IAT position. D change scores for participants instructed to fake in IAT2 ($M = .92$) were only slightly different from those instructed to fake in IAT3 ($M = .93$), $t(45) = -.02$, $p = .99$, $d = -.01$. Similarly, there was no difference between D change scores of nonfaking participants in IAT2 ($M = .12$) and those of nonfaking participants in IAT3 ($M = -.04$), $t(45) = 1.28$, $p = .21$, $d = -.37$. Consequently, mean D change scores for male and female participants in the gender identity study are combined across IAT2 and IAT3 in Figure 2. Both male and female participants changed their D scores when faking (relative to not faking): Mean D change score for faking male participants ($M = .90$, $SD = .68$) was statistically different from that for nonfaking male participants ($M = .06$, $SD = .51$), $t(44) = 4.78$, $p = 10^{-5}$, $d = 1.40$. Similarly, the mean D change score for faking female participants ($M = .95$, $SD = .68$) was statistically different from the mean D change score for nonfaking female participants ($M = .02$, $SD = .38$), $t(46) = 5.83$, $p = 10^{-7}$, $d = 1.69$. The difference in D change scores for faking male and faking female participants was not statistically significant, nor was the difference in D change scores for

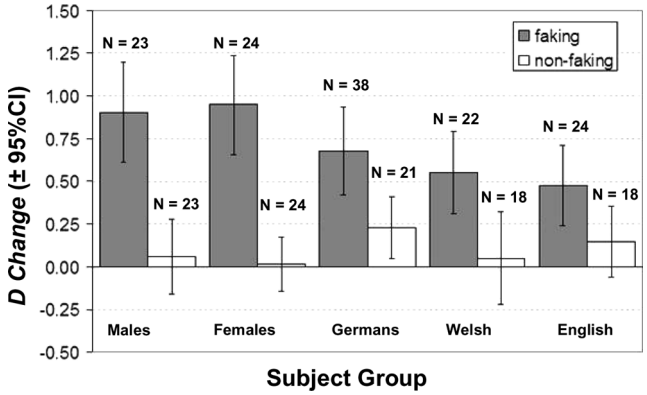

FIGURE 2 Measures of faking success (*D* change) for the three studies involving instructed faking (Gender Identity Implicit Association Test [IAT]; Pro-Turkish Attitude IAT and Welsh/English Attitude IAT). *Note. D* change is computed as the difference between the faked IAT *D* score and the immediately preceding nonfaked IAT *D* score. The researchers contributing the two data sets involving German and Welsh/English subjects are Fiedler and Bluemke (2005) and Brown (2005b), respectively.

nonfaking male and nonfaking female participants, both $ps > .73$, thus allowing us to collapse data across gender in the subsequent analyses involving comparison between faked and nonfaked data.

Combined task slowing. To quantify participants' slowing, average RTs in the combined task blocks from the faked IAT (IAT2 or IAT3) were examined relative to average RTs in combined task blocks from the immediately preceding nonfaked IAT (IAT1 or IAT2). Five candidate indexes were computed as RT differences between (a) slower combined task of the faked IAT and slower combined task of the preceding nonfaked IAT, (b) faster combined task of the faked IAT and the faster combined task of the preceding nonfaked IAT, (c) average of all combined tasks for the faked IAT and the average of single task blocks for IAT1, (d) slower combined task of the faked IAT and the faster combined task in the preceding nonfaked IAT, and (e) average of all combined tasks for the faked IAT and average of all combined tasks for the preceding nonfaked IAT. "Slower" and "faster" task in this context refer to the specific participant's slower and faster faked and nonfaked tasks rather than to the typically slower or faster faked or nonfaked task for the participant's gender group. All differences were computed so that positive values indicated slower response times in the task coming from the faked IAT performance.

A multiple regression analysis was conducted with all five indexes simultaneously entered as predictors and *D* change as the criterion. Index *d* was the strongest predictor of faking success in this simultaneous regression format, and this was true for both IAT2, $t(46) = 3.92$, $\beta = .79$, $p < .0001$, and IAT3, $t(46) = 1.43$, $\beta = .70$, $p = .16$.[1]

Next, separate two-step hierarchical regressions were conducted for IAT2 and IAT3. In each of these regressions, Index *d* was entered at Step 1 and the other three indexes at Step 2. For prediction of faking success in IAT2, Index *d* was significant at Step 1, $R^2 = .26$, $p < .0001$. There was a significant increase of prediction by the other three indexes at Step 2, $R^2 = .43$, $\Delta R^2 = .17$, $F(3, 42) = 4.16$, $p = .01$. Closer examination of the regression results at Step 2 revealed that Indexes *b* and *c* were both significant predictors at Step 2, as indicated by their partial correlations of $r = .39$, $p < .01$ and $r = -.37$, $p = .01$, respectively. For IAT3, the prediction of faking success by Index *d* at Step 1, $R^2 = .38$ $p < .0001$, was not increased by the other three indexes at Step 2, $\Delta R^2 = .003$, $p > .97$.

To further examine the performance of Index *d*, prediction of faked and nonfaked *D* change were examined in separate regression analyses. In the regression analyses of nonfaked *D* change, only Indexes *a* and *b* were statistically significant, $t(42) = -5.96$, $\beta = -.76$, $p < .0001$, and $t(42) = 2.54$, $\beta = .32$, $p = .02$, respectively. In the regression analyses of faked *D* change, Index *d* was the only statistically significant predictor, $t(42) = 2.21$, $\beta = 1.07$, $p = .03$; all other $ps > .31$. Because Index *d* was the strongest predictor of faking success in simultaneous regressions and was the only predictor of faked *D* change, this index, hereafter labeled combined task slowing (CTS), was the only index retained for use in further analyses.

To quantify the performance of the CTS index, cut-off scores for assigning faking status were varied across the range of CTS scores and hit rate (success in identifying fakers) was examined as a function of false-alarm rate (misclassification of nonfakers). The area under this receiver operating characteristic (ROC) analysis (Green & Swets, 1966) quantifies the success of the index in predicting faking versus nonfaking status.

Figure 3 gives the ROC for CTS in assigning faking status, combining results for those who faked IAT2 and IAT3. In this analysis, CTS produced an area under the ROC curve (AUC) of .80 ($SE = .05$), which differed significantly from the chance rate of .50 ($p = 10^{-8}$). Examined separately for each IAT, CTS produced an AUC of .82 ($SE = .06$) in IAT2, which differed significantly from .50 ($p = 10^{-4}$), and an AUC of 0.78

[1]Index *e* was excluded in these simultaneous regressions because of a linear dependency on a combination of the other four. The dependency could have been solved by dropping other indexes, but the results gave no compelling reason to retain Index *e*, so it was dropped.

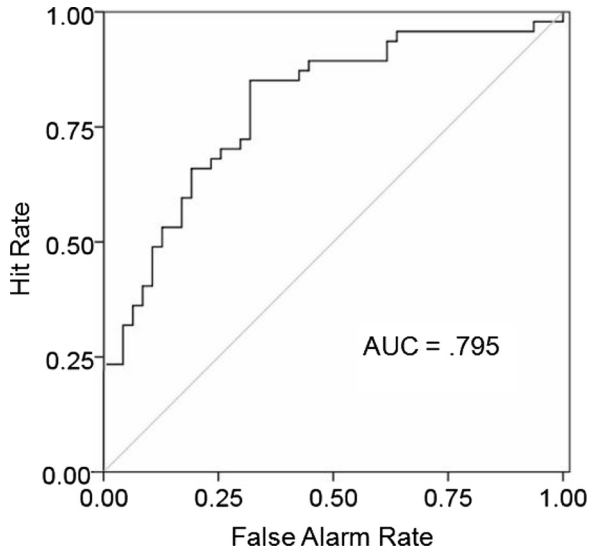

FIGURE 3 Receiver operating characteristic for use of the combined task slowing index to identify faking status in the gender identity study. *Note.* The hit rate (the proportion of faking participants correctly identified) is plotted against the false-alarm rate (the proportion of nonfaking participants incorrectly identified) as the combined task slowing criterion for identifying faked Implicit Association Tests is lowered. The diagonal line represents chance success. The data are from 47 participants who provided both a faked and a nonfaked Implicit Association Test performance. Area under the curve (AUC) corresponds to percentage correct in a two-alternative, forced-choice detection task.

($SE = .07$) in IAT3, which also differed significantly from .50 ($p = .001$).

REANALYSIS 1: GERMANS FAKING FAVORABLE IMPLICIT ATTITUDES TOWARDS TURKS

The present findings appear to contradict the results of Fiedler and Bluemke (2005), who recently reported that they and other skilled researchers were unable to find any indicators of faking in examination of IAT data produced by participants who were instructed to fake. To determine whether the CTS index could be applied successfully to Fiedler and Bluemke's data, we sought their data from three experiments in which German participants attempted to fake pro-Turkish attitudes on a German/Turkish Attitude IAT. With these data, the conditions were re-created as they existed for the experts whom Fiedler and Bluemke recruited to attempt to distinguish faked from nonfaked IAT protocols. The CTS index was applied without advance identification of which IATs were faked and which were not, only later using that knowledge to appraise the success of this use of the CTS index.

Each experiment by Fiedler and Bluemke (2005) included conditions with a nonfaked baseline IAT followed by a faked IAT. Between the two IATs participants received instructions to fake so as to appear nonprejudiced against Turks (*uninformed condition*). Some participants were additionally instructed (*implicitly informed*) that "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" (p. 309). Still other participants (*explicitly informed*) were told that "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" (p. 310). An additional *exploratory* condition was similar to the uninformed condition except that participants did not complete a preliminary baseline IAT. In the *control* condition that was used only in Fiedler and Bleumke's third experiment, participants did not receive any faking instructions prior to their second IAT.

Combined Task Slowing

For the Fiedler and Bluemke's IATs, the CTS index was computed as RT differences in parallel fashion to those from our gender identity IAT. More specifically, CTS was computed by subtracting the faster combined task in the baseline nonfaked IAT from the slower combined task of the faked IAT. An index of faking success (*D* change) was computed by subtracting the faked IAT score from the nonfaked baseline IAT score. Both IATs were scored so that positive values indicated the pro-Turkish attitudes. Consequently, positive *D change* values indicated successful faking in the pro-Turkish direction. Using a linear regression with CTS as predictor and faking success as criterion, CTS significantly predicted faking success in Fiedler and Bluemke's Study 1, $r = .72$, $t(49) = 7.19$, $p = 10^{-9}$; Study 2, $r = .68$, $t(34) = 5.37$, $p = 10^{-6}$; and Study 3, $r = .62$, $t(58) = 6.03$, $p = 10^{-7}$.[2]

ROC Analysis

Study 3 was the only one of Fiedler and Bluemke's three experiments for which the design included both faking (two conditions: *explicitly informed* and *exploratory*) and nonfaking conditions (*control*). The ROC analysis was therefore conducted only for Fiedler and Bluemke's Study 3. Figure 2 presents the mean *D* change scores for faking and nonfaking participants in Fiedler and Bluemke's Study 3. German participants were able to fake successfully when instructed to appear nonprejudiced against the Turks. Collapsed across the two faking

[2]More details for this reanalysis (and other reanalyses reported next) can be found in the original publications on which each reanalysis was based.

conditions, the mean D change score for faking Germans ($M = .68$, $SD = .78$) was statistically different from the mean D change score for nonfaking Germans ($M = .23$, $SD = .39$), $t(57) = 2.46$, $p = .02$, $d = .73$. Applying the same ROC method used in the Gender Identity study to assign faking status in Fiedler and Bluemke's Study 3, CTS correctly classified participants as fakers and nonfakers at levels above chance, as indicated by an AUC of .86 ($SE = .05$), which differed significantly from 0.50 ($p = 10^{-6}$).

REANALYSIS 2: WELSH AND ENGLISH FAKING NATIONAL ATTITUDES

In a study at Cardiff University by Brown, Gray, & Snowden (see Brown, 2005), groups of Welsh ($n = 40$) and English ($n = 42$) participants first completed a nonfaked baseline IAT measure of attitudes toward Wales and England. During the Welsh–English Attitude IAT, participants classified items representing four concepts: Welsh (pictures rated as representative of Wales), English (pictures rated as representative of England), pleasant words (*good, beautiful, health, honest, laugh, joke, lucky*, and *happy*), and unpleasant words (*accident, cancer, disaster, pollution, poverty, sickness, ugly*, and *vomit*). The experiment included a nonfaked baseline IAT followed by a second IAT of the same type, which was a faked IAT for a half of the participants. On both IATs, positive values indicated positive attitudes toward England. Between pretest and posttest, faking was manipulated explicitly for a half of the participants by instructing Welsh participants to appear English at retest and vice versa.

Faking Success

An index of faking success (D change) was computed as a difference between the faked IAT D score and the preceding nonfaked IAT D score. D change scores were reversed for English participants, so that, for all participants, positive values indicated successful faking in the opposite nationality direction. Figure 2 presents the mean D change scores for faking and nonfaking participants in the Welsh–English Attitude IAT. Participants were able to fake successfully when given the instructions to appear as a person of opposite nationality. More specifically, the mean D change score for faking Welsh ($M = .55$, $SD = .61$) was statistically different from the mean D change for nonfaking Welsh ($M = .05$, $SD = .48$), $t(38) = 2.85$, $p = .007$, $d = .91$. Similarly, the mean D change score for faking English ($M = .48$, $SD = .56$) was statistically different form the mean D score for nonfaking English ($M = .15$, $SD = .42$), $t(40) = 2.09$, $p = .04$, $d = .67$.

Combined Task Slowing and ROC Analysis

Combined Task Slowing values were computed as RT differences between the slower combined task of the faked IAT and the faster combined task in the baseline nonfaked IAT. Using a linear regression with CTS as predictor and faking success as criterion, CTS successfully predicted faking success, $r = .27$, $t(80) = 2.50$, $p = .01$. Using the ROC analysis to assign faking status as in the preceding two ROC analyses, CTS correctly classified participants as fakers and nonfakers in the Welsh–English Attitude IAT study, as indicated by an AUC of .62 ($SE = .06$), which was marginally significantly different from 0.50 ($p = .07$).

REANALYSIS 3: PEDOPHILES AND VIOLENT OFFENDERS

In the previous three ROC analyses, CTS successfully classified participants as fakers or nonfakers. However, a baseline IAT performance is necessary for the computation of the CTS. Using as a baseline an unrelated IAT for which there is no motivation to fake would be most desirable. A study comparing convicted pedophiles with nonpedophile prisoners (Brown, Gray, & Snowden, 2009) involved such a design.

The study by Brown, Gray, & Snowden (2009) used a baseline flower–insect attitude IAT, which was scored so that positive values indicated stronger association of *pleasant* with *flowers* than with *insects*. The control pretest was followed by a child–sex association IAT, during which participants classified items representing four concepts: adult (pictures rated as representative of adults), child (pictures rated as representative of children), sex (e.g., *suck, cock, lust, lick*), and nonsex (e.g., *eye, elbow, run, smile*; for a complete list of all items, see Brown, Gray, & Snowden, 2009, or contact the third author). Positive scores on the child–sex IAT indicated stronger association of *sex* with *adult* than with *child*.

The sample (all male; $N = 81$) was recruited from consecutive admissions to a medium secure prison. Some of the participants were convicted *pedophiles* ($n = 33$), whereas others had been convicted for a variety of serious offenses but never for a sexual offense against children (*nonpedophiles*; $n = 48$).[3] Fifteen of the convicted

[3]The sample consisted of another group, which was composed of offenders committing violent and sexual assaults against adolescents (*hebephiles*; $n = 14$), but had not been convicted of a sexual offense against children. Although there were no differences between *hebephiles* and *controls* in their child–sex IAT scores, $t(60) = 0.29$, $p = .77$, the difference between child–sex IAT scores of *hebephiles* and *pedophiles* was marginally significant, $t(45) = 1.94$, $p = .06$. Given this ambiguity about *hebephiles*' IAT scores, the *hebephile* offenders were omitted from further analyses.

pedophiles ($n = 15$) have denied their offense. All the control participants denied ever having sexually offended against children. Given the prison setting of the study and the number of offenders denying their offenses, one could suspect that at least some of the pedophiles were motivated to appear nonpedophile on the child–sex IAT.

Combined Task Slowing and ROC Analysis

As for the preceding reanalysis, CTS was computed as the RT difference between the slower combined task of the child–sex IAT and the faster combined task of the flower–insect IAT. Given the absence of experimentally manipulated faking, the effectiveness of CTS was evaluated using the ROC analysis to assign prisoner's offender status (instead of assigning faking status as in the previously reported ROC analyses). CTS correctly classified offenders as pedophiles and nonpedophiles at levels above chance in the child–sex IAT study, as indicated by an AUC of .65 ($SE = .06$), which was significantly different from 0.5 ($p = .02$). This success of CTS was comparable to the success of the child–sex IAT score: IAT score correctly classified offenders as pedophiles and nonpedophiles at levels above chance in the child–sex IAT study, as indicated by an AUC of .66 ($SE = .06$), which was also significantly different from 0.5 ($p = .02$).

GENERAL DISCUSSION

The present findings show that faking of the IAT can be detected statistically. Using an index of CTS, faking participants were detected correctly using the ROC analysis in our own two gender identity IATs and in two reanalyses of previous studies (Brown, 2005; Fiedler & Bluemke, 2005) with 75% accuracy (corresponding to the weighted average of the four AUCs reported previously for the studies involving instructed faking). This result also contrasts with Fiedler and Bluemke's (2005) published conclusion that faking of IAT protocols cannot be detected. The present reanalysis of Fiedler and Bluemke's data demonstrated that faking by their participants was detectable using the CTS index.

Basis for Success of CTS

Regression analyses showed that CTS was the best (and only consistent) predictor of faked *D* change scores but was not a predictor changes in the *D* measure from a first nonfaked IAT to a second nonfaked IAT. By its operational definition, the CTS index can reveal slowing in either the easier or the more difficult combined task of the preceding nonfaked IAT. Several indexes of slowing were tried, and it is not obvious to the authors why CTS performed better than others. Intuitively, it seemed more likely that the most successful slowing index would be one based on the difference between the slower combined task in a faked IAT and the slower combined task in a nonfaked IAT (this was Index *a*). Index *a* was a significant predictor of *D* change to a faked IAT but less so than was CTS, which was originally identified as Index *d*.

The superiority of the CTS over other indexes was also confirmed using the simultaneous regression format with all five indexes to predict faking success in other data sets (Brown, 2005; Fiedler & Bluemke, 2005). For these simultaneous regressions, *D* change was entered as a criterion and all available indexes as predictors (Index *c* could not be computed for all data sets). Across the four studies involving instructed faking, the partial correlation of the CTS index with *D* change (weighted average $r = .42$) was substantially higher than the partial correlation of the next-strongest predictor with *D* change (weighted average $r = -.10$), $t(224) = 5.72$, $p < .0001$. CTS also exhibited slightly higher zero-order correlations with *D* change (weighted average $r = .55$) than did the next-strongest predictor (weighted average $r = .48$).[4] However, this difference was not statistically significant, $t(224) = 1.26$, $p > .2$.

The hierarchical regression analysis predicting *D* change scores in IAT2 showed that Indexes *b* and *c* both predicted significant variance in *D* change at Step 2 over that already predicted by CTS at Step 1. However, the result was confirmed neither in the prediction of faking in IAT3 nor in the analysis that predicted faking for a combined data set with both faked IATs. This result nevertheless suggests that, in some cases, prediction of faking success may be improved by use of multiple predictor indexes.

The assumption that the respondent's most likely faking strategy for the IAT is to try to respond more slowly in the initially easier of the two combined tasks was based on Kim's (2003) observation that partially effective faking could be achieved by means of deliberately increasing response speed in one of the IAT's two combined tasks (see also Fiedler & Bluemke, 2005). Participants who rely on a different strategy than the one described here (e.g., Steffens, 2004) may not be classified correctly using the CTS index. However, even in a case in which the strategy used for faking is unknown (i.e., Fiedler & Bluemke, 2005) the CTS measure effectively predicted faking.

[4]In the three Fiedler and Bluemke's (2005) studies, this was Index *a*, and in the Brown, Gray, & Snowden study (see Brown, 2005) the next strongest predictor was Index *b*.

One issue closely related to the topic of faking strategies—an issue that was not examined directly in the present research—is the role of prior IAT experience. However, this issue was examined in more detail by Fiedler and Bluemke (2005), who have compared directly faked IAT performances of participants who did not have any prior IAT experience to those who have had IAT experience. Their results can be summarized as justifying the following two conclusions: First, at least one prior IAT seems to be necessary for participants to apply whichever faking strategy they may be using. Second, the pretest experience is not sufficient by itself: In Fiedler and Bluemke's Study 3, IAT scores were "only reversed when participants were instructed to fake intentionally" (p. 314).

The three reanalyses of previous studies showed that the CTS index can identify both those instructed to fake group identities (e.g., male, female, Welsh, English, etc.) and those who may have uninstructed motivation to fake (e.g., pedophiles wishing not to be identified by an IAT measure).

Examination of Possible Alternative Indexes

In addition to quantifying participants' slowing by computing the indexes described in the text, we also examined three other approaches to detect faking, none of which was directly related to slowing (a) differences in error rates between slower and faster combined tasks in the faked IATs relative to those in nonfaked IATs, (b) trial-to-trial changes in response latency in the faked IATs relative to those in nonfaked IATs (up to seven consecutive responses), and (c) trial-to-trial changes in error responses in the faked IATs relative to those in nonfaked IATs (up to seven consecutive responses). None of these approaches yielded information that could be interpreted as being systematically related to faking success.

Three alternatives to CTS that had some a priori justification were also examined. These were (a) a variant of CTS that used the divided Index *d* for each participant by the inclusive standard deviation of the nonfaked IAT used for the computation of CTS, (b) the difference between the slower combined task of the faked IAT and the faster combined task in nonfaked IAT1, and (c) the difference between slower combined task of the faked IAT and the faster combined task in second nonfaked IAT. The first two variants of CTS were not as successful as the one reported in the text, as evidenced by smaller cumulative AUCs (both .78, $SE = .05$). The third variant of CTS produced a larger AUC (.88, $SE = .04$) than that produced by CTS. However, because it required a three-IAT-format it will not be practical in most settings and, in particular, could not be used for the reanalyses described later in this article.

Statistically Adjusting Faked IAT Scores

Given its effective use to detect fakers, the CTS index's use might be extended to correcting faked IAT scores. This section evaluates an approach that computes an adjustment for the IAT scores by removing the component of that score that is predictable by CTS. The adjustment procedure was further designed so that the adjustment was expected to leave an IAT score unchanged for those who did not fake. Adjusted IAT scores were computed with this equation:

$$D_{\text{adj}} = (a * D_{\text{unadjusted}}) - (b * (\text{CTS} - c))$$

In this equation, coefficient *a* is the unstandardized slope of the regression of an unfaked *D* score on a previous baseline IAT *D* measure, reflecting the reliability of IAT measures. Coefficient *b* is the slope of the regression of the measure of faking success (*D* change) on the CTS index, indicating the expected distortion of *D* measures that is predictable from CTS. Constant *c* is the intercept of the regression of CTS on *D* change, indicating the CTS value associated with no change in IAT scores from a nonfaked to a faked IAT. Subtraction of *c* from CTS makes the expected adjustment zero for participants who are not faking.

Coefficients *a* and *b* and constant *c* were calculated separately for each study in which all three coefficients could be computed (i.e., slope *b* could not be computed for studies that did not include a nonfaking group). For studies involving groups of participants who were instructed to fake in two different directions (e.g., male and female participants instructed to fake in opposite gender direction in the Gender Identity study), coefficient computations were always conducted with samples limited to one group (e.g., female) not faking (i.e., high scores) and the other group (e.g., male) faking (i.e., also high scores). This way, coefficients were computed in a similar fashion across all studies (i.e., analyses based on all participants having scores in the same direction). Table 2 displays coefficients *a, b*, and *c* for four studies: IAT2 and IAT3 of the Gender Identity study, German/Turkish Attitude Study 3 (Fiedler & Bluemke, 2005) and the Welsh/English Attitude study (Brown, 2005). This presentation format allows evaluating the variability of each coefficient across samples and topics. A weighted average of all available estimates was computed for each of the three constants in the adjustment formula (*a, b*, and *c*). The German/Turkish Attitude Study 3 (Fiedler & Bluemke, 2005) involved a sample with very little variation on the IAT, which yielded a poor estimate of test–retest reliability ($r = .26$). Consequently, the reliability coefficient from this study was dropped as a basis for computing coefficient *a*. The strategy of averaging over estimates from different

TABLE 2
Values for Coefficients *a*, *b*, and *c* in the Four Studies Involving Faking and Nonfaking Groups

IAT	*Self = Female*[a]				*English = Positive*[b]		*Turkish = Positive*[c]	
Data Set	*IAT2*		*IAT3*		*Study 1*		*Study 3*	
Faking Group	*Female*	*Male*	*Female*	*Male*	*Welsh*	*English*	*German*	*Weighted Row Average*
Slope *a*	0.70	0.70	0.61	0.61	0.47	0.47	(0.26)	0.57
Slope *b*	2.94	1.47	1.04	1.02	1.22	0.86	1.49	1.39
Intercept *c*	0.15	0.20	−0.04	0.13	0.10	0.11	0.07	0.10
Sample size (*n*)	24	23	23	24	40	42	59	

Note. Coefficient *a* = Unstandardized slope of the regression of an unfaked *D* score (in *D* units) on a previous baseline Implicit Association Test (IAT; in *D* units), reflecting the reliability of IAT measures. Coefficient *b* = Slope of the regression of the measure of faking success (*D change;* in *D* units) on the combined task slowing (CTS) index (in seconds), indicating the expected distortion of *D* measures that is predictable from CTS. Constant *c* = Intercept of the regression of CTS on *D* change, indicating the values of CTS (measured in seconds) associated with no change in IAT scores from unfaked to a faked IAT. Sample sizes reflect the faking as well as the non-faking group used in each coefficient computation (see text for details). Parentheses indicate the Slope *a* value that was not used in the calculation of the weighted row average (see text for details).

[a]Present study.
[b]Brown (2005).
[c]Fiedler and Bluemke (2005).

studies was justifiable because, in practice, adjustments might be computed for data sets that did not have appropriate data to estimate these coefficients. With the weighted average coefficient values, the adjustment formula (used for all data sets) becomes:

$$D_{\text{adj}} = (.57 * D_{\text{faked}}) - (1.39 * (\text{CTS} - .10)).$$

Table 3 presents correlations of both unadjusted values and these computed adjustment values with known group membership criteria and with unfaked IAT scores for each study. Using weighted averages of correlations, the adjusted measure correlated more highly with both known group membership and with an unfaked IAT score.

This strategy for correcting IAT (*D*) scores on the basis of the CTS has three useful features:

1. The only information necessary for the calculation of CTS is the set of combined task latencies, without concern about what the specific combined task was: As evident from the results of Study 1, the CTS approach can be applied in situations in which respondents will show effects in different directions (e.g., a sample consisting of male and female individuals).

TABLE 3
Across Six Studies Involving Instructed Faking, Adjusted IAT Measures Outperformed the Unadjusted IAT Measures in Terms of Correlations with Known Group Membership and Unfaked IAT Scores

IAT and Data Set	*N*	*Correlations With Group Membership*						*Correlations With Unfaked IAT Score*			
		Unfaked IAT Score		*Unadjusted Faked IAT Score*		*Adjusted Faked IAT Score*		*Unadjusted Faked IAT Score*		*Adjusted Faked IAT Score*	
		r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Self = female IAT[a]											
IAT2	47	.79	10^{-8}	.01	.96	.52	10^{-4}	.01	.95	.42	.004
IAT3	47	.75	10^{-7}	−.09	.53	.35	.01	−.01	.94	.44	.002
English = pleasant IAT[b]	82	.67	10^{-12}	.05	.67	.16	.14	.16	.16	.19	.08
Turkish = positive IAT[c]											
Study 1	50							.14	.32	.39	.005
Study 2	35							.26	.13	.48	.004
Study 3	59							.11	.42	.25	.05
Weighted average correlation (*r*)		.72		.00		.32		.11		.34	

Note. IAT = Implicit Association Test.
[a]Present study.
[b]Brown (2005).
[c]Fiedler and Bluemke (2005).

2. The "nonfaked IAT" necessary for the CTS calculation can come from a different IAT: As evident from the results of the pedophile study, procedures other than a pretest measure of the same IAT can be used to obtain the data necessary for the computation of CTS.
3. The observed value of CTS for each respondent can be used in conjunction with the averaged values of coefficients *a* and *b* and constant *c* from other data sets (see Table 2).

When this approach was used in the pedophile study (Brown, Gray, & Snowden, 2009), the CTS-adjusted child–sex IAT measure correlated more highly with prisoners' offender status ($r = .41$, $p = .0002$) than did the unadjusted child–sex IAT measure ($r = .28$, $p = .01$). This finding displays the applicability of the adjustment procedure in settings for which there is no independent knowledge of whether a respondent is or is not faking. This result also indicates the potential of the adjustment formula to be used with respondents who may be motivated to fake, such as drug users (Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008), psychopathic murderers (Gray, MacCulloch, Smith, Morris, & Snowden, 2003), patients at risk for suicide (Nock & Banaji, 2007a, 2007b), or convicted pedophiles (Gray, Brown, MacCulloch, Smith, & Snowden, 2005; Steffens, Yundina, & Panning, 2008).

The strategy used to compute the CTS index could also be adapted to IAT task configurations different from the ones in this research. For example, in a recent study of faking, Agosta, Ghirardi, Zogmaister, Castiello, and Sartori (2010) reported an index of slowing relative to single task blocks. Across their four studies, their index classified faking participants with 88% accuracy. When the CTS approach was applied to Agosta et al.'s (2010) data, it correctly classified 75% of intentionally faking participants. This independent work establishes that the approach reported here is not unique in being able to detect faking in IAT data sets.

General Approach to Computing Adjustments

Because IAT procedures vary substantially across studies (e.g., as evident in the variability of fitted parameters presented in Table 2), the present correction formula cannot be claimed to be universally optimal. Subsequent research can use the same general strategy developed here to develop faking predictors and to compute adjustment formulas. The procedure to identify an indicator of faked IAT performances has four steps: First, participants complete a baseline IAT measure under standard (nonfaking) instructions. Second, a random subset of the participants is instructed to fake a subsequent IAT. Third, candidate indexes of faking are computed. Fourth, these candidate indexes are evaluated for ability to predict change in IAT scores on a faked IAT. Best performing indexes can then be used to compute adjusted IAT measures, as in the analyses of present Tables 2 and 3.

The present results come from studies involving well-established associations among categories self, gender, and nationality. Previous research has shown that participants can effectively fake novel associations (De Houwer et al., 2007). Applicability of the CTS-based approach to detecting faking should be considered, for the present, unknown as it pertains to detection of faking for novel associations. Future research might eventually suggest additional statistical indicators of faking (and corresponding methods for adjustment of faked IAT scores).

Such further research can assume practical significance in the context of clinical attempts to diagnose pathologies associated with criminal behavior using IAT measures (Gray et al., 2005; Gray et al., 2003). In addition, the Timed Antagonistic Response Alethiometer (Gregg, 2007) and the Autobiographical IAT (Sartori et al., 2008) are two recent adaptations of the IAT that have been successfully applied as lie detection techniques. The use of CTS might guide development of other indexes that can be computed from data obtained with TARA or an IAT to expand upon existing methods and provide both forensic and clinical fields with additional procedures that can be used as lie detection techniques.

Conclusion

Findings of the present experiment and reanalyses of three other experiments involving instructed or possibly motivated faking confirmed that an index of combined task slowing (CTS) can correctly classify faked and nonfaked IAT performances with an average 75% accuracy. This result contrasts with Fiedler and Bluemke's (2005) pessimistic conclusion that identification of fakers on IAT measures was "virtually impossible" (p. 315). The present CTS index was also shown to be effective in adjusting faked IAT scores, increasing correlations with known group membership and unfaked IAT score respectively from $r = .00$ and $r = .11$ to $r = .32$ and $r = .34$ with CTS-adjusted measures. In conclusion, faking of the IAT can not only be detected, but—to a useful extent—can be corrected.

ACKNOWLEDGMENTS

This research was supported in part by NIMH grants MH-57672 and MH-01533, as well as a grant from the

National Science Foundation (OMA-0835854) to the LIFE Science of Learning Center.

We thank Klaus Fiedler and Matthias Bluemke for providing the data used in Reanalysis 1.

REFERENCES

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2010). Detecting fakers of the *autobiographical* IAT. *Applied Cognitive Psychology*. Advance online publication. doi:10.1002/acp.1691

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.

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.

Brown, A. (2005). *Investigating faking on the streamlined IAT.* Unpublished doctoral dissertation, University of Cardiff, Cardiff, Wales.

Brown, A. S., Gray, N. S., & Snowden, R. J. (2009). Implicit measurement of sexual preferences in child sex abusers: Role of victim type and denial. *Sexual Abuse: A Journal of Research and Treatment*, *21*, 166–180.

Butcher, J. N., & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The *S* scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (pp. 25–50). Hillsdale, NJ: Erlbaum.

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology*, *43*, 972–978.

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., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, *27*, 307–316.

Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology*, *114*, 304–308.

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.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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.

Gregg, A. (2007). When vying reveals lying: The timed antagonistic response alethiometer. *Applied Cognitive Psychology*, *21*, 621–647.

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

Nock, M. K., & Banaji, M. R. (2007a). Assessment of self-injurious thoughts using a behavioral test. *The American Journal of Psychiatry*, *164*, 820–823.

Nock, M. K., & Banaji, M. R. (2007b). Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology*, *75*, 707–715.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*, 772–780.

Schnabel, K., Banse, R., & Asendorpf, J. B. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology*, *53*, 69–76.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology*, *51*, 165–179.

Steffens, M. C., Yundina, E., & Panning, M. (2008). Automatic associations with "erotic" in child sexual offenders: Identifying those in danger of reoffence. *Sexual Offender Treatment*, *3*(2). Available from http://www.sexual-offender-treatment.org/2-2008_03.html

APPENDIX

Instructions to Fake

This is a very important part of the experiment! We would appreciate your careful reading. THANK YOU. In the task you just completed, you may have noticed that it was easier for you to respond to one of the tasks than the other. Typically, women find it easier to give the same responses to female-self and male-other than to male-self and female-other, and men typically find the reverse pattern easier. These patterns are understandable in terms of psychological gender differences that have been demonstrated frequently in previous studies.

Of course, we can't know what your performance in the first part of the experiment will show until we later analyze the data. The tasks that you have already completed were intended to introduce you to the methods used in this research. The important part of this research is your next task.

Regardless of your performance in the previous task, please treat the following task as if YOU WERE A PERSON OF THE OPPOSITE GENDER. We are asking you to do this (and will give you suggestions of how to do it) to learn whether it is possible to successfully give an altered response pattern. Instructions for MALE subjects are on the next page.

If you are FEMALE:

1. Try to go deliberately slowly in the condition in which SELF and FEMALE get the left response (and OTHER and MALE get the right response).
2. Also try to respond rapidly for the condition in which OTHER and FEMALE get the left response (and SELF and MALE get the right response).

You will get reminders about this just before each block.

If you are MALE:

1. Try to go deliberately slowly for the condition in which OTHER and FEMALE get the left response (and SELF and MALE get the right response).
2. Also try to respond rapidly for the condition in which SELF and FEMALE get the left response (and OTHER and MALE get the right response).

You will get reminders about this just before each block.

Instructions to Stop Faking

The next task returns to the form of the task that you did at the beginning of this experiment. You should NO LONGER be trying to respond as a person of the opposite sex. Rather, for the remainder of the experiment, please do the tasks just trying to respond to each as well as you can. That is, you should try to respond rapidly for all tasks, while making few errors.

Just as it was important that you try to alter your response patterns in the previous tasks, it is now very important that, for the remainder of the experiment, you do your best to respond in normal fashion, just trying to perform the tasks as best you can.