RUNNING HEAD: DETECTING IAT FAKING
WORD COUNT: 6, 447 (abstract + main text + acknowledgments + footnotes)

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

Dario Cvencek and Anthony G. Greenwald

University of Washington

Anthony Brown

G4S Justice Services, UK

Robert Snowden and Nicola Gray

Cardiff University

Dario Cvencek
Department of Psychology
University of Washington
Box 351525
Seattle, Washington, 98195
Phone: (206) 543-8029
dario1@u.washington.edu

Anthony G. Greenwald
Phone: (206) 543-7227
agg@u.washington.edu

Anthony Brown
Interventions
G4S Justice Services, UK
Phone: +44(0)16 563 00200
anthony.brown@uk.g4s.com

Nicola Gray
School of Psychology
Cardiff University, UK
Phone: +44(0)29 208 76259
grayns@Cardiff.ac.uk

Robert Snowden
School of Psychology
Cardiff University, UK
Phone: +44(0)29 208 74937
snowden@cardiff.ac.uk

KEYWORDS: Implicit Association Test, faking, lie detection

Abstract (120 words)

Instructed male and female participants to fake an Implicit Association Test (IAT) 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 77% of intentionally faking subjects. Using the CTS index to adjust faked IAT scores increased the correlation with known group membership from r = .00 with faked measures to r = .31 with CTS-adjusted 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 MMPI-2, for example, faked profiles are identifiable, in part, by elevated scores on the Superlative Self-Assessment Scale (Butcher and 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, & Scwhartz, 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 (Schnabel, Banse, & Asendorpf, 2006; Steffens, 2004; Fiedler & Bluemke, 2005). Even when successful, faking of the IAT appears to be limited, comparatively smaller than on explicit tests (Steffens, 2004) and dependent upon prior IAT experience (Fiedler & Bluemke, 2005). In contrast, novel attitudes towards 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 (Kim, 2003; Fiedler & Bluemke, 2005). 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, in order 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 non-fakers with 84% 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. Lastly, 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; mean 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, 2006) 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 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 practiced sorting *female* items and *male* 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. During the combined tasks, two of the four categories were mapped onto the same response key. In one pairing, *self* items and *female* words shared a response key as did *other* and *male* items (*self/female* pairing). In the other pairing, these were reversed — *self* was paired with *male*, and *other* with *female* (*self/male* pairing). Before the second combined task, participants completed an additional single task which practiced the reversal of key assignments for the *self* and *other* words to create the second combined pairing. Initial assignment of the two pairings was counterbalanced across participants.

The first two single task blocks of IAT1 (practice of concept and attribute discrimination) consisted of 20 trials, whereas the third single task block (reversal of concept discrimination)

consisted of 30 trials. For each of the two combined task pairings, the first block consisted of 30 trials and the second block consisted of 40 trials. 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 non-faked 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 (four from each of three 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).

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 males (See Appendix A). 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

Manipulation check. 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 non-faked IAT performances, the *self/female* pairing (a *congruent* pairing for females) was expected to be the slower combined task for males, whereas the *self/male* pairing (a *congruent* pairing for males) was expected to be the slower combined task for females. Conversely, during the faked IAT performances, the *self/female* pairing (an *incongruent* pairing for males) was expected to be the faster combined task for males, whereas the *self/male* pairing (an *incongruent* pairing for females) was expected to be the faster combined task for females.

As expected, faking participants responded slower in congruent blocks and non-faking 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 non-faking 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 non-faking participants in both IAT2 and IAT3 (all *p*s < .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 down by increasing error rates. It should be noted that participants were advised only to slow responding. There was no consideration of complicating that by adding an instruction to increase errors.

Faking success: IAT score (D) difference. 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 non-faked 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 one non-faked IAT performance). For participants who did not fake IAT2, *D* change was calculated relative to IAT1, and for participants who did not fake IAT3 *D* change was calculated also relative to IAT1 (i.e. a *D* score difference between two non-faked IAT performances). The decline in response times across IATs that is visible in Figure 1 was, in part, the basis for not making both calculations relative to IAT1. 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 non-faking participants in IAT2 (M = .12) and those of non-faking participants in IAT3 (M = –.04), $t(45) = 1.28$, $p = .21$, d = –.37. Consequently, presentation of mean *D* change scores for faking and non-faking males and females in the gender identity study are combined across IAT2 and IAT3 in Figure 2. Mean *D* change score for faking males (M = .90, SD = .68) was statistically different from that for non-faking males (M =.06, SD = .51), $t(44) = 4.78$, $p = 10^{-5}$, $d = 1.40$. Similarly, the mean *D* change score for faking females (M = .95, SD =.68) was statistically different form the mean *D* change score for non-faking females (M = .02, SD = .38), $t(46) = 5.83$, $p = 10^{-7}$, $d = 1.69$.

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 non-faked IAT (IAT1 or IAT2). Five candidate indexes were computed as RT differences between (a) slower combined tasks of the faked IAT and the preceding non-faked IAT, (b) faster combined tasks of the faked IAT and the preceding non-faked IAT, (c) average of all combined tasks for the faked IAT and average of all combined tasks for the preceding non-faked IAT, (d) average of all combined tasks for the faked IAT and the average of single task blocks for IAT1 and (e) slower combined task of the faked IAT and the faster combined task in the preceding non-faked IAT. Conceptually these five indexes correspond to (a) the average latency *increase* from the slower combined task in a known-not-to-be-faked IAT to the same combined task in a possibly faked IAT, (b) the average latency *decrease* from the faster combined task in a known-not-to-be-faked IAT to the same combined task in a possibly faked IAT, (c) the average latency *increase* from the average of all combined task in a known-not-to-be-faked IAT to the same average in a possibly faked IAT, (d) the average latency *increase* from the average of all single tasks in a known-not-to-be-faked IAT

to the slower combined task in a possibly faked IAT, and (e) the average latency *increase* from the faster combined task in a known-not-to-be-faked IAT to the same combined task in a possibly faked IAT.

A multiple regression analysis was conducted with all five indexes simultaneously entered as predictors and *D* change as the criterion. Index (e) was the only one of the five predictors that uniquely predicted faking success in this simultaneous regression format, and did so for both IAT2, $t(46) = 3.92$, $\beta = .79$, $p = .0003$, and IAT3, $t(46) = 3.08$, $\beta = 1.05$, $p = .004$. In addition, four 2-step hierarchical regressions were conducted in which index (e) was entered at Step 1 and each of the other four indexes at Step 2. None of these analyses increased prediction of faking success at Step 2, all ts$(91) < 1.04$, βs $< .20$, ps $> .30$, beyond that predicted by index (e) at Step 1, $t(92) = 6.\ 27$, $\beta = 5.47$, $p < .0001$. Index (e), identified as combined task slowing (CTS), was therefore the only one retained for use in further analyses. [1]

CTS has an important advantage over *D* change when trying to assess faking: To calculate *D* change one needs to know which IAT performance was faked — something that would not be known when participants are engaging in uninstructed faking. To calculate CTS, one needs to know only which combined task was performed more slowly in each IAT — something that will always be knowable.

To quantify the performance of the CTS index, cut-off scores for assigning faking status were varied and CTS' success in correctly identifying fakers (i.e. hit rate) was examined relative to its success in excluding non-fakers (i.e. false-alarm rate). A plot of indexes' hit rate vs. false alarm rate is known as a receiver operating characteristic curve (ROC; Green & Swets, 1966). To quantify the performance of an index, the area under the curve (AUC) can be calculated as an equivalent to the percentage correct in a two-alternative, forced-choice detection task (see also

Swets, 1986). If CTS has no predictive validity, the hits should rise at the same rate as the false alarms and the AUC will be 0.5. If CTS could perfectly detect all the fakers without misidentifying non-fakers, the AUC would be 1.0.

Figure 3 plots the ROC for CTS in assigning faking status for all participants collapsed across the two IATs. Using the ROC analysis, CTS correctly classified participants as fakers and non-fakers at levels above chance in both IATs, as indicated by a cumulative AUC of 0.836 (*SE* = 0.042), which differed significantly from 0.50 ($p = 10^{-8}$). When examined separately for each IAT, CTS produced an AUC of 0.824 (*SE* = 0.064) in IAT2, which differed significantly from 0.50 ($p = 10^{-4}$), and an AUC of 0.868 (*SE* = 0.055) in IAT3, which also differed significantly from 0.50 ($p = 10^{-5}$).[2]

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 (2005) 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 non-faked 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 non-faked baseline IAT followed by a faked IAT. Between the two IATs participants received instructions to fake so as to appear non-prejudiced 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. Fiedler and Bluemke's IATs differed in block structure from the present gender identity IATs. Their combined tasks were administered in single blocks of 64 trials rather than in two blocks of trials as in our procedure. To compute the closest equivalent to our CTS index, each of their 64-trial blocks was treated as two 32-trial blocks, after which CTS could be 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 non-faked 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 non-faked baseline IAT score. Positive 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, $r = .62$, $t(58) = 6.03$, $p = 10^{-7}$.[3]

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 non-faking 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 non-faking participants in Fiedler and Bluemke's Study 3. German participants were able to fake successfully when instructed to appear non-prejudiced against the Turks. Collapsed across the two faking conditions, the mean *D* change score for faking Germans (M = .68, SD = .78) was statistically different from the mean *D* change score for non-faking Germans (M =.23, SD = .39), $t(57) = 2.46$, $p = .017$, $d = .73$. 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 non-fakers at levels above chance, as indicated by an AUC of 0.862 (SE = 0.046), 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 and Snowden (see Brown, 2005a) groups of Welsh (n = 40) and English (n = 42) participants first completed a non-faked baseline IAT measure of attitudes towards 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 non-faked baseline IAT followed by a second IAT of the same type, which was a faked IAT for a half of the participants.

Between pretest and post-test, faking was manipulated explicitly for a half of the participants by instructing Welsh participants to appear English at retest and vice versa.

An index of faking success (*D change*) was computed as a difference between the faked IAT *D* score and the preceding non-faked IAT *D* score. *D change* was scored so that positive *D* change values indicated successful faking in the opposite nationality direction. Figure 2 presents the mean *D* change scores for faking and non-faking 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 non-faking 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 non-faking English (*M* = .15, *SD* = .42), $t(40) = 2.09$, $p = .043$, $d = 67$.

Combined task blocks of the Welsh–English attitude IAT were administered as single blocks of 96 trials rather than in two separate blocks of trials. To compute an equivalent of the CTS index, each of these 96-trial blocks was treated as two 48-trial blocks, after which CTS values were computed as RT differences between the slower combined task of the faked IAT and the faster combined task in the baseline non-faked IAT. Using a linear regression with CTS as predictor and faking success as criterion, CTS successfully predicted faking success, $r = .27$, $t(80) = 2.50$, $p = .015$. Using the ROC analysis to assign faking status as in the preceding two ROC analyses, CTS correctly classified participants as fakers and non-fakers in the Welsh–English attitude IAT study, as indicated by an AUC of 0.617 (*SE* = 0.063), 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 non-fakers. 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 non-pedophile prisoners (Brown, 2005b) involved such a design.

The study by Brown (2005b) used a baseline flower–insect attitude IAT. 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 non-sex (e.g. *eye, elbow, run, smile;* for a complete list of all items see Brown, 2005b, or contact the third author).

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 (*non-pedophiles*; n = 45).[4] Fifteen of the convicted 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 non-pedophile on the child–sex IAT.

As for the preceding reanalysis, 96-trial blocks were split into two 48-trial blocks to permit computation of CTS. 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 non-pedophiles at levels above chance in the child–sex IAT study, as indicated by an AUC of 0.655 (*SE* = 0.063), which was significantly different from 0.5 ($p = .018$). This success of CTS was comparable to the success of the child-sex IAT score: IAT score correctly classified offenders as pedophiles and non-pedophiles at levels above chance in the child–sex IAT study, as indicated by an an AUC of 0.658 (*SE* = 0.062), which was also significantly different from 0.5 ($p = .016$).

Discussion

The present findings show that faking of the Implicit Association Test can be detected statistically. Using an index of combined task slowing (CTS), faking participants were detected correctly using the ROC analysis in our own two gender identity IATs and in two re-analyses of previous studies (Fiedler & Bluemke, 2005; Brown, 2005a) with 77% accuracy (corresponding to the weighted average of the four AUCs reported previously for the studies involving instructed faking).

The CTS index described in this paper measures the average latency increase from the faster combined task in a known-not-to-be-faked IAT to the same combined task in a possibly faked IAT. This measure is based on the assumption that the respondent's strategy in faking is to perform more slowly on the easier combined task.

The baseline IAT preceded the comparison IAT in the present research. In principle, the baseline performance might come from a subsequent IAT, although it is likely that the statistical adjustment using a CTS measure based on a subsequent IAT might differ from that for one based on a preceding IAT. The three re-analyses of previous studies showed that the CTS index is

usable both in identifying those instructed to fake group identities (e.g., *males* vs. *female*s, *Welsh* vs. *English* etc.) and those who may be motivated to fake identities (e.g., *non-pedophiles* vs. *pedophiles*).

Given its effectiveness to *detect* fakers, the use of CTS index can be extended to *correct* the faked IAT score. The section below evaluates one such approach. More specifically, an adjustment was computed for the faked IAT score by removing its faked component that was predictable from the CTS.

Adjusting IAT scores for possible faking. The adjustment procedure was designed so that the resulting adjustment would be (a) proportional to the unfaked IAT score, (b) larger for those who faked more and (c) zero for those who didn't fake. The adjusted IAT scores were computed with this formula:

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

in which coefficient *a* is the unstandardized slope of the regression of an unfaked *D* score on a previous baseline IAT, 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 and it indicated the values of CTS (measured in seconds) associated with no change in IAT scores from unfaked to a faked IAT. Subtraction of *c* from CTS makes the expected adjustment zero for participants who are not faking, thus effectively adjusting IAT scores *only* for those participants who are identified as possibly faking based on their CTS scores.

Coefficients *a* and *b* and constant *c* were calculated separately for the studies in which all three coefficients could be computed (e.g. Slope *b* could not be computed for studies that did not

include a non-faking group).[5] Table 1 displays coefficients *a*, *b*, and *c* for the following four studies: IAT2 and IAT3 of the gender identity study, Fiedler and Bluemke's Study 3 and the Welsh/English attitude study. This presentation format allows evaluating the variability of each coefficient across samples and topics. An unweighted average of all available estimates was computed for each of the three constants in the adjustment formula (*a*, *b*, and *c*). Applying these average values to the formula above resulted in the following adjustment that was used for all data sets:

$$D_{adj} = (.13 * D_{faked}) – (1.43 * (CTS – .52)).$$

Once the adjustment was computed for each of the studies, correlations with a known group membership criterion were examined separately for the adjusted and unadjusted IAT measures. Group membership criteria were: (a) participants' gender (*male* or *female*), and (b) participants' nationality (*Welsh* or *English*). Table 2 presents correlations of adjusted and unadjusted IAT measures with known group membership criteria and unfaked IAT scores for each study. Using weighted averages of correlations (as suggested by Hedges and Olkin, 1985), the adjusted measure was correlated higher with a known index of group membership (average r = .31) than was the faked measure (average r = .00). Similarly, the correlation with unfaked IAT score was higher for the adjusted measure (average r = .32) than for the faked measure (average r = .11). This increase in correlations with a known index of group membership and an unfaked IAT score indicates greater construct validity of the adjusted measure as a measure of association strength (Greenwald et al., 2003).

This strategy for correcting IAT (*D*) scores on the basis of the CTS has three noteworthy 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 *males* and *females*).

2. The "non-faked 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*, *b* and constant *c* from other data sets (see Table 1).

One practical implication of these features is that they allow computation of adjusted scores in situations in which there is no knowledge of whether somebody is faking or not. Consequently, the following 2-step adjustment procedure is recommended for IAT researchers who suspect that some participants might fake their IAT scores:

Step 1. Compute CTS as a difference between the slower combined task in the possibly faked IAT and the faster combined task in a known-not-to-be-faked IAT.

Step 2. Use the computed CTS index in the formula provided above, with the values of .13, 1.43, and .52 corresponding to coefficients *a*, *b* and *c* respectively.

When this 2-step approach was used in the Reanalysis 3, the CTS-adjusted child-sex IAT measure correlated higher with prisoner's offender status ($r = .37$, $p = .0003$) than the unadjusted child-sex IAT measure ($r = .27$, $p = .009$). These findings highlight the applicability of the 2-step adjustment procedure in settings for which there is no knowledge of whether somebody is faking or not (e.g., Reanalysis 3). Of course, these results do not provide a definitive answer to the question about the exact extent of faking that was taking place in the pedophile study. However,

the fact that adjusted child-sex IAT score produced an increased correlation with offender's known criminal status suggests that at least some of the pedophiles might have tried to distort their child-sex IAT scores.

The present approach depends on the assumption that participants are faking by slowing rather than via other strategies. The present slowing index was developed from the observation that partially effective faking could be achieved by means of deliberate slowing (Kim, 2003). 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 was effectively sensitive to faking. The strategy used to compute CTS and the adjustment index could also be applied to IAT task configurations different enough from the ones in this, so that there might be reason to question whether the values from this report should be used in such instances.

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, Brown, MacCulloch, Smith, & Snowden, 2005; Gray, MacCulloch, Smith, Morris, & Snowden, 2003). In addition, the Timed Antagonistic Response Alethiometer (TARA; Gregg, 2007) and the Autobiographical IAT (aIAT; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 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 aIAT 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 re-analyses of three other experiments involving instructed or possibly motivated faking confirmed that an index of deliberate combined task slowing (CTS) can correctly classify faked and non-faked IAT performances with an average 77% accuracy. This result contrasts with Fiedler and Bluemke's (2005) conclusion that it was "virtually impossible to identify IAT fakers" (p. 315). In addition, the CTS index was shown to be effective in adjusting faked IAT scores, increasing the correlations with known group membership and unfaked IAT score respectively from $r = .00$ and $r = .11$ with the faked measure to $r = .31$ and $r = .32$ with the CTS-adjusted measure. Taken together, these findings suggest that faking of the Implicit Association Test can be detected *and* corrected, thus highlighting resistance to faking as one of IATs advantages.

Appendix A

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.

Authors' note

Dario Cvencek, Anthony G. Greenwald. Department of Psychology, University of Washington.

Anthony Brown, Interventions, G4S Justice Services, UK.

Nicola Gray, Robert Snowden. School of Psychology, Cardiff University.

This research was supported in part by NIMH grants MH-57672 and MH-01533. The authors thank Klaus Fiedler and Matthias Bluemke for providing the data used in Renalysis 1.

Correspondence concerning this article should be addressed to Dario Cvencek or Anthony G. Greenwald, Department of Psychology, University of Washington, Box 351525, Seattle, Washington 98195. E-mails: dario1@u.washington.edu; agg@u.washington.edu.

Footnotes

[1] In addition to the indexes described in the text, we also examined (a) differences in error rates between slower and faster combined tasks in the faked IATs relative to those in non-faked IATs , (b) trial-to-trial changes in response latency in the faked IATs relative to those in non-faked IATs (up to seven consecutive responses), (c) trial-to-trial changes in error responses in the faked IATs relative to those in non-faked IATs (up to seven consecutive responses), (d) state-trace analysis (Bamber, 1979) of error responses and response latencies in the faked IATs. None of these approaches yielded information that could be interpreted as being systematically related to faking success.

[2] Alternatively, two variants of the CTS index could be computed as differences between (a) slower combined task of the faked IAT and the faster combined task in the non-faked IAT1 and (b) slower combined task of the faked IAT and the faster combined task in *following* non-faked IAT. The former variant of CTS was not as successful as the one reported in the text, as evidenced by a smaller cumulative AUC of 0.781 (*SE* = 0.048). The latter variant of CTS was more successful then the one reported in the text, as evidenced by a larger cumulative AUC of 0.877 (*SE* = 0.037). However, the disadvantage to both of these variants is that they require a three-IAT-format for their computation. Therefore, they were both dropped from further consideration, as neither of them could be applied in any of the re-analyses reported later in the text.

[3] More details for this reanalysis (and other re-analyses reported below) can be found in original publications on which each reanalysis was based.

[4] The sample consisted of another group which was comprised 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 *hebephiles* and *pedophiles* was marginally significant, $t(45) = 1.94$, $p = .059$. Given this ambiguity about *hebephiles*' IAT scores, the *hebephile* offenders were omitted from further analyses.

[5] Coefficient computations for the gender identity IAT, for example, were always conducted with samples limited to one gender group (e.g. females) not faking (i.e. high scores), that same gender group (e.g., females) faking as opposite gender group (i.e., low scores), and the other gender group (e.g., males) not faking (i.e., low scores). Similarly, for the computations of coefficients *a*, *b* and *c*, the *D change* scores were scored so that positive scores indicated change in the gender group direction that was associated with high scores (e.g. females in this example).

Figure Captions

Figure 1. Mean latencies for participants who were instructed to fake IAT2 (upper panel, *N* = 24) and participants who were instructed to fake IAT3 (lower panel, *N* = 23). Error bars = 95% Confidence Intervals.

Figure 2. Successful faking of varying magnitudes for the three studies involving instructed faking (Gender Identity IAT; Pro-Turkish Attitude IAT and Welsh/English Attitude IAT). An index of faking success (*D change*) is computed as a difference between the faked IAT *D* score and the immediately preceding non-faked IAT *D* score. Positive values indicate successful faking in the instructed direction. X-axis labels on the bottom identify the participant group for which scores are plotted. Error bars = 95% confidence interval for the mean. IAT = Implicit Association Test.

Figure 3. Receiver operating characteristic (ROC) for CTS in correctly assigning faking status in the gender identity study. The hit rate (the proportion of faking participants correctly assigned) is plotted against the false-alarm rate (the proportion of non-faking participants incorrectly assigned) as the cut-off scores for assigning faking status are varied. The diagonal line represents chance performance. ROCs are presented for 47 participants, each of whom contributed both a faked and a non-faked IAT performance. The area under the curve (AUC) is a measure of accuracy in classifying participants as faking or non-faking that corresponds to the percentage correct in a two-alternative, forced-choice detection task.

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.

Bamber, D. (1979). State trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181

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

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

Brown, A. (2005b). *Developing an Implicit Association Test for forensic use: Discriminating paedophiles from other offenders*. Unpublished doctoral thesis. Cardiff University, Wales, UK.

Butcher J. N., and Han, K.(1995). Development of inn MMPI-2 scale to assess the presentation of self in a superlative manner: The *S* scale. In J. N. Butcher and 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 measuring 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.

Gregg, A. I. (2007). When vying reveals lying: The Timed Antagonistic Response Alethiometer. *Applied Cognitive Psychology*, *21*, 621–647.

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

Greenwald, A. G., McGhee, D. E., & Schwartz, J. 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.

Hedges, L. V., Olkin I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Inquisit 2.0.60303 [Computer software]. (2006). Seattle, WA: Millisecond Software LLC.

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

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

Schnabel, K., Banse, R., & Asendorpf, J. (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.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, *99*, 100–117.

Figure 1.

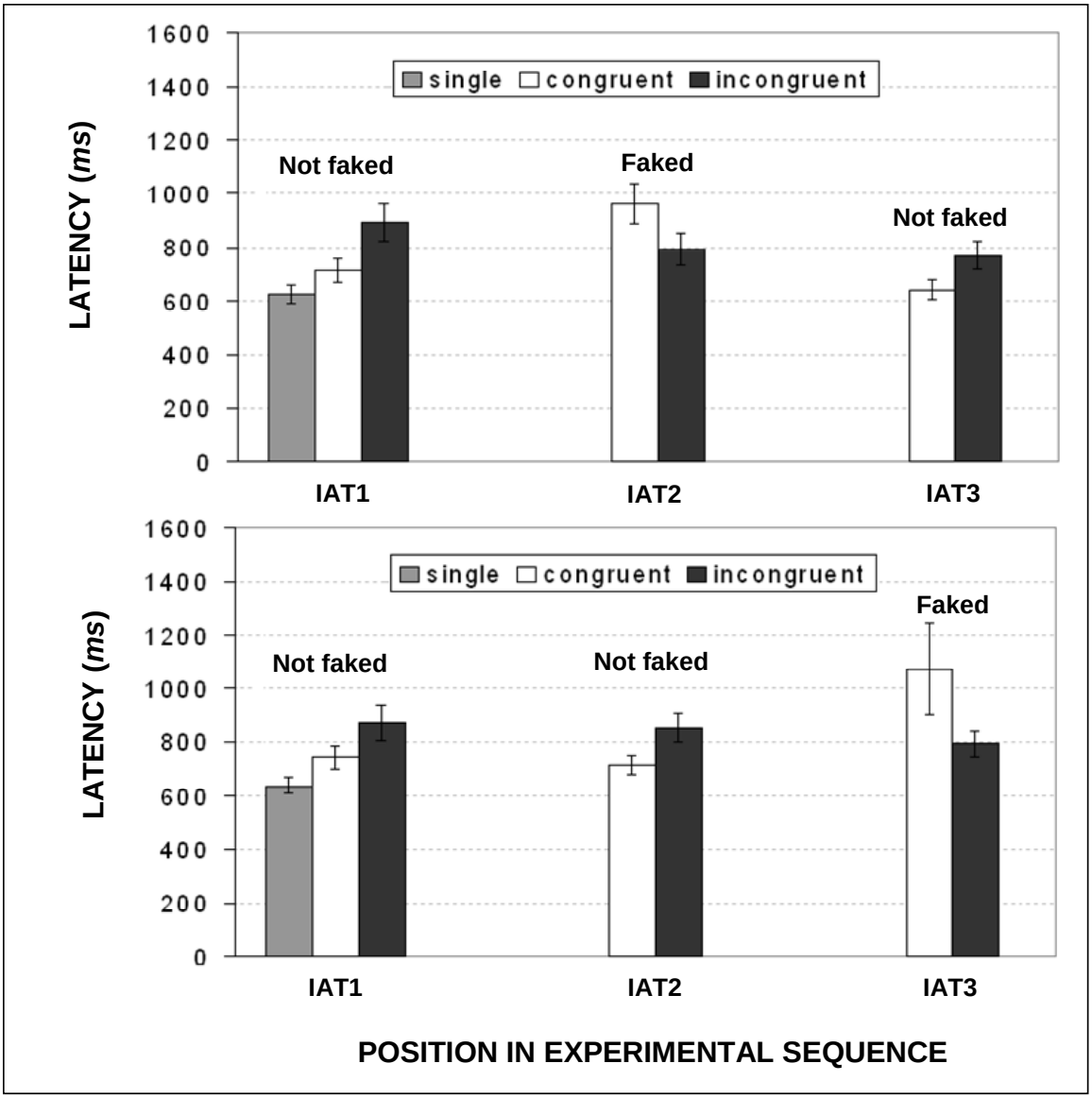
1600
1400
1200
1000
800
600
400
200
0
LATENCY (ms)
single
congruent
incongruent
Not faked
Faked
Not faked
IAT1
IAT2
IAT3
1600
1400
1200
1000
800
600
400
200
0
LATENCY (ms)
single
congruent
incongruent
Not faked
Not faked
Faked
IAT1
IAT2
IAT3
POSITION IN EXPERIMENTAL SEQUENCE

Figure 2.

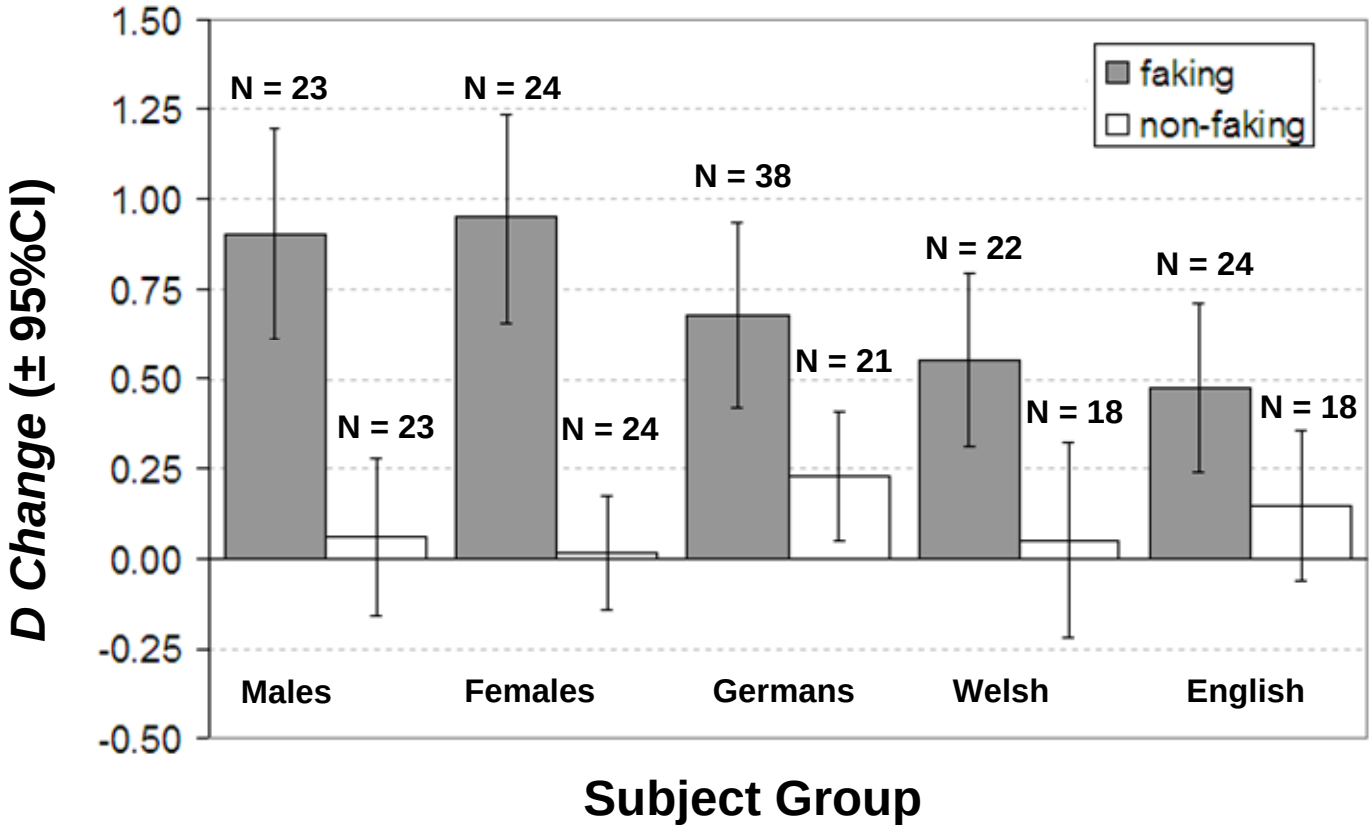
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
D Change (± 95%CI)
faking
non-faking
N = 23
N = 23
N = 24
N = 24
N = 38
N = 21
N = 22
N = 18
N = 24
N = 18
Males
Females
Germans
Welsh
English
Subject Group

Figure 3.

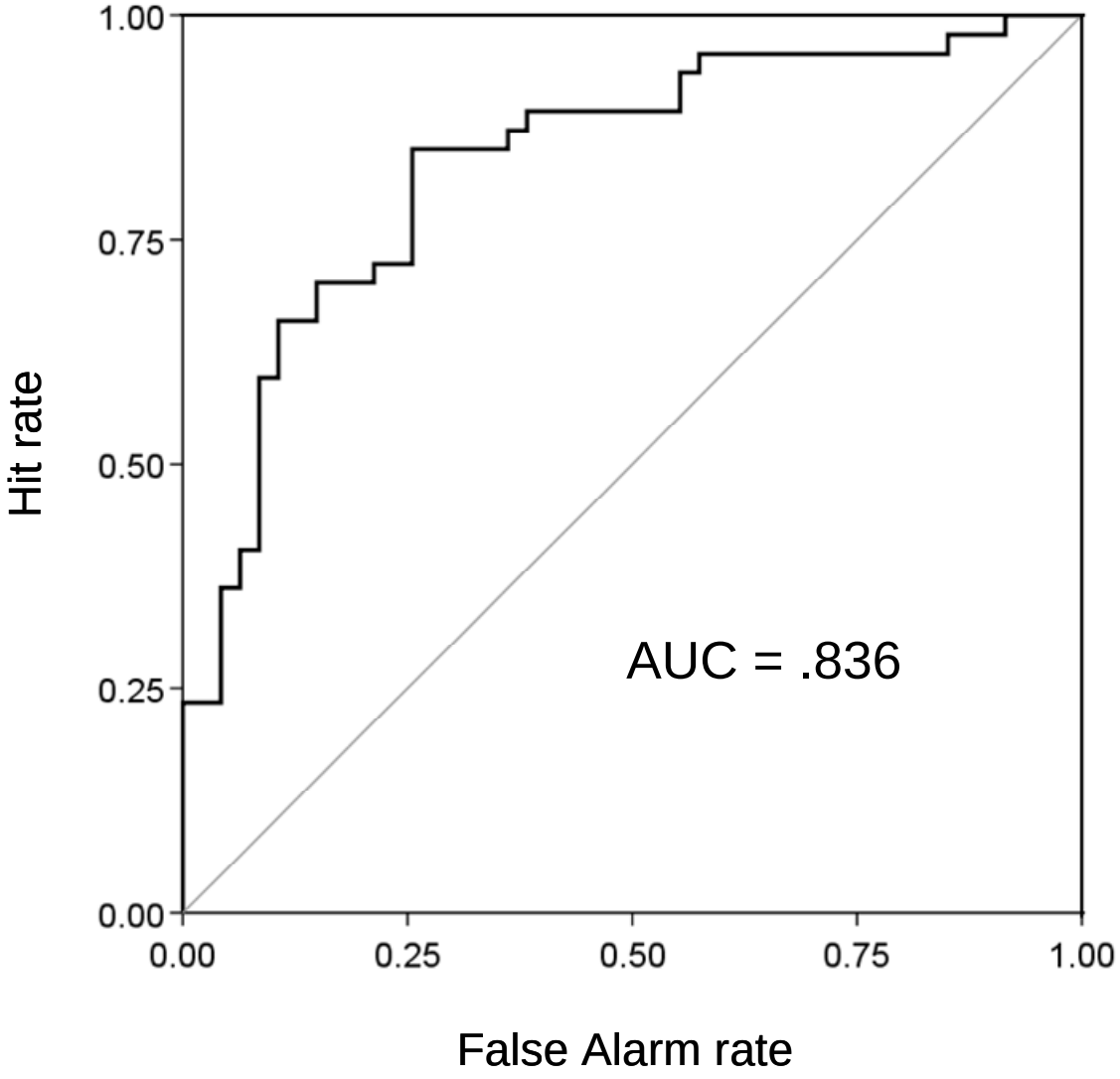
1.00
0.75
0.50
0.25
0.00
Hit rate
AUC = .836
0.00
0.25
0.50
0.75
1.00
False Alarm rate

Table 1. Values for Coefficients *a*, *b* and *c* in the Four Studies Involving Faking and Non-Faking Groups.

IAT	*Self = female* [a]				*English = positive* [b]		*Turkish = positive* [c]	Average Coefficients
Data Set	IAT2		IAT3		Study 1		Study 3	
Faking Group	Females	Males	Females	Males	Welsh	English	Germans	
Slope *a*	0.15	0.20	-0.04	0.13	0.10	0.11	0.26	0.13
Slope *b*	2.94	1.47	1.04	1.02	1.22	0.86	1.49	1.43
Intercept *c*	0.70	0.70	0.61	0.61	0.47	0.47	0.07	0.52

Note: Coefficient *a* = Unstandardized slope of the regression of an unfaked *D* score on a previous baseline IAT, reflecting the reliability of IAT measures. Coefficient *b* = 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* = 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. IAT = Implicit Association Test. The researchers contributing the four data sets for which coefficients are summarized above are: [a] Cvencek and Greenwald (2008). [b] Brown, Gray and Snowden (2005a). [c] Fiedler and Bluemke (2005).

Table 2. 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: Unfaked IAT score		Correlations with group membership: Unadjusted faked IAT score		Correlations with group membership: Adjusted faked IAT score		Correlations with unfaked IAT score: Unadjusted faked IAT score		Correlations with unfaked IAT score: Adjusted faked IAT score	
		r	p	r	p	r	p	r	p	r	p
Self = female IAT [a]											
IAT 2	*47*	.79	10^{-8}	.01	.96	.49	.001	.01	.95	.39	.006
IAT 3	*47*	.75	10^{-7}	– .09	.53	.32	.03	– .01	.94	.44	.002
English = pleasant IAT [b]	*82*	.67	10^{-12}	.05	.67	.18	.11	.16	.16	.19	.08
Turkish = positive IAT [c]											
Study 1	*50*							.14	.32	.39	.005
Study 2	*35*							.26	.13	.47	.004
Study 3	*59*							.11	.42	.19	.15
Weighted Average Correlation		r = .72		r = .00		r = .31		r = .11		r = .32	

Note: IAT = Implicit Association Test. The researchers contributing the six data sets for which correlations are summarized above are: [a] Cvencek and Greenwald (2008). [b] Brown, Gray and Snowden (2005a). [c] Fiedler and Bluemke (2005).