Applied Cognitive Psychology, Appl. Cognit. Psychol. (2010)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.1691

Detecting Fakers of the *autobiographical* IAT

SARA AGOSTA[1], VALENTINA GHIRARDI[1], CRISTINA ZOGMAISTER[2], UMBERTO CASTIELLO[1] and GIUSEPPE SARTORI[1]*

[1]*Department of Psychology, University of Padova, via Venezia 8, Padova 35131, Italy*
[2]*Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milano 20126, Italy*

Summary: Autobiographical memories might be identified using a variant of the implicit association test (IAT), or the autobiographical IAT (aIAT). The aIAT provides a measure of association between true sentences and sentences describing an autobiographical event. This tool might be used to evaluate whether specific autobiographical information is encoded within the respondent's mind/brain. This paper examines possible problems arising when the aIAT is used as a lie-detector technique. The results indicate that, when given previous instruction or training with an aIAT, examinees can alter their results and beat the 'memory-detector'. However, we have been able to detect successful fakers of aIAT on the basis of their specific response patterns. Our algorithm has the ability to spot the faker in a satisfactory manner. If, as demonstrated here, faking can be detected, then the real autobiographical event might also be identified when the examinee attempts to alter their results. Copyright © 2010 John Wiley & Sons, Ltd.

The autobiographical IAT (aIAT; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008) is a novel variant of the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which might be used to establish whether an autobiographical memory trace is encoded within a respondent's mind/brain. The aIAT is a reliable method, validated in both forensic and clinical settings (Sartori et al., 2008; Sartori, Agosta, & Gnoato, 2007), which has the ability to reveal factual knowledge regarding autobiographical events presented in a verbal format. More specifically, with the aIAT, it is possible to evaluate which of two alternative autobiographical events is true.

The aIAT includes stimuli belonging to four categories. Two categories are logical categories and are represented by sentences that are always true for the respondent (e.g. *I am in front of a computer)* or always false (e.g. *I am in front of a television*). Two other categories are represented by alternative versions of an autobiographical event (e.g. *I went to Paris for Christmas vs. I went to London for Christmas*); only one of the two being true. The true autobiographical event can be identified because it gives rise to faster response times (RTs) when it shares the same motor response with true sentences.

Lie-detection techniques fall within two categories: Methods for detecting veracity of assertions produced by an examinee, such as the control question technique (CQT; Office of Technology Assessment, 1983; Reid, 1947; Reid & Inbau, 1977) and methods that are intended to establish the existence of a specific autobiographical memory trace such as the Guilty Knowledge Test (GKT; Lykken, 1959, 1960). This latter method is also called the concealed information test (Elaad, 2009). The GKT measures deceit indirectly because it investigates the subject's knowledge regarding a particular event/crime, which can only be accessed by the offender. It is based on a series of questions regarding key elements of a case (e.g. the calibre of the gun used or some other facts pertinent to the case). The key elements are known only by the perpetrator, participating in, or witnessing the crime, or those persons informed of the details by someone who had access to the case facts (Ben-Shakhar & Elaad, 2003). Only the key question would draw a response from a person with 'guilty knowledge'. Typically, if the suspect's physiological responses (such as the galvanic skin response) to the relevant alternative are consistently larger than those to the neutral alternatives, (guilty) knowledge about the event (e.g. crime) is inferred. Because of its characteristics, the GKT might only be used in situations in which the innocent suspect was not exposed to the incriminating knowledge. The above-mentioned techniques can be used in conjunction with typical polygraphic measurements (e.g. heart-frequency; perspiration; skin conductance) as well as with functional magnetic resonance imaging (fMRI; Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003; Langleben et al., 2005) or even with non-invasive brain-stimulation such as the transcranial direct current stimulation (Priori et al., 2008).

The aIAT, like the GKT, investigates whether the examinee has a specific memory for some critical information. Used as a lie-detection technique, the aIAT has a number of unique features with respect to traditional psychophysiological techniques of lie-detection (e.g. Ben-Shakhar & Elaad, 2003) or the more recent fMRI-based lie-detection (e.g. Langleben et al., 2005). It can be administered quickly (10–15 minutes), it is based on unmanned analyses (no training for the user is necessary), it requires low-tech equipment (a standard personal computer is sufficient) and it can also be administered in a remote way to many participants (e.g. *via* the Web) for screening purposes. In contrast, other methods require a long testing session and a long session for data analysis (e.g. fMRI), a lot of training for interpreting the results (e.g. fMRI and polygraph) or require a high number of assumptions to be made to analyse the results (e.g. fMRI).

Lie-detectors may be used to screen suspects (e.g. terrorists at airports) or as a deterrent to reduce lapses in

*Correspondence to: Giuseppe Sartori, Department of Psychology, University of Padova, via Venezia 8, Padova 35131, Italy.
E-mail: giuseppe.sartori@unipd.it

safety or security regimes (e.g. in nuclear plants or other defence-sensitive plants). In an investigative or forensic setting, lie-detection systems play a key role in the defence of an innocent suspect, who may accept the test in order to prove their innocence. In this specific case, faking is a highly unlikely and irrational strategy. By contrast, guilty suspects have no interest in taking a test that is likely to prove their guilt. For this reason, they will be more likely to either refuse the test or, in the remote event of accepting the test, to be prone to faking the test.

The ideal lie-detector, for investigative and forensic applications, should minimize false positive errors, which make an innocent suspect appear guilty. This error is expected when the examinee is not faking in the test. By contrast, when examining a guilty suspect who may take advantage and fake the test, false negatives, which confuse the guilty subject for an innocent subject, have to be minimized.

Effective countermeasures are now known for almost every lie-detection technique. The first study investigating the efficacy and detection of polygraphic countermeasures goes back to Benussi (1914) who also introduced the first respiratory-based lie-detection technique. Countermeasures for the CQT have long been known. Most attempts to increase the response of a subject to the control questions have been to use physical (e.g. biting the tongue or pressing the toes to the floor) or mental (e.g. counting to seven, backwards) techniques (Honts, Raskin, & Kircher, 1994). Countermeasures against the GKT, when used with a polygraph, have also been demonstrated (Honts, Devitt, Winbush, & Kircher, 1996).

With respect to the aIAT countermeasures, Agosta (2005) and more recently Verschuere, Prati, and De Houwer (2009) have shown that properly trained participants may alter strategically the test outcome. Verschuere et al. (2009) instructed guilty participants in a mock-crime task to appear as innocent by slowing down their responses. Their results indicated that a great proportion of guilty participants, not previously exposed to the aIAT, succeeded in faking the test.

A critical aspect, however, which has not been fully investigated, is whether fakers could be detected on the basis of their response patterns. In fact, for the aIAT, detection of fakers should render countermeasures ineffective. A study by Fiedler and Bluemke (2005) showed that IAT experts are unable to identify faked IAT results on the basis of their expertise. Their experts were requested to evaluate the results of 24 participants (half were honest responders and half were dishonest responders) and to identify the dishonest responders. The results show that experts were unable to identify the dishonest responders on the basis of their IAT latencies. In contrast, recent evidence that IAT fakers can be detected comes from Cvencek, Greenwald, Brown, Gray, and Snowden (under review) who demonstrated that fakers can be uncovered by examining specific features of their response patterns.

Here, we report on a series of experiments aimed at both confirming and enhancing the validity of aIAT as a tool for evaluating autobiographical memories, even under circumstances in which faking is suspected. Although we confirm that the aIAT might be faked by appropriately instructing participants, we demonstrate that faking participants might be detected on the basis of their response patterns. We found that fakers leave a signature and, most importantly, this signature is valid for various unrelated aIATs. This signifies that this marker can be potentially used to check the authenticity of an aIAT, without a specific normative group of true and adulterated performances. Furthermore, we report on an algorithm specifically implemented for the identification of respondents who eventually succeed in faking the test.

FAKING THE *autobiographical* IAT

In this section, we describe four experiments aimed at evaluating whether participants, who were overtly instructed or simply trained previously in using an aIAT, can intentionally alter their aIAT outcome.

Methods and procedure were similar for all the experiments included in this section unless specified. The general methodology of the aIAT was the same as that described in Sartori et al. (2008). Here, we describe the general procedure for Experiments 1 and 2.

Sentences belonging to the logical category *true/false* and sentences describing two autobiographical events with only one of them being true (e.g. *Christmas in Paris vs. Christmas in London)* were used (true and false autobiographical events were specific for each participant and collected earlier using a questionnaire). The aIAT is accomplished by requiring the respondent to complete five blocks of speeded categorization trials. Participants are requested to classify sentences by pressing one of two labelled keys, one positioned on the left of the keyboard (e.g. 'A') and one situated on the right of the keyboard (e.g. 'L'). Sentences are presented in the centre of the monitor and two reminder labels are positioned, one on the left and one on the right of the monitor. These two labels show the name of the categories that must be used in order to classify each sentence. Two out of the five blocks (critical blocks) require the double categorization of an autobiographical event (e.g. *Christmas in Paris* or *Christmas in London*) with certainly-true events.

In Block 1 (20 trials), participants had to classify certainly-true or -false sentences, by pressing the left key to classify certainly-true sentences (five different sentences; e.g. *I am in front of a computer*) and the right key to classify certainly-false sentences (five different sentences; e.g. *I am in front of a television*). In Block 2 (20 trials), participants had to classify autobiographical sentences. They pressed the left key to classify real autobiographical-events sentences (five sentences; e.g. *I saw the Eiffel Tower*) and the right key to classify false autobiographical-event sentences (five sentences; e.g. *I saw Big Ben*). In Block 3 (60 trials), the left key was used to classify both certainly-true and real autobiographical-events sentences, whereas the right key was used to classify both false and false autobiographical-events sentences (congruent block). In Block 4 (40 trials), the left key was used to classify false autobiographical-events sentences, whereas the right key was used to classify real autobiographical-events sentences. Finally, in Block 5 (60 trials), participants had to classify with the left key both true and false autobiographical-events sentences, and with the right key, they had to classify false and real autobiographical-events sentences (incongruent block).

As the pairing of a true autobiographical event with certainly-true sentences should facilitate a specific response, the specific pattern of RTs in the two critical blocks (3 and 5) indicates which autobiographical event is true and which autobiographical event is false.

Following an aIAT training session, participants were randomly assigned to one of three groups that differed in terms of their instructions. *Non-faking* participants received the standard aIAT instructions (i.e. they were requested to categorize the sentences as indicated by the labels by pressing the appropriate keys as fast and as accurately as possible); *naïve-faking* participants were asked to do their best to hide their true autobiographical memory to the experimenter (Fiedler & Bluemke, 2005) but they were not instructed on how to fake the test. *Instructed-faking* participants were instructed to slow down in the congruent block and speed up in the incongruent block (Kim, 2003). Note that only participants taking part in Experiment 1 (i.e. the Christmas aIAT experiment) were not administered the preliminary aIAT training session.

In all studies, the order of the double categorization blocks was counterbalanced across subjects (congruent block first or congruent block after the incongruent block). In the next section, we describe the procedures for each of the four experiments; the findings for these experiments will be grouped and reported within the Results section.

Experiment 1: Faking without preliminary training in the aIAT

Forty-two participants (eight males and 34 females; age range 19–30 years) were randomly assigned to one of the three groups: 14 to the non-faking group; 14 to the naïve-faking group and 14 to the instructed-faking group. Participants were requested to complete a questionnaire regarding their last Christmas holiday (e.g. *Where were you on Christmas day?*) and a Christmas holiday they never had. For each participant, a specific aIAT was built with sentences describing the true holiday and the holiday they never had. Participants pressed one of two keys corresponding to the location where they spent the holiday (e.g. *Home* (real vacation) vs. *Mountain* (false vacation). For the congruent block, true sentences and real holiday sentences were assigned to the same response key. For the incongruent block, true sentences and real holiday sentences were assigned to different keys.

Experiment 2: Christmas holiday aIAT with previous aIAT experience

In order to investigate the effect of previous aIAT experience on the ability of participants to fake the test, and on our ability to detect fakers, we conducted a second study.

Fifty participants (14 males and 36 females; age range 19–30 years) took part in the experiment. Twenty participants were assigned to the non-faking group, ten participants to the naïve-faking group and 20 participants to the instructed-faking group. Participants were administered an aIAT pre-test session (a two-card aIAT). Then they received the Christmas aIAT, as in Experiment 1.

Experiment 3: Ten-card aIAT with the preliminary aIAT session

Experiments 3 and 4 were conducted in order to generalize the results in relation to short-term memory. In these experiments, participants were requested to respond to a previously-selected card and the procedure was similar to that of Experiments 1 and 2, except that sentences about the true vacation were substituted with sentences regarding the selected card; and sentences regarding the false vacation were substituted with sentences regarding a non-selected card.

Seventy-two participants (20 males and 52 females; age range 19–30 years) were randomly assigned to one of the three groups: 20 participants to the non-faking group; 18 participants to the naïve-faking group and 34 participants to the instructed-faking group. At the beginning of the experiment, participants were administered a preliminarily two-card aIAT, as training. Subsequently, they were requested to choose one among ten different playing cards. After a consolidation task, consisting of identifying the previously-selected card among other unrelated items (see Sartori et al., 2008), a subject specific 'ten-card' aIAT was administered to each participant. Here, the real autobiographical event was represented by the choice of the picked card (e.g. *I picked the card 2 of hearts*), whereas the false autobiographical event was represented by the choice of the other cards (e.g. *I picked the card 3 of clubs*). One of the two reminder labels corresponded with the selected card (e.g. *2 of hearts*), whereas the other label was referred to as 'other cards'.

Experiment 4: Two-Card aIAT with preliminary training in the aIAT

Thirty-six participants (12 males and 24 females; age range 19–30 years) were randomly assigned to the three groups: 12 to the non-fakers group, 12 to the naïve-fakers group and 12 to the instructed-fakers group. Procedures and stimuli were the same as for Experiment 1, reported in Sartori et al. (2008), except that participants were administered a preliminary aIAT training session (a cigarette aIAT) aimed at evaluating whether the respondent was a smoker or not. After the aIAT training, participants selected one of two playing cards (four of diamonds or seven of clubs) and were asked to memorize it during a consolidation task (see Experiment 3). After the consolidation task, participants performed the 'two-card' aIAT. Here, the real autobiographical event was represented by the actual choice of the card (e.g. *I picked the card number 4*), whereas the false autobiographical event was represented by the choice of the other card (e.g. *I chose the card number 7*).

RESULTS AND DISCUSSION

For all the experiments, the dependent measures were RT (between 150 and 10000 ms), D-IAT (D600 algorithm;

Greenwald, Nosek, & Banaji, 2003) and accuracy. For each experiment and for each group, we ran an analysis of variance (ANOVA) on mean RT with congruency (congruent *vs*. incongruent) as a within-subjects factor and order of presentation of the congruent block (congruent first in Block 3 *vs*. congruent second in Block 5) as a between-subjects factor. No significant main or interaction effect involving order of the presentation of the congruent block (first *vs*. second) emerged. Therefore, only the main effect of congruency will be discussed. Table 1 shows mean RT and accuracy percentage for each group.

Further, an ANOVA was conducted on accuracy with congruency (congruent *vs*. incongruent) as a within-subject factors, and order of presentation of the congruent block (congruent first in Block 3 *vs*. congruent second in Block 5) as between-subjects factor. No significant main or interaction effect involving order of the presentation of the congruent block (first *vs*. second) emerged for this second ANOVA on accuracy, except for the naïve-faking group in Experiment 1 ($F(1,12) = 7.594$, $p = .017$, $\eta 2 = 2.388$), and in Experiment 3 ($F(1,16) = 10.206$, $p = .006$, $\eta 2 = 0.2389$). In all experiments and in all groups, accuracy was higher in the congruent than in the incongruent block, except for the instructed faker of Experiment 3 ($F(1,32) = 20.700$, $p < .001$, $\eta 2 = 0.393$) and 4 ($F(1,10) = 0.647$, $p = .440$, $\eta 2 = 0.061$).

Non-faking groups

The non-faking groups, for all experiments, showed faster RTs for the congruent than for the incongruent block (Experiment 1: RTs, $F(1,12) = 21.657$, $p < .001$, $\eta 2 = 0.643$; Experiment 2: $F(1,18) = 33.862$, $p < .001$, $\eta 2 = 0.653$; Experiment 3: $F(1,18) = 71.012$, $p < .001$, $\eta 2 = 0.798$; Experiment 4: $F(1,10) = 11.539$, $p = .007$, $\eta 2 = 0.537$).

Results for the non-faking groups (Experiments 1, 2, 3 and 4) showed that the aIAT can detect the real autobiographical information with an accuracy rate above 92%. Experiments 1 and 2 were a replication of Experiment 4 of Sartori et al.'s (2008). The only difference here was that, in Experiment 2, participants had previously experienced another aIAT. This previous practice did not reduce the accuracy of the aIAT in detecting the real autobiographical event (90% in the original experiment and 95% in this replication). However, previous practice reduced the magnitude of the D-IAT index. This is clear, when comparing non-fakers from Experiments 1 and 2, who differed only because the second group had received previous training. As shown in Figure 1, average D-IAT for

Table 1. Experiments 1, 2, 3 and 4: Summary table for the main results (rts and accuracy). In Experiments 1, 2 and 3, a positive D-IAT index indicates the correct identification of the autobiographical information. Instructed-fakers are successful in transforming the positive score into a negative score. In Experiment 4, positive D-IAT indicates the autobiographical information 'card four' whereas the negative D-IAT indicates 'card seven'. Both naïve-fakers and instructed-fakers reduce the difference between the two D-IATs when compared with non-fakers. Accuracy is generally higher for the congruent than for the incongruent block; in Experiment 3, only instructed-fakers do not show this pattern

Experiment 1	no. 42	Congruent	Incongruent	D-IAT	Correct
Non-fakers	14	1149 ms 96.90%	2104 ms 84.88%	1.06	14/14
Naïve-fakers	14	1360 ms 95%	2045 ms 92.26%	.78	14/14
Instructed-fakers	14	2381 ms 98.45%	1781 ms 88.21%	−0.45	5/14
Experiment 2	**no. 50**	**Congruent**	**Incongruent**	**D-IAT**	**Correct**
Non-fakers	20	1163 ms 95.16%	1608 ms 92.16%	0.64	19/20
Naïve-fakers	10	1520 ms 88.5%	1692 ms 83.83%	0.24	6/10
Instructed-fakers	20	1967 ms 96.5%	1535 ms 86.33%	−0.42	7/20
Experiment 3	**no. 72**	**Congruent**	**Incongruent**	**D-IAT**	**Correct**
Non-fakers	20	1068 ms 96.42%	1752 ms 88.67%	1.13	20/20
Naïve-fakers	18	1024 ms 95.19%	1545 ms 91.57%	0.82	18/18
Instructed-fakers	34	1976 ms 82.94%	1313 ms 91.76%	−0.81	4/34

Experiment 4	no. 36	Congruent	Incongruent	D-IAT (card four)	D-IAT (card seven)	Correct
Non-fakers	12	1081 ms 97.77%	1317 ms 99.30%	0.37	−0.52	11/12
Naïve-fakers	12	1241 ms 93.61%	1315 ms 92.91%	0.03	0.06	6/12
Instructed-fakers	12	1155 ms 93.47%	1248 ms 94.44%	0.27	0.15	7/12

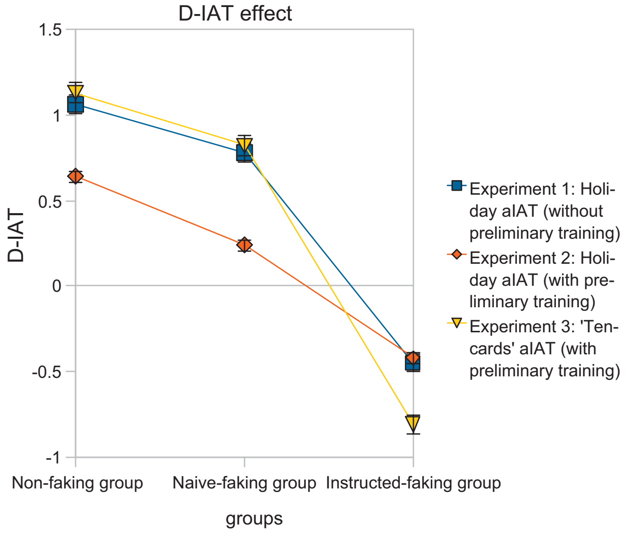

Figure 1. Experiments 1, 2 and 3: D-IAT index for the three groups; non-fakers, naïve-fakers and instructed-fakers. The D-IAT, which is positive for the non-fakers indicating a correct classification, becomes negative for the instructed-fakers, indicating an erroneous classification. Please note instructed participants can reverse the direction of the effect, therefore, succeeding in falsifying the test outcome. Naïve fakers fall in between these two groups

non-fakers in Experiment 1 was 1.06, whereas average D-IAT for non-fakers in Experiment 2 was .64, $F(1,32) = 9.790$, $p = .004$, $\eta2 = 0.234$.

The high accuracy of the procedure was also confirmed in Experiment 3. In this experiment, participants were requested to select and memorize one of ten playing cards. In such circumstances, no classification error was observed (i.e. 100% accuracy). In Experiment 4, which was a replication of Experiment 1 of Sartori et al.'s (2008), accuracy reached 92%.

In sum, we confirmed that when participants are not instructed to alter intentionally the outcome of an aIAT, the actual autobiographical memory can be detected with high accuracy. Memory detection was equally effective for short-term memories (selected cards) and for longer term real-life memories (last Christmas vacation).

Naïve-faking groups

When participants were instructed to modify intentionally the outcome of the aIAT without receiving explicit instructions from the experimenter, the results were less consistent (Experiment 1: $F(1,12) = 27.423$, $p < .001$, $\eta2 = 0.696$; Experiment 2: $F(1,8) = 1.549$, $p = .248$, $\eta2 = 0.162$; Experiment 3: $F(1,16) = 29.604$, $p < .001$, $\eta2 = 0.649$; Experiment 4: $F(1,10) = .514$, $p = .490$, $\eta2 = 0.049$). On some occasions, about 45% of the participants succeeded in faking (Experiment 2 = 40% and Experiment 4 = 50%). On other occasions, they were unable to fake the results (Experiment 1 = 0% and Experiment 3 = 0%).

Results do not seem to be related to the preliminary training with the aIAT, given that participants did not succeed in faking the test in both Experiment 1 (without the preliminary aIAT) and Experiment 3 (with the preliminary aIAT).

Instructed-fakers

When participants were instructed explicitly to slow down on congruent trials, and speed up on incongruent trials, most of them were able to modify their performance in the instructed direction (Experiment 1: $F(1,12) = 3.705$, $p = .078$, $\eta2 = 0.263$; Experiment 2: $F(1,18) = 4.583$, $p = .046$, $\eta2 = 0.203$; Experiment 3: $F(1,32) = 27.802$, $p < .001$, $\eta2 = 0.465$; Experiment: $F(1,10) = 0.165$, $p = .693$, $\eta2 = 0.016$). Therefore, participants could reverse the aIAT effect and consequently falsify the outcome of the test. The percentage of successful fakers ranges from a minimum of 42% in Experiment 4 to a maximum of 88% in Experiment 3. To sum up, we found that the aIAT is vulnerable to faking, at least when participants are explicitly instructed to slow down their responses on congruent trials. Figure 1 depicts the observed inversion of the D effect in this group of participants.

DETECTING FAKERS

Findings from the series of experiments reported above show that a high proportion of participants can revert their results when instructed to do so. This result renders the aIAT vulnerable to countermeasures if used in real-life forensic settings, unless an effective procedure for detecting fakers is developed.

To address this issue, we further analysed the performance of a subgroup of participants from those taking part in Experiments 2, 3 and 4. A total of 50 non-fakers (19 from the holiday experiment, 20 from the ten-cards experiment and 11 from the two-card experiment) were contrasted with 58 successful fakers (17 for the Christmas experiment: 4 naïve-fakers and 13 instructed-fakers, 30 from the ten-card experiment: 0 naïve-fakers and 30 instructed-fakers and 11 from the two-card experiment: 6 naïve-fakers and 5 instructed-fakers). Non-fakers and fakers were selected using the D-IAT. In order to validate faking indexes, which were not task dependent, relevant data from the different experiments were collapsed.

Description of the indexes

As reported above, an efficient strategy to faking effectively consists of slowing down the congruent trials. For this reason, we analysed the difference in the RTs between the simple blocks and the double blocks in the two groups. An analysis comparing the successful fakers and correctly classified non-fakers, showing that the difference between single blocks (1, 2 and 4) and double blocks (3, 5) is larger in fakers than in non-fakers, $F(1,106) = 26.927$, $p < .001$, $\eta2 = 0.203$.

We, therefore, developed and analysed the candidate indexes to detect faking on the basis of this information. The two best indexes are reported in Table 2, in which their accuracy in classifying fakers and non-fakers is also reported. The methods that most efficiently discriminate between the two groups are variants of a comparison between RTs in double blocks (Blocks 3 and 5) and single blocks (Blocks 1, 2 and 4). Specifically, the efficiency of the best index was calculated using a cut-off = 1.08 (Table 2). We considered values above the cut-off, indexing fakers and values below the cut-off, indexing non-fakers. The cut-off

Table 2. Here we report the two most efficient indexes that discriminate between fakers and non-fakers. The logic captured by these indexes is robust as we have found ten other indexes with AUC above 0.8. These non-reported indexes are variants of the two reported ones, with changes in the cut-off and penalties. The indexes encode the selective increase in rts for the double classification blocks with respect to the single classification block that characterize successful fakers. The procedures for eliminating extremely fast and slow responses and the penalty for the errors were inspired by Greenwald et al. (2003). The more efficient algorithm for detecting fakers consisted of three steps: (i) First eliminate all responses below 150 and above 10000 ms; (ii) substitute errors with the mean of the block added with a penalty of 600 ms; (iii) calculate the ratio between the average RT of the fastest block (between 3 or 5) and single tasks that are directly connected to the fastest task, in terms of motor response (1 and 2 or 1 and 4, respectively). If the result exceeds 1.08 then the respondent is faking. Threshold was identified as the cut-off, which yielded the maximal classification accuracy (average index is for non-fakers 0.94 and for fakers 1.3)

Index	Description	AUC	Classification accuracy using D-IAT
Ratio 150–10000 with penalty	(i) Use only the RTs between 150 and 10000 ms; (ii) substitute errors with the mean of the corresponding block with an added penalty of 600 ms; (iii) divide the average RT of the fastest block by the average RTs of the corresponding single blocks (1, 2 or 1, 4).	0.88	82.4
Slow down 150–10000 with penalty	(i) as above; (ii) as above; (iii) subtract to average RT of the fastest block (3 or 5) the average RTs of the corresponding single blocks (1, 2 or 1, 4).	0.88	83.3

was calculated using Binary Logistic Regression under the assumption that false alarms and missed responses had equal costs, which corresponds to the value of 50% probability of the respondent being a faker. In other words, higher values indicate that the probability of being a faker is more than 50% and lower values indicate that the probability of being a faker is fewer than 50%. Of course, in many cases, costs for the two types of mistakes are different; hence, higher or lower cut-off values should be used accordingly. We also analysed the best index (see Table 2) by using median RTs rather than average RTs. The median-based faking-detection index yielded an area under the curve (AUC[1]) of 0.87 and the classification accuracy based on a Binary Logistic Regression was 81.5%.

The previous results were derived from participants who were analysed in their second aIAT, after the aIAT training. In order to evaluate if the same procedure was effective in detecting fakers who did not carry out a preliminary aIAT, we analysed a further subgroup of 14 non-fakers and nine fakers from Experiment 1 (Christmas aIAT, without training). In this case, the classification accuracy was calculated using the index AUC of ROC analysis (AUC = 0.88) and 19/23 participants were correctly classified using Binary Logistic Regression on the D-IAT. Therefore, this index seems to be quite robust as it classifies very reliably participants from different aIATs, and also participants with and without previous aIAT training.

One might argue that an efficient countermeasure to faking detection might also imply a generalized slowing down for all blocks. Indeed, this strategy would invalidate the faking detection strategy, based on comparing single *versus* double blocks. However, this countermeasure would be quite easy to detect given that participants should manifest abnormally longer reaction times for simple blocks. By contrast, the analysis of simple blocks for all our experiments indicated that instructed fakers are not prone to slow down on single blocks, but surprising, they are slightly faster (Non-fakers = 1184 ms, sd = 270; Naïve fakers = 1124 ms, sd = 414; Instructed fakers = 1026 sd = 245; $F(2,188) = 5.146$, $p = .007$, $\eta^2 = 0.050$).

The results reported above, refer to an in-sample analysis. In order to evaluate whether the proposed detection strategy generalizes across tasks, we performed a cross-validation analysis. We used data of fakers and non-fakers for Experiments 1 and 2 (i.e. autobiographical) in order to calculate the cut-off, which was then used to evaluate classification accuracy of participants (fakers and non-fakers) for Experiments 3 and 4 (i.e. cards). In-sample classification accuracy of participants for Experiments 1 and 2 was 75% (using the cut-off of 1.13), whereas out-of-sample classification of participants of Experiments 3 and 4 was 79.8% (using the same cut-off). We also calculated the cut-off using Experiments 3 and 4 (cards) and used this to classify participants for Experiments 1 and 2. In-sample classification accuracy of participants for Experiments 3 and 4 was 84.7% (cut-off 1.06), whereas out-of-sample accuracy of participants for Experiments 1 and 2 was 75.3%. These data indicate that the index used for detecting fakers may be generalized across differing tasks.

Including errors in the analysis did not improve faking detection and, even if fakers produce more classification errors than non–fakers, the pattern is not very efficient in discriminating the two groups across experiments. Classification accuracy for fakers and non-fakers, on the basis of the ratio between accuracy in double blocks and single blocks, yielded an AUC of 0.79. These results were observed when classifying fakers and non-fakers from Experiments 2, 3 and 4, in which participants had preliminary training. Accuracy analysis does not efficiently classify participants for Experiment 1 in which there was no preliminary training (AUC = 0.46).

CONCLUSIONS

The autobiographical IAT might be used as a memory-detection technique in forensic setting in which guilty

[1] The area under the curve (AUC) is a measure of accuracy in classifying subjects as faking or non-faking and corresponds to the percentage that is correct in a two-alternative, forced-choice detection task.

suspects may be prone to faking. We conducted four experiments comparing non-faking participants with naïve-faking participants and instructed participants.

In these experiments, the aIAT correctly identified the autobiographical event in non-faking participants (correct identification average over all experiments = 96.7%). We also showed that a significant proportion of naïve-fakers succeeded in faking the test using spontaneously-developed strategies (22.5%) and this proportion was much higher when participants were trained to use an optimal faking strategy (65% of them succeeded in making the experimenter believe what was not true; see also Verschuere et al., 2009).

Therefore, the aIAT could be faked using self-discovered strategies or, much more efficiently, using coached strategies. We studied the effectiveness of the self-discovered strategies of naïve-faking participants who might autonomously develop a procedure so as to alter the results. In these cases, participants were instructed to fake, but they were not explicitly told how. Under these circumstances, previous experience with an aIAT facilitated the development of a self-discovered faking strategy. Indeed, when comparing the holiday aIAT without training and the holiday aIAT with training, the percentage of naïve-fakers, who succeeded in faking the test, increased from 0 to 40%.

Instructed-fakers, by contrast, were explicitly taught an optimal strategy consisting of slowing down during the congruent block and speeding up during the incongruent block. Most of them were successful in faking the test, and previous exposure to an aIAT did not increase the percentage of successful fakers. In fact, for the holiday aIAT with and without previous training, participants who succeed in faking the aIAT were 65 and 64.2%, respectively. With respect to the speeding-up observed for incongruent trials, we found that only in Experiment 1 participants were faster for incongruent trials when faking (Non-fakers = 2104 ms; Instructed fakers = 1781 ms). Speeding up for incongruent trials was not found in Experiments 2, 3 and 4. This difference, between Experiment 1 and the other experiments, might be due to the fact that in Experiment 1, we did not administer a practice aIAT prior to the experimental task. Non-faking, practiced participants for Experiments 2, 3 and 4 were presumably responding at their maximum possible speed for the incongruent trials, having previously been trained with a practice aIAT.

We did not investigate whether participants were aware of their success in faking the test, but Kim (2003) specifically tested awareness of strategies in naïve and instructed fakers of a classical attitude IAT. This author reported that only 3/24 participants, who received explicit instructions, believed they were successful in faking the test.

Noticeably, when faking, participants left behind their signature: They did not alter their RTs in single blocks and they were abnormally slow in double blocks as compared to single blocks. We showed that this feature might be used to detect fakers with reliable accuracy. Ideally, the system for detecting fakers should generalize across subjects and conditions. We, therefore, tested the algorithm accuracy on participants, who were not involved in the model's development phase. Therefore, accuracy was tested with an out-of-sample procedure. Furthermore, the algorithm is equally effective for participants who did have, or who did not have a preliminary aIAT. Finally, it is important to note that the algorithm does not require previous knowledge regarding the congruent block, given that this aspect would not be known in in-field applications.

The present research sought to identify an indicator of the deliberate slowing of responses that might be considered as an index of faking the aIAT outcome. The rationale underlying the development of our marker of faking was based on the observation of typical IAT results and our aIAT studies. In particular, Fiedler and Bluemke (2005) showed that participants were not able to speed up responses for the incongruent block. Furthermore, Greenwald et al. (1998), for the IAT, and Sartori et al. (2008), for the aIAT, found that latencies for non-fakers in congruent blocks were comparable to those of single blocks.

In addition, Cvencek et al. (under review) reported an alternative and effective faking detection procedure, which might complement the procedure presented here. These authors compared RTs for the double blocks of two IAT, which were administered to the same participant. Their faking detection index, the combined task slowing (CTS), consists of the difference between 'the slower combined task for the faked IAT and the faster combined task for the preceding non-faked IAT'. In our Experiments 2, 3 and 4, we administered to participants a preliminary practice aIAT (which was different from the IAT analysed in full) and therefore, we can calculate the CTS on our data sets. We applied the CTS, contrasting non-fakers ($n = 50$; belonging to the non-faking group) with participants from the instructed-faking group, with practice aIAT, who successfully faked the test ($n = 48$). In this case, the CTS yielded an AUC equal to 0.75. Differences in the experimental design do not permit us to apply our index to Cvencek at al.'s data. In fact, they collected responses to single blocks (which were used in our algorithm to distinguish non-fakers from fakers) only in the first non-faked IAT. In their subsequent faked IAT, only double blocks were administered.

In sum, although Cvencek et al.'s faking detection index and the index reported here are based on differing logics, they both allow us to detect fakers of IAT efficiently. Indeed, they might complement each other: Whereas Cvencek et al.'s faking indicator might be applied when two different IATs are administered to the same subject, a first non-faked IAT and a second suspected-faked IAT; our indicator might be used when a single aIAT is administered to a suspected faker.

One might argue that the laboratory experiments reported here are very different from in-field lie-detection applications in which participants might be expected to be very anxious about the results of their performance. If high anxiety is reflected in an increase in reaction times in double blocks, then a non-faker could, in such situations, be misclassified as a faker. In order to evaluate this hypothesis we re-analysed the data for Experiment 5, which was originally reported in Sartori et al. (2008). Participants for the experimental group (25 participants) had their driving license suspended for driving with an excessive blood alcohol level. They were examined as part of a medico-legal screening and were let to believe that the aIAT outcome would determine whether or not their driving license would be reinstated. By contrast, control participants were never caught by the police with excessive blood alcohol level and they were tested in the laboratory. The drunk drivers group can be considered a high anxiety and low-faking group for the following reasons: (i) They have no advantage in faking the test (i.e. responding as if they had not driven while drunk) because drunk driving was already established with incontrovertible evidence; (ii) the setting was anxiety-prone, as drunk drivers knew that the reinstatement of their driving license depended on their results of the reaction time test. Results showed that in the field, anxiety did not cause a slowdown during the double blocks (average RT on double blocks for the drunk drivers = 1984 ms and for the control group = 1995 ms), supporting the generalization of the present results to more stressful and anxious situations.

In conclusion, we confirm that the aIAT is a simple, but powerful procedure for evaluating autobiographical memories. When used as a lie-detection technique, it can be faked, but fakers can be identified. The indexes that we have developed are quite robust, given that minor changes in the algorithm did not cause significant reductions of classification accuracy. Further, they provided similar classification accuracy when analysing participants with and without preliminary IAT practice.

In sum, our results confirm for the aIAT, the findings of Cvencek et al. (under review) on the IAT who concluded that 'faking of the Implicit Association Test can be detected and corrected, thus highlighting the resistance to faking as one of IATs advantages'.

REFERENCES

Agosta, S. (2005). A new lie-detector based on the IAT. *Master Degree Thesis*.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology*, *88*, 131–151. DOI: 10.1037/0021-9010.88.1.131

Benussi, V. (1914). Die Atmungssymptome der Lüge. *Archiv für die gesamte Psychologie*, *31*, 244–273. English translation printed in 1975. Poligraph, 4(1), 52–76.

Cvencek, D., Greenwald, A. G., Brown, A., Gray, N. S., & Snowden, R. J. (under review). Faking of the implicit association test is statistically detectable and partly correctable.

Elaad, E. (2009). Effects of context and state of guilt on the detection of concealed crime. *International Journal of Psychophysiology*, *71*, 225–234. DOI: 10.1016/j.ijpsycho.2008.10.001

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association test. *Basic and Applied Social Psychology*, *27*, 307–316. DOI: 10.1207/s15324834basp2704_3

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex*, *13*, 830–836.

Greenwald, A. G., McGhee, D. E., & Schwarz, J. L. K. (1998). Measuring individual difference in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. DOI: 10.1037/0022-3514.74.6.1464

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. DOI: 10.1037/0022-3514.85.3.481

Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology*, *7*, 10–14. DOI: 10.1111/j.1469-8986. 1996.tb02111

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of the polygraph test. *Journal of Applied Psychology*, *79*, 252–259. DOI: 10.1037/0021-9010.79.2.252

Kim, D. Y. (2003). Voluntary controllability of the implicit association test. *Social Psychology Quarterly*, *66*, 83–96.

Langleben, D. D., Loughead, J. W., Bilker, W. B., Ruparel, K., Childress, A. R., Busch, S. I., et al. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping*, *26*, 262–272. DOI: 10.1002/hbm.20191.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*, 385–388.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*, 258–262.

Office of Technology Assessment. (1983). Scientific validity of polygraph testing: A research review and evaluation. A technical memorandum, *Report No. OTA-TM-H-15*, Washington, DC: Office of Technology Assessment.

Priori, A., Mameli, F., Cogiamanian, F., Marceglia, S., Tiriticco, M., Mrakic-Sposta, S., et al. (2008). Lie-specific involvement of dorsolateral prefrontal cortex in deception. *Cerebral Cortex*, *18*, 451–455. DOI: 10.1093/cercor/bhm088

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *The Journal of Criminal Law, Criminology, and Police Science*, *37*, 542–547.

Reid, J. E., & Inbau, F. E. (1977). *Truth and deception: The polygraph ('lie-detector') technique* (2nd edition), Baltimore, MD: Williams & Wilkins.

Sartori, G., Agosta, S., & Gnoato, F. (2007). *High accuracy detection of malingered whiplash syndrome*. Paper presented at the International Whiplash Trauma Congress, Miami, FL.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*, 772–780. DOI: 10.1111/j.1467-9280.2008.02156

Verschuere, B., Prati, V., & De Houwer, J. (2009). Cheating the lie-detector: Faking the autobiographical IAT. *Psychological Science*, *20*, 410–413. DOI: 10.1111/j.1467-9280.2009.02308