Combating Automatic Autobiographical Associations: The effect of instruction and training in strategically concealing information in the Autobiographical Implicit Association Test

Journal:	*Psychological Science*
Manuscript ID:	PSCI-11-2478
Manuscript Type:	Research report
Date Submitted by the Author:	21-Nov-2011

Complete List of Authors:	Hu, Xiaoqing; Northwestern University, Psychology Department Rosenfeld, J Peter; Northwestern University, Psychology Bodenhausen, Galen; Northwestern Univ., Psychology

Keywords:	Associative Processes, Autobiographical Memory

Running head: BEATING the aIAT

Combating Automatic Autobiographical Associations:

The effect of instruction and training in strategically concealing information in the Autobiographical Implicit Association Test

Xiaoqing Hu[CA], J. Peter Rosenfeld & Galen V. Bodenhausen

Department of Psychology, Northwestern University

Corresponding Author:

Xiaoqing Hu

Department of Psychology, Northwestern University, 2029, Sheridan Rd., Evanston, IL 60208

e-mail: xiaoqinghu@u.northwestern.edu

Word count: 2498 (main body)

Abstract

One of the most heavily debated questions in implicit social cognition is the extent to which implicit measures can be voluntarily controlled. This experiment is the first to show that people can intentionally control their performance in the autobiographical implicit association test (aIAT) in a novel way without being detected. Specifically, when explicitly instructed to do so, participants were able to speed up their responses in the incongruent blocks of the aIAT, allowing them to beat the test. This effect obtained whether or not the experimental instruction was followed by practice in speeding responses. A process-dissociation analysis suggested that the effect was likely due to reductions in the ability of the automatic bias to influence responses when instructions to speed up were provided. The present experiment provides new insight into the potential for strategic control in the performance of implicit measures.

Word Count: 141

Key words: implicit measures, automatic associations, memory detection, voluntary control

Utilizing implicit measures, researchers have claimed that people's basic beliefs and attitudes can be assessed based on their split-second responses (Fazio & Olson, 2003, Nosek, Hawkins, & Frazier, 2011). Implicit measures are capable of capturing automatic reactions that people may not want to reveal, may consciously disavow, or may not even be aware of (Gawronski & Bodenhausen, 2011; Greenwald & Banaji, 1995; Nosek et al., 2011). They also have been used to predict future behaviors that self-report measures cannot predict (Nock, Park, Finn, Deliberto, Dour, & Banaji, 2010; Galdi, Arcuri, & Gawronski, 2008). Implicit measures have been used in the measurement of stereotype activation and intergroup bias (Amodio & Devine, 2006), self-esteem (Greenwald, Nosek, & Banaji, 2003), decision making processes (Galdi et al., 2008) and mental health state (Nock et al., 2010). Recently, an implicit measure, called the autobiographical implicit association test (aIAT), has been used to detect autobiographical memory (Sartori, Agosta, Zogmaister, Ferrar, & Catiello, 2008).

The aIAT was developed to detect autobiographical and especially crime-relevant memory (Sartori, et al., 2008). Based on a structure similar to the original IAT (Greenwald et al., 1998), the test contains four categories of sentences: 1) logically true sentences (I am taking a test, 2) logically false sentences (I am climbing a mountain), 3) actual crime sentences (I stole a laptop) and 4) innocent sentences (I read an article). These four types of sentences are combined to form two diagnostic (double-classification) blocks: 1) a true-crime/false-innocent block in which participants press one key for both true and crime sentences, and another key for both false and innocent sentences; and 2) a true-innocent/false-crime block in which participants press one key for both false and crime sentences, and another key for both true and innocent sentences. It is hypothesized that for a guilty examinee, given that the crime sentences should be associated with truth, the RTs from the true-crime (congruent) block should be faster than the RTs from the false-crime (incongruent) block. The reverse should be true for an innocent examinee. In the original report, high diagnostic accuracy (>90%) was obtained across six different experiments (Sartori et al., 2008).

Unlike other methods in memory detection that require polygraphs (Ben-Shakhar & Elaad, 2003), electroencephalographs (Rosenfeld, 2011), or functional MRI (Gamer, 2011), the aIAT simply requires a standard computer. The test can be finished within 10~20 minutes, can be conducted either individually or in groups, and can even be administered remotely via internet. Despite these apparent advantages and satisfactory diagnostic accuracy, one issue requiring greater scrutiny is the extent to which examinees can intentionally control aIAT performance.

Verschuere, Prati and De Houwer (2009) found that when instructed to deliberately slow down their responses in the congruent blocks, participants could distort the aIAT results (see also Fiedler & Bluemke, 2005, who also observed a speed up in the incongruent blocks). However, this previously used faking strategy was detectable, based on the abnormally prolonged RTs in the double-classification blocks compared with the single-block RTs (Agosta, Ghiradi, Zogmaister, Castiello & Sartori, 2011; Cvencek, Greenwald, Brown, Gray, & Snowden, 2010).

A novel alternative strategy for beating the aIAT would consist of only reducing RTs in the incongruent blocks. Currently, it is unknown 1) whether respondents can successfully implement this strategy, and 2) whether pursuit of this strategy can be detected or not. These are important questions not only because claims of an undefeatable memory test must be strictly examined, but also because the ability to reduce response latencies in the incongruent blocks, if it can be achieved, could reveal new insights into people's capacity to control automatic associations more generally. Automatic associations may influence peoples' interpersonal perception, decision making, and behavior (Galdi et al., 2008; Hugenberg & Bodenhausen, 2003), so investigating the capacity to control these associations could have important implications for self-regulation across multiple psychological domains (e.g. Sherman,Gawronski, Gonsalkorale, Hugenberg, Allen, & Groom, 2008).

Here we test 1) whether or not participants can speed up responses in the incongruent blocks so as to distort aIATs; 2) whether or not the automatic/control process that underlies the aIAT can be changed via this strategy. We examined performance across four different instructional groups. We explicitly instructed one group of participants to speed up their responses in the incongruent blocks in the aIAT (instruction group). We had a second group of participants additionally practice speeding up their responses in the incongruent blocks of the aIAT (training group). A practice group without instruction, and a mere repetition group were also run as controls to ensure that any observed effect is specific to intentional control instead of just practice or repetition. If a conscious intention is sufficient to produce fast, yet accurate responses in the incompatible block, then the first two groups should be able to beat the aIAT. However, because IAT performance involves response conflict and executive control (De Houwer, 2003; Klauer, Schimitz, Teige-Mocigemba, & Voss, 2010), it may be that practice is required in order for respondents to be able to speed their responses (Shiffrin & Schneider, 1977). In that case, only the training group may show the capacity to beat the test. Meanwhile, we employed the process dissociation procedure to decompose the performance of the aIAT so as to investigate the extent to which the underlying processes (i.e. automatic vs. control processes) can be changed (Jacoby, 1991; Payne, 2005).

Method

Participants

Sixty-four participants (28 males,19-24 years old) were recruited for monetary compensation for their participation. They were randomly assigned to one of four groups (each with *N*=16): *repetition, practice, instruction,* and *training*.

Procedure

After signing consent forms, participants were randomly assigned to an *exam* or an *article* scenario. In both scenarios, participants were told to take either an exam copy or a research article copy from a mailbox in the main office of the department. After the task, participants were seated in front of a monitor for the aIAT test. The aIAT was similar to that of Sartori et al. (2008): the first block (20 trials) required participants to classify sentences presented on the monitor based on whether they were true (e.g., I am in a lab) or false (e.g., I am in a shop). The second block (20 trials) required participants to classify sentences based on whether they belonged to the *exam* category (e.g., I took an exam copy) or *article* category (I took an article). The third block (60 trials) required participants to press one key for both true and *exam* sentences, and the other key for both false and *article* sentences. The fourth block (40 trials) required participants to reverse the button press from the second block for *exam* and *article* sentences. The fifth block (60 trials) required participants to press one key for true and *article* sentences, and the other key for false and *exam* sentences. Thus, the true-*exam* /false-*article* block is a congruent block for the *exam* participants, but an incongruent block for *article* participants. The order of congruent and incongruent block was counterbalanced between participants in each scenario.

Following the first aIAT, participants completed the task again under one of the four possible instructional conditions. Participants in the *repetition* group simply completed the aIAT for a second time. This group was to control for the possible effect of task repetition. In the *practice* group, after the first aIAT, the participants were instructed to repeat the incongruent blocks three times, with a cover story that the experimenter was interested in the influence of time passage on the participants' performance. Thus, participants in this group were simply repeating the incongruent block without the intention to control their performance. In the *instruction* group, after learning how the test worked, participants were explicitly instructed to speed up in the incongruent block in the second aIAT so as to show a stronger association between true and non-autobiographical sentences. Thus, only instruction but no practice was given to this group. In the *training* group, the participants were provided with instructions to speed up in the incongruent block as in the instruction group, and then they practiced with the incongruent blocks for the same length of time as in the practice group.

Results

Manipulation check

Participants' mean RTs across conditions are presented in Table 1. It is clear that participants all reduced their RTs from the first to second aIATs in both blocks. Thus, participants in the instruction and the training groups did follow the instruction to speed up their responses in the incongruent block. Specifically, the effect size of the speed up in incongruent blocks was largest in the training group ($p<.001$, Cohen's $d=2.79$), followed by instruction ($p<.001$, $d=1.34$) and practice groups ($p<.05$, $d=0.45$). There was no significant effect in the repetition group ($p>.2$). No speed-accuracy tradeoff was observed.

D-score Analysis

We calculated the D600 score as an indicator of the association strength between true and the autobiographical events (for detailed algorithm, see Greenwald et al. (2003) and Sartori, et al.

(2008)). A positive D score suggests a stronger association between autobiographical sentences and truth than between non-autobiographical sentences and truth.

Since neither scenario (exam vs. article) nor block order (congruent first vs. second) had a significant effect on D-scores (*F*s<1, *p*s>.1), these two factors were not considered in the following analysis. A mixed ANOVA using group as a between-subject variable (repetition vs. practice vs. instruction vs. training) and time as a within-subject variable (the first vs. the second aIAT) was conducted on the D-score (Figure 1A). This analysis showed that the D-score changed dramatically from the first to the second aIAT (0.49 versus 0.04; $F(1, 60)= 67.36$, $p<.001$, $\eta^2=.53$). Moreover, a significant time by group interaction was seen, $F(1, 60)= 10.95$, $p<.001$, $\eta^2=.35$, suggesting that time exerted different effects over different groups. Post-hoc tests showed that the D-score was reduced significantly only in the instruction group, $t(15)=4.36$, $p<.001$, and in the training group, $t(15)=7.38$, $p<.001$.

Classifying autobiographical memory

Participants were classified as to either *exam* or *article-scenarios* based on their D-scores. Classification efficiency of the test was measured by receiver operating characteristic (ROC) curves. The area under the curve (AUC) indexes the accuracy with which a given participant's autobiographical memory could be identified correctly (0.5=chance performance; 1.0=perfect performance). Results (Figure 1B) showed that the test successfully discriminated participants from *exam*- and *article*-scenarios in the first aIATs (AUCs > .90, *p*s<.01). However, the AUC was reduced to chance-level in the second aIAT in the instruction group (AUC=.60, $p>.3$) and in the training group (AUC=.57, $p>.4$) but not in the other groups (AUCs>.85, *p*s<.01).

Furthermore, we investigated that whether or not fakers could be differentiated from non-fakers. Confirming that participants were not beating the test in a conventionally detectable manner (e.g. Agosta et al., 2011; Cvencek, et al., 2010), results showed that the differences between double-classification and single-classification blocks could not distinguish fakers from non-fakers (AUCs :0.47-0.56, *p*s>.5, see also additional analyses).

An exploratory process dissociation (PD) analysis

Given there are multiple on-going cognitive processes interacting to influence task performance (Jacoby, 1991; Conrey, Sherman, Gawronski, Hugenberg & Groom, 2005), we conducted a process-dissociation analysis to estimate the automatic versus controlled processes underlying the performance of the aIAT. The control process here serves to detect and execute the correct responses even when there is stimulus-response interference. The automatic process here reflects the automatic associations between autobiographical sentences and true sentences that drive responses when control fails (Sherman et al., 2008).

By analyzing accuracy from congruent and incongruent blocks, a control parameter was calculated as C = P(correct/congruent)-P(incorrect/incongruent) while an automatic parameter is calculated as A = P(incorrect/incongruent)/(1-C). (P is the probability; see Payne, 2005; Stewart, von Hippel, & Radvansky, 2009).

We conducted separate group (repetition vs. practice vs. instruction vs. training) × time (first vs. second aIAT) mixed ANOVAs on the control and automatic estimates (Figure 2). No effect was observed for the control estimates (Fs<1, ps>.4), probably due to the ceiling effect across all conditions (control estimates: 0.86-0.93, Figure 2A). However, the automatic estimates (Figure 2B) were reduced from the first (0.663) to the second aIAT (0.558, $F(1,60)=12.261$, $p<.01$, $\eta^2=.17$), yet the interaction was not significant ($p>.5$). Given that the D-score changes were significant only in the instruction and the training groups; we further conducted a regression analysis using the controlled and automatic estimates to predict their D-score change.

Results showed that the change of D-score could be predicted with the reduction of the automatic estimates (β=0.536, p<.01), but not with the change of the control estimates (β=-0.115, p>.3).

Discussion

The present study found that although the aIAT can perfectly detect autobiographical memory in naïve participants, participants can successfully change their aIAT outcome in both the instruction and the training groups. This was achieved by only speeding up RTs in the incongruent blocks. Underlying this behavior change, the process dissociation (PD) analysis suggested that the faking participants may effectively resist the response tendencies to associate autobiographical statements and 'true'.

Previous aIAT studies supported it as a promising tool for forensic investigation (Agosta, et al., 2011; Agosta et al., 2011; Sartori et al., 2008). The results here, however, showed that people can control their aIAT performance by speeding up the RTs in the incongruent block. Importantly, this speed up is achieved without concomitant response slowing in the congruent blocks (cf. Fiedler & Bluemke, 2005). In particular, instruction itself is sufficient to change the aIAT's outcome, producing reductions in the ability of automatic autobiographical associations to bias responses. Given that people could easily obtain the relevant information about IATs from the internet, further studies are warranted to develop a test that resists such motivated faking.

Despite the fact that the participants could control their performance with mere instruction, one should be cautious about generalizing the present findings to IATs in other domains. It is possible that since the mental associations measured in this study were recently acquired autobiographical events, they were relatively easy for motivated participants to strategically influence (De Houwer, Beckers & Moors, 2007). However, it remains an open question about whether or not one could similarly control associations established via long-term socialization, such as intergroup biases (Baron & Banaji, 2006).

The training condition provided us with an opportunity to examine whether or not automatic autobiographical associations can be controlled via intentional practice. Indeed, participants from this group exhibited a larger behavioral change compared with other groups. This is also consistent with previous studies that showed that training could effectively decrease automatically activated racial stereotypes (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000; Plant & Peruche, 2005). Our exploratory PD analysis provided evidence that practice could effectively limit the ability of automatic associations to influence responses, but only when participants are intentionally trying to control their performance (Sherman et al., 2008). Since automatic associations exert considerable influence over people's judgment and behavior (Galdi, et al., 2008), the training provided us with a novel way to influence automatic associations, and thus may shed light on attitude formation/change and intergroup bias regulation (Gawronski & Bodenhausen, 2011).

Acknowledgement: We thank Niels Birbaumer, Steve Franconeri, Bertram Gawronski and Jeff Sherman for comments on earlier versions of this manuscript.

References

Agosta, S., Ghiradi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology, 25*, 299-306.

Agosta, S., Mega, A., & Sartori, G. (2011). Detrimental effects of using negative sentences in the autobiographical IAT. *Acta Psychologica*, *136,* 269-275.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluations in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652-661.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6, 10 & adulthood. *Psychological Science, 17,* 53-58.

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

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, *89*, 469-487.

Cvencek, D., Greenwald, A.G., Brown, A.S., Gray, N.S., & Snowden, R.J. (2010). Faking the implicit association test is statically detectable and partly correctable. *Basic and Applied Social Psychology, 32,* 302-314.

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.

De Houwer, J. (2003). A structural analysis of indirect measures of attitudes. In J. Musch & K.C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219-244). Mahwah, NJ: Lawrence Erlbaum.

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

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.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science*, *321,* 1100-1102.

Gamer, M. (2011). Detecting of deception and concealed information using neuroimaging techniques. In B. Versuchere, G. Ben-Shakhar, & E. Meijer, (Eds.), *Memory Detection: Theory and application of the Concealed Information Test (*pp.90-113*).* Cambridge University Press.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127.

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.

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

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*, 640-643.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513-541.

Kawakami, K., Dovidio, J.F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): Effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology, 78*, 871-888.

Klauer, K.C., Schimitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology, 63,* 595-619.

Nosek, B. A., Hawkins, C. B., & Frazier, R.S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Science, 15,* 152-159.

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive function modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology, 89,* 488-503.

Plant, E.A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science, 16,* 180-183.

Rosenfeld, J. P. (2011). P300 in detecting concealed information. In B. Versuchere, G. Ben-Shakhar, & E. Meijer, (Eds.), *Memory Detection: Theory and application of the Concealed Information Test (*pp.63-89*).* Cambridge University Press.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S.D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, 18, 772-780.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*, 314-335.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing II: Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127-190.

Stewart, B. D., von Hippel, W., & Radvansky, G. A. (2009). Age, race, and implicit prejudice: Using process dissociation to separate the underlying components. *Psychological Science*, *20*, 164-168.

Verschuere, B., Prati, V., & De Houwer, J. (2009). Cheating the lie detector: Faking in the Autobiographical IAT. *Psychological Science, 20*, 410-413.

Table 1: Participants' Mean RTs (in milliseconds) and standard errors (in parentheses) as a function of time, groups and response blocks. RT-difference is calculated as the RT from the second aIAT subtracted from the first aIAT, with the corresponding Cohen's *d* values for effect sizes. Error rates were given in percentages. The C stands for congruent blocks, i.e. true-*autobiographical event*/false-*non-autobiographical event* block; the IC stands for incongruent blocks, i.e. true-*non-autobiographical event*/false-*autobiographical event* block.

	Groups							
	Repetition		Practice		Instruction		Training	
	C	IC	C	IC	C	IC	C	IC
1st aIAT	828.63 (31.15)	949.75 (33.55)	808.44 (28.63)	922.25 (42.36)	799.25 (25.22)	902.50 (30.19)	763.69 (15.59)	857.94 (21.63)
2nd aIAT	813.50 (41.35)	905.19 (48.60)	788.50 (34.01)	841.75 (47.33)	759.69 (21.27)	734.31 (32.55)	711.94 (22.51)	642.88 (16.55)
RT-Difference	12.63	44.56	19.94	80.50*	39.56	168.19***	51.75*	215.06***
Cohen's *d*	0.09	0.27	0.16	0.45	0.42	1.34	0.67	2.79
1st aIAT Errors (%)	2.91	6.20	2.18	5.21	4.37	8.13	3.49	7.19
2nd aIAT Errors(%)	2.55	5.00	3.59	4.74	6.98	7.55	5.10	5.31

Note: *, p<.05;

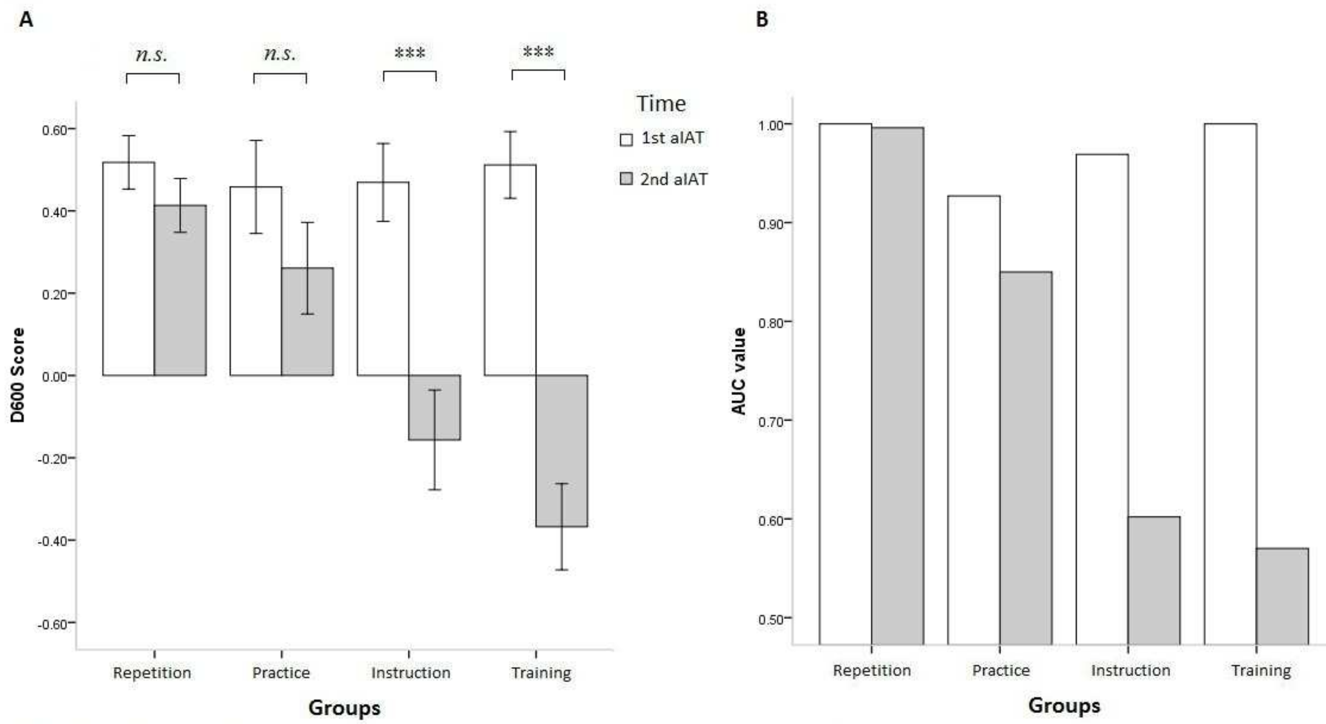
***, p<.001

Figure Captions:

Figure1: The D-600 score (left panel, positive score means a stronger associations between crime sentences and truth sentences), and AUC (area under curve, right panel. The AUC indicates the classification efficiency of the test. It varies from 1.00, which means perfect classification, to 0.50, which means chance-level classification) of the first and second aIAT across the four groups. The error bar stands for one standard error.

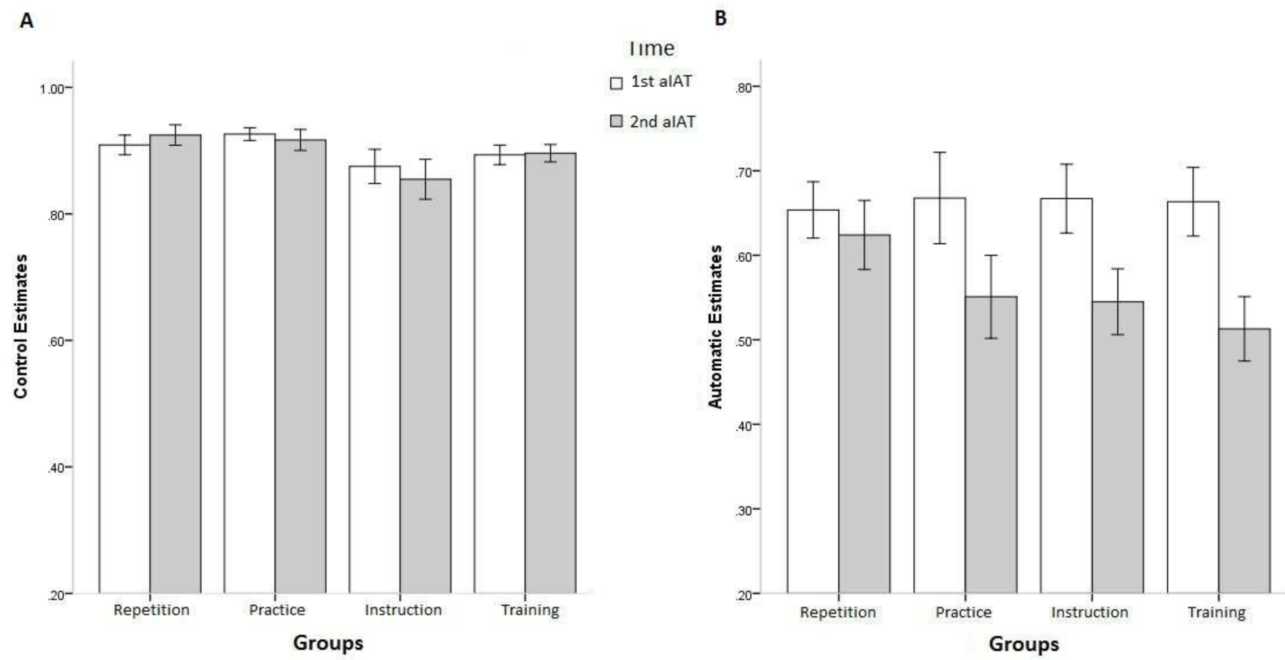
Figure 2: The control (left panel) and automatic (right panel) estimates from the process dissociation analysis in the four groups. The error bar stands for one standard error.

Figure 1:



n.s. not significant *** p<.001

Figure 2:


A
Time
1st aIAT
2nd aIAT
Control Estimates
1.00
.80
.60
.40
.20
Repetition
Practice
Instruction
Training
Groups
B
Automatic Estimates
.80
.70
.60
.50
.40
.30
.20
Repetition
Practice
Instruction
Training
Groups