Dhammapeera, Phot, Hu, Xiaoqing and Bergström, Zara M (2020) ***Imagining a false alibi impairs concealed memory detection with the autobiographical Implicit Association Test.*** **Journal of Experimental Psychology: Applied, 26 (2). pp. 266-282. ISSN 1076-898X.**

Accepted manuscript, to be published in *Journal of Experimental Psychology: Applied*

Imagining a false alibi impairs concealed memory detection with the autobiographical Implicit Association Test

Phot Dhammapeera[1,2], Xiaoqing Hu[3,4], Zara M. Bergström[1]

[1]School of Psychology, University of Kent, UK

[2]Faculty of Psychology, Chulalongkorn University, Thailand

[3]Department of Psychology, The State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China

[4]HKU-Shenzhen Institute of Research and Innovation, Shenzhen, China

Correspondence should be addressed to Zara Bergström, School of Psychology, Keynes College, University of Kent, Canterbury, CT2 7NP, UK. E-mail: z.m.bergstrom@kent.ac.uk

Acknowledgements: We thank Jessica Amos, Ellie Anslow, Ashley Bailey, Chloe Brunskill, Rhiannon Chappell, Amber Gardner, Lucy Hendleman, Catalina Marin, Chloe Walker, and Eleanor Webster for help with data collection. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Declaration of interests: none.

Abstract

Imagining counterfactual versions of past events can distort memory. In three experiments, we examined whether imagining a false alibi for a mock crime would make suspects appear less guilty in a concealed memory detection test, the autobiographical Implicit Association Test (aIAT), which aims to determine which of two autobiographical events are true. First, “guilty” participants completed a mock crime, whereas “innocent” participants completed an innocent act. Next, some of the guilty participants were asked to imagine a false alibi that corresponded to the innocent act. Finally, all groups completed the aIAT. Across experiments, we varied the type of aIAT used and also compared the effectiveness of the false alibi countermeasure when only imagined once, versus when it was repeatedly imagined over a week long period. The aIAT accurately detected the mock crime as true for guilty participants without a false alibi, but was consistently less able to detect the mock crime as true for guilty participants who had imagined a false alibi. The findings suggest that if guilty suspects fabricate an alibi, this may create a memory for the alibi that appears to be true based on the aIAT, which is problematic for its real-life applications in concealed memory detection.

Keywords: Memory; Imagination; autobiographical Implicit Association Test; Truth

Public significance statement: We found that rehearsing a false alibi can impair truth detection with a computerized test, the autobiographical Implicit Association Test. This finding is important because it suggests the test is vulnerable to faking, and that real life applications of this test are premature.

Introduction

Forensic memory detection aims to determine if a criminal suspect has concealed information stored in their memory that is indicative of guilt. Guilty suspects are expected to have unique knowledge of the crime that would not be known by innocent suspects. Therefore, non-verbal markers of memory, such as memory-related brain activity (e.g. Allen, Iacono, & Danielson, 1992; Gamer, Klimecki, Bauermann, Stoeter, & Vossel, 2012; Rosenfeld, Angell, Johnson, & Qian, 1991; Van Hooff, Brunia, & Allen, 1996), autonomic activity (Gamer, 2011; Lykken, 1959) or reaction times and accuracy on indirect memory tests (Sartori et al., 2008; Verschuere & De Houwer, 2011), can be measured to detect if a suspect is concealing incriminating knowledge. Many of these methods can very accurately detect concealed information, at least in cooperative research participants with little motivation to hide their guilt (Granhag, Vrij & Vershuere, 2015; Verschuere, Ben-Shakar & Meijer, 2011). However, one prominent concern is that real criminals may use countermeasure strategies to attempt to hide their guilt (e.g. Bergström, Anderson, Buda, Simons, & Richardson-Klavehn, 2013; Hu, Bergström, Bodenhausen, & Rosenfeld, 2015; Verschuere, Prati & De Houwer, 2009; for a review, see Ben-Shakhar, 2011), threatening the validity of these tests in real-life settings. Considering the important societal, legal and ethical implications of forensic memory detection, it is therefore critical to evaluate whether memory detection tests are susceptible to countermeasures. It is also important to assess which types of countermeasures are likely to be successful in order to ensure that memory detection tests are optimally designed to withstand evasion attempts.

The autobiographical Implicit Association Test (aIAT, Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008), is a computerised task that bears high promise in assessing the implicit truth value of autobiographical statements, which can therefore be used to detect concealed autobiographical memories. The aIAT measures reaction times and accuracy in a simple sentence classification task as markers of whether an autobiographical event is true or false for an individual, and is thus considerably easier and cheaper to implement than physiology and brain activity-based techniques that necessitate specialist equipment and highly trained administrators. In a criminal context (e.g. Sartori et al., 2008), the aIAT involves presenting suspects four different types of statements that suspects have to classify on two dimensions: logically true versus false, or crime-related versus innocent-related, by pressing two different buttons. Sentences for the first dimension are true or false for everyone taking the test (e.g. true: “I am in front of a computer” vs. false: “I am in a restaurant”), whereas the truth of sentences for the second dimension depend on whether the suspect has committed the crime or not (e.g. true if guilty/false if innocent: “I stole a ring” (a crime-related sentence) vs. false if guilty/true if innocent: “I bought a ring”(an innocence-related sentence). In guilt congruent blocks, logically true and crime-related statements share one button, whilst logically false and innocent-related statements share another button. In guilt incongruent blocks, logically false and crime-related statements share one button, whilst logically true and innocent-related statements share another button. Guilty suspects are expected to respond faster and more accurately in guilt congruent than incongruent blocks due to crime-related sentences having implicit and automatic associations with the truth. Innocent suspects are expected to show the opposite pattern.

Many studies have shown very accurate memory detection using the aIAT (reviewed in Agosta & Sartori, 2013; however see Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017; for evidence that the aIAT may be less effective than other RT-based memory detection paradigms). Moreover, the aIAT is not only able to detect which of two autobiographical events is more strongly associated with truth, but is also better at detecting true memories than false memories that the participant believes are true (Marini, Agosta, Mazzoni, Dalla Barba, & Sartori, 2012). Because of such promising results, the aIAT has already been applied in at least one real court case in Italy, where it was used by the defense team as part of a battery of tests to suggest that the defendant had memory impairments, which was accepted by the judge as evidence of diminished culpability and contributed to a reduced penalty for a convicted murderer (Sirgiovanni, Corbellini, & Caporale, 2016). In contrast, other research has shown that the aIAT may be susceptible to relatively simple countermeasures that guilty suspects can apply during the test, such as slowing down responses in the guilt congruent blocks (Verschuere, Prati, & Houwer, 2009) or speeding up responses in the guilt incongruent blocks (Hu, Rosenfeld & Bodenhausen, 2012), especially when participants are allowed to practice in advance of the test. However, suspects who used such strategies may be caught out by selectively modifying their response times only during critical blocks but not during other, non-critical blocks (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011). Thus, trying to beat the aIAT by directly altering response times may not be a particularly effective countermeasure, since such faking attempts may be detectable by unusual patterns of response times across different blocks (although see Hu et al., 2012).

An alternative strategy that guilty suspects could use for evading forensic memory detection is to intentionally modify their memories in advance of the test, in order to make these memories more consistent with innocence. A large body of evidence shows that memories for experienced events remain malleable after encoding and can be updated or inhibited at a later stage (e.g. Anderson & Hanslmayr, 2014; Dudai, 2012). Indeed, in several experiments we have found that by intentionally suppressing memories of committing a mock crime, guilty suspects were able to significantly reduce retrieval-related ERPs thus increasing the likelihood of appearing innocent on an EEG-based memory detection test (Bergström, et al., 2013; Hu et al., 2015). Furthermore, suppression of mock crime memories weakened the associative strength between the crime and the truth so that guilty suspects also appeared more innocent on a later aIAT, even without engaging any intentional strategies during the aIAT itself (Hu et al., 2015). Thus, modifying memories in advance of a memory detection test may be an effective countermeasure strategy that is less detectable than on-line faking attempts during the test itself.

Whereas previous research showed that suspects can intentionally weaken incriminating memories to evade detection, another strategy by which guilty suspects could appear innocent is to intentionally store false information in memory that suggests innocence. It is well established that people can hold vivid memories for events that they have never experienced in real life (Loftus & Pickrell, 1995; Schacter, Guerin & St Jacques, 2011). Such memories can be created simply by imagining a novel event (Loftus, 2003) that becomes encoded as a memory representation with similar perceptual and conceptual features as a memory based on an experienced event, making true and false memories similar in terms of their neural and behavioural characteristics (Mitchell & Johnson, 2009). Consistent with this view, imagining performing simple actions (such as picking a specific card from a deck of playing cards) enhances implicit associations between the imagined event and the truth when contrasted with non-imagined events in an aIAT. Some research found this to be the case particularly when participants misremembered imagined actions as previously performed (Takarangi, Strange, Shortland, & James, 2013), whereas in other studies, aIAT truth detection of imagined actions was enhanced even when participants knew the imagined event did not occur in real life (Shidlovski, Shul, & Mayo, 2014; see also Mangiulli, Lanciano, Curci, et al., 2018; Takarangi, Strange & Houghton, 2015; Vargo, Petróczi, Shah, & Naughton, 2014). Furthermore, in a mock criminal context, asking people to deliberately memorise a hypothetical alternative version of a mock crime can weaken skin conductance responses associated with a true mock crime, and thereby impair memory detection with autonomic measures (Gronau, Elber, Satran, Breska & Ben-Shakar, 2015).

However, to our knowledge, no previous research has investigated whether guilty suspects can intentionally memorise false information indicative of innocence as a countermeasure strategy for evading guilt detection with the aIAT. In real life, guilty suspects may fabricate an untrue version of what they were doing at the time of the crime to use as a false alibi, and by doing so, they may encode this information into memory in a form that may share some characteristics with a true memory, which may potentially also distort or impair their memory for the true crime event (Otgaar & Baker, 2018). Recent research has shown that adopting a false alibi can impair identification of guilty suspects in deception detection paradigms (Foerster, Wirth, Herbort, Kunde, & Pfister, 2017; Suchotzki, Berlijn, Donath, & Gamer, 2018), but this issue has not been investigated with the aIAT. We addressed these issues in three experiments that used the aIAT to investigate whether imagining a false alibi impaired guilt detection by enhancing the implicit truth value of an alibi and/or decreasing the implicit truth value of a committed mock crime. We also investigated whether the alibi countermeasure was more effective when applied repeatedly over an extended time period compared to just in one brief session. To preempt the results, we found a consistent pattern across studies whereby the false alibi significantly impaired guilt detection with the aIAT, which seemed to be primarily driven by the alibi being detected as true rather than a substantial impairment of the original mock crime memory.

Experiment 1

The first experiment was conducted in three stages. First, “guilty” participants carried out a mock crime which involved stealing a ring from a bag in a University staff office area, whereas “innocent” participants carried out an innocent act that involved going to the same office area but instead writing their email address on a paper slip on a staff member's door. Next, half of the guilty participants were instructed to imagine performing the innocent act with the explicit intention of using this as a false alibi in order to appear innocent. The other half of guilty participants and the innocent group performed an unrelated filler task. Finally, all three groups undertook an aIAT where the relative truth value of the mock crime and innocent/false alibi events were compared in all three groups.

We hypothesised that imagining a false alibi would create a memory for the imagined act, which may have some implicit associations with the truth even though participants knew their alibi was fake at an explicit level (Shidlovski, et al., 2014). Imagining a fake alibi would thus lead to lower aIAT discrimination between the objectively true mock crime and the objectively false innocent act when this group was compared to the guilty group who did not imagine the alibi. If imagining an alibi as a countermeasure was completely successful at making guilty suspects appear innocent, aIAT performance for these guilty participants would be indistinguishable from the innocent group who actually conducted the innocent act in real life.

Methods

Participants

The design was based on our previous experiment which included 78 participants divided across three groups and found a large effect size (Cohen's *d*=0.78) for reduced aIAT memory detection in a suppression countermeasure group compared to a standard guilty group (Hu et al., 2015). That prior experiment was designed to have 0.8 power to detect a *d*=0.8 effect size, and we increased our sample size in the current study to further enhance statistical power, and therefore recruited 108 participants who were split into three groups, resulting in >0.9 power to detect a *d*=0.8 effect size, or 0.8 power to detect a *d*=0.7 effect size (we decided a priori that we were primarily interested in detecting large effects of the alibi countermeasure on the aIAT, as only large countermeasure effects have substantial implications for practical applications involving guilt classification at the individual level). The participants were undergraduate students at the University of Kent who took part via a research participation scheme in return for course credits. Participants were randomly assigned to three experimental groups (N = 36 in each); the Guilty-Alibi group (30 female and 6 male), the Guilty-Standard group (29 female and 7 male), and the Innocent group (28 female and 8 male). Twenty additional participants were replaced due to technical problems or not following the instructions during the mock crime/innocent act (such as stealing the wrong object, or going to the wrong part of the building). Participants' age ranged from 18-28 (M = 19.83, SD = 1.62). The groups did not significantly differ in terms of age ($F(2,104) = .80$, $p = .451$, $\eta_p^2 = .02$) nor gender ($\chi^2(2) = .36$, $p = .837$, $\varphi = .84$). All participants had English as their first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethics committee.

Materials, design and procedure

First, participants in the two Guilty groups were required to go to a kitchen adjacent to staff offices in a university building, find a bag, and steal a box from inside the bag. They were explicitly asked to look and take note of what was inside the box (a ring), and then return with the box and its content to the experimental room. The word ring was not mentioned in the instructions so that the memory of the ring was gained solely from enacting the crime. Innocent participants were required to go to the same area in the building, but instead they were told to write their email address on an appointment sign-up sheet on the door of a lecturer's office. Thus, Innocent participants were unaware of the mock crime.

Next, participants in the Guilty-Alibi group were provided with a fake alibi scenario, which was designed to help them appear innocent on the aIAT. Participants were told that they would soon take part in a test designed to detect their guilt, however they should aim to appear innocent by adopting the alibi. Participants were instructed that it was essential that they try to imagine the scenario as if it were true and that their memory for scenario details would later be tested. The alibi scenario was a short verbal description of the innocent act: “You were on your way to find your lecturer. On their door, there was a sheet of paper specifying that you could leave your email address for the lecturer to get back to you. So you tore off a bit of paper and wrote your email address and left it in the envelope provided and came back here. The envelope has since been destroyed so there is no evidence that your alibi is false”. Participants were told to close their eyes and vividly imagine the alibi for two minutes. Next, they were asked to describe the scenario in detail and answer a few questions about it. If they gave incorrect answers, the alibi story was repeated and the questions asked again until the correct answers were given. Participants in the Guilty-Standard and Innocent groups were instead required to carry out a filler task of solving Sudoku puzzles. They were given two puzzles as well as written instructions and told to do the best they could while they were timed for 5 minutes.

In the final stage, all participants took part in a seven-block computerised aIAT (Hu et al., 2015; Sartori et al., 2008). Participants were instructed that multiple sentences would appear on the screen and they would need to classify them as either logically true or false, or ring-related or email-related by pressing left or right buttons on the keyboard. To avoid on-line attempts to modify the test result, they were not informed regarding how the test worked or how to alter their responses to appear innocent (*cf*. Agosta et al., 2011; Hu et al., 2012; Verschuere et al., 2009). The first block (20 trials) was a simple classification block that required participants to classify 5 true and 5 false sentences, with each sentence repeated twice in random order. Participants were instructed to press the left key ‘Z’ for logically true sentences (e.g., “I am a research participant”) and the right key ‘M’ for logically false sentences (e.g., “I am playing football”), based on what they were doing at that time. The labels “True” and “False” were displayed on the left and right sides of the screen respectively, to remind participants of the response-key mapping. The second block (20 trials) was a simple classification block that required participants to classify 5 sentences related to the guilty act (e.g., “I took a ring”) and 5 sentences related to the innocent act/alibi scenario (e.g., “I wrote my email”). Participants were asked to press the left key ‘Z’ for ring-related sentences and the right key ‘M’ for email-related sentences, and the labels “Ring” and “Email” were displayed on the left and right sides of the screen respectively. Blocks three (20 trials) and four (40 trials) were critical double classification blocks which tested participants’ responses to guilt congruent sentence pairings, because logically true and autobiographically true sentences for the Guilty groups were paired to the same response button. Participants were instructed to press ‘Z’ if the sentence was logically true or ring-related and ‘M’ if the sentence was logically false or email-related, and the labels “True/Ring” and “False/Email” were displayed on the left and right sides of the screen respectively. Block five (20 trials) was a practice reverse simple classification block, which reversed the key assignments for ring and email-related sentences (‘Z’ for email-related and ‘M’ for ring-related sentences, with the left label changed to “Email” and the right label changed to “Ring”). The final blocks six (20 trials) and seven (40 trials) were also critical double classification blocks with the reversed keys, thus testing participants’ responses to guilt incongruent sentence pairings, because logically false and autobiographically true sentences for the Guilty groups were paired to the same response button. Participants were instructed to press ‘Z’ if the sentence was logically true or email-related and ‘M’ if the sentence was logically false or ring-related, and the labels

“True/Email” and “False/Ring” were displayed on the left and right sides of the screen respectively. Faster RT and higher accuracy for guilt congruent blocks than guilt incongruent blocks indicate an association between the crime and the truth, whereas the reverse pattern indicate an association between the innocent act and the truth.

Half of the participants within each group conducted the blocks in the order described above, while blocks 2-4 and 5-7 were swapped for the other half of participants in order to counterbalance the order of guilt congruent and guilt incongruent blocks. Thus, counterbalancing formats were balanced within groups and matched across groups. For all blocks, sentences were presented on the screen in random order, and stayed on the screen until participants pressed a button. Participants were instructed to respond as quickly and accurately as possible, and if they pressed the incorrect button a red ‘X’ appeared on the screen until the pressed the correct button.

Data analysis

The main measure of guilt in the aIAT is the D-score, which combines accuracy and RT into a single, standardized measure (Greenwald et al., 2003; Sartori et al., 2008). We used the same formula to calculate D as in the most relevant previous studies (Hu et al., 2012; 2015). First, extreme RTs (<100ms or >10,000ms) were deleted. As in prior research, incorrect responses were given a 600ms penalty, and the mean RTs were calculated for the guilt congruent and guilt incongruent blocks separately, including the incorrect responses with the applied penalties. Finally, the mean RT difference between guilt congruent and guilt incongruent blocks was divided by the standard deviation of the RT distribution for correct trials only, from both blocks combined, in order to obtain the D-score. In the Experiment 1 version of the aIAT, a positive D-score indicated guilt because it suggests that participants associated sentences describing the mock crime with the truth, whereas a negative D-score indicated innocence because it suggests that participants associated sentences describing the innocent act with the truth.

Potential group differences in D-scores were analysed with commonly used frequentist inferential tests from the GLM (ANOVA, t-tests). Effect sizes were estimated using partial eta-squared for ANOVAs, and Cohen's *d* for t-tests. Cohen's *d* for both paired and independent t-tests was calculated as the difference between means divided by the pooled standard deviation rather than from the t-values to avoid inflating effect size estimates for paired t-tests (Dunlap, Cortina, Vaslow, & Burke, 1996). As the key hypotheses relied on testing whether D-scores were above or below zero within each group and whether there were pairwise group differences in D-scores, frequentist t-tests for such differences were supplemented with Bayes factors (BF_{10}) to evaluate the relative support for a difference (H_1) versus no difference (H_0). These were calculated with Bayesian t-tests in JASP (JASP Team, 2017) using default priors (a Cauchy distribution with centre = 0, r = 0.707). The Bayes Factors is a ratio that contrasts the likelihood that the data would occur under the alternative (H_1) versus null (H_0) hypotheses, with values over 1 indicating support for H_1 and values below 1 indicating support for the H_0. Values close to 1 are only considered weakly/anecdotally supportive of one hypothesis over the other, whereas BF_{10} >3 are typically interpreted as substantial evidence in support of H^1 over H^0, and BF_{10} <0.33 are interpreted as substantial evidence in support of H^0 over H^1 (see Wagenmakers et al., 2011).

The aIAT was developed to diagnose guilt or innocence at the individual level, which is typically done by classifying individuals with positive D-scores as "guilty" and individuals with negative D-scores as "innocent" when contrasting a guilty vs. innocent event in this way (Sartori et al., 2008). However, because such classification rates are dependent on choosing specific cut-offs and the optimal cut-off may vary across samples and experimental designs, we instead conducted a threshold-independent receiver operating characteristic (ROC) analysis to evaluate classification performance using Areas Under the Curve (AUCs; following e.g. Hu et al., 2015, but see Supplementary Information for threshold-based classification). The AUCs reflect the accuracy with which a randomly chosen participant can be classified into the correct group (Guilty or Innocent), where .5 reflects chance classification and 1.0 reflects perfect classification.

In addition to analysing the D-score, we also analysed the raw RT and accuracy rates separately for the guilt-congruent versus incongruent blocks for each group. However, since these analyses only revealed patterns that were consistent with the main D-score findings, they are presented in the supplementary file. Furthermore, in a final analysis, we also calculated a "faking index" (Agosta et al., 2011) that has been proposed as a method for detecting whether participants are showing unusual reaction time patterns that indicates countermeasure use. Therefore, we used the faking index to assess whether rehearsing a false alibi resulted in unusual reaction time patterns across aIAT blocks that could function as signals of guilt even when the main guilt measure (i.e. D-score) is disrupted by countermeasures. However, this analysis revealed that the faking index did not discriminate well between the groups, so these results are also presented in the supplementary file. Individual level data for this project is available at: https://osf.io/wumdy/

Results

Mean D-scores were in the expected direction, with the highest scores in the Guilty Standard group and the lowest scores in the Innocent group, and were significantly different between the three groups ($F(2, 105) = 9.46$, $p < .001$, $\eta_p^2 = 0.15$; Fig. 1). The innocent participants, who undertook the innocent act but did not have any knowledge of the mock crime, elicited D-scores below zero ($t(35) = -2.48$, $p = .018$, $d = 0.41$; BF_{10}

=2.55). Guilty-Standard participants, who committed the mock crime but did not have any knowledge of the innocent act, elicited D-scores above zero ($t(35) = 3.25$, $p = .003$, $d = 0.54$; $BF_{10} = 13.70$). The Guilty-Alibi participants, who committed the mock crime and were also provided with an alibi scenario consistent with the innocent act, elicited D-scores non-distinguishable from zero ($t(35) = 0.17$, $p = .87$, $d = 0.03$; $BF_{10} = 0.18$). D-scores were higher in the Guilty-Standard group than the Innocent group, strongly supported by both frequentist and Bayesian statistics ($t(70) = 4.06$, $p < .001$, $d = 0.96$; $BF_{10} = 179.99$). However, there was only a non-significant trend for higher D-scores in the Guilty-Alibi compared to the Innocent group, and the Bayes Factor was very close to 1 and thus inconclusive ($t(70) = 1.80$, $p = .076$, $d = 0.43$; $BF_{10} = 0.97$). Importantly, D-scores were significantly reduced in the Guilty-Alibi group compared to the Guilty-Standard group, and the Bayes Factor indicated substantial evidence in favour of a difference (H_1) compared to no difference (H_0) between groups ($t(70) = 2.66$, $p = .010$, $d = 0.62$; $BF_{10} = 4.55$). These results indicate that, as expected, imagining a fake alibi consistent with innocence impaired memory detection with the aIAT.

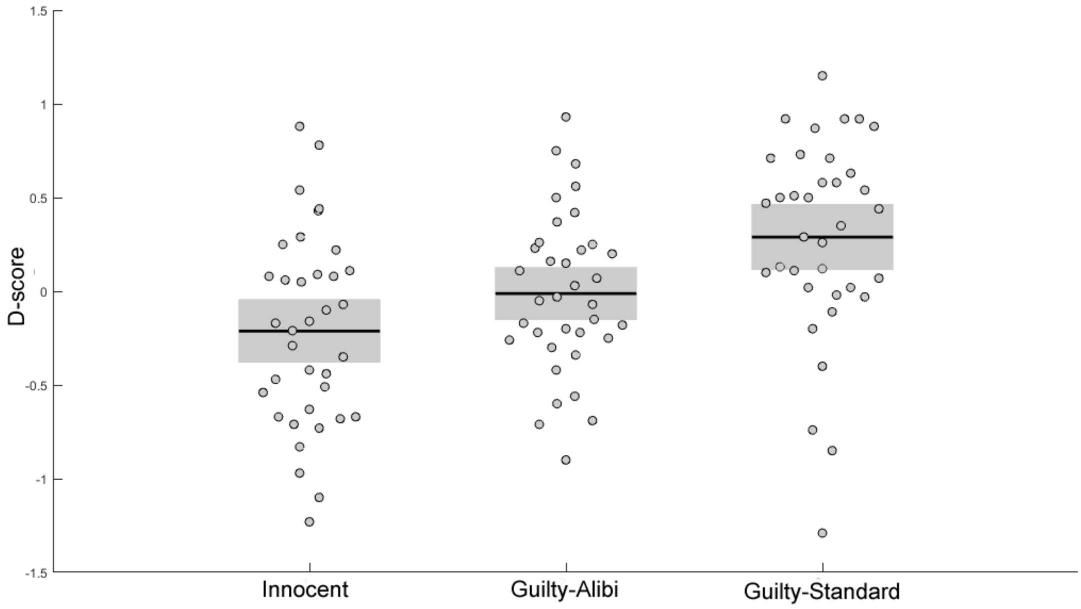

Figure 1. D-scores for the three groups from the Mock Crime/Innocent event aIAT in Experiment 1. Each dot indicates an individual score. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the mock crime-related sentences are associated with the truth) and D-scores below zero suggest innocence (that the innocent-related sentences are associated with the truth). Scores are jittered along the x-direction for display purposes.

Because applied uses of the aIAT involves classifying individual suspects as guilty or innocent, we also conducted a ROC analysis to evaluate how accurately our participants could be classified based on their D-scores. This analysis showed that when comparing Guilty-Standard and Innocent groups, D-score classification was significantly better than chance (*AUC* = .70, *SE* = .06, *p* = .004), but comparing Guilty-Alibi and Innocent groups, D-score classification was less accurate and not significantly different than chance (*AUC* = .62, *SE* = .07, *p* = .093). Thus, individual classification rates also supported our prediction that imagining a false alibi would impair memory detection.

Experiment 1 Discussion

In Experiment 1, the aIAT showed relatively good discrimination between guilt and innocence in participants who did not employ countermeasures, consistent with previous findings (e.g. Sartori, et al., 2008; Agosta & Sartori, 2013). However, the false alibi countermeasure reduced memory detection when compared to a standard guilty group who were not trying to evade the test, consistent with our predictions. Performance in the Innocent group showed a stronger relative association between the innocent act and the truth than the mock crime and the truth, whereas performance in the Guilty-Standard group indicated the opposite relative association. Performance in the Guilty-Alibi group however was equivocal as to which scenario was truthful. This pattern indicates that imagining a fake alibi scenario likely created a memory for the imagined alibi act that had some implicit associations with the truth, even though participants knew their alibi was false at an explicit level (*cf.* Shidlovski, et al., 2014; Takarangi, Strange & Houghton, 2015; Takarangi, Strange, Shortland, & James, 2013). This account is consistent with more general findings that imagining an event can create a memory for that event that has similar perceptual and behavioural characteristics as memories based on true experiences (e.g. Loftus, 2003; Loftus & Pickrell, 1995; Mitchell & Johnson, 2009; Schacter, Guerin & St Jacques, 2011). Presumably, because both the mock crime and the imagined alibi act had some associations with the truth, neither of the critical aIAT blocks were truly congruent or incongruent with their memories, leading to similar performance in both blocks.

The results are consistent with the explanation that imagining a false alibi increased the implicit truth value of that scenario, which thereby disrupted aIAT discrimination between the alibi and the mock crime. However, imagining a counterfactual version of an event may also interfere with the veridical memory of the event and decrease its implicit truth value (*cf.* Otgaar & Baker, 2018). Gronau et al. (2015) asked participants to learn a hypothetical crime scenario with various details that were different from a mock crime they had actually conduced. Results showed that learning a false version of the mock crime impaired explicit recall of true crime details, and furthermore, reduced skin-conductance markers of true crime memories. They argued that true crime memories may have become inhibited as a result of retrieval competition between true and false crime details, similarly to the retrieval-induced forgetting phenomenon (Anderson, Bjork, & Bjork, 1994, 2000; Anderson & Levy, 2007), or alternatively, that the memory for alibi information interfered with and blocked access to the memory for the true mock crime (see Anderson & Neely, 1996, for review). Because the aIAT in Experiment 1 measured the relative truth of the false alibi versus mock crime scenarios, we can conclude that these scenarios had similar implicit truth values in the alibi countermeasure group. However, we cannot determine whether the lack of a difference was due to increased implicit truth value of the false alibi, or reduced implicit truth value of the mock crime, or a combination of both. This issue was addressed in the next experiment.

Experiment 2

Experiment 2 used exactly the same false alibi manipulation, materials and procedure as in Experiment 1, with the only change being that the final test involved a different aIAT design that contrasted the mock crime with an unexperienced event that was clearly different from the learned false alibi. Thus, this study investigated whether imagining a false alibi would still impair detection of the mock crime regardless of which other scenario it is compared to. If such a pattern was found, it would indicate that the implicit truth value of the original crime-related memory was weakened by rehearsing an alibi, since any reduction in mock crime detection in this aIAT could not be due to inflated implicit truth value of the imagined alibi event as this scenario was not used as a contrast in the test.

We hypothesised that if the alibi manipulation was successful at reducing the implicit truth value of the true mock crime memory, perhaps by reducing access to this memory through inhibition or an interference “blocking” mechanism (Anderson et al., 1994; Anderson & Levy, 2007; Gronau et al., 2015), then rehearsing an alibi should reduce detection of guilty suspects on the aIAT by lowering their D-scores when compared to guilty suspects who did not rehearse an alibi after committing the mock crime. As a consequence, the D-scores for guilty suspects who rehearsed an alibi should be more similar to the Innocent group than to the Guilty-Standard group. Alternatively, if our previous finding was caused only by an increase in implicit truth value of the alibi scenario due to an imagination inflation-related process (e.g. Loftus & Pickrell, 2995; Shidlovski et al., 2014), then there should be no difference in aIAT performance between the Guilty-Alibi and Guilty–Standard groups as guilt detection rates in both groups should be equal, but both groups should have higher D-scores and be more likely to be detected as guilty than the Innocent group.

Method

Participants

The final sample consisted of 108 undergraduate students from the University of Kent who took part via a research participation scheme in return for course credits (M_{age} = 18.94 years, *SD* = 1.98, age range = 18-36 years), maintaining the same statistical power as in Experiment 1. Twelve additional participants were excluded due to technical errors or failures to follow instructions. Participants were randomly assigned to three experimental groups (N = 36 in each group): the Guilty-Alibi group (31 female and 5 male), the Guilty-Standard group (33 female and 3 male), and the Innocent group (30 female and 6 male). The groups did not differ in age ($F(2, 105) = 0.78$, $p = .461$, $\eta_p^2 = 0.02$), nor gender ($\chi^2(2) = 1.15$, $p = .563$, $\varphi = 0.10$). All participants had English as their first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethics committee.

Materials, design and procedure

The materials, design and procedure was identical to Experiment 1 with one exception; the aIAT version was different. As in Experiment 1, the study was conducted in three stages. First, participants in the two guilty groups carried out a mock crime in which they required to go to an office block and steal a ring from a bag, whilst innocent participants carried out an innocent act, involving writing their email address on a paper in the same area as the guilty participants. Next, half of the guilty participants were instructed to imagine performing the innocent act as a fake alibi with the explicit intention to use it as a strategy to appear innocent. The rest of participants performed a filler task. Finally, all three groups took an aIAT, which assessed which of two events had a stronger relative association with the truth. Importantly, instead of contrasting the mock crime and innocent act/false alibi directly, the aIAT in Experiment 2 contrasted the mock crime with a completely novel unexperienced event involving entering a lecturer’s office and stealing a CD with exam questions on (henceforth referred to as the “exam” event, adapted from Sartori et al., 2008) that should not be associated with any truth value for any of the groups. All aspects of the aIAT task design and instructions were the same as in Experiment 1, with the only change being that sentences related to the alibi/innocent act were replaced with sentences related to the unexperienced event. As in Experiment 1, the order of the guilt congruent vs. incongruent blocks was counterbalanced across participants, and an equal number of participants within each group received each order.

After the main experiment, all participants completed a questionnaire where they rated how they had experienced and conducted the different tasks. They rated how nervous they had been while conducting the mock crime/innocent act (as applicable; on a 0-6 scale where 0 indicated not nervous at all; 6 extremely nervous), and how often they were thinking about the mock crime/innocent act during the aIAT (with 0 indicating not at all; 6 indicating all the time). The two guilty groups also rated their motivation to beat the aIAT (with 0 indicating not motivated at all; 6 indicating extremely motivated), and answered open-ended questions on whether they used any strategy to intentionally distort the test. There were also two additional questions for guilty-alibi participants: how vividly they had been able to imagine the alibi (with 0 indicating not vivid at all; 6 indicating extremely vivid) and how often they were thinking about the alibi during the aIAT (with 0 indicating not at all; 6 indicating all the time).

Results

The mean standardized D-score indices of guilt (Greenwald, et al., 2003; Hu et al., 2015) were significantly different between the groups ($F(2, 105) = 6.73$, $p = .002$, $\eta_p^2 = 0.11$; see Figure 2). Innocent participants, who had no knowledge of neither the mock crime nor the novel “exam” event, obtained a D-score that was not significantly different from zero as expected, and the Bayes Factor showed relatively stronger evidence for no difference than a difference ($t(35) = -0.57$, $p = .569$, $d = 0.10$; $BF_{10} = 0.21$). Guilty-Standard participants, who committed the mock crime and did not have any knowledge of the exam event, elicited D-scores significantly above zero, strongly supported by a very large Bayes Factor ($t(35) = 4.10$, $p < .001$, $d = 0.68$ $BF_{10} = 115.39$). The Guilty-Alibi participants, who also committed the mock crime and did not have any knowledge about the exam event, also elicited D-scores significantly above zero, but with only anecdotal support for a difference from the Bayes Factor ($t(35) = 2.28$, $p = .029$, $d = 0.38$; $BF_{10} = 1.75$). D-scores were significantly lower in the Innocent group than Guilty-Standard ($t(70) = 3.59$, $p <$

.001, d = 0.85; BF_{10} = 46.66) and Guilty-Alibi groups ($t(70) = 2.06$, $p = .043$, d = 0.49; BF_{10} =1.48). There was also a non-significant trend towards lower D-scores in the Guilty-Alibi than Guilty-Standard group, but for this test the Bayes Factor was weakly more supportive of no group difference than a difference ($t(70) = 1.67$, $p = .099$, d = 0.39; BF_{10} =0.80). These results indicate that imagining a false alibi does not abolish the implicit truth value of the true crime memory since the mean D-score was significantly above zero in the Guilty-Alibi group, and there was now only a weak, non-significant tendency, and no Baysian support for reduced aIAT memory detection in this group compared to the standard guilty condition.

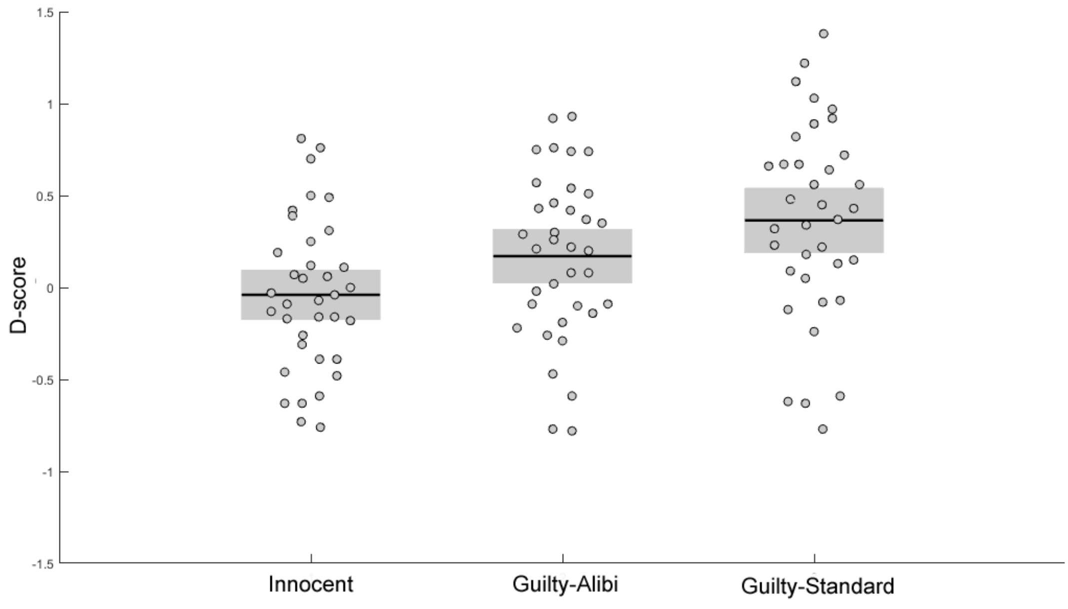

Figure 2. D-scores for the three groups from the Mock Crime/Unexperienced event aIAT in Experiment 2. The black lines shows the mean scores and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "innocent" event, innocence cannot be detected in this aIAT version.

A threshold-independent ROC analysis to evaluate classification performance showed that when comparing Guilty-Standard and Innocent groups, D-score classification was significantly better than chance (*AUC* = .73, *SE* = .06, *p* = .001). Comparing Guilty-Alibi and Innocent groups, D-score classification was lower, but also better than chance (*AUC* = .64, *SE* = .07, *p* = .043). The D-score classification results thus indicated that rehearsing an alibi did not fully impair the original memory of the mock crime because these participants could still be detected as guilty, yet there was a subtle numerical reduction in guilt classification for Guilty-Alibi participants.

Post-Experiment Questionnaire Analysis

Ten participants (4 Innocent, 3 Guilty-standard and 3 Guilty-Alibi) were excluded from the questionnaire analysis due to missing responses. The results revealed no differences between Guilty-Standard (*M* = 2.76, *SD* = 1.60) and Guilty-Alibi (*M* = 2.60, *SD* = 1.46) groups in nervousness during the mock crime ($t(64) = 0.40$, *p* = .689, *d* = 0.10) and the extent to which they thought about the mock crime during the aIAT (*M* = 3.21, *SD* = 1.53; *M* = 3.52, *SD* = 1.17, respectively; $t(64) = 0.90$, *p* = .372, *d* = 0.23). However, there was a significant difference between guilty groups in their motivation to beat the test: the Guilty-Alibi (*M* = 4.15, *SD* = 1.18) group was more motivated to appear innocent than the Guilty-Standard group (*M* = 3.45, *SD* = 1.35; $t(62) = 2.24$, *p* = .029, *d* = 0.56). The Innocent group reported being significantly less nervous while conducting the innocent task than the Guilty groups were while conducting the mock crime (Innocent *M* = 1.78, *SD* = 1.60; Innocent vs. Guilty-Alibi: ($t(63) = 2.17$, *p* = .033, *d* = 0.55; Innocent vs. Guilty-Standard: ($t(63) = 2.46$, *p* = .017, *d* = 0.62). They also thought less about the innocent act during the aIAT than the two Guilty groups thought about the mock crime during the aIAT (Innocent *M* = 1.00, *SD* = 1.50; Innocent vs. Guilty-Alibi: ($t(63) = 7.53$, $p < .001$, *d* = 1.90; Innocent vs. Guilty-Standard:

($t(63) = 5.87$, $p < .001$, $d = 1.48$), as would be expected since there were no sentences related to the innocent act in this aIAT version. Exploratory correlation analyses were also conducted to investigate whether any of the self-report measures correlated with performance in the aIAT, but there were no significant correlations.

Experiment 2 Discussion

Experiment 2 assessed whether imagining a false alibi reduces the implicit truth value of the true crime memory, in line with previous findings that have shown that learning counterfactual details after a mock crime can impair true memories of the crime (Gronau, et al., 2015). In Experiment 1, the results showed that the aIAT was unable to determine whether an experienced mock crime or an imagined false alibi was true. However, the aIAT design did not permit us to test whether this lack of discrimination was caused by increased truth value of the imagined alibi or decreased truth value of the mock crime, or a combination of both. In Experiment 2, we therefore contrasted the mock crime with a novel event that had been neither experienced nor imagined in an aIAT, in order to assess the implicit truth value of the mock crime memory independent of the alibi memory. In this study, the mock crime was still detected despite participants previously imagining a false alibi, suggesting that the alibi had not impaired the true memory of the crime to a substantial extent.

As expected in Experiment 2, the mean D-score of innocent participants was close to zero, suggesting that neither event was strongly associated with the truth in this group. Both guilty groups scored above zero, indicating that they associated the mock crime with the truth more than the unexperienced event. Therefore, it appears that the low discrimination between the experienced mock crime and imagined alibi in Experiment 1 was mainly driven by the alibi manipulation increasing the implicit truth value of the imagined scenario, rather than a reduction of implicit truth value of the mock crime memory. This finding contrasts with other research that has suggested that rehearsing a false alibi can cause it to become a default response such that when a cue triggers a memory about a crime, that memory is automatically inhibited to facilitate a false alibi response (Foerster et al, 2017), and that thinking counterfactually can impair memories for the event that actually occurred (Petrocelli & Crysel, 2009; see also Otgaar and Baker, 2018).

One possible reason why the true mock crime memory was unimpaired in Experiment 2 might be that the alibi manipulation was only implemented through one brief rehearsal and imagination phase. Thus, the effect of the alibi manipulation may not have been as strong as in real life situations where suspects may prepare and imagine an alibi repeatedly and over a long-time period before the interrogation. If participants were able to rehearse/imagine the alibi in this way, it may be more likely to impair the true memory of the mock crime, either by increased retroactive interference or by inhibition of the crime memory representation itself (e.g. Gronau et al., 2015). Previous research has suggested that when multiple memories are associated to the same cue, repeatedly retrieving one memory in the face of competitive activation of another memory can cause the non-selected memory to become inhibited (Anderson, Bjork & Bjork, 1994). Likewise, repeatedly pushing an unwanted memory out of mind by thinking of a substitute thought may interfere with (Bergström et al., 2009) retrieval of the original memory, or even inhibit it (Anderson & Benoit, 2012). The literature on motivated forgetting suggests that such impairments of unwanted memories are gradual and increase with repetition (e.g. Anderson & Green, 2001), predicting that a true crime memory might only become impaired if a false alibi is *repeatedly* retrieved. Likewise, the retroactive interference theory suggests that repeatedly rehearsing one memory associated to a cue may strengthen that association, which can block access to other associated memories without those memories being inhibited (see Anderson & Neely, 1996). Thus, multiple theoretical accounts suggest that repeated and temporally extended imagination of an alibi should be more likely to impair access to the original crime memory, as addressed in the next experiment.

Experiment 3

Experiment 3 was designed to replicate and extend on findings from the previous studies, with particular focus on whether repeated rehearsal of a false alibi over an extended time period might be more effective at impairing the true memories compared to a single brief alibi intervention just before the aIAT. In the previous two experiments all experimental phases were conducted in the same session; participants first conducted a mock crime, then immediately learned and imagined the false alibi, which was followed by the aIAT. We therefore added a time delay of one week between the mock crime and test, which made the design more realistic and enabled us to investigate the effect of repeated and distributed false alibi rehearsal on aIAT memory detection.

The experimental design was similar to the previous studies, except that it was conducted in two sessions one week apart, and included an additional experimental group. Furthermore, in the second session, all participants completed three versions of the aIAT that contrasted the mock crime vs. the innocent/alibi event (same aIAT as in Experiment 1), the mock crime vs. an unexperienced event (same aIAT as in Experiment 2), and the alibi vs. the unexperienced event (a new aIAT version to assess the implicit truth value of the innocent act/alibi independently of the mock crime). Similarly to previous experiments, participants first conducted either an innocent act or a mock crime, depending on which group they were assigned to. All participants then came back for the aIAT session a week later. In one countermeasure group ("Guilty-Alibi"), participants conducted a mock crime during the first session, then left and returned a week later at which point they learned and imagined the false alibi immediately before the aIATs. In the other countermeasure group ("Guilty-Alibi with home training"), participants learned and imagined the false alibi during the first session immediately after conducting the mock crime, and were also required to repeat this imagination task at home once a day for a week before returning to complete the aIATs. These two countermeasure groups were compared against Innocent and Guilty-Standard groups, as in the previous two studies.

We expected that participants who carried out an innocent act should be detected as innocent and participants who committed a mock crime without learning an alibi should be detected as a guilty across the relevant aIAT versions. However, participants who learned the false alibi would be less likely to be detected as guilty than the standard guilty group. If imagining a false alibi leads to gradual strengthening of the false alibi information in memory and/or gradual impairment of the true memory with repetition, then extended rehearsal of a false alibi for a week before the test should be particularly effective at reducing detection of guilty suspects.

Methods

Participants

The final sample consisted of 144 undergraduate students from the University of Kent who took part via a research participation scheme in return for course credits (M_{age} = 19.13 year, *SD* = 1.57, age range = 18-34 years). Twenty-eight additional participants were excluded due to technical errors, failures to follow instructions, or failure to attend both sessions. Participants were randomly assigned to one of the four groups (*N* = 36 in each group): Innocent (30 female, 6 male), Guilty-Standard (30 female, 6 male), Guilty-Alibi (27 female, 9 male), and Guilty-Alibi with Home Training (HT; 31 female, 5 male). Thus, this experiment maintained the same statistical power as the previous two experiments for pairwise comparisons between groups. The groups did not differ in terms of age (F(3,140) =

0.74, p = .531, η_p^2 = .02) nor gender ($\chi^2(3)$ = 1.69, p = .639, φ = 0.11). All participants had English as their first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethics committee.

Materials, Design, and Procedure

First, participants in all three Guilty groups committed a mock crime involving going to a staff office area and stealing a ring, whereas participants in the Innocent group completed an innocent task involving writing their email address on a sign-up sheet in the same area (both these tasks were kept identical to Experiments 1 and 2). Next, all participants were dismissed and asked to come back to the laboratory after a week, except the Guilty-Alibi with HT group. The latter group were given instructions to perform an extra task after completing the mock crime. They first learned and imagined a false alibi, which described the innocent act, using the same materials and procedure as in Experiments 1 and 2. Next, they were given a home training task, which required them to access an internet link in order to rehearse the false alibi once every day in the intervening six days until the test day. When they accessed the link, they were asked to read a description of the alibi (using the same text as used on the first day) and imagine themselves completing the described actions as vividly and accurately as possible. After that, they were asked to write down a detailed description of the scenario they had imagined and rate how vivid their imagination of the alibi had been. Participants were only included in the final sample if they had completed this task as instructed.

After a week, all participants came back to the lab to complete the rest of the study. Participants in Innocent and Guilty-Standard group were asked to complete a filler task (solving Sudoku puzzles), while the two Alibi groups rehearsed the alibi (describing the innocent act). For the Guilty-Alibi group, this was the first time they learned that they needed to use an alibi to appear innocent and found out the details of the alibi/innocent act, whereas for the Guilty-Alibi with HT group it was another chance to rehearse the alibi they had learned and repeatedly rehearsed during the preceding week. Finally, all participants completed three versions of the aIAT: 1) contrasting the mock crime vs. the innocent/alibi event (same aIAT as in Experiment 1); 2) contrasting the mock crime vs. the unexperienced event involving stealing an exam (same aIAT as in Experiment 2); and 3) contrasting the innocent/alibi vs. the unexperienced event (a novel aIAT version used to assess whether the innocent event would be detected as true after rehearsing a false alibi). The aIAT task design, sentences and instructions were identical to those used in the previous studies, with the only changes being the added new version 3, and that all participants undertook all three versions. The order of aIAT congruent/incongruent blocks and order of versions was fully counterbalanced across participants to prevent order effect confounds, and counterbalancing formats were equally distributed within each of the four groups.

After the experiment, participants were asked to complete a questionnaire, which was similar to the one used in Experiment 2 with a few additional questions about details of the innocent act or mock crime. For the Innocent group, participants were required to give answers relating to details of the innocent act and give ratings on a scale from 0 to 6 regarding their behaviour and experience during the initial act and the aIAT (e.g. in how much detail they could remember the act, their motivation to beat the aIAT, and the extent to which they thought about the act during the aIAT). The Guilty groups were asked to provide answers regarding details of the mock crime and provide various ratings on a 0-6 scale regarding their nervousness during the mock crime, their motivation to beat the aIAT, the extent to which they thought about the mock crime during the aIAT, and whether they had intentionally used any strategy to distort the test, including the extent to which they thought about the alibi scenario during the aIAT and how vividly they had imagined an alibi (for the Guilty Alibi groups only).

Results

Mock Crime/Innocent event aIAT

The Mock Crime/Innocent version of the aIAT directly contrasted the mock crime (ring) with the innocent/alibi (email) event, and was identical to the aIAT used in Experiment 1. In this test, positive D-scores (Greenwald, 2003; Hu et al., 2015) are indicative of guilt because they suggest participants associate the mock crime with the truth whereas negative D-scores are indicative of innocence because they suggest participants associate the innocent event with the truth. The mean D-scores were significantly different between the groups ($F(3, 140) = 6.78$, $p < .001$, $\eta_p^2 = 0.13$; see Figure 3). The mean D-score of the Innocent group was not significantly different from zero, with the Bayes Factor indicating (weak) relative support of no difference over a difference ($t(35) = -1.30$, $p = .20$, $d = 0.22$; $BF_{10} = 0.39$), inconsistent with the predictions and suggesting that the innocent event was on average not detected as true in this group. The Guilty-Standard group however did obtain a D-score that was significantly above zero with strong supporting evidence from the Bayes Factor ($t(35) = 3.75$, $p < .001$, $d = 0.63$; $BF_{10} = 47.32$) indicating successful guilt detection in this group. The Guilty-Alibi group who committed a mock crime and learned a false alibi just prior to the test however had a mean score significantly *below* zero ($t(35) = -2.06$, $p = .047$, $d = 0.34$; $BF_{10} = 1.18$), thus appearing more innocent than guilty, although the Bayes Factor was only weakly supportive of a difference from zero in this group. In contrast, the Guilty-Alibi with HT group, who committed a mock crime and then repeatedly rehearsed a false alibi for a week before the test, did not have a mean D-score that differed from zero ($t(35) = 1.01$, $p = .320$, $d = 0.17$; $BF_{10} = 0.29$). Independent t-tests revealed that the mean D-score of the Innocent group was significantly lower than in the Guilty-Standard group ($t(70) = 3.54$, $p < .001$, d = 0.83; BF_{10} = 40.34), while there were no differences between the Innocent and either of the Alibi groups (Innocent vs. Guilty-Alibi: $t(70) = 0.54$, $p = .59$, $d = 0.13$, BF_{10} =0.28; Innocent vs. Guilty-Alibi with HT: $t(70) = 1.64$, $p = .10$, $d = 0.39$; BF_{10} =0.76). However, the mean D-score of the Guilty-Standard group was significantly higher than in the Guilty-Alibi group, with strong support for a difference from the Bayes Factor ($t(70) = 4.08$, $p < 0.001$, $d = 0.96$; BF_{10} =194.05), but only trend level higher than in the Guilty-Alibi with HT group with only anecdotal Bayesian support for a difference ($t(70) = 1.95$, $p = .056$, $d = 0.46$; BF_{10} =1.21). Surprisingly, the mean D-score of the Guilty-Alibi with HT group was significantly higher than the Guilty-Alibi group with anecdotal Bayesian support for a difference between the two alibi groups ($t(70) = 2.19$, $p = 0.03$, $d = 0.52$; BF_{10} = 1.84), suggesting that extended training with the alibi actually made it a *less* effective strategy for appearing innocent on this aIAT version.

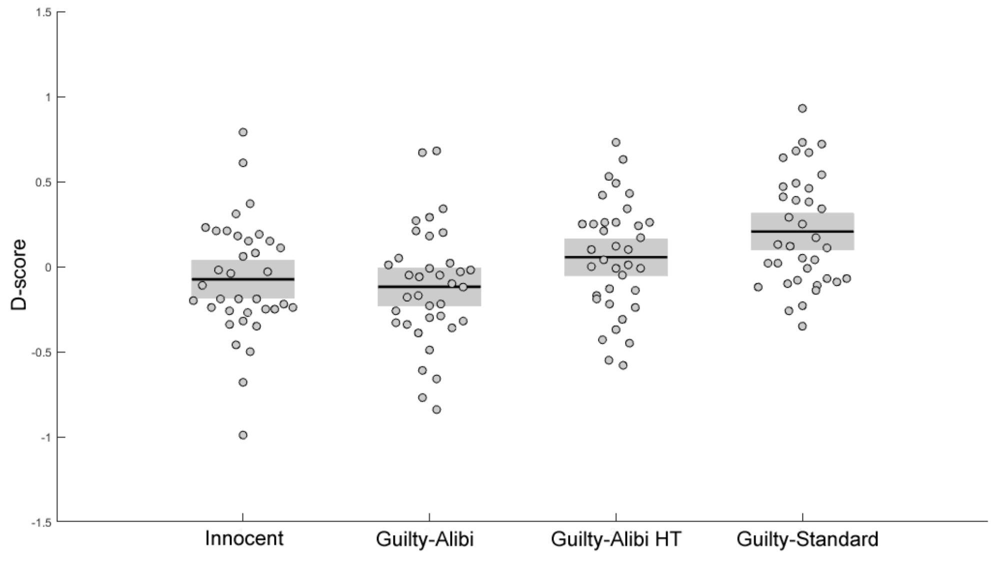

Figure 3. D-scores for the four groups from the Mock Crime/Innocent event aIAT in Experiment 3. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth) and D-scores below zero suggest innocence (that the email-related sentences are associated with the truth).

A threshold-independent ROC analysis to evaluate classification performance showed that when comparing Innocent and Guilty-Standard groups, D-score classification was significantly better than chance (*AUC* = .72, *SE* = .060, *p* = .001). However, D-score classification was not accurate when comparing Innocent and Guilty-Alibi groups (*AUC* = .54, *SE* = .069, *p* = .581), nor when comparing Innocent and Guilty-Alibi with HT groups, although the latter was at trend-level (*AUC* = .62, *SE* = .067, *p* = .073).

So in sum, the Mock Crime/Innocent aIAT largely replicated the findings from Experiment 1; guilty participants who did not use countermeasures could be detected as guilty, whereas imagining a false alibi led to lower detection rates. However, this countermeasure was most effective when applied only once immediately before the aIAT, contrary to our predictions that extended and repeated alibi rehearsal would enhance the effectiveness of this strategy. Also somewhat surprising was that detection of innocent participants was relatively poor compared to Experiment 1.

Mock Crime/Unexperienced event aIAT

The Mock Crime/Unexperienced event version of the aIAT contrasted the mock crime (ring) with an event that none of the groups had experience nor knowledge of (exam), and was identical to the aIAT version used in Experiment 2. In this test, positive D-scores are indicative of guilt because they suggest that participants associate the mock crime with the truth, whereas D-scores around zero suggest that participants associate both events equally strongly with the truth (i.e. they associate either both, or neither event with the truth). Because none of the two events is indicative of innocence there is no result that would be diagnostic of innocence in this aIAT version, and no groups were predicted to show negative D-scores. In this test, there was only a trend towards differences between the groups in mean

D-scores ($F(3, 140) = 2.50$, $p = .062$, $\eta_p^2 = 0.05$; see Figure 4), suggesting that this aIAT version did not discriminate between the groups as well as the Mock Crime/Innocent event aIAT (as would be expected since there should be less variability between groups when the test is designed to only produce scores either around zero or above, and no negative scores). The mean D-scores of Guilty-Standard ($t(35) = 3.99$, $p < .001$, $d = 0.67$; BF_{10} =87.79) and Guilty-Alibi with HT group ($t(35) = 3.07$, $p = .004$, $d = 0.51$; BF_{10} =8.97) were significantly above zero, supported by large Bayes Factors. However, the mean D-scores for Innocent ($t(35) = -0.17$, $p = .868$, $d = 0.03$; BF_{10} =0.18) and Guilty-Alibi groups ($t(35) = 1.91$, $p = .064$, $d = 0.32$; BF_{10} =0.91) were not significantly different from zero, with the Bayesian evidence more in favour of no difference than a difference. Independent t-tests revealed that the mean D-score of the Innocent group was significantly lower than in the Guilty-Standard group ($t(70)$ =2.59, $p = .01$, d =0.61; BF_{10} =4.04) and the Guilty-Alibi with HT groups ($t(70)$ =2.19, $p = .03$, d =0.52; BF_{10} =1.84), however no significant differences between the groups emerged from the other pairwise comparisons (Innocent vs. Guilty-Alibi: $t(70)$ =1.49, $p = .14$, d =0.35, BF_{10} =0.63; Guilty-Standard vs. Guilty-Alibi: $t(70)$ =0.89, $p = .38$, d =0.21, BF_{10} =0.34; Guilty-Standard vs. Guilty-Alibi with HT: $t(70)$ =0.27, $p = .79$, d =0.06, BF_{10} =0.25; Guilty-Alibi vs. Guilty-Alibi with HT: $t(70)$ =0.61, $p = .55$, d =0.14, BF_{10} =0.29).

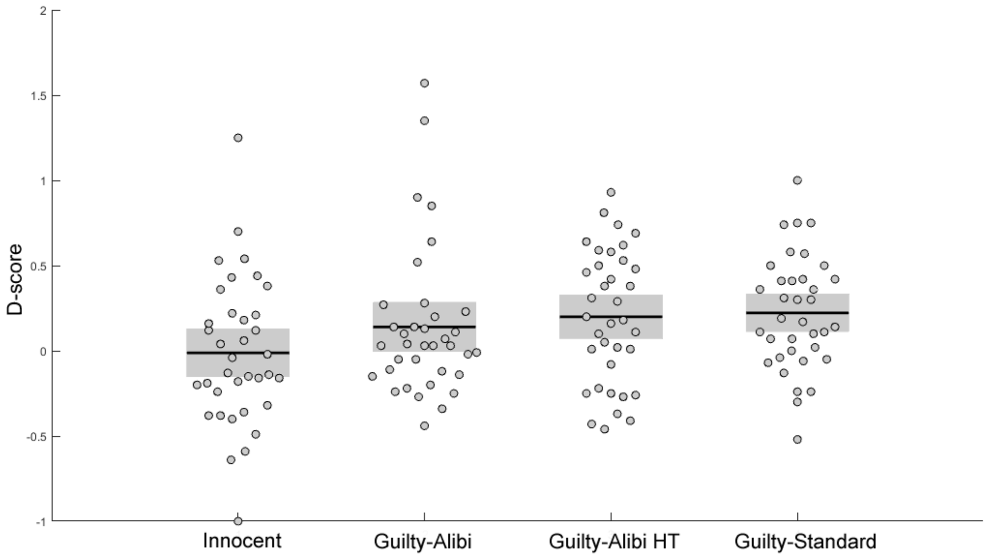

Figure 4. D-scores for the four groups from the Mock Crime/Unexperienced event aIAT in Experiment 3. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "innocent" event, innocence cannot be classified in this aIAT version.

Threshold independent ROC analyses showed that D-score classification performance was above chance when comparing the Innocent and Guilty-Standard groups (*AUC* = .68, *SE* = .064, *p* = .009) and when comparing the Innocent and Guilty-Alibi with HT groups (*AUC* = .62, *SE* = .065, *p* = .037). However, classification performance was not accurate when comparing the Innocent and Guilty-Alibi groups (*AUC* = .59, *SE* = .068, *p* = .207).

To summarise, results of the Mock Crime/Unexperienced event aIAT in Experiment 3 replicated the findings from Experiment 2 that guilty participants who did not use countermeasures could be detected as guilty when compared to an innocent group. Consistent with results from the Mock Crime/Innocent event version in Experiment 3, the Mock Crime/Unexperienced event aIAT also indicated that whereas the Guilty-Alibi with HT group could be detected as guilty, the Guilty-Alibi group without home training appeared less guilty (they were not significantly different from the Innocent group in any analysis). This pattern again suggests that the false alibi countermeasure was most effective when applied only once immediately before the aIAT, contrary to our predictions. However, the effects of the alibi manipulation were weaker on this version of the aIAT compared to the Mock Crime/Innocent event aIAT, since the Guilty-Alibi group did not show a significant reduction in D-score compared to the Guilty Standard group. Thus, consistent with Experiments 1 and 2, the alibi manipulation was most effective when the mock crime and alibi were directly contrasted, and was less effective when the mock crime was contrasted with an unexperienced event.

Innocent/Unexperienced event aIAT

The Innocent/Unexperienced event version of the aIAT contrasted the innocent/alibi event (involving writing an email) with an event that none of the groups had experience nor knowledge of (stealing an exam) in order to assess whether the innocent/alibi event would be detected as true for any of the groups. That is, would learning and rehearsing a false alibi lead that scenario to be detected as true, or would it only be detected as true for the Innocent group who had actually conducted the act? In this test, positive D-scores are indicative of innocence because they suggest that participants associate the email event with the truth, whereas D-scores around zero suggest that participants associate both events equally strongly with the truth (i.e. they associate either both, or neither event with the truth). Because neither of the two events is indicative of guilt there is no result that would be diagnostic of guilt in this aIAT version, and no groups were predicted to show negative D-scores. In this test, the mean D-score of the Guilty-Standard group was not different from zero ($t(35) = 0.09$, $p = .928$, $d = 0.02$; $BF_{10} = 0.18$) as expected, since this group had no knowledge of either event. In contrast, the Guilty-Alibi ($t(35) = 2.28$, $p = .029$, $d = 0.38$; $BF_{10} = 1.73$) and Guilty-Alibi with HT groups ($t(35) = 2.23$, $p = .033$, $d = 0.37$; $BF_{10} = 1.58$) did score significantly above zero, suggesting that the alibi was detected as if true on average in these groups (although with only weak support from the Bayes Factor). Surprisingly however, the Innocent group's mean D-score was not significantly above zero and the Bayes Factor indicated relative support for no difference from zero (t(35) = 0.40, p = .687, d = 0.07; BF_{10} =0.19), showing a failure of the test to detect the innocent event even though it was actually true for that group. There was also no overall significant difference between the groups in mean D-scores (F(3, 140) = 1.95, p = .124, η_p^2 = 0.04; see Figure 5), suggesting that this aIAT version did not discriminate between the groups well. Comparing differences in mean D-score between groups using independent t-tests, there were non-significant trends towards more positive D-scores in the two Alibi groups than in the Guilty-Standard group (Guilty-Alibi vs. Guilty-Standard: t(70) =1.81, p = .08, d =0.43, BF_{10} =0.98; Guilty-Alibi with HT vs. Guilty-Standard: t(70) =1.79, p = .08, d =0.42; BF_{10} =0.95) but none of the other differences approached significance and the Bayesian analysis indicated relatively more support for no difference than a difference for all comparisons (Innocent vs. Guilty-Standard: t(70) =0.37 p = .72, d =0.09, BF_{10} =0.26; Innocent vs. Guilty-Alibi: t(70) =1.36, p = .18, d =0.32, BF_{10} =0.54; Innocent vs. Guilty-Alibi with HT: t(70) =1.35, p = .18, d =0.32, BF_{10} =0.53; Guilty-Alibi vs. Guilty-Alibi with HT: t(70) =0.01, p = .99, d <0.01,BF_{10} =0.24).

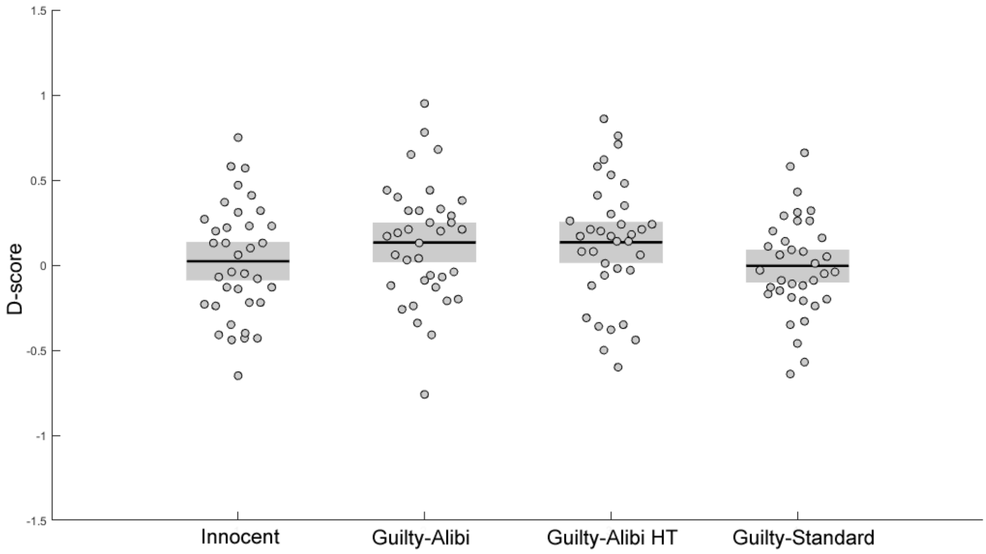

Figure 5. D-scores for the four groups from the Innocent/Unexperienced event aIAT in Experiment 3. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest innocence (that the email-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "guilty" event, guilt cannot be classified in this aIAT version.

Threshold-independent ROC analyses revealed that D-score classification based on the Innocent/Unexperienced event aIAT was inaccurate. Comparing the Innocent group with the Guilty-Standard group, classification performance was at chance (*AUC* = .52, *SE* = .069, *p* = .787), and it was only slightly better but still not significant when comparing Innocent participants to Guilty-Alibi (*AUC* = .59, *SE* = .067, *p* = .177) and Guilty-Alibi with HT (*AUC* = .59, *SE* = .068, *p* = .169).

Thus, in this aIAT version, we found very poor detection of the participants who had actually performed the innocent act, whereas imagining a false alibi seemed to have slightly increased detection of this false scenario as true in the two Alibi groups. However, since the groups were not significantly different from each other in mean D-scores or classification rates, this slight increase in the Alibi groups was not reliable.

Post-Experiment Questionnaire

Results from the final questionnaire are shown in Table 1. The Innocent group rated their memory of the innocent act as less vivid than the three Guilty groups rated their memory for the mock crime act (Innocent vs. Guilty-Standard: $t(70) = 3.46$, $p = .001$, $d = 0.83$; Innocent vs. Guilty-Alibi: $t(70) = 3.39$, $p = .001$, $d = 0.81$; Innocent vs. Guilty-Alibi with HT: $t(70) = 4.45$, $p < .001$, $d = 1.06$) and they also reported that they remembered fewer details of the act (Innocent vs. Guilty-Standard: ($t(70) = 4.42$, $p < .001$, $d = 1.06$; Innocent vs. Guilty-Alibi: $t(70) = 5.20$, $p < .001$, $d = 1.24$; Innocent vs. Guilty-Alibi with HT: $t(70) = 4.93$, $p < .001$, $d = 1.18$). The Innocent group also reported having been less nervous during the innocent act than the three Guilty groups were when they committed the mock crime (Innocent vs. Guilty-Standard: $t(70) = 2.80$, $p = .007$, $d = 0.67$; Innocent vs. Guilty-Alibi: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Innocent vs. Guilty-Alibi with HT: $t(70) = 3.83$, $p < .001$, $d = 0.92$), and reported thinking about the innocent act less during the aIATs than the three Guilty groups thought about the mock crime during the aIATs (Innocent vs. Guilty-Standard: $t(70) = 3.85$, $p < .001$, $d = 0.92$; Innocent vs. Guilty-Alibi: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Innocent vs. Guilty-Alibi with HT: $t(70) = 3.95$, $p < .001$, $d = 0.94$). There were no significant differences between the three Guilty groups on any of those questions (all ps >0.14).

Table 1. Mean and standard deviations of self-reported ratings on the final questionnaire for the four groups. The scale had seven points (0-6), and lower scores always indicate less of the item being measured (e.g. less vividness/nervousness/motivation, etc.) and higher scores always indicate more of the item being measured (e.g. more vividness/nervousness/ motivation, etc.).

Questionnaire item	Innocent	Guilty-Alibi	Guilty-Alibi with HT	Guilty-Standard
Remember detail of the act	3.39 (1.25)	4.64 (0.72)	4.64 (0.87)	4.53 (0.91)
Vividness of the act memory	3.50 (1.76)	4.36 (0.83)	4.69 (0.98)	4.44 (1.03)
Nervousness during the act	1.67 (1.29)	2.33 (1.37)	3.05 (1.76)	2.69 (1.79)
Thinking about the act during aIAT	1.58 (1.56)	2.50 (1.68)	3.11 (1.71)	3.08 (1.75)
Motivation to beat the aIAT	3.86 (1.50)	4.14 (1.22)	4.14 (1.50)	3.11 (1.58)
Imagine detail of the alibi	-	3.94 (1.33)	4.57 (0.70)	-
Vividness of the alibi imagination	-	3.89 (1.47)	4.57 (0.88)	-
Thinking about the alibi during aIAT	-	2.83 (1.68)	3.14 (1.78)	-

Note: the “act” refers to the act conducted in the first session (i.e. either mock crime or innocent act, depending on group).

The Alibi groups and the Innocent group were all more motivated to appear innocent on the aIATs than the Guilty-Standard group (Guilty-Standard vs. Innocent: $t(70) = 2.04$, $p = .045$, $d = 0.49$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 3.09$, $p = .003$, $d = 0.74$; Guilty-Standard vs. Guilty-Alibi with HT: $t(70) = 2.83$, $p = .006$, $d = 0.68$), but did not differ between each other in levels of motivation (all *p*s >0.39). With regards to the alibi-specific questions, there were no differences between the Albi groups in terms of how much they were thinking of the alibi during the aIATs ($t(70) = 0.75$, $p = .46$, $d = 0.18$), but the Guilty-Alibi with HT group reported being able to imagine the alibi scenario in more details ($t(70) = 2.48$, $p = .016$, $d = 0.59$) and more vividly than the Guilty-Alibi group ($t(70) = 2.36$, $p = .021$, $d = 0.56$). Exploratory correlation analyses were also conducted to investigate whether any of the self-report measures correlated with performance in the aIAT, but there were no significant correlations.

So in sum, the questionnaire data from Experiment 3 suggested that the Innocent group had poorer memory of the innocent act than the Guilty groups’ memory of the mock crime, whereas repeated and extended rehearsal of the alibi scenario in the Guilty-Alibi with

HT group led to improved ability to imagine the alibi scenario when compared to the Guilty-Alibi group. Furthermore, the Innocent and Alibi groups were more motivated to appear innocent on the aIATs than the Guilty-Standard group.

Experiment 3 Discussion

The aim of this study was to further investigate the effect of rehearsing alibi as a countermeasure on the aIAT (Agosta & Sartori, 2013; Sartori et al., 2008). Previous research suggested that rehearsing an counterfactual scenario to what actually happened during a mock crime can impair access to the true memory (Gronau et al., 2015). In Experiment 3, we investigated whether learning and imagining a false alibi prior to the aIAT would impair the original memory for a mock crime and/or increase the implicit truth value of the alibi itself, and whether these effects would be particularly enhanced when the alibi was repeatedly rehearsed and imagined over an extended time period, in line with theoretical accounts of retrieval interference and inhibition (see e.g. Anderson & Green, 2001; Anderson & Hanslmayr, 2014; Anderson & Neely, 1996). Such extended and repeated practice of an alibi might be expected to occur in real life, since a guilty criminal might adopt a false alibi and then practice it extensively prior to an investigation several days, weeks or months later.

The results indicated that in the aIAT that tested the relative strength of the mock crime vs. innocent act/alibi, the mock crime was possible to detect after a week delay in Guilty-Standard participants. However, this aIAT could not distinguish which of the two events were true for Innocent participants, nor for the Guilty-Alibi with HT groups. Interestingly, in the Guilty-Alibi group that did not receive home training, the test result was more indicative of innocence than guilt. In the aIAT that tested the relative strength of the mock crime vs. an unexperienced event, results suggested that the mock crime was possible to detect in Guilty-Standard and Guilty-Alibi with HT groups, while it was undetectable in

Innocent and Guilty-Alibi groups. In the aIAT that tested the relative strength of the innocent/alibi act vs. an unexperienced event, none of the groups showed strong evidence of innocence and this aIAT showed poor discrimination between all groups.

Our findings thus indicate that the strongest effect of the alibi countermeasure was in the Guilty-Alibi participants who learned and imagined a fabricated alibi one week *after* the mock crime and just prior to the test, without repeated rehearsal. In this group, the results suggested that they associated the imagined false alibi event more with the truth relative to the objectively true mock crime event. Moreover, the aIAT that contrasted the mock crime with an unexperienced event was not able to distinguish which of the two events was true for these guilty participants, suggesting that access to the mock crime memories may have been impaired in this group. Thus, the effect of the alibi countermeasure in this group was even stronger than the findings in Experiments 1 and 2, where the alibi group did not show significant associations between the alibi and truth (Experiment 1) and they also showed evidence of associating the mock crime with truth when contrasted with the unexperienced event (Experiment 2). These differences across studies may be due to differences in the relative strength of the memory representations for the alibi information versus the mock crime. Mental simulation of the alibi event just before the aIAT may have caused this imagined event memory to be more vivid or salient than the true memory of the mock crime, which may have been weaker in this experiment than in the previous two studies due to the longer time delay between the event and the test. Because of the relatively weak memory for the mock crime, the alibi countermeasure may have been more effective at obscuring detection of that memory than in the previous two studies (*cf.* Gronau et al., 2015, for related findings with psychophysiological memory detection).

Surprisingly, a different result pattern was observed in the guilty participants who received repeated alibi training for a week before the aIATs. We predicted that extended rehearsal of an imagined alibi would be particularly effective at inducing blocking by retroactive interference or competitive inhibition of the true memory (e.g. Anderson & Hanslmayr, 2014, Anderson & Neeley, 1996), and that this group would therefore be more likely to appear innocent compared to a group who only imagined the alibi once just before the test. However, we found the opposite result – although the extended alibi training did reduce memory detection on the aIAT version that directly contrasted the mock crime with the alibi, the magnitude of this reduction was smaller than in the alibi group without extended training. Furthermore, in the aIAT version that contrasted the mock crime with an unexperienced event, the mock crime was still detected as true in the extended training group. These results suggest that extensive and repeated rehearsal of the false alibi did not impair the original mock crime memory, rather, it may have actually strengthened that memory. The home training task may have had an ironic effect of reminding participants of the mock crime and leading the memory for the crime to become strengthened as a result, consistent with prior findings that repeated reminders can enhance automatic influences of memories, which can produce ironic effects when such enhancement affects behaviour in unwanted ways (Jacoby, 1999). Future research should assess whether alibi-induced ironic strengthening of the true crime memory can be avoided by explicitly training participants to suppress thoughts of the mock crime while completing the alibi imagination task, which might be an effective strategy for reducing mock crime memory strength whilst simultaneously strengthening memory for the alibi (*cf.* Anderson & Green, 2001; Bergström et al., 2013; Hu et al., 2015).

Another surprising finding in Experiment 3 was that none of the aIAT versions detected the innocent act as true for participants in the Innocent group despite them actually having conducted the act in real life. In contrast, the mock crime could be detected in the guilty participants who did not use countermeasures. This difference may be related to the one week delay that we introduced between the initial act and the aIATs, which may have weakened innocent participants' memory of the innocent act more than it weakened guilty participants' memory of the mock crime. In line with this suggestion, the innocent participants rated their memories of their act as less vivid and detailed than the guilty participants' ratings of the mock crime memories, and also reported that they had been less nervous while conducting the act than the guilty participants were when conducting the mock crime. This pattern of results suggest that the mock crime memories were associated with higher emotional arousal, which is known to enhance the subjective vividness of memories and their durability over time (Kensinger, 2009). This finding is interesting as it converges with other evidence that memories of recent, familiar events are more detectable in the aIAT than memories of distant, less familiar events (Takarangi et al., 2015) in pointing towards a role of subjective memory quality in aIAT accuracy – the test may only be able to detect memories that are subjectively detailed and vivid, and any factors that reduce memory quality may also reduce the test's effectiveness. It also suggests general limitations with laboratory studies that investigate memory detection with mock crimes, since memories of mock crimes may differ substantially from real criminal memories in terms of emotional arousal. Future research should investigate whether countermeasures can be used against aIAT memory detection of real autobiographical memories that are emotionally arousing.

Overall, the results of Experiment 3 support our hypothesis that rehearsing a false alibi before an aIAT may distort the test results, but they also show that the effectiveness of this strategy depends on how the alibi countermeasure is used, and also on how the aIAT is designed.

General Discussion

The aIAT has been promoted as an accurate tool for determining which of two autobiographical events are true, with promising applications in forensic memory detection

(Sartori et al., 2008; Agosta & Sartori, 2013). However, a growing body of research has revealed potential countermeasures that guilty suspects can adopt to make themselves appear innocent, such as intentionally altering their responses during the test itself (Agosta et al., 2011; Hu et al., 2012; Verschuere et al., 2009), or suppressing their incriminating memories in advance of the test (Hu et al., 2015). We tested whether a novel countermeasure that has recently been applied in physiological memory detection (Gronau et al., 2015) and deception detection paradigms (Foerster et al., 2017; Suchotzki et al., 2018) would also be effective at reducing detection using the aIAT. Specifically, we assessed whether instructing guilty suspects to intentionally store false information in memory would enable those suspects to appear innocent on the test. In line with our predictions, imagining a false alibi impaired memory detection with the aIAT so that the test could no longer distinguish between the objectively true mock crime memory and the objectively false alibi, and this finding was replicated with a large effect size in two experiments. Consistent with previous research (e.g. Sartori et al., 2008, Agosta & Sartori, 2013), our results showed relatively good discrimination between guilt and innocence in participants who did not employ countermeasures. However, the false alibi countermeasure significantly reduced memory detection when compared to a standard guilty group who were not trying to evade the test.

Across experiments, the strongest and most consistent effect of the alibi manipulation occurred on the aIAT version that directly contrasted the mock crime with the alibi to assess their relative truth value, whereas there were only weaker, less consistent effects on the aIAT that contrasted the mock crime with an unexperienced novel event to detect the truth value of the mock crime itself. This pattern indicates that the effectiveness of the alibi strategy was primarily driven by increased detection of the alibi as true, rather than decreased detection of the mock crime as true. Imagining a false alibi may have created a memory for the alibi scenario that had some implicit associations with the truth, even though participants knew their alibi was false at an explicit level. This account converges with more general findings that imagining an event can create a memory for that event that has similar perceptual and behavioural characteristics as memories based on true experiences (e.g. Loftus, 2003; Loftus & Pickrell, 1995; Mitchell & Johnson, 2009; Schacter, Guerin & St Jacques, 2011), and previous findings that imagining simple actions can increase detection of those actions as true in the aIAT, either by inducing misremembering that imagined actions were actually performed (Takarangi et al., 2013) or sometimes even despite participants knowing the imagined action did not actually happen (Shidlovski, et al., 2014).

Our findings thus converge with other research that have found dissociations between explicit and implicit measures of truth (Shidlovski, et al., 2014). It has been suggested that these dissociations occur because people can make contrary implicit and explicit evaluations of truth, which may help them deceive both themselves and others (Shidlovski et al., 2014). However, an alternative and more parsimonious explanation is that the aIAT does not actually measure implicit associations between events and the truth, but instead is simply sensitive to the relative salience of different events. In line with this view, asking participants to rehearse and imagine the alibi may have increased the relatively salience of this event compared to the mock crime or unexperienced event (*cf.* Rothermund & Wentura, 2004). Regardless of which account is correct, this uncertainty regarding what the aIAT measures is in our view a fundamental problem for using the aIAT in real criminal cases (see Sirgiovanni, et al., 2016) – if researchers do not know what the test is measuring, how can using the test be justified when a false result may have direct real life consequences? Clearly, practical applications of the aIAT are premature until further research has clarified what the test actually measures, and in what situations it will produce accurate results.

Although our key finding that the false alibi countermeasure reduced the aIAT’s ability to discriminate between a true mock crime and a false alibi was strong and robust, our sample sizes and designs were not optimized to detect more subtle changes in guilt detection between groups. For example, there were non-significant trends towards differences between groups in several other comparisons (e.g., alibi vs. innocent groups in Experiment. 1) that could have been informative if we had increased the statistical power of the design. Likewise, these other group comparisons sometimes produced inconclusive Bayes Factors that were not clearly supportive of the alternative nor the null hypothesis, which indicates that the sample sizes were too small to discriminate between these competing hypotheses using Bayesian analyses (see Lakens et al., 2018 for discussion). This limitation should be addressed by employing larger sample sizes in future research to better understand variations in truth detection of autobiographical events with the aIAT.

To conclude, we show that imagining a false alibi impaired memory detection with the aIAT since it was unable to distinguish between a true mock crime and a false alibi. This finding raises serious concerns for potential real life applications of this test as a forensic tool with lying, uncooperative suspects. In real life, guilty suspects may spontaneously fabricate false alibis, and investigators may want to use the aIAT to compare the truth value of a suspect's alibi with the crime they are accused of. Our results suggest that such real life applications may be unsuccessful due to suspects inadvertently modifying their memories by fabricating a false alibi. Furthermore, memories of unethical behaviour such as crimes may be particularly susceptible to modification because forgetting immoral acts allow people to maintain a positive self-concept (Kouchaki & Gino, 2016; although see Stanley, Wang & De Brigard, 2018). Thus, guilty suspects may have several strong motivations to change their memories for self-serving reasons, which in turn may enable them to appear innocent on forensic memory detection tests.

Supplementary material

See supplemental file.

References

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*(2), 299-306. doi:10.1002/acp.1691

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00519

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The Identification of Concealed Memories Using the Event-Related Potential and Implicit Behavioral Measures: A Methodology for Prediction in the Face of Individual Differences. *Psychophysiology, 29(5),* 504-522. doi: 10.1111/j.1469-8986.1992.tb02024.x

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410,* 366–369. doi:10.1038/35066572

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, *18*(6), 279-292. doi: 10.1016/j.tics.2014.03.002

Anderson, M. C., & Levy, B. J. (2007). Theoretical issues in inhibition: Insights from research on human memory. In D. Gorfein & C. Mac Leod (Eds.), *Inhibition in cognition* (pp. 81-02). Washington, DC: American Psychological Association.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review*, *7*(3), 522-530. doi:10.3758/BF03214366

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063. doi:10.1037/0278-7393.20.5.1063

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memoryretrieval. In E. L. Bjork, & R. A. Bjork (Eds.), Memory: *Handbook of perception and cognition* (2nd ed, pp. 237–313). San Diego, CA, US: Academic Press.

Benoit, R.G., & Anderson, M.C. (2012). Opposing Mechanisms Support the Voluntary Forgetting of Unwanted Memories. *Neuron*, *76*, 450-460. https://doi.org/10.1016/j.neuron.2012.07.025

Ben-Shakar, G. (2011). Countermeasures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer. *Memory detection: Theory and application of the Concealed Information Test* (pp. 200-214). Cambridge, UK: Cambridge University Press.

Bergström, Z. M., Anderson, M. C., Buda, M., Simons, J. S., & Richardson-Klavehn, A. (2013). Intentional retrieval suppression can conceal guilty knowledge in ERP memory detection tests. *Biological Psychology*, *94*(1), 1-11. doi: 10.1016/j.biopsycho.2013.04.012.

Bergström, Z. M., De Fockert, J., & Richardson-Klavehn, A. (2009). ERP and behavioural evidence for direct suppression of unwanted memories. *Neuroimage, 21*, 1280–1301.

Dudai, Y. (2012). The restless engram: consolidations never end. *Annual Review of Neuroscience*, *35*, 227–247. doi: 10.1146/annurev-neuro-062111-150500

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177.

Foerster, A., Wirth, R., Herbort, O., Kunde, W., & Pfister, R. (2017). Lying upside-down: Alibis reverse cognitive burdens of dishonesty. *Journal of Experimental Psychology: Applied*, *23*(3), 301. doi: 10.1037/xap0000129

Gamer, M. (2011). Detecting concealed information using autonomic measures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer. *Memory detection: Theory and application of the Concealed Information Test* (pp. 27–45). Cambridge, UK: Cambridge University Press.

Gamer, M., Klimecki, O., Bauermann, T., Stoeter, P., & Vossel, G. (2012). fMRI-activation patterns in the detection of concealed information rely on memory-related effects. *Social Cognitive and Affective Neuroscience*, *7*(5), 506–515. doi: 10.1093/scan/nsp005

Granhag, P. A., Vrij, A., & Verschuere, B. (2015). *Detecting deception: Current challenges and cognitive approaches*. West Sussex, UK: John Wiley & Sons.

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*(2), 197. doi:10.1037/0022-3514.85.2.197

Gronau, N., Elber, L., Satran, S., Breska, A., & Ben-Shakhar, G. (2015). Retroactive memory interference: A potential countermeasure technique against psychophysiological knowledge detection methods. *Biological Psychology, 106*, 68-78. doi:10.1016/j.biopsycho.2015.02.002

Hu, X., Bergström, Z. M., Bodenhausen, G. V., & Rosenfeld, J. P. (2015). Suppressing unwanted autobiographical memories reduces their automatic influences evidence from electrophysiology and an implicit autobiographical memory test. *Psychological Science*, *26*, 1098-1106. doi: 10.1177/0956797615575734

Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating automatic autobiographical associations: the effect of instruction and training in strategically concealing information in the Autobiographical Implicit Association Test. *Psychological Science*, *23*(10), 1079-1085. doi: 10.1177/0956797612443834

Jacoby, L.L. (1999). Ironic effects of repetition: measuring age-related differences in memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25,* 3-22.

JASP Team (2019). JASP (Version 0.10.2). [Computer software]

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emotion Review, 1,* 99-113.

Kouchaki, M., & Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, *113*(22), 6166-6171. doi: 10.1073/pnas.1523586113

Lakens, D., Mclatchie, N., Isager, P.M., Roekel, E.V., & Dienes, Z. (2018). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B.* doi: 10.1093/geronb/gby065

Loftus, E. F. (2003). Make-Believe Memories. *American Psychologist*, *58*(11), 867-873. doi:10.1037/0003-066X.58.11.867

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric annals, 25*(12), 720-725. doi: 10.3928/0048-5713-19951201-07

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385. doi: 10.1037/h0046060

Mangiulli, I., Lanciano, T., Curci, A., Jelicic, M., van Oorsouw, K., Battista, F., & Curci, A. (2018). Can implicit measures detect source information in crime-related amnesia. *Memory, 26,* 1019-1029. DOI: 10.1080/09658211.2018.144142 1

Marini, M., Agosta, S., Mazzoni, G., Dalla Barba, G., & Sartori, G. (2012). True and false DRM memories: Differences detected with an implicit task. *Frontiers in Psychology, 3.* doi:10.3389/fpsyg.2012.00310

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: what have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin, 135*(4), 638–677. doi: 10.1037/a0015849

Otgaar H., & Baker, A. (2018). When lying changes memory for the truth, *Memory, 26,* 2-14, DOI: 10.1080/09658211.2017.1340286

Petrocelli, J. V., & Crysel, L. C. (2009). Counterfactual thinking and confidence in blackjack: A test of the counterfactual inflation hypothesis. *Journal of Experimental Social Psychology, 45,* 1312-1315. https://doi.org/10.1016/j.jesp.2009.08.004

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J. H. (1991). An ERP-Based, Control-Question Lie Detector Analog: Algorithms for Discriminating Effects Within Individuals' Average Waveforms. *Psychophysiology, 28(3),* 319-335. doi: 10.1111/j.1469-8986.1991.tb02202.x

Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test (IAT): Dissociating salience from associations. *Journal of Experimental Psychology: General, 133,* 139–165

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772-780. doi:10.1111/j.1467-9280.2008.02156.x

Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467-474. doi: 10.1016/j.tics.2011.08.004

Shidlovski, D., Schul, Y., & Mayo, R. (2014). If I imagine it, then it happened: The implicit truth value of imaginary representations. *Cognition*, *133*(3), 517-529.

Suchotzki, K., Berlijn, A., Donath, M., & Gamer, M. (2018). Testing the applied potential of the Sheffield Lie Test. *Acta Psychologica, 191,* 281–288. doi:10.1016/j.actpsy.2018.10.011

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143(4),* 428–453. doi:10.1037/bul0000087

Sirgiovanni. E., Corbellini, C., & Caporale, C. (2016). A recap on Italian neurolaw: epistemological and ethical issues. *Mind and Society.* doi:10.1007/s11299-016-0188-1

Stanley, M.L., Yang, B.W., & De Brigard, F., (2018). No evidence for unethical amnesia for imagined actions: A failed replication and extension. *Memory & Cognition. 46,* 787-795.

Takarangi, M. K., Strange, D., & Houghton, E. (2015). Event familiarity influences memory detection using the aIAT. *Memory*, *23*(3), 453-461. doi:10.1080/09658211.2014.902467

Takarangi, M. K., Strange, D., Shortland, A. E., & James, H. E. (2013). Source confusion influences the effectiveness of the autobiographical IAT. *Psychonomic Bulletin & Review*, *20*(6), 1232-1238. doi:10.3758/s13423-013-0430-3

Van Hooff, J. C., Brunia, C. H., & Allen, J. J. (1996). Event-related potentials as indirect measures of recognition memory. *International Journal of Psychophysiology*, *21*(1), 15-31. doi: 10.1016/0167-8760(95)00043-7

Vargo, E.J., Petróczi, A., Shah, I., and Naughton, D.P. (2014). It is not just memory: Propositional thinking influences performance on the autobiographical implicit association test. *Drug and Alcohol Dependency, 145,* 150–155.

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (Eds.). (2011). *Memory detection: Theory and application of the Concealed Information Test*. Cambridge, UK: Cambridge University Press.

Verschuere, B., & Houwer, J. D. (2011). Detecting concealed information in less than a second: response latency-based measures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer. *Memory detection: Theory and application of the Concealed Information Test* (pp. 46-62). Cambridge, UK: Cambridge University Press.

Verschuere, B., Prati, V., & Houwer, J. D. (2009). Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science*, *20*(4), 410-413. doi:10.1111/j.1467-9280.2009.02308.x

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, 100, 426–432.

Supplementary information

Reaction times and Accuracy

As reported in the main paper, the key measure of aIAT performance is the D-score, which combines RT and accuracy into a standardized summary score**.** However, we also analysed these measures separately to gain further insight into exactly how the Alibi manipulation affected performance.

Experiment 1 RT and Accuracy

For RT (Fig. S1B), a 3 (Group) x 2 (Block) mixed ANOVA showed a significant interaction between Group and Block ($F(2, 105) = 5.46$, $p = .006$, $\eta_p^2 = 0.09$). However, there was no main effect of neither Block ($F(1, 105) = 0.01$, $p = .932$, $\eta_p^2 < 0.001$), nor Group ($F(2, 105) = 1.47$, $p = .234$, $\eta_p^2 = .03$). Follow-up paired t-tests showed no significant RT difference between guilt congruent and guilt incongruent blocks in the Guilty-Alibi group ($t(35) = 0.47$, $p = .639$, $d = 0.08$). The Innocent group had significant slower RTs in the guilt congruent than the guilt incongruent block ($t(35) = 2.13$, $p = .040$, $d = 0.40$), whereas the Guilty-Standard group showed the opposite pattern ($t(35) = 2.27$, $p = .029$, $d = 0.47$).

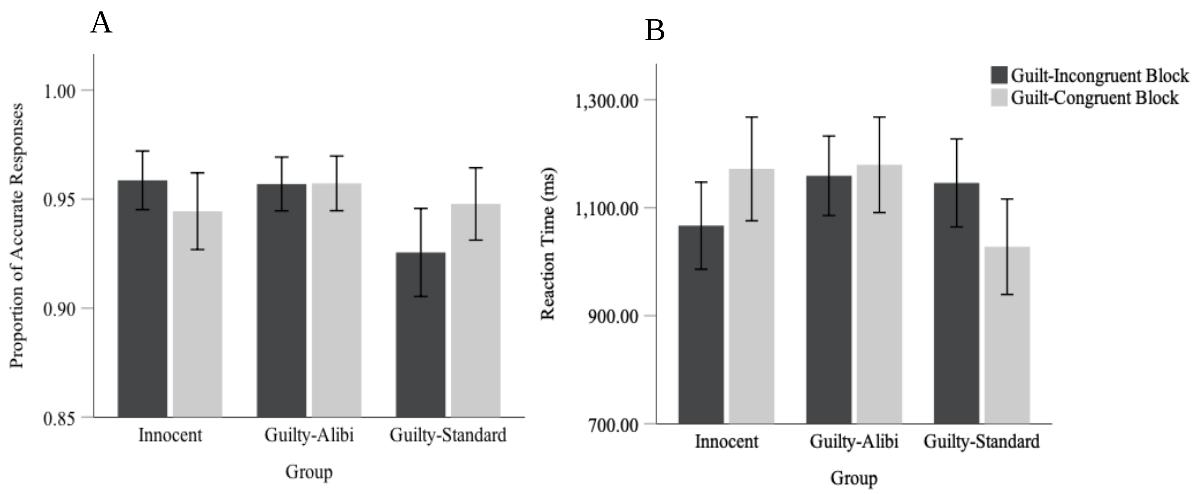

Figure S1. *Proportion accurate responses (A) and mean reaction time (B) from the Guilty-Incongruent (True+Email/False+Ring) and Guilty-Congruent (True+Ring/False+Email) blocks of the Mock Crime/Innocent event aIAT in Experiment 1. Error bars denote 95% confidence intervals.*

Comparing the groups directly within each block separately revealed that the Guilty-Standard group responded significantly faster than the Innocent group in the guilt congruent block ($t(70) = 2.24$, $p = .028$, $d = 0.53$) as predicted. The Guilty-Standard group also responded significantly faster than the Guilty-Alibi group in the guilt congruent block ($t(70) = 2.46$, $p = .016$, $d = 0.58$). However, there was no reaction time difference between Innocent and Guilty-Alibi group ($t(70) = 0.12$, $p = .905$, $d = 0.03$). There were no significant RT differences between the groups during the guilt incongruent block (Innocent vs. Guilty-Alibi: $t(70) = 1.72$, $p = .09$, $d = 0.43$; Innocent vs. Guilty-Standard: $t(70) = 1.40$, $p = .16$, $d = 0.33$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 0.25$, $p = .81$, $d = 0.06$).

For accuracy (Fig S1A), a 3 (Group) x 2 (Block) mixed ANOVA showed a significant interaction between Group and Block ($F(2, 105) = 3.65$, $p = .029$, $\eta_p^2 = 0.07$). However, there was no main effect of neither Block ($F(1, 105) = 0.252$, $p = .617$, $\eta_p^2 < 0.001$) nor Group ($F(2, 105) = 3.02$, $p = .053$, $\eta_p^2 = 0.05$). Paired t-tests revealed no significant difference in accuracy between guilt congruent and guilt incongruent blocks in the Innocent group ($t(35) = 1.38$, $p = .176$, $d = 0.30$), and Guilty-Alibi group ($t(35) = 0.04$, $p = .971$, $d = 0.01$). However, the Guilty-Standard group were more accurate in the guilt congruent block than guilt incongruent block ($t(35) = 2.09$, $p = .044$, $d = 0.41$).

Comparing the groups directly within each block separately revealed that the Innocent group was significantly more accurate than the Guilty-Standard group in the guilt incongruent block, ($t(70) = 2.77$, $p = .007$, $d = 0.67$). The Guilty-Alibi group was also significantly more accurate than the Guilty-Standard group in the guilty-incongruent block ($t(70) = 2.69$, $p =.009$, $d = 0.65$). However, there was no difference between Innocent and Guilty-Alibi groups in the guilt incongruent block ($t(70) = 0.19$, $p = .853$, $d = 0.04$). There were no significant Accuracy differences between groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70) = 1.20$, $p = .23$, $d = 0.28$; Innocent vs. Guilty-Standard: $t(70) = 0.28$, $p = .780$, $d = 0.07$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 0.92$, $p = .360$, $d = 0.27$).

Thus, this analysis showed that raw reaction times and accuracy on the critical guilt congruent and incongruent blocks only distinguished between the Guilty-Standard and the other two groups, whereas there were no significant differences between the Guilty-Alibi and Innocent groups on either measure in either block. Therefore, Guilty-Alibi participants managed to appear innocent also when analysing raw RTs and Accuracy separately in Experiment 1.

Experiment 2 RT and Accuracy

For RT (Fig. S2B), a 3 (group: Innocent vs. Guilty-Standard vs. Guilty-Alibi; between subjects) x 2 (block: congruent vs. incongruent; within subjects) mixed ANOVA showed a significant main effect of Block ($F(1, 105) = 18.30$, $p < .001$, $\eta_p^2 = 0.15$) and interaction between Group and Block ($F(2, 105) = 6.98$, $p = .001$, $\eta_p^2 = 0.12$). However, there was no main effect of Group ($F(2, 105) = 0.82$, $p = .443$, $\eta_p^2 = 0.02$). Follow-up paired t-tests showed significantly faster RTs in the guilt congruent than incongruent blocks for both Guilty-Alibi ($t(35) = 2.48$, $p = .018$, $d = 0.38$) and Guilty-Standard groups ($t(35) = 4.76$, $p < .001$, $d = 0.70$), but no significant RT differences between blocks in the Innocent group ($t(35) = 0.39$, $p = .699$, $d = 0.05$).

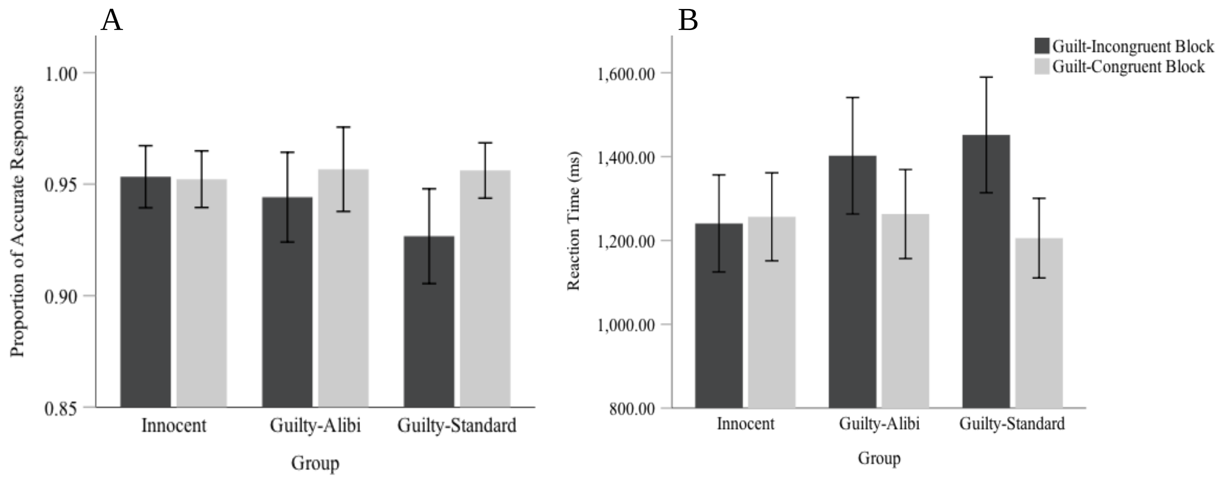

Figure S2. *Mean response times (A) and proportion accurate responses (B) from the guilt-incongruent (True+Exam/False+Ring) and guilt-congruent (True+Ring/False+Exam) blocks of the Mock Crime/Unexperienced event aIAT in Experiment 2. Error bars denote 95% confidence intervals.*

Comparing the groups directly within each block separately revealed that the Guilty-Standard group responded significant slower than the Innocent group in the guilt incongruent block ($t(70) = 2.49$, $p = .015$, $d = 0.60$), and there was a trend in the same direction for the Guilty-Alibi group compared to the Innocent group ($t(70) = 1.81$, $p = .074$, $d = 0.43$), but no significant RT differences between Guilty-Alibi vs. Guilty-Standard groups in the guilt incongruent block($t(70) = 0.60$, $p = .552$, $d = 0.15$). There were no significant RT differences between any groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70) = 0.090$, $p = .929$, $d = 0.02$; Innocent vs. Guilty-

Standard: $t(70) = 0.73$, $p = .468$, $d = 0.17$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 0.82$, $p = .415$, $d = 0.20$).

For accuracy (Fig. S2A), a 3 (group: Innocent vs. Guilty-Standard vs. Guilty-Alibi; between subjects) x 2 (block: congruent vs. incongruent; within subjects) mixed ANOVA showed a significant main effect of Block ($F(1, 105) = 5.50$, $p = .021$, $\eta_p^2 = .05$). However, there was no main effect of Group ($F(2, 105) = .812$, $p = .447$, $\eta_p^2 = .02$) and the interaction was at trend-level ($F(2, 105) = 2.32$, $p = .104$, $\eta_p^2 = .042$). Paired t-tests revealed no significant difference in accuracy between guilt congruent and guilt incongruent blocks in the Innocent group ($t(35) = 0.14$, $p = .890$, $d = 0.07$), nor the Guilty-Alibi group ($t(35) = 1.27$, $p = .211$, $d = 0.22$). However, the Guilty-Standard group were more accurate in the guilt congruent block than guilt incongruent block ($t(35) = 2.46$, $p = .019$, $d = 0.57$).

Comparing the groups directly within each block separately revealed that the Guilty-Standard group was significantly less accurate than the Innocent group in the guilty incongruent block, but there were no other group differences in that block (Innocent group vs. Guilty-Standard group: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Guilty-Alibi vs. Guilty-Standard: $t(70) = 1.21$, $p = .229$, $d = 0.29$; Innocent vs. Guilty-Alibi: $t(70) = 0.76$, $p = .450$, $d = 0.18$). There were also no significant accuracy differences between groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70) = .396$, $p = .693$, $d = 0.09$; Innocent vs. Guilty-Standard: $t(70) = 0.60$, $p = .548$, $d = 0.14$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 0.08$, $p = .936$, $d = 0.02$).

Thus, these results suggest suggests that in Experiment 2, manipulation effects on accuracy were rather limited and the main D-score findings were mostly driven by group differences in speed at responding during the guilt incongruent block where the guilty groups were slower than the innocent group, presumably due to increased response conflict.

Experiment 3 RT and Accuracy

Mock Crime/Innocent event aIAT

For RT (Fig. S3B), a 4 (Groups: Innocent, Guilty-Standard, Guilty-Alibi, and Guilty-Alibi with HT; between group) x 2 (Block: congruent and incongruent; within subject) mixed ANOVA showed that there were no main effect of neither group (F(3,140) = 1.587, p = .195, η_p^2 = .033) nor block (F(1, 140) = 0.031, p = .861, η_p^2 < .001). However, there was a significant group x block interaction (F(3, 140) = 5.91, p = .001, η_p^2 = .112).

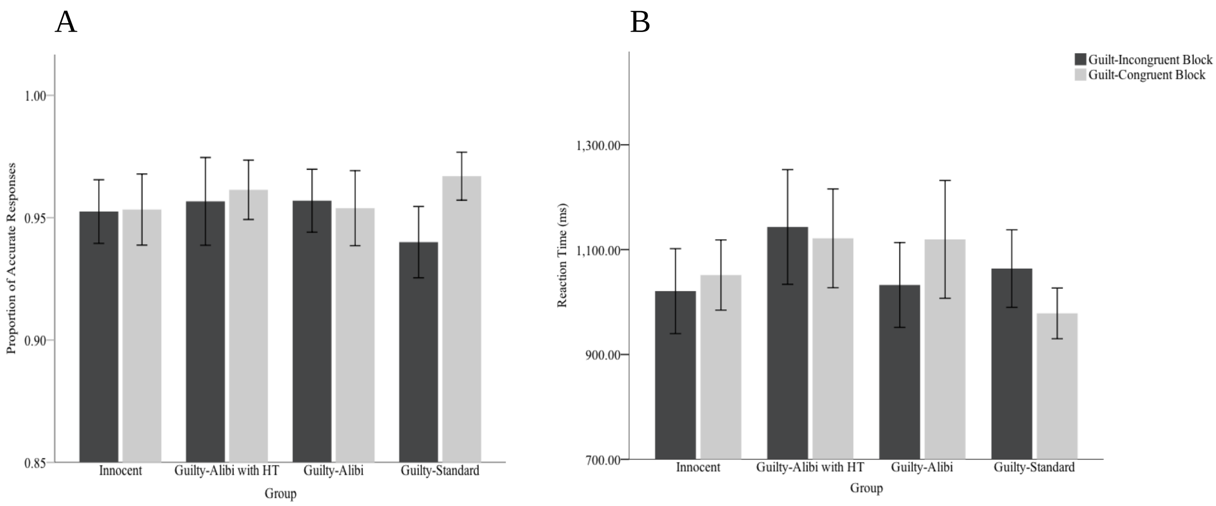

Figure S3. *Proportion accurate responses and mean response times for guilt-incongruent (True+Email/False+Ring) and guilt-congruent (True+Ring/False+Email) blocks of the Mock Crime/Innocent event aIAT in Experiment 3. Error bars denote 95% confidence intervals.*

Follow-up paired t-tests showed significantly faster RTs in the guilt congruent than the guilt incongruent block for the Guilty-Standard group (t(35) = 3.06, p = .004, d = 0.46), whereas the Guilty-Alibi group showed the reverse pattern with significant faster RTs in the guilt-incongruent compared to the guilt-congruent block (t(35) = 2.52, p = .016, d = 0.30). There were no RT differences between the two blocks in Innocent (t(35) = 1.22, p = .231, d = 0.14), and Guilty-Alibi with HT groups ($t(35) = 0.67$, $p = .510$, $d = 0.11$). Independent t-tests (Table S1) were conducted to compare the groups within each block. The results showed that the Guilty-Standard group was significantly faster in the guilt congruent block than the Guilty-Alibi and Guilty-Alibi with HT groups, whilst the other comparisons were not significant. There were also no significant RT differences between any groups for the guilt incongruent block.

Table S1. Independent t-test results comparing performance across groups during the Mock Crime/Innocent event aIAT in Experiment 3.

	RT						ACC					
	Guilt-congruent			Guilt-incongruent			Guilt-congruent			Guilt-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent vs. Guilty-Standard	1.80	.08	0.42	0.80	.43	0.19	1.58	.12	0.37	1.30	.20	0.31
Innocent vs. Guilty-Alibi	1.06	.30	0.25	0.21	.84	0.05	0.05	.96	0.01	0.49	.62	0.12
Innocent vs. Guilty-Alibi with HT	1.23	.22	0.29	1.83	.07	0.43	0.86	.39	0.20	0.38	.70	0.09
Guilty-Standard vs. Guilty- Alibi	**2.34**	**.02**	**0.55**	0.58	.56	0.14	1.46	.15	0.34	1.77	.08	0.42
Guilty-Standard vs. Guilty-Alibi with HT	**2.75**	**.01**	**0.65**	1.22	.23	0.29	0.72	.47	0.17	1.46	.15	0.35
Guilty-Alibi vs. Guilty-Alibi with HT	0.03	.98	0.01	1.65	.10	0.39	0.78	.44	0.18	0.03	.98	<0.001

Note: Significant differences are marked in bold. $Df = 70$. All t-values and d-values are absolute.

For accuracy (Fig. S3A), a 4 (Groups: Innocent, Guilty-Standard, Guilty-Alibi, and Guilty-Alibi with HT; between groups) x 2 (Block: guilt congruent and guilt incongruent; within subject) mixed ANOVA revealed a significant main effect of block ($F(1, 140) = 4.15$, $p = .044$, $\eta_p^2 = .029$) and interaction between group and block ($F(3, 140) = 3.46$, $p = .018$, $\eta_p^2 = .069$). However, there was no main effect of group ($F(3, 140) = 0.22$, $p = .880$, $\eta_p^2 = .005$).

Follow-up paired t-tests indicated that Guilty-Standard group was more accurate in the guilt congruent block compared to the guilt incongruent block ($t(35) = 3.78$, $p = .001$, $d = 0.73$), whist there were no differences between blocks in Innocent ($t(35) = 0.108$, $p = .915$, $d = 0.02$), Guilty-Alibi ($t(35) = 0.38$, $p = .704$, $d = 0.07$), or Guilty-Alibi with HT group ($t(35) = 0.804$, $p = .427$, $d = 0.10$). There were no significant accuracy differences between groups in either guilt congruent or guilt incongruent blocks (see Table S1). Thus, similar to previous experiments, the strongest effects of the manipulation were on reaction times rather than accuracy, and the Standard Guilty group showed the expected effects on both measures most clearly (slower RT and lower accuracy in guilt incongruent than congruent blocks).

Mock Crime/Unexperienced event aIAT

For RTs (Fig. S4B), a 4 (group: innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed significant main effect of block ($F(1, 140) = 13.70$, $p < .001$, $\eta_p^2 = .089$) and group x block interaction ($F(3, 140) = 3.30$, $p = .022$, $\eta_p^2 = .066$). However, there was no main effect of group ($F(3,140) = 0.54$, $p = .652$, $\eta_p^2 = .012$).

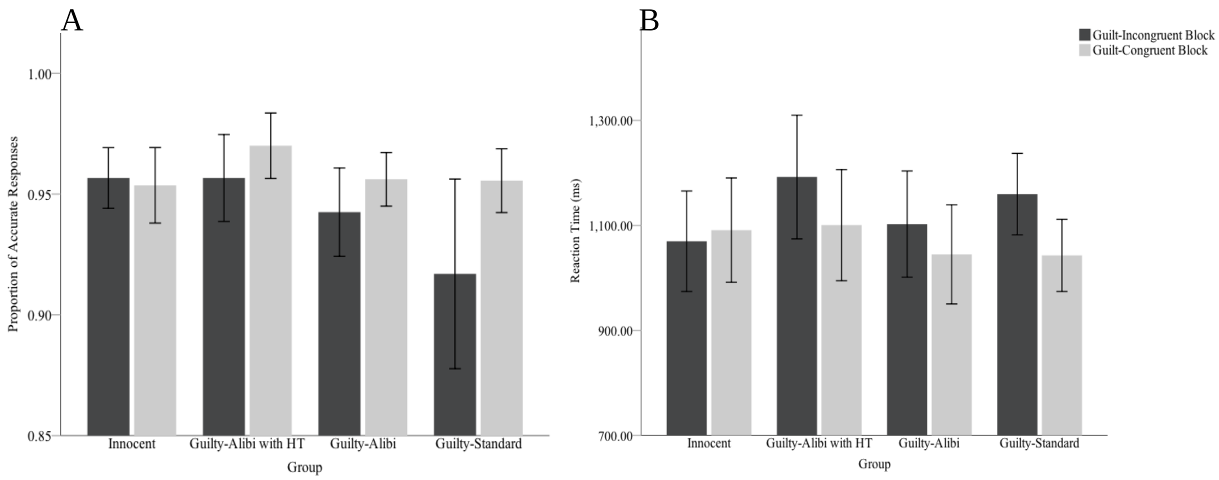

Figure S4. *Proportion accurate responses and mean response times from guilt-incongruent (True+Exam/False+Ring) and guilt-congruent (True+Ring/False+Exam) blocks of the Mock Crime/Unexperienced event aIAT in Experiment 3. Error bars denote 95% confidence intervals.*

Follow up paired t-tests revealed significant faster RTs in the guilt congruent than the incongruent block in Guilty-Standard ($t(35) = 4.09$, $p < .001$, $d = 0.54$) and Guilty-Alibi with HT groups ($t(35) = 2.53$, $p = .016$, $d = 0.28$). However, there were no significant RT differences between blocks in Innocent ($t(35) = 0.635$, $p = .530$, $d = 0.07$) and Guilty-Alibi group ($t(35) = 1.71$, $p = .096$, $d = 0.20$). When comparing the groups directly, results showed no significant differences between groups in neither congruent nor incongruent blocks (Table S2).

Table S2. Independent t-test results comparing performance across groups during the Mock Crime/Unexperienced event aIAT in Experiment 3.

	RT						ACC					
	Guilt-congruent			Guilt-incongruent			Guilt-congruent			Guilt-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent vs. Guilty-Standard	0.81	.42	0.19	1.48	.14	0.35	0.19	.85	0.05	1.96	.054	0.47
Innocent vs. Guilty-Alibi	0.68	.50	0.16	0.48	.64	0.11	0.27	.79	0.06	1.30	.20	0.31
Innocent vs. Guilty-Alibi with HT	0.14	.89	0.03	1.64	.11	0.39	1.61	.11	0.38	<0.001	.99	<0.001
Guilty-Standard vs. Guilty- Alibi	0.04	.97	0.01	0.91	.37	0.22	0.07	.95	0.02	1.20	.24	0.29
Guilty-Standard vs. Guilty-Alibi with HT	0.93	.36	0.22	0.47	.64	0.11	1.55	.13	0.37	1.87	.07	0.45
Guilty-Alibi vs. Guilty-Alibi with HT	0.80	.43	0.19	1.17	.25	0.28	1.61	.11	0.38	1.12	.27	0.27

Note: $Df = 70$. All t-values and d-values are absolute.

A 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: guilt-congruent, guilt-incongruent; within subject) mixed ANOVA was also conducted to examine accuracy (Fig. S4A). There was a significant main effect of block ($F(1, 140) = 7.52$, $p = .007$, $\eta_p^2 = .051$), but no main effect of group ($F(3, 140) = 2.129$, $p = .099$,

η_p^2 = .044) and no interaction effect (*F*(3, 140) = 2.275, *p* = .083, η_p^2 = 046). Paired t-tests showed significant accuracy differences between blocks in the Guilty-Standard group (*t*(35) = 2.13, *p* = .040, *d* = 0.45), but not in Innocent (*t*(35) = 0.415, *p* = .681, *d* = 0.07), Guilty-Alibi (*t*(35) = 1.66, *p* = .106, *d* = 0.30), nor Guilty-Alibi with HT groups (*t*(35) = 1.61, *p* = .116, *d* = 0.28). Independent t-tests were conducted to investigate accuracy differences between groups in guilt-congruent and guilt-incongruent blocks (Table S2). These showed only a trend towards a difference (*p* = .054) in the guilt-incongruent block when comparing Innocent and Guilty-Standard groups, and no other differences between groups in guilt-congruent nor guilt-incongruent blocks. Thus, as in the Mock Crime/Innocent event aIAT version and the previous experiments, the strongest and most consistent effects on RT and accuracy were in the standard guilty group.

Innocent/Unexperienced event aIAT

For RTs (Figure S5B), a 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed no main effect of block (*F*(1, 140) = 2.47, *p* = .118, η_p^2 = .017), no main effect of group (*F*(3, 140) = 1.00, *p* = .394, η_p^2 = .021), nor a block x group interaction (*F*(3, 140) = 1.62, *p* = .188, η_p^2 = .034). Follow up paired t-tests comparing the blocks within each groups showed no significant differences in RT between innocence-congruent and innocence-incongruent blocks in Innocent (*t*(35) = 0.05, *p* = .960, *d* = 0.01), Guilty-Standard (*t*(35) = 0.64, *p* = .527, *d* = 0.09), Guilty-Alibi (*t*(35) = 1.61, *p* = .116, *d* = 0.19), or Guilty-Alibi with HT group (*t*(35) = 1.83, *p* = .075, *d* = 0.19).

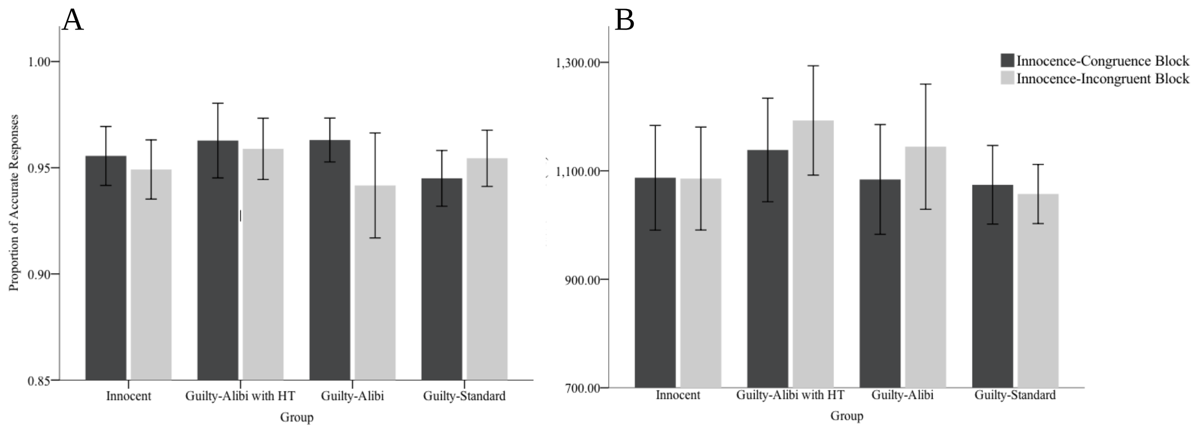

Figure S5. *Proportion accurate responses and mean response times from innocence-incongruent (True+Exam/False+Email) and innocence-congruent (True+Email/False+Exam) blocks of the Innocent/Unexperienced event aIAT in Experiment 3. Error bars denote 95% confidence intervals.*

Independent t-tests also showed no differences between groups across either congruent or incongruent blocks, except RT in the innocence-incongruent block, where the Guilty-Alibi with HT group was significantly slower than the Guilty-Standard group (see Table S3). For accuracy (Fig. S5a), 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with

Table S3. Independent t-test results comparing performance across groups during the Innocent/Unexperienced event aIAT in Experiment 3.

	RT						ACC					
	Innocence-congruent			Innocence-incongruent			Innocence-congruent			Innocence-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent vs. Guilty-Standard	0.22	.83	0.05	0.53	.60	0.13	1.12	.27	0.27	0.56	.58	0.13
Innocent vs. Guilty-Alibi	0.04	.97	0.01	0.80	.43	0.19	0.88	.38	0.21	0.54	.59	0.13
Innocent vs. Guilty-Alibi with HT	0.77	.45	0.18	1.57	.12	0.38	0.66	.52	0.16	0.98	.33	0.24
Guilty-Standard vs. Guilty- Alibi	0.16	.87	0.04	1.39	.17	0.33	**2.20**	**.03**	**0.52**	0.93	.36	0.22
Guilty-Standard vs. Guilty-Alibi with HT	1.09	.28	0.26	**2.40**	**.02**	**0.57**	1.65	.11	0.39	0.46	.65	0.11
Guilty-Alibi vs. Guilty-Alibi with HT	0.79	.43	0.19	0.64	.52	0.15	0.03	.98	0.01	1.22	.23	0.29

Note: Significant differences are marked in bold. *Df* = 70. All t-values and d-values are absolute.

HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed no main effect of block (F(1, 140) = 1.77, p = .185, η_p^2 = .012), no main effect of group (F(3, 140) = 0.55, p = .647, η_p^2 = .012), nor block x group interaction (F(3, 140) = 2.29, p = .081, η_p^2 = .047). When comparing each block within the groups, there was no differences in accuracy between blocks in Innocent (t(35) = 1.00, p = .326, d = 0.16), Guilty-Standard (t(35) = 1.19, p = .241, d = 0.24), Guilty-Alibi (t(35) = 1.61, p = .116, d = 0.19), and Guilty-Alibi with HT group (t(35) = 0.65, p = .519, d = 0.08). When comparing the groups within each block, there were no differences between groups in either innocence-congruent or innocence-incongruent blocks, except that the Guilty-Alibi group was more accurate than the Guilty-Standard group in the innocence-congruent block (see Table S3). Thus, RT and accuracy differences between blocks and groups were very small and mostly non-significant in the email/exam version of the aIAT, consistent with the main D-score analysis.

So, in sum, the RT and Accuracy analysis across experiments converged with the D-score analysis in suggesting that standard guilty participants who did not use a countermeasure could generally be distinguished from the Innocent group. This seemed to be driven both by RT and Accuracy differences. However, the Alibi countermeasure had the effect of making guilty participants appear more similar to innocent participants across measures and tests.

Threshold based individual classification

As previously mentioned, the aIAT was developed to diagnose guilt or innocence at the individual level, which is typically done by classifying individuals with positive D-scores as “guilty” and individuals with negative D-scores as “innocent” when contrasting a guilty vs. innocent event in this way (Sartori et al., 2008). Although such threshold based classification can be problematic (see main paper), in a supplementary analysis we also tested group differences in classification rates this way in order to more directly compare our results to the previous literature. We first excluded participants scoring too close to zero (absolute D-scores between 0-0.2) as inconclusive based on previous recommendations (Agosta & Sartori, 2013), and then classified individuals with scores larger or smaller than this as guilty or innocent, as appropriate.

In Experiment 1, after excluding inconclusive scores (excluded N for Guilty-Standard = 11, Guilty-Alibi =13, Innocent = 10), we classified individuals as guilty (positive scores above 0.2) or innocent (negative scores below -0.2, see Agosta & Sartori, 2013) and found that a significantly larger proportion of the Guilty-Standard Group were classified as guilty than in the Guilty-Alibi and Innocent groups, which were not significantly different from each other (see Table S4 for statistical results for all experiments).

Table S4. Individual classifications as guilty or innocent across the three experiments with a D-score threshold at >0.2.

	Percent of participants detected				Pairwise differences in detection rates: χ^2 (φ)					
	I	*GA*	*GS*	*GA HT*	*I vs. GA*	*I vs. GS*	*GA vs. GS*	*I vs. GAHT*	*GA vs. GAHT*	*GS vs. GAHT*
Experiment 1: Mock Crime/Innocent aIAT guilt detection	31	48	84	-	1.50 (0.18)	14.72*** (0.54)	7.05** (0.38)	-	-	-
Experiment 2: Mock Crime/Unexperienced aIAT guilt detection	50	74	82	-	4.11* (0.30)	7.24** (0.39)	0.53 (0.10)	-	-	-
Experiment 3: Mock Crime/Innocent aIAT guilt detection	32	29	84	64	0.05 (0.04)	11.36*** (0.53)	12.48*** (0.56)	4.46* (0.32)	5.31* (0.35)	2.20 (0.23)
Experiment 3: Mock Crime/Unexperienced aIAT guilt detection	48	59	81	65	0.47 (0.11)	5.08* (0.35)	2.24 (0.24)	1.50 (0.18)	0.19 (0.07)	1.41 (0.17)
Experiment 3: Innocent/Unexperienced aIAT innocence detection	50	74	56	70	2.84 (0.25)	0.15 (0.06)	1.33 (0.18)	1.87 (0.20)	0.11 (0.05)	0.72 (0.14)

Note: *= p<0.05, **= p<0.01, ***= p<0.001; I = Innocent; GA = Guilty-Alibi; GS = Guilty-Standard; GAHT = Guilty-Alibi with Home Training. Percentages were calculated after first excluding inconclusive scores (between 0.02 and -0.02), and for all tests indicate *guilt* detection, except Experiment 3 Innocent/Unexperienced test, where they indicate *innocence* detection.

In Experiment 2, individuals who elicited a positive D-score were classified as "guilty" and these classification rates were compared between groups, after first excluding participants who scored too close to zero. Since this version of the aIAT is designed to elicit scores close to zero for innocent participants, this criterion led to a high number of exclusions for innocent participants (excluded N for Guilty-Standard = 8, Guilty-Alibi = 9, Innocent = 16). Threshold-based guilt classification in the Guilty-Standard Group was similar to Experiment 1 and significantly higher than in the Innocent group. However in contrast to Experiment 1, the Guilty Alibi group also showed higher levels of guilt classification than the Innocent group, and no reduction compared to the Guilty-Standard group.

In Experiment 3, we conducted separate classification analysis for the three different aIAT versions. For the mock crime/innocent act version, we classified individuals with a positive score as "guilty" and individuals with a negative score as "innocent" (Agosta et al., 2013), after first excluding participants who scored too close to zero (excluded N for Innocent = 14, Guilty-Standard = 17, Guilty-Alibi = 15; and Guilty-Alibi with HT = 14). In this threshold-based analysis, guilt classification for the Guilty-Standard group was similar to Experiments 1 and 2, and significantly higher than for the Innocent group (which also showed a similar rate as in Experiment 1). In the Guilty-Alibi with HT group, guilty classification was not significantly different from the Guilty-Standard group, but it was significantly higher than in the Innocent group. Guilt classification for the Guilty-Alibi group was however significantly lower than both Guilty-Standard and Guilty-Alibi with HT groups, but not different from the Innocent group.

For the mock crime/unexperienced event aIAT, we classified individuals as guilty if they scored above 0.2, and compared guilt detection proportions across groups after excluding participants who obtained D-scores too close to zero (excluded N: Innocent = 15, Guilty-Standard = 15, Guilty-Alibi = 19, Guilty-Alibi with HT = 10). Threshold-based guilt detection in the Guilty-Standard group was similarly high as in the previous experiments and the Mock Crime/Innocent aIAT in this Experiment, and again significantly higher than in the Innocent group. Guilt detection in the Guilty-Alibi and Guilty-Alibi with HT groups was lower, but not significantly different from each other or from the detection rates of the other groups.

Finally, in the innocent/unexperienced aIAT version, individuals scoring higher than 0.2 were classified as innocent after excluding participants who scored too close to zero (excluded: Innocent = 11, Guilty-Standard = 20, Guilty-Alibi = 13, and Guilty-Alibi with HT = 13; see Table S4). This analysis showed that around 2/3 of the Guilty-Alibi with HT and Guilty-Alibi groups were erroneously identified as associating the innocent-related sentences more with the truth than the unexperienced sentences, whereas surprisingly, only half of the Innocent group was correctly identified as such. This detection rate was similar to in the Guilty-Standard group, for which a 50% detection rate was predicted since they had no knowledge of either of the events. However, there were no significant differences among the groups in threshold-based classification rates.

To summarise, the results from the threshold-based individual classification analysis supported the conclusions from the main group level and ROC analyses that guilt detection was reduced after participant rehearsed a false alibi, and that in Experiment 3, this countermeasure was most effective when the alibi was learned and imagined once just before the test. However, applying these thresholds to our data was problematic because many participants had to be excluded due to inconclusive D-scores (scores close to zero) rendering the remaining group sizes small. Furthermore, even when excluding scores close to zero, many innocent participants were erroneously classified as guilty, which would be a serious problem in real life applications. We believe these patterns illustrates practical problems with using the aIAT D-score to classify individuals – depending on the design of the aIAT and the characteristics of the suspects, the test may produce inconclusive results rather than unambiguously guilty or innocent classifications, and even unambiguous results may be subject to many errors.

Faking analysis

In a final analysis, we calculated a "faking index" (Agosta et al., 2011) to assess whether rehearsing a false alibi would result in unusual reaction time patterns across aIAT blocks, since such patterns may function as signals of guilt even when the main guilt measure (i.e. D-score) is disrupted by countermeasures. The faking index is based on calculating the ratio between the mean RT in whichever double classification block is fastest for a particular person (which presumably reflects the truth-congruent block for that person) with the mean RT in the corresponding single classification blocks, based on the logic that suspects who are trying to beat the test may be slowing down more in the critical double classification blocks than in the non-critical single classification blocks. Thus, the higher this index, the more that person is slowing down in the critical compared to non-critical blocks. To calculate the index, first all RTs below 150ms and above 10000ms were excluded. Next, any errors were replaced with the average RT of the block plus a penalty of 600ms. Finally, the ratio between the average RT of the fastest block (between 3 or 5) and single tasks that are directly connected to the fastest block in terms of motor response (1 and 2 or 1 and 4, respectively) was calculated (see Agosta et al., 2011, for more information).

Experiment 1 faking results

In Experiment 1, the average faking index was higher in the Guilty-Alibi group (*M*=1.05, *SD*=0.20) than the Guilty-Standard (*M* = 0.97, *SD* = 0.15; $t(70)=2.06$, $p=.043$, $d=0.49$) and Innocent groups (*M* = 0.95, *SD* = 0.15; $t(70) = 2.43$, $p=.02$, $d = 0.58$), who did not differ from each other ($t(70) = 0.40$, $p = .69$, $d = 0.10$). Using a cut-off value of 1.08 on the index (as suggested by Agosta et al., 2011), around 47% of the Guilty-Alibi group but only 19% of the Guilty-Standard group were classified as faking, and these rates were significantly different ($\chi^2(1) = 6.25$, $p = .012$, $\varphi = .30$). Faking classification was also higher in the Guilty-Alibi group than in the Innocent group (25%; $\chi^2(1) = 3.85$, $p = .050$, $\varphi = 0.23$), however classification rates did not differ between Guilty-Standard and Innocent groups ($\chi^2(1) = 0.32$, $p = .570$, $\varphi = 0.07$).

Similar to the D-score analysis, we also conducted a threshold-independent ROC analysis to evaluate faking classification performance. This analysis is appropriate because the most suitable threshold to use for detecting faking may differ across studies. The ROC analyses showed that when comparing Guilty-Alibi and Innocent groups, faking classification was significantly better than chance (*AUC* = .65, *SE* = .07, $p = .027$). When comparing Guilty-Standard and Innocent groups, faking classification was not different from chance (*AUC* = .55, *SE* = .07, $p = .480$). Thus, the faking analysis showed that guilty suspects who rehearsed a false alibi may reveal themselves by unusual reaction time patterns across aIAT blocks, although classification performance based on the faking-index was fairly poor. With only a 65% probability of classifying an individual correctly, this index would not be suitable to apply in practice.

Experiment 2 faking results

In Experiment 2, there were no difference between Innocent (*M* = 1.01, *SD* = .18) and Guilty-Standard groups (*M* = 0.94, *SD* = 0.16; *t*(70) = 1.78, *p* = .080, *d* = 0.42), Innocent and Guilty-Alibi groups (*M* = 0.97, *SD* = 0.16; *t*(70) = 1.00, *p* = .320, *d* = 0.24), nor Guilty-Standard and Guilty-Alibi groups (*t*(70) = 0.80, *p* = .427, *d* = 0.19) in the average faking index. Using the 1.08 cut-off as suggested by Agosta and colleagues (2011), 25% of the Guilty-Alibi group and 17% of the Guilty-Standard group were classified as faking, which was not significantly different ($\chi^2(1) = 0.76$, $p = .384$, $\varphi = 0.103$). There was also no difference between Innocent (31%) and Guilty-Alibi groups ($\chi^2(1) = 0.28$, $p = .599$, $\varphi = 0.06$) nor between Innocent and Guilty-Standard groups ($\chi^2(1) = 1.93$, $p = .165$, $\varphi = 0.16$) in faking classification proportions.

The ROC analysis showed that faking classification was not different from chance when comparing Innocent and Guilty-Alibi groups (AUC = .56, *SE* = .07, *p* = .368), nor when comparing Innocent and Guilty-Standard groups (AUC = .60, *SE* = .07, *p* = .128), nor when comparing Guilty-Alibi groups and Guilty-Standard groups (AUC = .55, *SE* = .07, *p* = 454). Thus, the faking analysis in Experiment 2 showed that rehearsing an alibi did not cause any unusual reaction time patterns across aIAT blocks when the aIAT contrasted the mock crime with an unexperienced event, because faking classification was relatively low and similar across all groups.

Experiment 3 faking results

Mock Crime/Innocent event aIAT

In the aIAT version that contrasted the mock crime and alibi/innocent act directly in Experiment 3, there were no significant differences between Innocent (*M* = 1.14, *SD* = 0.16) and Guilty-Standard groups (*M* = 1.10, *SD* = 0.19; *t*(70) = 0.93, *p* = .354, *d* = 0.22), Innocent and Guilty-Alibi groups (*M* = 1.15, *SD* = 0.19; $t(70) = 0.41$, $p = .683$, $d = 0.10$), Innocent and Guilty-Alibi with HT groups (*M* = 1.21, *SD* = 0.20; $t(70) = 1.57$, $p = .121$, $d = 0.37$), Guilty-Standard and Guilty-Alibi groups ($t(70) = 1.25$, $p = .215$, $d = 0.30$), or Guilty-Alibi and Guilty-Alibi with HT groups ($t(70) = 1.11$, $p = .272$, $d = 0.26$) in the average faking index. However, the average faking index was lower in the Guilty-Standard than Guilty-Alibi with HT group ($t(70) = 2.31$, $p = .024$, $d = 0.54$). Using the 1.08 cut-off as suggested by Agosta and colleagues (2011), 53% of Guilty-Standard, 67% of Guilty-Alibi, 69% of Innocent and 72% of Guilty-Alibi with HT group were classified as faking in the ring/email classification aIAT. These classification rates were not different (Guilty-Standard vs. Guilty-Alibi group: $\chi^2(1) = 1.44$, $p = .230$, $\varphi = 0.142$; Guilty-Standard vs. Guilty-Alibi with HT group: $\chi^2(1) = 2.90$, $p = .088$, $\varphi = 0.201$; Guilty-Alibi vs. Guilty-Alibi with HT: $\chi^2(1) = 0.262$, $p = .61$, $\varphi = .060$; Innocent vs. Guilty-Standard group: $\chi^2(1) = 2.10$, $p = .147$, $\varphi = 0.171$; Innocent vs. Guilty-Alibi group: $\chi^2(1) = 0.064$, $p = .800$, $\varphi = .030$; Innocent vs. Guilty-Alibi with HT group: $\chi^2(1) = .067$, $p = .795$, $\varphi = 0.031$)

The ROC analyses showed that the classification was not different from chance when comparing Innocent with Guilty-Standard group (*AUC* = .54, *SE* = .07, $p = .612$), when comparing Innocent with Guilty-Alibi group (*AUC* = .56, *SE* = .07, $p = .386$), or when comparing Innocent with Guilty-Alibi with HT group (*AUC* = .60, *SE* = .068, $p = .128$). There were also no differences in classification performance between Guilty-Standard and Guilty-Alibi group (*AUC* = .57, *SE* = .068, $p = .290$) or between Guilty-Alibi and Guilty-Alibi with HT (*AUC* = .56, *SE* = .07, $p = .356$). However, the classification performance was just significantly better than chance when comparing Guilty-Standard with Guilty-Alibi with HT group (*AUC* = .63, *SE* = .065, $p = .050$). Thus, faking analyses showed that when the aIAT contrasted the mock crime to the innocent/alibi event, rehearsing an alibi repeatedly over a week may cause unusual response patterns in the aIAT blocks, but this effect was rather weak and only significant when compared to a guilty standard group, and not compared to the other groups.

Mock Crime/Unexperienced event aIAT

In the mock crime vs. unexperienced event aIAT version, there were no differences between Innocent (*M* = 1.12, *SE* = .21) and Guilty-Standard (*M* = 1.14, *SE* = 0.19; *t*(70) = 0.475, *p* = .636, *d* = 0.11), Innocent and Guilty-Alibi (*M* = 1.18, *SE* = 0.22; *t*(70) = 1.14, *p* = .259, *d* = 0.27), or Innocent and Guilty-Alibi with HT groups (*M* = 1.13, *SE* = 0.18; *t*(70) = 0.22, *p* = .827, *d* = 0.05) in the average faking index. There were also no differences in faking index between Guilty-Standard and Guilty-Alibi (*t*(70) = 0.716, *p* = .476, *d* = 0.17), Guilty-Standard and Guilty-Alibi with HT (*t*(70) = 0.279, *p* = .781, *d* = 0.07), or Guilty-Alibi and Guilty-Alibi with HT groups (*t*(70) = 0.992, *p* = .325, *d* = 0.23).

Using the 1.08 classification cut-off as suggested by Agosta et al. (2011), 56% of Guilty-Standard group and 64% of Guilty-Alibi group were classified as faking and this was not significantly different (χ^2(1) = 0.52, *p* = .471, φ = 0.085). There was also no difference between Guilty-Standard and Guilty-Alibi with HT (72%; χ^2(1) = 2.17, *p* = .141, φ = 0.173), and Guilty-Standard and Innocent group (also 56%, so both groups were the same). There was also no significant difference when comparing Innocent to Guilty-Alibi group (χ^2(1) = 0.52, *p* = .471, φ = 0.085), Innocent to Guilty-Alibi with HT group (χ^2(1) = 2.17, *p* = .141, φ = 0.173), and Guilty-Alibi and Guilty-Alibi with HT group (χ^2(1) = 0.58, *p* = .448, φ = 0.089) in faking classification at this threshold.

Threshold independent ROC analysis showed that faking classification was not different from chance when comparing Innocent and Guilty-Alibi group (*AUC* = .57, *SE* = .068, *p* = .280), Innocent and Guilty-Alibi with HT (*AUC* = .54, *SE* = .069, *p* = .551), Innocent and Guilty-Standard (*AUC* = .53, *SE* = .069, *p* = .693) and Guilty-Alibi and Guilty-

Alibi with HT (*AUC* = .55, *SE* = .07, *p* = .471). When compared to Guilty-Standard group, the classification of Guilty-Alibi group (*AUC* = .56, *SE* = .068, *p* = .375) and Guilty-Alibi with HT (*AUC* = .52, *SE* = .069, *p* = .787) as fakers was also at chance. Thus, according to the faking index all of the groups showed equal amounts of unusual slowing in double classification blocks in this aIAT version.

Innocent/Unexperienced event aIAT

In the innocent vs. unexperienced event version of the aIAT, there was no difference between Innocent (*M* = 1.15, *SD* = .14) and Guilty-Standard (*M* = 1.15, *SD* = 0.16; *t*(70) = .03, *p* = .979, *d* = 0.01), Innocent and Guilty-Alibi (*M* = 1.18, *SD* = 0.17; *t*(70) = 0.66, *p* = .509, *d* = 0.16), and Innocent and Guilty-Alibi with HT (*M* = 1.11, *SD* = 0.19; *t*(70) = 1.11, *p* = .270, *d* = 0.26) in mean faking score. There were also no differences in mean faking score between Guilty-Standard and Guilty-Alibi (*t*(70) = 0.66, *p* = .513, *d* = 0.15), Guilty-Standard and Guilty-Alibi with HT (*t*(70) = 1.04, *p* = .301, *d* = 0.25), and Guilty-Alibi and Guilty-Alibi with HT (*t*(70) = 1.60, *p* = .114, *d* = 0.38).

Using the 1.08 cut-off (Agosta et al., 2011), 64% of Innocent group and 67% of Guilty-Standard group were classified as faking and these rates were not significantly different ($\chi^2(1) = 0.06$, $p = .804$, $\varphi = 0.03$), neither were Innocent and Guilty-Alibi groups (83%; $\chi^2(1) = 3.50$, $p = .061$, $\varphi = 0.22$), nor were Innocent and Guilty-Alibi with HT group (64% also). There were also no differences in faking classification between Guilty-Standard and Guilty-Alibi group ($\chi^2(1) = 2.67$, $p = .102$, $\varphi = 0.19$), Guilty-Standard and Guilty-Alibi with HT ($\chi^2(1) = 0.06$, $p = .804$, $\varphi = 0.03$), and Guilty-Alibi and Guilty-Alibi with HT ($\chi^2(1) = 3.50$, $p = .061$, $\varphi = 0.22$).

The ROC analyses showed that faking classification was not different from chance when comparing Innocent and Guilty-Standard group (*AUC* = .53, *SE* = .069, *p* = .719), Innocent and Guilty-Alibi group (*AUC* = .54, *SE* = .069, *p* = .547), and Innocent and Guilty-Alibi with HT (*AUC* = .55, *SE* = .069, *p* =.451). Classification performance was also at chance when comparing Guilty-Standard to Guilty-Alibi (*AUC* = .58, *SE* = .069, *p* = .270), Guilty-Standard to Guilty-Alibi with HT (*AUC* = .54 *SE* = .069, *p* = .558), and Guilty-Alibi to Guilty-Alibi with HT (*AUC* = .59, *SE* = .068, *p* = .188).

So, in sum, although the faking analysis was able to detect a proportion of Guilty-Alibi participants as fakers in Experiment 1, this measure did not detect faking in Experiment 2 nor in any of the three aIAT versions in Experiment 3 (apart from one difference between Guilty-Standard and Guilty-Alibi with HT groups). In Experiment 3, all groups had higher faking scores than in Experiments 1 and 2, which is likely because all groups completed three versions of the aIAT and therefore had more practice at the tasks. Practice would be expected to produce different RT patterns across blocks compared to when participants only completed one aIAT as in the prior two experiments (and also in Agosta et al., 2011). The results thus suggest that the faking index only has limited usefulness against a false alibi countermeasure, which is likely because the alibi participants were not trying to beat the test by intentionally altering their response times, which is what they were doing in the previous studies that have shown better detection of fakers (Agosta, et al., 2011). Rather, the alibi countermeasure produced RT patterns that were fairly similar to a truly innocent group, meaning that these participants were not likely to be detected as faking.