How vulnerable are reaction time based lie detection tests to faking?

Kristina Suchotzki*[1], Bruno Verschuere[2] and Matthias Gamer[1]

[1]Department of Psychology, University of Würzburg, Marcusstr. 9-11, 97080 Würzburg, Germany

[2] Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 128, 1018 WS Amsterdam, The Netherlands

***Corresponding author:**

Kristina Suchotzki

Experimental Clinical Psychology

Department of Psychology

University of Würzburg

Marcusstr. 9-11

97070 Würzburg, Germany

Tel: +49-(0)931-31-82861

Fax: +49-(0)931-31-82733

kristina.suchotzki@uni-wuerzburg.de

Word count total: 10037

Word count introduction: 1353

Word count discussion: 2248

Abstract

Despite few available research, it is often argued that Reaction time (RT)-based deception test are easily faked. In the current two experiments, guilty and innocent mock crime participants took an RT-Concealed Information Test (CIT) and an autobiographical Information Test (aIAT) and received instructions how to fake one of the tests. Experiment 1 showed that the CIT was highly effective in naïve participants, but also vulnerable to faking (*n*=83). The aIAT was ineffective for naïve and faking participants (*n*=85). Experiment 2 showed that when using a response deadline, the CIT remained effective, even when participants received faking instructions (*n*=54). The aIAT was again ineffective for naïve and now even more for faking participants (*n*=86), perhaps due to our suboptimal aIAT design. Faking algorithms performed poorly in both experiments. Results support the validity of the CIT and suggest that a response deadline may make faking more difficult in this test.

Keywords: Deception, Lying, Reaction Times, Concealed Information Test, autobiographical Implicit Associations Test

General Audience Summary

Reaction time (RT)-based test are easy to apply and provide a valid means for the detection of concealed information and deception. While it is often argued that they are easily faked, research on their vulnerability to faking is scarce. In the current two experiments, we investigated to what extent participants who received instructions about the test principle and strategies how to obtain an innocent test outcome succeeded in passing two different RT-based deception detection tests (the RT-based Concealed Information Test [RT-CIT] and the autobiographical Implicit Associations Test [aIAT]). Results of our first experiment revealed that while test validity of the RT-CIT was very high in a naive group, it dropped significantly in a faking group. Test validity of the aIAT was low in both groups. Results of the second experimented revealed that when participants are given a response deadline in both tests (i.e., they receive a message that they are too slow after 800 or 1600 ms in the RT-CIT and the aIAT, respectively), participants do not seem to be able to successfully implement their faking strategies in the RT-CIT. Faking was successful in the aIAT in our second experiment. Our results support the validity of the RT-CIT and suggest that a response deadline may make faking more difficult in this test.

1. Introduction

Faking, that is participants using systematic strategies to influence their test outcome, poses a problem for all deception detection tests. Faking has been shown to reduce classification accuracy for autonomic (Ben-Shakhar, 2011), electrophysiological (Rosenfeld et al, 2004) and neural measures (Ganis, Rosenfeld, Meixner, Kievit, & Schendan, 2011). But although has often been argued that reaction times are particularly vulnerable to faking, here the empirical evidence is not conclusive.

A recent meta-analysis revealed a high average effect size of *d* = 1.05, 95% CI [0.93, 1.17] across different reaction time (RT) measures of deception (Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017). Of the four paradigms included in the meta-analysis, the two that so far show the most promise for a use in applied deception detection contexts are the RT Concealed Information Test (RT-CIT; Seymour, Seifert, Shafto, & Mosmann, 2000) and the autobiographical Implicit Associations Test (aIAT; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). In the RT-CIT, participants have to classify which of different items they recognize. The RT-CIT effect describes the observation that deceptively denying recognition of actually recognized items prolongs RTs compared to honestly denying recognition of not-recognized items. In the aIAT, participants have to categorize sentences referring to generally true and false sentences and sentences referring to true and false events under investigation (e.g., a crime and an alibi). The aIAT effect describes the observation that categorizing true sentences and the actually experienced event with the same response button results in shorter RTs compared to categorizing true sentences and the not-experienced event with the same response button (for a more detailed description of both paradigms see also the method section).

Although with *d* = 0.82, 95% CI [0.54, 1.11] and *d* = 1.30, 95% CI [1.06, 1.54], moderator analysis revealed a smaller average effect for the aIAT compared to the RT-CIT, both paradigms produced large effects (Suchotzki et al., 2017). In their meta-analysis, Suchotzki et al. (2017) also attempted to quantify the effect of faking on RT deception measures. A second smaller meta-analysis including only studies in which participants were instructed how to beat the tests only revealed a small non-significant effect of d = 0.13, 95% CI [-0.17, 0.43]. Yet, there are several reasons why more research on the effects of faking on the RT-CIT and the aIAT is promising. First, the faking meta-analysis relied on a very heterogeneous sample: about 85% of the observed variance between effect sizes was caused by systematic differences between studies. Second, the faking meta-analysis included also studies in which RTs were not the primary measure of interest. Parameters in those studies may not have been optimal for the measurement of RTs (e.g., pace of the stimulus presentation or speed instructions; see also Verschuere, Suchotzki, & Debey, 2014). And indeed, when reviewing only studies that used RT versions of the CIT and the aIAT (and thus with RTs as primary measure), results so far seem less clear.

For the RT-CIT, no published studies systematically comparing its effectiveness in faking and non-faking participants exist. However, the few studies that did instruct participants to fake either the RT- or the very similar event-related potential (ERP)-based CIT showed a surprising resistance against countermeasures. In a paper by Seymour et al. (2000), two experiments still revealed significant CIT-effects, when participants were either given a warning about the lie-detection intent of the experiment and were told to avoid responding differently to probe and irrelevant items or even when they were given concrete instructions on what the expected data pattern was and how to avoid this. Sample sizes in those studies were small (n = 11 and n = 14) and effect sizes for the crucial CIT-effect are unfortunately not given, so no information is available whether effect sizes differed in comparison to the control study in the paper, in which no warning or faking instructions were given. Importantly, in all three studies, Seymour et al. used a response deadline of 1000 ms, ensuring that participants could not slow down their responses beyond this deadline. A study of Huntjens, Verschuere, & McNally (2012) whose primary aim was to study memory effects in patients with Dissociative Identity Disorder also revealed that giving a group of actors detailed instructions on how to influence their response patterns by not responding any faster or slower to the different categories of words did not result in an elimination of the CIT-effect. With *d* = .45 (the effect size that was reported by the authors), the effect was however considerably smaller than the average RT-CIT-effect reported in the meta-analysis. Note that also here, a response deadline was used, this time of 800 ms. The conclusion that even if effects may be smaller, faking may not eliminate the RT-CIT-effect is also supported by a study of Mertens & Allen (2008). Although they did not measure RTs as primary measure, the CIT used in combination with ERPs is in its design very similar to the RT-CIT. Also those authors found that different faking strategies applied to irrelevant items (like thinking about being slapped or applying pressure to the toe) did numerically reduce the RT-CIT effect, yet did not eliminate it.

For the aIAT, published evidence clearly suggests its vulnerability to faking. In three experiments in which participants were instructed to slow down their performance in the test block in which confessions of a crime were paired with generally true sentences, Verschuere, Prati, and De Houwer (2009) found that participants who were guilty of performing a mock crime largely succeeded in avoiding a guilty test outcome and succeeded in being classified as innocent, with even a reversal of the aIAT effect on a group level. Importantly here, also a response deadline of 1200 ms did not prevent faking in an additional experiment. Note here, that all participants not only received faking instructions but also the chance to practice the aIAT or the similar IAT before (yet without the faking instructions). In a similar vein, Agosta, Ghirardi, Zogmaister, Castiello, & Sartori (2011), found in four experiments that compared to participants who did not receive any instructions, participants who were simply asked to hide their true memory succeeded in two of four experiments to reverse their test outcome. Participants who received specific instructions to slow down in one block and speed up in the other succeeded to lower or reverse their outcomes in all four experiments, even in the first experiment in which they had no opportunity to practice the aIAT before. Also in two experiments by Hu, Rosenfeld, & Bodenhausen (2012), participants who received the instruction to speed up their response in one particular block (the one pairing the denial of a mock crime with the true category) succeeded in reversing their outcomes. This reversal was present independent of the opportunity to practice (with and without faking instructions), yet even stronger for participants who also received the opportunity to practice this faking strategy.

Thus, while the available evidence consistently shows that the aIAT is vulnerable to faking, there is no direct evidence regarding the fakeability of the RT-CIT. The aim of our current study therefore was to test the fakeability of the RT-CIT and the aIAT within one experiment, with the faking and the non-faking conditions being as closely matched as possible. We employed a mock crime paradigm and tested both guilty and innocent participants. Based on the above reviewed previous literature, we expected both test to be fakeable, yet the aIAT possibly to a stronger extent than the RT-CIT. Whereas the first of our two experiments should address the general question of the fakeability of both tests, the second experiment aimed to have a closer look at the potential merit of a response deadline to prevent faking. In both experiments, different algorithms to detect faking were employed and compared regarding their efficacy. In case faking can not prevented, such algorithms may help to at least detect it. Such algorithms have been proposed for the aIAT (Agosta et al., 2011) and we evaluate whether they would also prove useful in our experiments. For the RT-CIT, no such algorithms are available and our aim was to develop and test two such algorithms.

2. Experiment 1

2.1. Method

2.1.1. Participants

In total, 88 participants volunteered to take part in the study. The study conformed to the principles expressed in the Declaration of Helsinki. All participants provided written informed consent. Data of three participants were excluded because they indicated orally or in the questionnaire afterwards that they did not complete the task they were assigned to (mock crime or alibi activity).

Using the common exclusion criteria as reported in the literature, two additional participants were excluded for the CIT analysis, as they had less than 50% trials for one item type after exclusion of error trials and RT outliers (see e.g., Noordraven & Verschuere, 2013). The mean age of the remaining 83 participants for the CIT was 24.45 (*SD* = 5.45; 63 female, 20 male), with 44 participants in the mock crime group and 39 participants in the alibi group. Of those, 25 from the mock crime group and 22 from the alibi group received faking instructions for the CIT (see the explanation of the procedure below).

No additional participants were excluded for the aIAT analysis (Sartori et al., 2008), which left a sample of 85 participants with a mean age of 24.89 (*SD* = 6.47; 85 female, 20 male), with 46 participants in the mock crime group and 39 participants in the alibi group. Of those, 20 from the mock crime group and 17 from the alibi group received faking instructions for the aIAT.

2.1.2. Procedure

Participants arrived in the lab in pairs of two.[1] First, the experimenter explained to participants that by choosing one of two envelopes, one of them would be randomly determined to having to commit a small mock theft in the building, whereas the other participants would have to perform a simple everyday activity. Participants were told that after they fulfilled their task, they would return to the lab and perform two RT-based lie detection test that would aim to detect whether they were guilty or innocent of performing the theft. It was also stressed that during the whole experiment, participants should reveal neither to the experimenter (who was blind to the condition assignment) in which condition they were, nor talk to the other participant about their activity. One participant was then asked to pick one of two envelopes, and the remaining envelope was given to the other participant. Participants were given around five minutes to carefully read the instructions provided for each of the two activities (see below). When both participants indicated that they had finished reading and had memorized their tasks, they were instructed to leave the lab, perform their tasks and return to the lab. After the return of participants to the lab, it was explained to them that a theft had taken place in the building, and that they were both suspects. Participants were told that now they would have to perform the two deception detection tests that aim to detect their guilt or innocence. They were given instructions for the first test (either a CIT or an aIAT) and then performed the first test that took about ten minutes. Then they were given instructions for the second test (the respective other test) and performed it (again taking about ten minutes). Test order was counterbalanced across participants, with faking instructions always being given for the second test. Afterwards, participants were asked to reveal which activity they had performed. They were given a questionnaire assessing demographic variables, how nervous and motivated they were, how difficult they found the test, whether

[1] In case only one participant showed up, it was explained to participants that there should have been two present (so the experimenter could try to find out who of the two was guilty), but that the experiment could also be conducted with one participant and the experimenter trying to find out whether this participant was in the guilty or innocent condition.

they tried to apply the faking strategy and how difficult they found this. We also asked them to repeat the faking strategy and to honestly indicate all five items that were used as probe items in the CIT (to ensure memory of those).

2.1.3. Mock crime and alibi activity

The instruction for the mock crime told participants to steal exam questions from a professor’s office. Therefore, they should leave the lab and go to the second floor of the building. They were told that there they would find the office of *Prof. Meykenhorst*, which was not closed due to construction works. Prof. Meykenhorst was known for always saving his exam questions on a *USB-stick* hidden under a *flower pot*. They should take the USB-stick and hide it in their clothes. After they had finished this, they should return to the lab. Not announced in the instructions, participants would find a six-pack of *beer* in the office that they had to pass and the flower pot was standing behind a picture with *stones*. Words marked in *italics* refer to the probe items used in the CIT.

The instruction for the alibi activity told participants that they had a couple more minutes until the experiment began and they should use it to make some tea. Therefore, they should leave the lab and go to the student pool in the basement of the building. They should enter it, take a cup and a package of tea from the cupboard (marked with the words “Lie To Me”) and heat water in the water boiler. They should make a tea and wait a couple of moments until the tea was finished. They should then take the paper bag that had contained the tea bag, and return to the lab.

Both participants were told not to reveal their activity to the experimenter, who was blind to which participant was in which condition. They should only reveal their activity at the end of the experiment, when they were also asked to return the respective item they had taken.

2.1.4. Concealed Information Test

The pre-test procedure and the CIT were presented with Inquisit 4. During the pre-test procedure, participants were presented with five details belonging to a not committed theft. These details served as target details during the test and were “Stolen item: key”, “Name of the professor it belonged to: Schuffenhauer”, “Where it was placed: Shelf”, “What beverage was there: liquor”, “What was on the picture: Eiffel Tower”. Participants were told to remember those details thoroughly, so that during the CIT, they could successfully admit recognition of those details. A screen with all five details was presented three times for 30 seconds. After each disappearance from the screen, participants had to type in all five details via the keyboard. In the CIT, participants were then instructed that they would see a number of details and that they had to indicate whether they recognized them or not. Recognition should only be acknowledged for the previously learned target details (by pressing the “yes” key) and be denied for all other details, including the details referring to the theft that had taken place (by pressing the “no” key). The “a” and the “l” key of a standard QWERTZ keyboard were used, with the assignment of “yes” and “no” responses being counterbalanced between participants.

In total, 30 different details (the five aforementioned target details, five probe details consisting of the actual theft details and 20 matched neutral details) were each presented six times in completely randomized order (180 trials in total). Reminder labels for “yes” and “no” responses appeared on the left and right lower part of the screen. Participants were instructed to respond as fast and correct as possible. The details disappeared as soon as a response was given. No error feedback was given. The inter trial interval was set to vary randomly between 500, 600, 700, 800, 900 and 1000 ms. After 90 trials, participants could take a self-paced break.

2.1.5. Autobiographical Implicit Associations Test

The aIAT was presented with Inquisit 4 and in its design we aimed to closely follow the recommendations of Agosta and Sartori (2013). In the aIAT, participants were presented with four types of statements: true statements (e.g., “I’m in a room with computers”), false statements (e.g., “I’m on top of a mountain”), statements referring to the event under investigation (e.g., “I stole property of the professor”), and statements referring to the alibi activity (e.g., “I heated water and made tea”). They were instructed that each statement they saw had to be classified with a left (“s”) or right (“k”) button press (on a standard QWERTZ keyboard). The test then started with a practice block in which only true and false statements had to be classified, followed by a second practice block in which only statements referring to the theft and statements referring to the alibi had to be classified. The third block was one of the two relevant test blocks, in which all four statements had to be classified in all four categories. The fourth block was again a practice block, in which again only statements referring to the theft and statements referring to the alibi had to be classified (yet in switched positions). And in the fifth (test) block again all four statements had to be classified. Reminder labels showing the titles of the four categories “true”, “false”, “theft”, and “alibi” were presented on the upper left and right side of the screen. The assignment of the “true” and “false” categories to the left and right side of the screen was determined randomly for each participant and remained the same during the whole experiment. The assignment of the “theft” and “alibi” category was also determined randomly, and then switched between block three and four.

In total, 20 different statements (five generally true statements, five generally false statements, five positive statements referring to the theft and five positive statements referring to the alibi) were used. In block one, each true and false statement was presented twice in completely randomized order (20 trials in total). In block two, each theft and alibi statement was presented twice in completely randomized order (20 trials in total). In block three, each of the four different statements was presented three times in completely randomized order (60 trials in total). In block four, each theft and alibi statement was presented four times in completely randomized order (40 trials in total; with switched locations of the category labels). In block five, each of the four different statements was again presented three times in completely randomized order (60 trials in total). Participants were instructed to respond as fast and correct as possible. The statements disappeared as soon as a response was given. The inter trial interval was set to vary randomly between 500, 600, 700, 800, 900 and 1000 ms. If participants made a mistake, a red “X” was presented in the middle of the screen for 400 ms. After each block, participants could take a self-paced break.

2.1.6. Faking instructions

Faking instructions were presented on the screen after the general test instructions and before the participants completed the second test. Thus, participants starting with the aIAT received faking instructions for the CIT just before starting the CIT and were included in the aIAT no faking group and the CIT faking group. Participants starting with the CIT received faking instructions for the aIAT, just before starting the aIAT and were included in the CIT no faking group and the aIAT faking group. For the CIT, the faking instructions were (translated from German):

“To increase your chances to influence the test result in your favor and be classified as ‘innocent’, we now give you some information about the test principle! In the test you will see three different categories of words. 1. Words of the list, that you just learned by heart (to which you should reply with ‘Yes’). 2. Words that are meaningless for you (to which you should reply with ‘No’). 3. Words that are related to the theft and which you will only recognize in case you committed the theft (to which you should reply with ‘No’). Important for your test result is your reaction time on the meaningless words and the words that are related to the theft. If you respond faster to the meaningless words than to the theft-related words, you will be classified as guilty. An effective strategy, in case you are guilty, is therefore to try to deliberately respond slower to the words that are unknown to you than to the theft-related words. Of course, this should happen without it attracting too much attention. In case you are innocent, you should not recognize the theft-related words. In this case you can always try to respond as fast as possible. Thus, in case you are guilty: Try to react slower to the words that are unknown to you and try to react faster to the words that are related to the theft."

For the aIAT, the faking instructions were (translated from German):

"To increase your chances to influence the test result in your favor and be classified as 'innocent', we now give you some information about the test principle! In total, the test consists of five different test phases. Test phases 1, 2, and 4 are simply practice phases, here you can react normally. The important phases are phase 3 and 5. In these phases, you have to classify all four types of statements. In one of the two phases (3 OR 5), you will have to push the same button for true statements and for theft statements (and for false statements and for alibi statements). In the other phase (3 OR 5), you will have to push the same button for false statements and for theft statements (and for true statements and for alibi statements). The test will classify you as guilty if you are faster in the phase in which you have to press the same button for true statements and theft statements. An effective strategy to influence your test result is therefore to deliberately react slower in this phase. Of course, this should happen without it attracting too much attention. Thus, independent of whether you are guilty or innocent: Try to be slower in the phase in which you have to press the same button for true statements and theft statements and try to be faster in the phase in which you have to press the same button for false statements and theft statements."

Faking instructions were given irrespective of guilt or innocence of the participants for two reasons. First, as we found it important that the experimenter would be blind to the guilt or innocence of the participants, this would have complicated the experimental procedure. Second, it is not unlikely that in high stakes situations, also innocent suspects would inform themselves about faking strategies. Note that the two tests differ in the sense that in the aIAT, both guilty and innocent participants can in principle apply the faking strategy whereas in the CIT, only knowledgeable participants can sensibly apply the faking strategy.

2.1.7. Data analysis

Data were analyzed with R and raw data as well as the analysis script can be accessed on https://osf.io/t9y5d/?view_only=5b32017f1be042f0a8161bd7d16cdaf2. For the CIT analysis, error trials (8.84 %) and RT outliers (2.48%; RTs > 2.5 *SD*s from the mean per subject and item type) were removed. Individual Cohen's *d*s were computed by subtracting the mean RT for irrelevant items from the mean RT for probe items for each participant and dividing this number by the respective pooled standard deviations (i.e., the square root of (SD_{probe}^2 + $SD_{\text{irrelevant}}^2$)/2; Noordraven & Verschuere, 2013). Those were then analyzed with a 2 x 2 ANOVA with the between-subject factors Guilt (Guilty vs. Innocent) and Faking (No vs. Yes). Positive *d* should be an indication of memory for crime-related details, whereas values of *d* around zero should be indication of no such memory and innocence.

For the aIAT analysis, we calculated the D6 values according to Greenwald Nosek, & Banaji, (2003; see also Sartori et al., 2008). Therefore, we first deleted latencies below 400 ms and above 10000 ms. We then replaced error trials by the mean of the respective block plus 600 ms. Individual D6 values were computed by subtracting the mean corrected RT for the block in which true sentences where paired with crime sentences (and false sentences with alibi sentences) from the mean corrected RT for the block in which true sentences where paired with alibi sentences (and false sentences with crime sentences) and dividing this number by the respective standard deviation of each participant. Positive aIAT D scores should be an indication of guilt, whereas negative D6 values should be indication of innocence.

To follow up on significant interaction effects in the ANOVAs, the differences between the guilty and innocent group were compared separately in the faking and the no-faking condition using Welch's *t*-tests (Delacre, Lakens, & Leys, 2017).

As index of classification accuracy, we followed the recommendation of the National Research Council (2003) and calculated a measure that is independent of any specific cut-off value. We computed the area (*a*) under the Receiver Operator Curve (ROC). In this method, *a* depicts the overall classification accuracy across all possible cut-off points, with values ranging from 0 to 1. Values of .50 and 1 reflect chance classification and perfect classification of guilty and innocent examinees, respectively. ROC curves and the corresponding *a* values were computed for the no faking and the faking groups, separately.

Finally, we explored the possibility to detect fakers. In the aIAT, we used the method proposed by Agosta et al. (2011). We therefore first eliminated all responses below 150 and above 10000 ms in the aIAT. We then substituted errors in the test blocks with the mean of each block plus a penalty of 600 ms. Finally, we calculated the ratio between the average RT of the fastest of the two test blocks and the respective practice blocks. We then evaluated the performance of the faking detection algorithm by quantifying the accuracy of the classification of participants as fakers or non-fakers on the basis of the cut-off proposed by Agosta et al. (2011; classification as faker with values > 1.08) and by directly contrasting the ratio values of faking and not faking participants using a ROC curve. As no faking detection algorithm has been proposed for the RT CIT so far, we explored two options. First, we followed a recommendation of Lykken (1960) and examined the possibility that reversed CIT effects (i.e., longer RTs for irrelevants compared to probes) could be used as an indication for faking. Therefore, we classified all participants with *d* values smaller than -.2 as fakers and participants with *d* values equal or larger than 1.2 as non-fakers and also calculated the corresponding area under the ROC curve contrasting the *d* values of fakers and non-fakers. Second, similar to the logic of Agosta et al. (2011) that faking should manifest itself in the crucial test items but not in practice or items that are irrelevant for the test result, we calculated a ratio of the mean RT for irrelevants and the mean RT for (task irrelevant) target items. To get an idea how well such a ratio may help to distinguish between fakers and non-fakers, we computed the ROC curve and the corresponding *a* for the separation of the distribution of faking and not faking participants.

2.2. Results

2.2.1. Questionnaire

Using the larger sample that had been used for the aIAT analysis (n = 85) revealed that participants rated their average nervousness during the experiment with M = 2.18 (SD = 1.08) on a scale from 1 to 5. Not surprisingly, nervousness of guilty participants (M = 2.61; SD = 0.95) exceeded the nervousness of innocent ones (M = 1.67; SD = 1.01), t(79.08) = 4.40, p < .001, d = 0.96. Participants' general motivation to pass the tests was with M = 4.18 (SD = 0.82) very high on a scale from 1 to 5. General test difficulty was perceived as medium, with M = 2.40 (SD = 0.86) on the same 5-point scale. Comparing the motivation ratings of the CIT and the aIAT in the guilty and the innocent groups separately with a 2 (Test: CIT vs. aIAT) x 2 (Guilty: Guilty vs. Innocent) ANOVA revealed neither significant main effects of Test, F(1, 83) = 1.05, p = 0.308, n_p^2 = 0.01, and Guilt, F(1, 83) = 0.92, p = .342, n_p^2 = 0.01, nor a significant interaction of both factors, F(1, 83) = 0.01, p = .933, n_p^2 < 0.01. The same ANOVA of the difficulty revealed a significant main effect of Test, F(1,83) = 8.83, p = .004, n_p^2 = 0.10 with the aIAT (M = 2.58; SD = 1.05) being rated as more difficult than the CIT (M = 2.22; SD = 1.00). Also the main effect of Guilt was significant, F(1, 83) =

17.72, p < .001, n_p^2 = 0.18, with guilty participants (M = 2.73; SD = 0.77) perceiving both tests as more difficult than innocent participants (M = 2.01; SD = 0.79). The interaction of Test x Guilt was not significant, F(1, 83) = 1.93, p = .169, n_p^2 = 0.02. In total, 38 of the 85 participants indicated that they applied the faking strategy they were instructed to apply, with 16 of 46 of the guilty participants (34.78 %) and 22 of 39 innocent participants (56.41 %), χ^2(1) = 3.99, p = .046. There were no differences depending on which test needed to be faked, χ^2(1) = 0.46, p = .498. Comparing the rating of the perceived difficulty of the faking in participants that indicated that they did apply the faking strategy with a 2 (Test: CIT vs. aIAT) x 2 (Guilt: Guilty vs. Innocent) ANOVA also revealed that the aIAT (M = 3.00, SD = 0.95) was rated as more difficult to fake than the CIT (M = 2.63, SD = 1.09), F(1,71) = 4.30, p = .042, n_p^2 = 0.06. There was no significant main effect of Guilt, F(1,71) = 2.80, p = .099, n_p^2 = 0.04. The significant interaction of F(1,71) = 11.46, p = .001, n_p^2 = 0.14 showed that the aIAT was only rated as more difficult to fake in the innocent group, t(28.95) = 3.21, p = .003, d = 1.13, but not in the guilty group t(35.24) = 1.12, p = .269, d = 0.35.

2.2.3. ANOVAs

The mean d and D6 values for all four conditions for the CIT and the aIAT are depicted in Figure 1. The 2 x 2 ANOVA on the d values in the CIT revealed no significant main effect of Faking F(1,79) = 2.06, p = .155, $\mathrm{n_p}^2$ = .03. There was a significant main effect of Guilt, F(1, 79) = 38.15, p < .001, $\mathrm{n_p}^2$ = .33, and a significant interaction of Guilt x Faking, F(1,79) = 12.03, p = .001, $\mathrm{n_p}^2$ = .13. T-tests showed a larger (and positive) mean d for the guilty compared to the smaller and negative mean d for the innocent group in the no faking condition, t(33.10) = 7.50, p < .001, d = 2.51, and the same difference being smaller yet still significant in the faking condition, t(43.97) = 2.19, p = .034, d = 0.64.

The 2 x 2 ANOVA on the D6 values in the aIAT revealed no significant main effect of Faking $F(1,81) = 0.23$, $p = .633$, $\eta_p^2 < .01$. There was a significant main effect of Guilt, $F(1,81) = 14.07$, $p < .001$, $\eta_p^2 = .15$, with significantly more negative D6 values in the innocent (*M* = -0.60; *SD* = 0.52) compared to the guilty group (*M* = -0.15; *SD* = 0.58). There was no significant interaction of Guilt x Faking, $F(1,81) = 1.90$, $p = .172$, $\eta_p^2 = .02$.

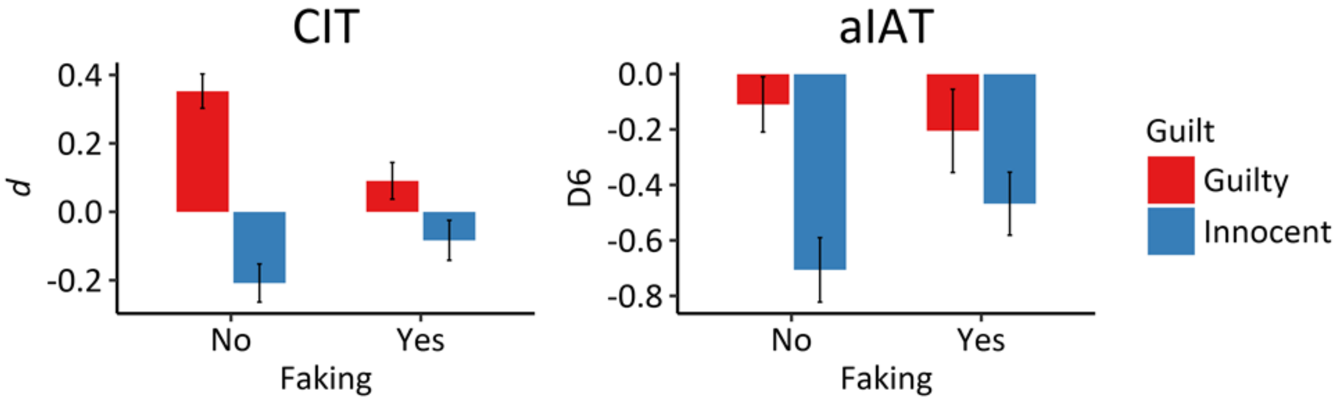

Figure 1. Mean *d* and D6 values in the CIT and the aIAT in the no faking and the faking group in Experiment 1. Error bars represent the standard errors of the mean.

Computing the ROC curves revealed an area under the curve of *a* = .97, 95% CI [0.92, 1.00] in the CIT no faking condition, *a* = .68, 95% CI [0.52, 0.84] in the CIT faking condition, *a* = .82, 95% CI [0.69, 0.95] in the aIAT no faking condition, and *a* = .61, 95% CI [0.42, 0.80] in the aIAT faking condition.

2.2.4. Faking detection

Classifying all participants with *d* values smaller than -.2 in the CIT as fakers resulted in a not above average correct detection of fakers and non-fakers of 45.78 %. Accordingly, calculation the corresponding ROC curve did not reveal an *a* above average (*a* = 0.57, 95% CI [0.44, 0.70].). Calculating the ratio of the mean RT for irrelevants and targets, however, revealed an *a* = .69, 95% CI [0.57, 0.80] under the ROC curve contrasting fakers and non-fakers. As predicted, faking participants showed a higher irrelevant/target ratio (*M* = 0.94, *SD* = 0.14) than not faking participants (*M* = 0.86, *SD* = .08), *t*(74.13) = 3.28, *p* = .002, *d* = 0.68.

Using the algorithm as described by the Agosta et al. (2011) resulted in a detection of fakers (vs. non-fakers) not above average both when using the proposed cut-off of 1.08 (44.71 % correct classification rate of fakers vs. non-fakers). Correspondingly, the area under the ROC curve contrasting both groups did not differ from chance discrimination (*a* = 0.60, 95% CI [0.48, 0.73].)

2.2. Discussion

The results of Experiment 1 revealed the CIT to be fakeable, yet faking did not eliminate the CIT effect. As can also be seen in the ROC analysis, classification accuracy, which was very high in the no faking group, significantly decreased (there was no overlap of the two confidence intervals), yet still stayed significantly above chance level. Of the two newly investigated faking detection algorithms for the RT-CIT, the ratio of mean RT of irrelevant and target items showed some promise as it did result in a statistically significant discrimination of fakers and non-fakers above chance level.

Results did not reveal a faking effect on the aIAT yet were in general not in accordance with expected response pattern as also guilty participants showed an average negative D6 value. Whereas the two distributions of mean values of guilty and innocent participants still were sufficiently distinct to produce a large *a* value, this negative mean value is against the theoretical predictions of the aIAT and would have resulted in poor classification accuracies when using fixed cut-off scores as proposed in the literature (e.g., 0 or 0.2; Agosta, & Sartori, 2013). Using the proposed faking detection algorithm of Agosta et al. (2011) did not result in a detection of fakers above chance level. Next, to investigate to what degree faking could be prevented, we ran a second experiment which was identical to the first one except that response deadlines were used in both tests (800 ms for the CIT and 1600 ms for the aIAT).

3. Experiment 2

3.1. Method

3.1.1. Participants

In total, 92 participants volunteered to take part in the study. The study conformed to the principles expressed in the Declaration of Helsinki. All participants provided written informed consent. Data of two participants were excluded because they indicated orally or in the questionnaire afterwards that they did not complete the task they were assigned to (mock crime or alibi activity).

Using the common exclusion criteria as reported in the literature and as used in Experiment 1, 35 additional participants were excluded for the CIT analysis, as they had less than 50% trials for one item type after exclusion of trials exceeding the response deadline, error trials and RT outliers (see e.g., Noordraven & Verschuere, 2013). One more participant was excluded as his/her subject number was erroneously assigned twice. The mean age of the remaining 54 participants for the CIT was 25.83 (*SD* = 6.27; 35 female, 19 male), with 20 participants in the mock crime group and 34 participants in the alibi group. Of those, 8 from the mock crime group and 13 from the alibi group received faking instructions for the CIT.

Four additional participants were excluded for the aIAT analysis (Sartori et al., 2008), as their subject number was assigned twice. This left a sample of 86 participants with a mean age of 25.66 (*SD* = 6.77; 61 female, 25 male), with 45 participants in the mock crime group and 41 participants in the alibi group. Of those, 23 from the mock crime group and 22 from the alibi group received faking instructions for the aIAT.

3.1.1. Procedure, tests and faking instructions

The procedure, mock crime and alibi instructions were identical to Experiment 1. Also the CIT and the aIAT were identical, except for the use of response deadlines in both tests. For the CIT, a response deadline of 800 ms was used. This value was chosen in accordance with the mean RTs in the CIT in Experiment 1, the response deadlines used in previous studies (e.g., Verschuere, Crombez, Degrootte, & Rosseel, 2010) and after the feasibility had been established in 7 pilot participants. If participants did not react after 800 ms, the message “Too slow!” was presented for 1000 ms in red in the center of the screen. For the aIAT, a response deadline of 1600 ms was used. This response deadline was again deducted from the mean RT in the aIAT in Experiment 1, and initially set to 1400. As pilot participants indicated problems, the response deadline was set 200 ms higher, still being under the 1861 ms response deadline used in the aIAT study of Verschuere et al., (2009). In the aIAT, if participants did not react after 1600 ms, the message “Too slow!” was presented for 400 ms in red in the center of the screen. Faking instructions were identical to the ones provided in Experiment 1. Except that the sentence about not attracting too much attention was modified to “Of course this should happen without it attracting too much attention and in a way that you stay within the given timeframe.”

3.1.1. Data analysis

Data were analyzed with R and raw data as well as the analysis script can be accessed on https://osf.io/t9y5d/?view_only=5b32017f1be042f0a8161bd7d16cdaf2. For the CIT analysis, trials exceeding the response deadline (4.18 %), error trials (11.58 %) and RT outliers (1.20 %; RTs > 2.5 *SD*s from the mean per subject and item type) were removed. For the aIAT analysis, trials exceeding the response deadline (4.20 %) were initially removed. All further analysis steps were identical to the ones of Experiment 1.

3.2. Results

3.2.1. Questionnaire

Using the larger sample that had been used for the aIAT analysis (n = 86) revealed that participants rated their average nervousness during the experiment with M = 2.38 (SD = 1.09; scale from 1 to 5). Not surprisingly, nervousness of guilty participants (M = 2.64; SD = 1.07) exceeded the nervousness of innocent participants (M = 2.10; SD = 1.04), t(83.59) = 2.40, p = .019, d = 0.52. Participants' general motivation to pass the tests was with M = 4.26 (SD = 0.68) on a 5-point scale from 1 to 5 very high. General test difficulty was perceived as medium, with M = 2.95 (SD = 0.84) on a similar scale. Comparing the motivation ratings of the CIT and the aIAT in the guilty and the innocent groups separately with a 2 (Test: CIT vs. aIAT) x 2 (Guilt: Guilty vs. Innocent) ANOVA revealed no significant main effects of Test, F(1, 84) = 0.23, p = .630, $n_p^2 < 0.01$, and Guilt, F(1, 84) = 0.16, p = .691, $n_p^2 < 0.01$, and no significant interaction, F(1, 84) = 0.03, p = .857, $n_p^2 < 0.01$. The same ANOVA on the difficulty ratings revealed a significant main effect of Guilt, F(1, 84) = 7.31, p = .008, n_p^2 = 0.08, with guilty participants (M = 3.18; SD = 0.80) perceiving both tests as more difficult than innocent participants (M = 2.71; SD = 0.81). There was no significant main effect of Test, F(1,84) = 2.09, p = .152, n_p^2 = 0.02, and also the interaction of Test x Guilty was not significant, F(1, 84) = 1.39, p = .242, n_p^2 = 0.02. In total, 42 of the 86 participants indicated that they applied the faking strategy, with 16 of 45 of the guilty participants (35.56 %) and 26 of 41 innocent participants (63.41 %), $\chi^2(1)$ = 6.66, p = .010. There were no differences depending on which test needed to be faked, $\chi^2(1)$ = 2.95, p = .086. Comparing the rating of the perceived difficulty of the faking in participants who indicated that they did apply the faking strategy with a 2 (Test: CIT vs. aIAT) x 2 (Guilt: Guilty vs. Innocent) ANOVA also revealed that that guilty participants rated faking more difficult (M = 3.98; SD = 0.82) than innocent participants (M = 3.06; SD = 1.14) on a scale ranging from 1 to 5, F(1, 75) = 19.02, p < .001, n_p^2 = 0.20. There was no significant main effect of Test, F(1,75) = 0.09, p = .763,

$n_p^2 < 0.01$ and no significant interaction between Test and Guilt, $F(1,75) = 3.03$, $p = .086$, $n_p^2 = 0.04$.

3.2.2. ANOVAs

The mean *d* and D6 values for all four conditions for the CIT and the aIAT are shown in Figure 2. The 2 x 2 ANOVA on the *d* values in the CIT revealed no significant main effect of Faking $F(1,50) = 1.03$, $p = .315$, $\text{n}_\text{p}^2 = .02$. There was a significant main effect of Guilt, $F(1,50) = 17.17$, $p < .001$, $\text{n}_\text{p}^2 = .26$, with a significantly more positive average *d* value in the guilty ($M = 0.17$; $SD = 0.29$) compared to the average negative *d* value in the innocent group ($M = -0.14$; $SD = 0.24$). There was no significant interaction of Guilt x Faking, $F(1,50) = 0.00$, $p = .997$, $\text{n}_\text{p}^2 < .01$.

The 2 x 2 ANOVA on the D6 values in the aIAT revealed no significant main effect of Faking $F(1,82) = 2.45$, $p = .121$, $\text{n}_\text{p}^2 = .03$. There was a significant main effect of Guilt, $F(1,82) = 21.29$, $p < .001$, $\text{n}_\text{p}^2 = .21$, with significantly more negative D6 values in the innocent ($M = -0.69$; $SD = 0.43$) compared to the guilty group ($M = -0.15$; $SD = 0.65$). There was also a significant interaction of Guilt x Faking, $F(1,82) = 4.39$, $p = .039$, $\text{n}_\text{p}^2 = .05$. T-tests showed a larger (and slightly positive) mean D6 for the guilty compared to the smaller and negative mean D6 for the innocent group in the no faking condition, $t(36.72) = 5.29$, $p < .001$, $d = 1.61$. The same difference was not significant in the faking condition, $t(39.49) = 1.78$, $p = .082$, $d = 0.53$ (with also the mean D6 for the guilty group being negative).

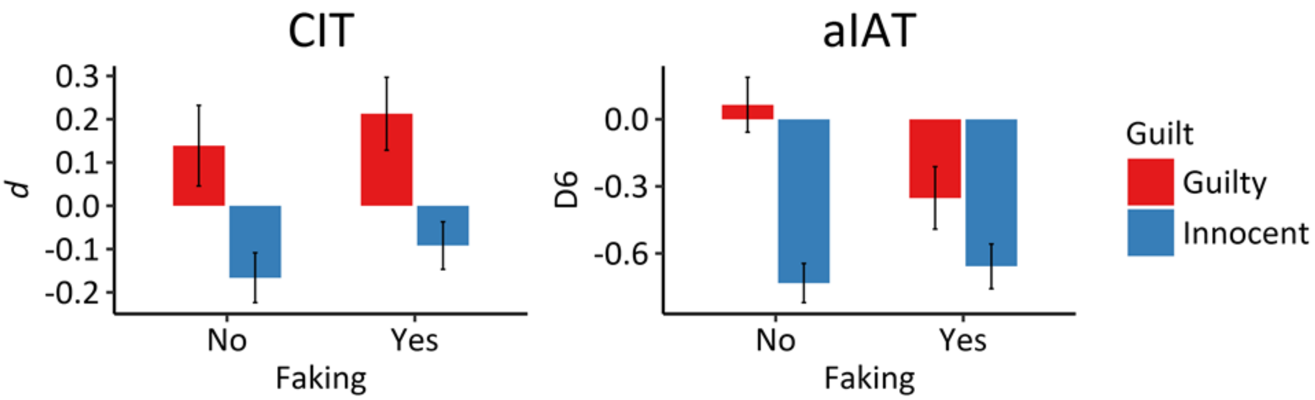

Figure 2. Mean *d* and D6 values in the CIT and the aIAT in the no faking and the faking group in Experiment 2. Error bars represent the standard errors of the mean.

Computing the ROC curves revealed an area under the curve of *a* = .76, 95% CI [0.57, 0.94] in the CIT no faking condition, *a* = .85, 95% CI [0.68, 1.00] in the CIT faking condition, *a* = .89, 95% CI [0.79, 0.99] in the aIAT no faking condition, and *a* = .63, 95% CI [0.46, 0.80] in the aIAT faking condition.

Due to the large exclusion rate, we recalculated the CIT analyses on an exploratory basis, now including the additional 35 participants. These results closely mirrored the results with the smaller sample. The 2 x 2 ANOVA on the *d* values revealed no significant main effect of Faking *F*(1,85) = 1.58, *p* = .213, n_p^2 = .02. There was a significant main effect of Guilt, *F*(1,85) = 12.94, *p* = .001, n_p^2 = .13, with a significantly more positive average *d* value in the guilty (*M* = 0.09; *SD* = 0.29) compared to the average negative *d* value in the innocent group (*M* = -0.11; *SD* = 0.24). There was no significant interaction of Guilt x Faking, *F*(1,85) = 0.19, *p* = .663, n_p^2 < .01. Computing the ROC curves revealed an area under the curve of *a* = .70, 95% CI [0.54, 0.86] in the CIT no faking condition (with 24 guilty and 23 innocent participants), and an *a* = .75, 95% CI [0.59, 0.91] in the CIT faking condition (with 22 guilty and 20 innocent participants).

2.2.4. Faking detection

Classifying all participants with d values smaller than *d* = -.2 in the CIT as fakers resulted in a correct detection of fakers and non-fakers of 50.00 %. Accordingly, calculation the corresponding ROC curve did not reveal an *a* above average (*a* = 0.59, 95% CI [0.44, 0.74].). Calculating the ration of the mean RT for irrelevants and targets revealed an *a* = .60, 95% CI [0.44, 0.77] under the ROC curve.

Using the algorithm as described by the Agosta et al. (2011) resulted in a detection of fakers (vs. non-fakers) not above average both when using the proposed cut-off of 1.08 (44.19 % correct classification rate of fakers vs. non-fakers) and when computing the ROC curve (*a* = 0.60, 95% CI [0.48, 0.72].)

3.3. Discussion

In contrast to Experiment 1, results now revealed no effect of faking on the CIT effect. Although such a null effect may of course not be taken as indication for the absence of an effect, it may serve as an indication that in the CIT, a response deadline may serve to hinder or reduce faking attempts at least to a certain degree. This is also in line with previous CIT studies who often employed response deadlines and showed no or small effects of faking. It should be noted, however, that the use of the response deadlines resulted in a considerable drop in classification accuracy also in the no faking group in comparison with Experiment 1 (even though confidence intervals still slightly overlap) and, if one uses the common exclusion criteria for the CIT, a considerable number of participants being excluded (i.e., not producing meaningful results). In contrast to Experiment 1, the aIAT effect now showed a significant reduction in the faking group. Whereas the aIAT results in the no faking group in this experiment showed the expected pattern of a slightly positive mean D6 value, this same mean D6 value became negative in the faking group and classification accuracy dropped (although confidence intervals still slightly overlapped). None of the proposed faking detection algorithms was capable of reliably detecting faking attempts.

4. General discussion

The aim of the present experiments was to investigate the fakeability of two of the most promising RT-based deception detection tests: the RT-CIT and the aIAT. Although meta-analytic estimates revealed promising high effect sizes for both paradigms, it has often been argued that RT-based measures are under voluntary control and therefore very easy to fake. The current paper aimed to add to the so far insufficient and partly inconsistent results and to directly compare the faking vulnerability of both tests within the same experimental design.

4.1. Faking effects in the RT-CIT

Results of Experiment 1 revealed a faking effect in the CIT, resulting also in a considerable drop in classification accuracy of guilty and innocent test subjects. In line with previous literature, however, the CIT effect did not completely vanish and classification accuracy still stayed above chance (Huntjens et al., 2012; Mertens & Allen, 2008; Seymour et al., 2000). The aim of Experiment 2 was to investigate whether the use of a response deadline may prevent faking. In the RT-CIT, using a response deadline of 800 or 1000 ms is quite common (e.g., Seymour et al., 2000, Verschuere et al., 2010). And our results showed that indeed, a response deadline of 800 ms eliminated the faking effect in the CIT. Those results are encouraging as they implicate that modifying the structural properties of the CIT and restricting the time participants have to respond may prevent them from successfully implementing faking strategies at least to a certain degree. Considering that the rate of participants indicating that they applied faking strategies did not differ strongly between both experiments (44.71 % vs. 49.43 %), the response deadline may prevent the successful implementation of faking strategies rather than preventing faking attempts in general. This is promising, as in applied contexts the motivation of suspects to implement faking strategies will be considerable higher than in our experimental context (even though participants rated their motivation to pass the tests in both our experiments also as relatively high). Note that similar to previous CIT studies (Huntjens et al., 2012; Mertens & Allen, 2008; Seymour et al., 2000), participants in the current study did not have the opportunity to practice the particular test or even the specific faking strategies before (except for the study of Huntjens et al., 2012, in which participants were given a very short practice of 21 trials). We would argue that not giving participants the chance to practice the particular test before best resembles real-life forensic situations, as there the chances are also small that suspects get the chance to practice the test with the very specific set of items. Our design nevertheless does not answer the question to what degree practice may enable suspects to fake the CIT, even with a response deadline.

Unfortunately, however, the prevention of faking in our study by using a response deadline apparently came at a cost. First, using exclusion criteria common in the literature (e.g., Noordraven & Verschuere, 2013; Kleinberg & Verschuere, 2015, 2016; Verschuere, Kleinberg, & Theocharidou, 2015), resulted in a considerable loss of included participants (38.89 % in Experiment 2). Of course, the question is whether similar exclusion criteria would be used in applied forensic contexts (in the sense that data of such suspects would be considered as inconclusive). Interestingly, including those participants did not change the pattern of our results in Experiment 2, yet resulted in a slightly (yet non-significantly) reduced classification accuracy compared to the reduced sample. Second, using a response deadline of 800 ms resulted in a considerable drop of test validity from Experiment 1 (*a* = .97) to Experiment 2 (*a* = .76) in the no faking group. This is surprising as the use of response deadlines is quite common in the RT-CIT, and several examples can be found in which such

CITs still produced larger effects and better classification accuracies (Verschuere & Kleinberg, 2015; Visu-Petra, Miclea, & Visu-Petra, 2012; Visu-Petra, Varga, Miclea, & Visu-Petra, 2013). Difficulty ratings of the CIT revealed higher values in Experiment 2 compared to Experiment 1 (M = 2.84 vs. M = 2.22; $t(179) = 3.66$, $p < .001$, $d = 0.56$), yet confidence intervals of both *a* values still slightly overlap, so further research is necessary to determine whether this drop in accuracy represents a genuine drop or random variation.

4.2. Faking effects in the aIAT

Results of the aIAT showed a different pattern. Surprisingly, in Experiment 1, we found a negative mean D6 value not only as expected in the innocent group but also in the guilty group. This result is in contrast with theoretical predictions and would considerably lower classification accuracies when using the cut-offs usually proposed in the literature (for a review see Agosta et al., 2013). Some small changes of our aIAT design compared to the original design proposed by Sartori et al. (2008) may have lowered the aIAT validity in our study. First, instead of using category labels referring to the active involvement of the test subjects in the respective activity (e.g., “I committed the theft” and “I went to make tea”), we used category labels that simply referred to the activity as “theft” and “alibi”. Second, as some participants still had to perform the CIT after the aIAT, we could not use any of the probe items in the aIAT crime-related statements (to not reveal those to the innocent participants). This may have resulted in statements that were less salient in their connection to the crime. Our results nevertheless indicate that apparently small changes to the aIAT procedure may already result in considerable lower validities, which may be critical in circumstances where for instance also no salient crime details are available.

Note that different from some previous aIAT studies (e.g., Sartori et al., 2008; Verschuere & Kleinberg, 2016), we motivated participants even in the no faking condition to try to beat the test in our instructions. This is in line with the data of Agosta et al. (2011) who found that participants who were simply asked to hide their true memory already succeeded in two of four experiments to reverse their test outcome. Interestingly, this was the case in our experiment even without offering participants the chance to practice an aIAT or an IAT before. Although speculative at this point, this may be taken as indication that the aIAT is more susceptible to spontaneous faking attempts, probably due to its structural properties. As also argued by Suchotzki et al. (2017), faking in the aIAT requires a strategic slowdown during an entire block, whereas faking in the RT-CIT requires slowing down only on specific items of one block. Applying a faking strategy to an entire block might be easier than having to determine on each trial whether to apply such a strategy. This may be reinforced by the fact that for all participants, the instruction to classify sentences admitting the crime under investigation as "true" may be counterintuitive with the aim of hiding their involvement in the crime, making it even more intuitive that faking needs to be applied in this part of the test. Note, however, that this explanation is not supported by our questionnaire data, as in Experiment 1, innocent participants faking the aIAT rated faking as more difficult than innocent participants faking the CIT, whereas no difference had been indicated by guilty participants and by all participants in Experiment 2. Despite the implementation of a response deadline of 1600 ms, our results did show a significant faking effect in Experiment 2. This faking effect also resulted in a reduced classification accuracy, although confidence intervals here still slightly overlapped. This is in line with the results of Verschuere et al. (2009), who also found that a response deadline did not prevent faking in the aIAT. The only very small positive mean average in the guilty no faking group again would be in line with the idea that the aIAT is also fakeable by naïve yet motivated test subjects. Yet of course it should be mentioned that although our response deadline was shorter than the one used by Verschuere et al. (2009), it was considerably longer than the one used in the CIT. The reason why we decided to use such longer deadline is that the aIAT is more difficult than the CIT (as supported by ratings of Experiment 1) or at least more complex, as it uses more complex stimulus material (sentences instead of words) and requires more task and response switching. Piloting a shorter response deadline showed that participants made too many errors and indicated that they had problems performing the test. But of course, this leaves the possibility that the response deadline we have chosen may not have been optimal and that with a prolonged practice phase, participants would have learned to perform the test even faster and that a shorter response deadline would have more efficiently helped to prevent faking.

4.3. Future research

This directly points to an interesting challenge for future research, the search for better ways to determine optimal response deadlines in the different tests. In general, it is unlikely that a single optimal response deadline can be found for each test, as each test differs depending on the facts under investigation. In the CIT, those determine the items that will be presented and in the aIAT, they affect the complexity of the statements that are used. Furthermore, in the CIT, different choices can be made regarding the presentation modality of the items (e.g., written words, pictures, auditory presentation). Also, individual differences between test subjects will influence the feasibility of different response deadlines. One may, for instance, ask whether response deadlines used in a student sample would also be feasible in a forensic sample. Another interesting avenue therefore is the exploration of the use of individual response deadlines, that may be determined based on the performance in the first trials and maybe even adapted throughout the whole test. Another point that merits more exploration is our finding that many participants indicated in the questionnaire afterwards that they did not apply the faking strategy. Unfortunately, we did not follow up on this question by asking what the reasons for this were. Several reasons are possible. First, participants may not have understood the faking instructions. Second, they may have attempted to apply those, but may have feared that they were not successful in doing this. Third, they may have lacked the motivation to do so (although the general high motivation ratings speak against this possibility). In would be very interesting to learn these reasons in future research, as those may also be informative regarding what may hinder participants to apply faking strategies in general. Also, obtaining results using only the participants that indicated that they did apply the faking strategy would be very informative, yet to this means a larger sample would be needed as in our sample, this resulted in too small sample sizes for the different conditions. A final suggestion would be to adapt the structural properties of the aIAT to make it more resistant to faking. As explained above, faking the aIAT simply requires a slow down during an entire block. Switching category labels in the aIAT on a trial by trial basis instead may make faking considerably more difficult (see e.g., the Implicit Relational Assessment Procedure for a comparable paradigm; Barnes-Holmes et al., 2006).

4.4. Faking detection

A different approach than preventing faking is to accept that faking cannot be completely prevented and to aim at detecting such attempts (Verschuere & Meijer, 2014). Using algorithms that were proposed in the literature, none proved very successful in our study. Using the algorithm as proposed and successfully implemented by Agosta et al. (2011) did not result in a valid classification of fakers vs. non-fakers in our two samples. Note that this was not only the case when using the proposed cut-off (which may be sample-dependent), but also when calculating the cut-off independent ROC curve. This was also the case for the approach of using the individual *d* values in the CIT – which are also used for the classification of guilty vs. innocent – for the classification of fakers vs. non-fakers. Results looked a bit more promising for our approach to use the irrelevant/target ratio. This approach is based on the logic that participants who use faking strategies will most likely slow down their responses on irrelevant items, whereas they will still try to perform well on target items.

Thus, using a longer RT on irrelevant compared to targets items as indication of faking resulted in a valid classification of fakers vs. non-fakers in Experiment 1. Note, however, that this result could not be replicated in Experiment 2. Also, it should be noted that the use of faking detection algorithms even if successful in detecting fakers still suffers from the drawback that it does not allow any conclusions on the guilt of fakers. As it is not unlikely that also innocent suspects may apply faking strategies to ensure their (correct) test outcome, faking detection must always result in an inconclusive verdict.

4.5. Conclusion

To sum up, our two experiments show that both the RT-CIT and the aIAT seem fakeable to a certain degree. But whereas the CIT still produced classification accuracies above chance level and the implementation of a response deadline seems promising, the aIAT results in general were less clear in our study. Algorithms to detect faking did not prove reliable. By providing several suggestions for avenues to explore other faking prevention methods, we hope to stimulate further research in this area.

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

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? developing the implicit relational assessment procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, (32)*7, 169-177

Ben Shakhar, G. (2011). Countermeasures. In B. Verschuere, G. Ben Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 200-2015). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511975196.012

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *International Review of Social Psychology, 30*(1).

Ganis, G., Rosenfeld, P. J., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *Neuroimage, 55*(1), 312–319. doi:10.1016/j.neuroimage.2010.11.025

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.

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

Huntjens, R. J., Verschuere, B., & McNally, R. J. (2012). Inter-identity autobiographical amnesia in patients with dissociative identity disorder. *PloS one, 7*(7), e40580. doi: 10.1371/journal.pone.0040580

Kleinberg, B., & Verschuere, B. (2015). Memory Detection 2.0: The First Web-Based Memory Detection Test. *PLoS One, 10*(4). doi:10.1371/journal.pone.0118715

Kleinberg, B., & Verschuere, B. (2016). The role of motivation to avoid detection in reaction time-based concealed information detection. *Journal of Applied Research in Memory and Cognition, 5*(1), 43-51. doi:10.1016/j.jarmac.2015.11.004

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44*(4), 258. doi:10.1037/h0044413

Mertens, R., & Allen, J. J. B. (2008). The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology, 45*(2), 286-298. doi:10.1111/j.1469-8986.2007.00615.x

Noordraven, E., & Verschuere, B. (2013). Predicting the Sensitivity of the Reaction Time-based Concealed Information Test. *Applied Cognitive Psychology, 27*(3), 328-335. doi:10.1002/acp.2910

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

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge". *Journal of Applied Psychology, 85*(1), 30-37. doi:10.1037/0021-9010.85.1.30

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

Verschuere, B., Crombez, G., Degrootte, T., & Rosseel, Y. (2010). Detecting concealed information with reaction times: Validity and comparison with the polygraph. *Applied Cognitive Psychology, 24*(7), 991-1002.

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition, 4*(1), 59-65. doi:10.1016/j.jarmac.2015.01.001

Verschuere, B., & Meijer, E. (2014). What´s on your mind? Detecting concealed information. *European Psychologist. 19(*3), 162-171.

Verschuere, B., Prati, V., & De Houwer, J. (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

Verschuere, B., Suchotzki, K., & Debey, E. (2015). Detecting deception through reaction times. In P. A. Granhag, A. Vrij & B. Verschuere (Eds.), *Deception detection: Current challenges and new approaches*. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9781118510001.ch12

Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Applied Cognitive Psychology, 26*(3), 342-351. doi:10.1002/acp.1827

Visu-Petra, G., Varga, M., Miclea, M., & Visu-Petra, L. (2013). When interference helps: increasing executive load to facilitate deception detection in the concealed information test. *Frontiers in Psychology, 4*, 146. doi:10.3389/fpsyg.2013.00146

Author Contributions

Kristina Suchotzki was involved in the study conception and the design, the acquisition of the data, the analysis and interpretation of the data and the writing of the manuscript. Bruno Verschuere and Matthias Gamer were involved in the study conception, the interpretation of the data and the critical revision of the manuscript.

Acknowledgements

We thank Max Berg, Elisa Kniesburges, Michael Strunz und Markus Tomzig for their help with the data collection.