How Vulnerable is the Reaction Time Concealed Information Test to Faking?

Suchotzki, K.; Verschuere, B.; Gamer, M.

DOI
10.1016/j.jarmac.2020.10.003

Publication date
2021

Document Version
Final published version

Published in
Journal of Applied Research in Memory and Cognition

Journal of Applied Research in Memory and Cognition 10 (2021) 268–277
Accepted by the outgoing editorial team

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Empirical Article

How Vulnerable is the Reaction Time Concealed Information Test to Faking?

Kristina Suchotzki*
Institute of Psychology, University of Mainz, Germany
Department of Psychology, University of Würzburg, Germany

Bruno Verschuere
Department of Clinical Psychology, University of Amsterdam, The Netherlands

Matthias Gamer
Department of Psychology, University of Würzburg, Germany

The reaction time-based Concealed Information Test (RT-CIT) can be used to detect information a suspect wishes to conceal. While it is often argued that it is easily faked, empirical research on its vulnerability to faking is scarce. In three experiments, we tested whether receiving faking instructions enables guilty participants to fake an innocent test outcome in an RT-CIT. In Experiment 1, when not using a response deadline, we found the RT-CIT to be vulnerable to faking ($d = 1.06$). Experiment 2 showed that when using a response deadline, faking was ineffective ($d = -0.25$). Critically, Experiment 3 replicated these findings within one between-subject design, showing again a faking effect when no response deadline was used ($d = 1.08$) that vanished with the use of a response deadline ($d = -0.56$). By providing suggestions for the development of a faking detection algorithm, we hope to stimulate further research in this area.

Keywords: Deception, Lying, Reaction times, Concealed information test, Response deadline, Faking, Countermeasures

General Audience Summary

It is widely believed that reaction time-based deception test can easily be faked. The reaction time-based Concealed Information Test (RT-CIT) is an example for such a test that uses response times to uncover whether someone deceptively denies knowledge of incriminating information. Empirical research on its vulnerability to faking is scarce. In three 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 the RT-CIT. Our results show that while in principle being fakeable, a response deadline—that pushes people to respond rapidly—seems to hinder effective faking in this test.

* Correspondence concerning this article should be addressed to Kristina Suchotzki, Social and Legal Psychology, Institute of Psychology, University of Mainz, Mainz, Germany.
Contact: kristina.suchotzki@uni-mainz.de (K.S.).

Faking is a great challenge to any deception detection test. With faking (also known as countermeasures), we refer to systematic strategies applied by suspects to obtain a favorable test outcome. Such strategies differ depending on the particular deception tests and measures used, ranging from attempts to artificially increase arousal in control conditions (e.g., with hidden movements or mental images) to attempts to decrease responding to the questions of interest (e.g., with distraction or relaxation). Faking has been shown to reduce classification accuracy of tests relying on autonomic (Ben-Shakhar, 2011), electrophysiological (Rosenfeld, Soskins, Bosh, & Ryan, 2004) and neurovascular measures (Ganis, Rosenfeld, Meixner, Kievit, & Schendan, 2011). However, being under voluntary control, reaction time (RT) measures are believed to be particularly vulnerable to faking (Farwell & Donchin, 1991). It has been argued that simple strategic slowing and speeding up of responses suffices to shift a test outcome into the desired direction. Interestingly enough, the empirical evidence on the effects of faking of RT-based deception tests is inconclusive.

A recent meta-analysis revealed a high average effect size across four different reaction time (RT) measures of deception (d = 1.05, 95% CI [0.93, 1.17]; Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017, d referring to the standardized difference between the deceptive vs. truthful responses within-individuals). Suchotzki et al. (2017) also attempted to quantify the effect of faking on RT deception measures and reported a second meta-analysis including only studies in which participants were instructed how to beat the tests. This analysis revealed a small non-significant effect of $d = 0.13$, 95% CI [−0.17, 0.43]. At first glance, these data fuel the concern that RT based lie tests are easily faked. Yet, there are several reasons why such a conclusion is premature. 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. Unfortunately, separate analyses for each experimental paradigm were not possible due to the overall small number of faking studies. Second, the faking meta-analysis included 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, 2015). Indeed, when reviewing only studies that used RTs as primary measure, results seem less clear. Third, comparing effects of the two separate meta-analyses (that for the larger part contained studies either giving faking instructions or not giving any) suffers from all disadvantages inherent to between-study comparisons.

Surprisingly, for one of the most promising paradigms for applied contexts, the RT-based Concealed Information Test (RT-CIT; Seymour, Seifert, Shafto, & Mosmann, 2000), not a single published study directly tested the effects of faking. 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. However, the few studies that did instruct participants to fake the RT-CIT showed a surprising resistance against countermeasures. In a paper by Seymour et al. (2000), two experiments revealed large CIT effects even when participants 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 no information was available on whether effect sizes differed from another reported experiment that served as (between-study) control condition. A study of Huntjens, Verschuere, and McNally (2012) 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 RT-CIT effect. With d = .45, the effect was however considerably smaller than the average RT-CIT effect reported in the meta-analysis. 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 and Allen (2008). Although these authors did not measure RTs as primary measure, the CIT used in combination with event-related brain potentials is in its design very similar to the RT-CIT. Also those authors found that different faking strategies (like thinking about being slapped or applying pressure to the toes) applied to irrelevant items whose recognition should be honestly denied did numerically reduce the RT-CIT effect, but did not eliminate it.

Using a mock crime paradigm, in the current study we examined how vulnerable the RT-CIT is to faking. Note that originally, our experiments also included an investigation of the fakeabilty of another lie detection test, the so-called autobiographical Implicit Associations Test (aIAT; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). In the aIAT, participants have to categorize sentences referring to generally true and false events 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 non-experienced event with the same response button. After the first two experiments, we realized, however, that the currently used procedure for the aIAT might have been suboptimal, thus potentially restricting the generalizability of results. We therefore decided to focus our investigation on the RT-CIT and conducted Experiment 3 accordingly. Methods, results, raw and analyzed data as well as a short discussion of our aIAT findings can be found on the OSF page of the current paper on https://osf.io/t9y5d/.

Experiment 1

Method

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% valid

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 years (*SD* = 5.45 years; 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).

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 participant 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 also informed that at the end of data collection, the two participants who performed best (as innocent) would receive an extra 25 Euro. 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 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).

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.

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 s. 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 QWERTY 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 irrelevant 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,

[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.

800, 900 and 1000 ms. After 90 trials, participants could take a self-paced break.

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.

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.

Data analysis. Data were analyzed with R and raw data, the analysis script as well as a table with the aggregated data can be accessed on https://osf.io/t9y5d/. 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 deviation (i.e., the square root of $(SD_{probe}^2 + SD_{irrelevant}^2)/2$; Noordraven & Verschuere, 2013). Those were then analyzed with a 2 × 2 ANOVA with the between-subject factors guilt (guilty vs. innocent) and faking (no vs. yes). Positive *d* values 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.

To follow up on significant interaction effects in the ANOVAs, the differences between the faking and no faking condition were compared separately in the guilty and innocent group using Welch's *t*-tests (Delacre, Lakens, & Leys, 2017). Partial eta square (η_p^2) was calculated as measure of effect size in the ANOVAs and the standardized mean difference *d* was calculated as measure of effect size in the follow-up t-tests (Cohen, 1988). Note that in the follow-up tests, positive standardized mean differences *d* indicate an effect of faking (i.e., larger CIT-effect in the no faking than the faking condition), whereas negative standardized mean differences *d* indicate the reversed effect.

To gain more information about those comparisons and the informative value of null results, we also calculated the Bayesian independent sample t-tests and reported the resulting Bayes Factors (BFs). Those were calculated using the open source program JASP (https://jasp-stats.org/). A JZS-prior (Cauchy prior of 0.707) was used for the alternative Hypothesis. BFs can be reported either as evidence for the alternative Hypothesis (BF_{10}) or as evidence for the Null-Hypothesis (BF_{01}) and in the manuscript, for the ease of interpretation, we will always report the one that exceeds 1.

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 separately computed for the no faking and the faking groups.

Results

Questionnaire. Analysis of the questionnaire revealed that on a scale from 1 to 5, participants rated their nervousness during the experiment with $M = 2.15$ ($SD = 1.07$) as low, their motivation with $M = 4.12$ ($SD = 0.90$) as high, and test difficulty with $M = 2.59$ ($SD = 1.06$) as low to medium. For a more detailed report on the comparison of those ratings between experimental conditions, see the supplementary material on https://osf.io/t9y5d/. With regard to probe recognition, guilty participants were better able to report the probe items than innocent participants (88% vs. 27%), $t(81) = 21.93$, $p < .001$, $d = 4.82$. Of the 25 guilty participants receiving faking instructions for the CIT, 16 indicated that they applied the faking strategy. Eleven of the 25 could correctly repeat the faking strategy at the end of the experiment. A Pearson chi-square test did not reveal a significant relationship between the application and the correct report of the faking strategy, $\chi 2(2) = 2.71$, $p = .258$.

ANOVAs. The mean *d* values of all four conditions for the CIT are depicted in Figure 1. The 2 × 2 ANOVA on the *d* values in the CIT revealed no significant main effect of faking $F(1, 79) = 2.06$, $p = .155$, $n_p^2 = .03$. There was a significant main effect of guilt, $F(1, 79) = 38.15$, $p < .001$, $n_p^2 = .33$, and a significant

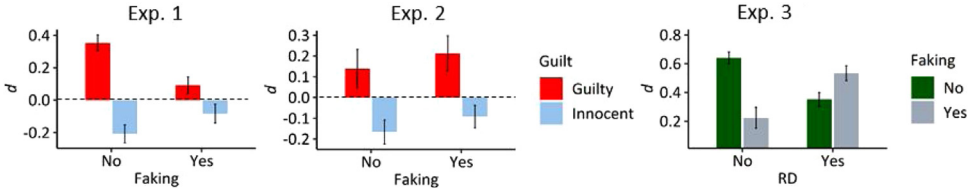

Figure 1. Mean *d* values for the Guilty and the Innocent Group in the No Faking and the Faking group in Experiments 1 (left) and 2 (middle). Mean *d* values in the No Response Deadline and the Response Deadline CITs for the No Faking and the Faking Group in Experiment 3 (right). Error bars represent standard errors of the mean.

Guilt × Faking interaction, $F(1, 79) = 12.03$, $p = .001$, $\eta_p^2 = .13$. A follow-up *t*-test in the guilty group showed that faking instructions significantly reduced the RT-CIT effect, $t(41.77) = 3.58$, $p = .001$, $d = 1.06$, $BF_{10} = 26.80$. In the innocent group, the faking instructions did not lead to significant changes, $t(36.72) = 1.55$, $p = .130$, $d = -0.49$, $BF_{01} = 1.31$.

The ROC curves showed detection efficiency was close to perfection in the no faking condition, $a = .97$, 95% CI [.92, 1.00], yet reduced to modest detection efficiency in the faking condition, $a = .68$, 95% CI [.52, .84].

Discussion

The results of Experiment 1 revealed the CIT to be fakeable. Interestingly, faking did not completely eliminate the CIT effect. Experiment 1 did not, however, use a response deadline. As response deadlines promote fast responding and have been proposed to diminish faking effects (Degner, 2009; Verschuere, Prati, & De Houwer, 2009), we tested the influence of such deadline on the fakeability of the CIT in Experiment 2.

Experiment 2

Method

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 same exclusion criteria as in Experiment 1, 35 additional participants were excluded for the CIT analysis, as they had less than 50% valid trials for one item type after exclusion of trials exceeding the response deadline, error trials and RT outliers. 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 years ($SD = 6.27$ years; 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.

Procedure, tests, and faking instructions. The procedure, mock crime and alibi instructions were identical to Experiment 1. Also the CIT was identical, except for the use of a response deadline. 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., Huntjens et al., 2012; Kleinberg & Verschuere, 2015; Suchotzki, De Houwer, Kleinberg, & Verschuere, 2018; Verschuere, Crombez, Degroote, & Rosseel, 2010; Verschuere, Kleinberg, & Theocharidou, 2015) 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. 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."

Data analysis. Data were analyzed with R and raw data, the analysis script as well as a table of the aggregated data can be accessed on https://osf.io/t9y5d/. 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. All further analysis steps were identical to Experiment 1.

Results

Questionnaire. On a scale from 1 to 5, participants rated their nervousness during the experiment with $M = 2.43$ ($SD = 1.16$) as low, their motivation with $M = 4.31$ ($SD = 0.80$) as high, and test difficulty with $M = 2.56$ ($SD = 1.11$) as low to medium. With regard to probe recognition, guilty participants were better able to report the probe items than innocent participants (93% vs. 25%), $t(51) = 18.73$, $p < .001$, $d = 5.36$. Of the 8 guilty participants receiving faking instructions for the CIT, 6 indicated that they applied the faking strategy. Four of the 8 could correctly repeat the faking strategy at the end of the experiment. A Pearson chi-square test could not be performed due to the small sample size.

ANOVAs. The 2 × 2 ANOVA on the *d* values in the CIT (see Figure 1) revealed no significant main effect of faking, $F(1, 50) = 1.03$, $p = .315$, $\eta_p^2 = .02$. There was a significant main effect of guilt, $F(1, 50) = 17.17$, $p < .001$, $\eta_p^2 = .26$, with a significantly more positive average *d* value in the guilty group compared to the average negative *d* value in the innocent group. There was no significant Guilt × Faking interaction, $F(1, 50) = 0.00$, $p = .997$, $\eta_p^2 < .01$. For sake of completion, we report the results of the same follow-up tests as used in Experiment 1. Neither for the guilty $t(17.72) = 0.59$, $p = .565$, $d = -0.25$, $BF_{01} = 2.23$, nor for the innocent group, the effect of faking was significant, $t(30.67) = 0.93$, $p = .358$, $d = -0.31$, $BF_{01} = 2.23$.

The ROC curves showed moderate to high detection efficiency in both the no faking condition, a = .76, 95% CI [.57, .94], as well as in the faking condition, a = .85, 95% CI [.68, 1.00]. Due to the large exclusion rate, we recalculated the CIT analyses on an exploratory basis, now including all participants. These results closely mirrored the results reported above. Critically, the 2 × 2 ANOVA on the d values again revealed no significant main effect of faking, $F(1, 85) = 1.58$, $p = .213$, $n_p^2 = .02$, nor a Guilt × Faking interaction, $F(1, 85) = 0.19$, $p = .663$, $n_p^2 < .01$.

Discussion

Experiment 2 showed that, with a response deadline, there was no significant effect of faking. Such a null effect may not be taken as indication for the absence of an effect and the Bayesian independent sample t-test revealed that the data were (only) 2.23 times more likely under the hypothesis that faking does not affect the RT-CIT than under the hypothesis that it does. Combined with the observed faking effect in Experiment 1, where no response deadline was used, these data point to the possibility that a response deadline reduces faking effects in the CIT. To put this idea to the test, we ran a third experiment. Aside from providing a direct comparison by comparing RT-CITs with and without response deadline within one experiment, Experiment 3 also served to overcome three serious limitations of Experiments 1 and 2. First, in both experiments faking instructions were always given for the second test. Thereby, in addition to being given faking instructions, faking participants also had a certain degree of additional practice compared to naïve participants, although this practice was obtained in another RT-based test. Second, not all participants in the guilty faking group indicated to have applied the faking strategies (i.e., 16 of 25 participants in Experiment 1, 6 of 8 participants in Experiment 2). To increase the proportion of participants actually implementing the faking strategy, in Experiment 3 we therefore added an extra reminder to do so in the break. Third, especially in Experiment 2, sample size was very small. In Experiment 3 were therefore preregistered a minimum of included participants.

Experiment 3

To examine whether a response deadline reduces faking effects in the CIT, Experiment 3 used a 2 × 2 between-subject design, providing guilty participants with faking instructions (or not), and using a response deadline (or not). As faking is typically implemented by guilty participants to avoid detection and since faking cannot influence the detection of innocent participants in a properly constructed CIT, Experiment 3 only concentrated on guilty examinees. Consequently, we did not perform any ROC curve analyses but concentrated on the individual Cohen's ds as dependent measures of the CIT-effect and thus the probe minus irrelevant difference in each individual. As in Experiment 1 and 2, large positive individual Cohen's ds thereby indicate probe recognition and can thus be used to identify guilty examinees. This experiment, including our power calculation and sample size was preregistered on https://osf.io/hq9ru.

Deviations from Preregistration

We preregistered to stop data collection after obtaining, for each cell of the design, a minimum of 42 included and a maximum of 60 tested participants (+-2) (see Method section). We did, however, not monitor the maximum tested participants per group during data collection, ending up with 70 tested participants in the response deadline/faking group (42 included participants). Excluding the data of these additional 10 participants does not change the pattern of results and in the interest of increased power, we therefore opted to include those in the analysis.

We did not preregister the Bayesian analyses, but we think they are informative. To keep the description of the 3 studies consistent, we report them also in the main analyses of Experiment 3. We also did not preregister the specifics of the questionnaire analyses, but report them as additional information comparable to Experiments 1 and 2.

Method

Participants. To inform our power analysis, we pooled the data sets of the guilty participants of Experiment 1 and 2. On this pooled dataset we ran a 2 (faking: yes vs. no) by 2 (response deadline: yes vs. no) ANOVA and found the effects size of the interaction to be Cohen's f = .28 for the RT CIT-effect. With a targeted power of .95 and an α = .05, this resulted in a required sample size of 168 participants (42 per group; calculated with G*Power 3.1). This served as the targeted group size (thus 42 included participants / group), and exclusions were monitored throughout data collection. Data collection was terminated once 42 included participants were available for each group.

In total, 217 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 mock crime. Forty-seven additional participants were excluded, 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). The final sample consisted of 168 participants (M_{age} = 23.49 years, SD = 5.29 years; 131 female, 37 male), with 42 participants in each group.

Procedure, tests, and faking instructions. The recruitment and instruction procedures were identical to Experiment 1 and 2, except that both envelopes contained mock crime instructions. Participants were nevertheless told the same as in the first two experiments that is that one of them would have to perform a mock crime and the other one a simple everyday activity. Both mock crime instructions were kept similar, except that the mock crimes had to be executed on different floors of the building (so that participants would again not see what the other participant was doing).

The instruction for the mock crime told participants to steal exam questions on a USB-stick. Therefore, they should leave the lab and go to the first floor/basement of the building. They were told that there they would find a cupboard in which *Prof.*

Magert kept all her old documents, including her previous exam questions. She always kept them on a *USB-stick* in a *folder* in the cupboard that was covered with a poster depicting a *dog*. In addition to taking and hiding the USB-stick, they should also remember the password (*oak tree*) written in the folder, as it was necessary to access the data on the stick. After they had finished this, they should return to the lab. Words marked in *italics* refer to the probe items used in the CIT.

Aside from the replacement of the items with the ones belonging to the new mock crime, the RT-CIT remained unchanged, with one version without the response deadline being identical to the one used in Experiment 1 and one version with the response deadline of 800 ms being identical to the one used in Experiment 2. Faking instructions were identical to the ones provided in Experiment 1 and 2, respectively, except that participants were now given a reminder to employ them in the self-paced break. At the end of the experiment, in addition to the questions employed in Experiment 1 and 2, participants in the Faking group were asked for their reasons why they may have had problems implementing the faking strategy (i.e., ratings on a scale from 1 to 5 of whether they forgot it, it was too tiring, too difficult, they were not motivated enough or they did not understand the strategy). Participants in the No Faking group were asked whether they did use any strategies to obtain an innocent test outcome and if so, which those were. Also, in order to check whether participants understood the faking instructions, at the end they were given an additional short quiz with 4 questions regarding the test principle and what the instructed faking strategies were.

Data analysis. Data were analyzed with R and raw data as well as the analysis script can be accessed on https://osf.io/t9y5d/. For the RT-CIT with response deadline, trials exceeding the response deadline (39.29%) were removed. For both RT-CITs, error trials (7.73%) and RT outliers (1.58%; RTs > 2.5 *SD*s from the mean per subject and item type) were removed. As in Experiment 1 and 2, individual Cohen's *d*s served as dependent measure. Those were analyzed with a 2 × 2 between-subject ANOVA with the factors faking (no vs. yes) and response deadline (no vs. yes). Interactions were followed up with Welch's *t*-tests.

Results

Questionnaire. On a scale from 1 to 5, participants rated their nervousness during the experiment with $M = 2.85$ ($SD = 1.12$) as low to medium, their motivation with $M = 4.45$ ($SD = 0.70$) as high, and test difficulty with $M = 3.11$ ($SD = 1.06$) as medium. The percentage of remembered items was high with 99% correct. Of the 84 guilty participants receiving faking instructions for the CIT, 76 indicated that they applied the faking strategy (90%; 39 in the CIT without response deadline, 37 in the CIT with response deadline). The ratings (all scales 1–5) revealed that participants who did not follow the faking instructions indicated little forgetting of the strategy ($M = 2.23$, $SD = 1.36$), some amount of strain (i.e., the strategy was too tiring, $M = 3.13$, $SD = 1.46$) and difficulty ($M = 3.13$, $SD = 1.89$) and little problems with motivation ($M = 2.13$, $SD = 1.59$) and understanding of the strategy ($M = 2.13$, $SD = 1.55$). Interestingly, participants not following the faking instructions did not differ significantly on any of those ratings from participants who indicated that they did implement them (all $ps > 0.155$), yet those comparisons are hindered by the large differences in group size. Sixty-one of the 84 participants could correctly repeat the faking strategy at the end of the experiment. A Pearson chi-square test did not reveal a significant relationship between not implementing the strategy and not being able to correctly repeat it, $\chi^2(1) = 0.16$, $p = .693$. The analysis of the sum of correct answers in the quiz (maximum of 4) did not reveal any difference between the faking ($M = 2.68$, $SD = 0.50$) and the no faking group ($M = 2.73$, $SD = 0.50$), $t(166.0) = 0.62$, $p = .536$, $d = .10$, and indicated that also the no faking group showed some understanding of the test principle.

ANOVAs. The mean *d* values of all four conditions are shown in Figure 1. We first conducted two one sample t-tests, testing the mean *d* values in both no faking groups against zero. As expected, both tests revealed positive and significant CIT effects with $t(41) = 15.50$, $p < .001$, $d = 2.39$ in the no response deadline CIT and $t(41) = 7.21$, $p < .001$, $d = 1.11$ in the Response Deadline CIT.[2]

The 2 × 2 ANOVA on the *d* values revealed a significant main effect of faking $F(1, 164) = 4.46$, $p = .036$, $n_p^2 = .03$, with higher *d* values in the no faking compared to the faking group. There was no significant main effect of response deadline, $F(1, 164) = 0.03$, $p = .856$, $n_p^2 < .01$. Critically, there was a significant Faking × Response Deadline interaction, $F(1, 164) = 29.55$, $p < .001$, $n_p^2 = .15$. As expected, Welch's *t*-tests showed a large faking effect when no response deadline was used, $t(64.87) = 4.96$, $p < .001$, $d = 1.08$, $BF_{10} = 4141$. There was no faking effect (in the expected direction) when a response deadline was used, $t(81.60) = 2.56$, $p = .012$, $d = -0.56$, $BF_{10} = 3.77$. Of note, this effect was significant, but in the opposite direction, revealing larger CIT-effects in the faking compared to the no faking group.

Discussion

Experiment 3 served the purpose of testing the robustness of the results of Experiment 1 and 2. It again suggested that the CIT can be faked when no response deadline is used, but also that a response deadline reduces the effect of faking. By tackling the issues of test order, actual implementation of faking strategies and statistical power as well as by comparing versions of the RT-CIT with and without response deadline within one experiment, Experiment 3 therefore was crucial to establish the validity of the findings of Experiment 1 and 2.

General Discussion

Although meta-analytic estimates revealed promising high effect sizes for the RT-CIT (Meijer, Verschuere, Gamer, Merckelbach, & Ben-Shakhar, 2016; Suchotzki et al., 2017), it has often been argued that RT-based measures are under volun-

[2] The effect size *d* calculated as the mean individual *d* values divided by the standard deviation of the individual *d* values.

tary control and therefore very easy to fake (Farwell & Donchin, 1991). The current paper aimed to add to the so far insufficient and partly inconsistent previous results by systematically investigating the fakeabilty of the RT-CIT and a potential faking prevention via a response deadline.

Experiment 1, where no response deadline was used, 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 or at least reduce the effects of 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. Experiment 3 also showed that comparing RT-CITs without and with response deadline directly, faking instructions did indeed enable guilty participants to obtain more innocent test outcomes in the former, yet not in the latter. Those results are encouraging as they implicate that modifying the structural properties of the CIT and restricting the time participants have to respond may hinder them to successfully implement faking strategies. This is promising, as in applied contexts, the motivation of suspects to attempt to implement faking strategies will be considerable higher than in our experimental context.

Unfortunately, however, using a response deadline came at a cost. First, excluding participants who did not strictly adhere to the response deadline resulted in a considerable loss of participants (Experiment 2: 39%; Experiment 3: 22%; compared to 2% in Experiment 1 where no response deadline was used). This exclusion seems critical, as an exploratory analysis in Experiment 3 showed that in the faking group with response deadline, excluded participants ($n = 28$; $M\ d_{CIT} = 0.32$, $SD = 0.39$) showed smaller CIT effects than included participants ($n = 42$; $M\ d_{CIT} = 0.53$, $SD = 0.34$), $t(52.09) = 2.35$, $p = .022$, $d = 0.58$, $BF_{10} = 2.90$. So while being effective in hindering faking, a response deadline restricts the utility of the RT-CIT. This may be particularly problematic when using the RT-CIT in forensic samples. In a sample of prisoners, Suchotzki, Kakavand, and Gamer (2018) found a very large RT-CIT effect for items encountered during an imagined mock crime but also observed relatively high mean RTs in the range of 750 ms to over 1000 ms for irrelevant and probe trials, respectively. Of note, in the respective study, a very lenient deadline of 2.5 s was used, leaving it unclear what the utility and validity of a 800 ms deadline is in a forensic sample.

Second, using a response deadline of 800 ms resulted in a considerable drop of test validity in the no faking groups. This was apparent both when comparing classification accuracy from Experiment 1 to Experiment 2 ($a = .97$ vs. $a = .76$) as well as comparing the effect sizes of the CIT effects in Experiment 3 (without response deadline: $d = 2.39$; with response deadline: $d = 1.11$). This is surprising as the use of response deadlines is common procedure in the RT-CIT, and several examples can be found in the literature where larger effects and higher classification accuracies of such CITs were reported than those observed in the current study (Kleinberg & Verschuere, 2015; Visu-Petra, Miclea, & Visu-Petra, 2012; Visu-Petra, Varga, Miclea, & Visu-Petra, 2013). But of course, none of those studies directly compared validity with and without a response deadline and following up this surprising aspect of our results would be an important task for future research. Such research could then also concentrate on calibrating response deadlines more carefully, with the goal of identifying one that hinders faking attempts without impairing validity.

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 or could not repeat the strategy correctly. Ratings in Experiment 3 indicated that applying them being too tiring and difficult are possible reasons for not applying the faking strategies, yet the informative value of this is restricted as those ratings were based on data of only 8 participants. In would be very interesting and important to investigate these reasons in future research. This is even more important, as participants were usually not asked whether they actually applied instructed faking strategies in previous research. Of note, excluding participants who indicated that they did not implement the faking strategies or that did not understand them weakened—rather than strengthened—the faking effect (which may also be due to reduced power), yet did not affect the crucial interaction of faking and response deadline in Experiment 3 (see supplementary material on https://osf.io/t9y5d/).

An interesting further avenue for future research is the investigation of faking detection algorithms. While such algorithms have been proposed for other lie detection tests (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011), none have been developed the RT-CIT. For the aIAT, Agosta et al. (2011) proposed that faking should manifest itself in the crucial test items but not in practice trials or items that are irrelevant for the test result. Our proposal would be to transfer this logic to the CIT by calculating a ratio of the mean RT for irrelevants and the mean RT for (task irrelevant) target items, expecting this value to be larger for guilty faking participants (due to their intentional slowing for irrelevant items) compared to innocent non-faking participants.

It is noteworthy that our faking instruction (i.e., deliberate slowing and speeding of responses) as well as the proposed faking algorithm concentrate on faking strategies at the level of participants' responses. Other faking strategies may influence performance already at an earlier stage. For instance, asking participants to suppress their memory of certain events or to vividly imagine having performed another action (e.g., the alibi) has been shown effective in RT-based tests and in tests using other measures (Dhammapeera, Hu, & Bergström, 2020; Gronau, Elber, Satran, Breska, & Ben-Shakhar, 2015; Hu, Bergström, Bodenhausen, & Rosenfeld, 2015; Suchotzki, Berlijn, Donath, & Gamer, 2018). Thereby, such strategies may affect the memories themselves rather than just participants behavior during the test and may therefore be more difficult to detect. Future research should directly compare the effectiveness of early faking strategies targeting the crime memory and late strategies targeting test performance via deliberate response adjustments as well as potential safeguards against the former strategies.

Despite its strengths, the current study is not without limitations. First, participants were recruited via the recruitment system of the University of Wuerzburg and were for the biggest part either undergrads or participants from the normal population. To gain more information on the applied use and fakeabilty of the RT-CIT, more studies with forensic populations are needed. Second, as mentioned above, exploratory analyses point towards a smaller CIT effect in included than excluded participants in the faking group that performed the CIT with response deadline. That reinforces the validity of the exclusion criterion. At the same time, we cannot exclude the possibility that the effect of the response deadline may also (at least partly) be attributed to selective exclusions. For instance, that the response deadline screened out inattentive participants or participants who managed to implement the faking strategy of a strategic slowdown, yet at the cost of too many trials exceeding the response deadline. This explanation is of course post-hoc and speculative, yet certainly deserves more attention in future research.

To sum up, the RT-CIT seems fakeable when no response deadline is given. It did, however, even then still produce classification accuracies above chance level. Most importantly, a response deadline hindered effective faking. By providing suggestions for the development of a faking detection algorithm, we hope to stimulate further research in this area.

Author Contributions

K.S. 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. B.V. and M.G. were involved in the study conception, the interpretation of the data and the critical revision of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

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

References

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

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. https://doi.org/10.1017/CBO9780511975196.012

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Lawrence Erlbaum.

Degner, J. (2009). On the (un-) controllability of affective priming: Strategic manipulation is feasible but can possibly be prevented. *Cognition and Emotion*, *23*, 327–354.

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), 92–101.

Dhammapeera, P., Hu, X., & Bergström, Z. M. (2020). Imagining a false alibi impairs concealed memory detection with the autobiographical Implicit Association Test. *Journal of Experimental Psychology Applied*, *26*, 266–282.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, *28*, 531–547.

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*, 312–319. https://doi.org/10.1016/j.neuroimage.2010.11.025

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.

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.

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

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PLoS One*, *10* https://doi.org/10.1371/journal.pone.0118715

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, *53*, 593–604.

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*, 286–298. https://doi.org/10.1111/j.1469-8986.2007.00615.x

National Research Council. (2003). *The polygraph and lie detection*. National Academies Press.

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test. *Applied Cognitive Psychology*, *27*, 328–335. https://doi.org/10.1002/acp.2910

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*, 205–219.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*, 772–780. https://doi.org/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 ğuilty knowledge. *Journal of Applied Psychology*, *85*, 30–37. https://doi.org/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*, 428–453. https://doi.org/10.1037/bul0000087

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

Suchotzki, K., De Houwer, J., Kleinberg, B., & Verschuere, B. (2018). Using more different and more familiar targets improves the detection of concealed information. *Acta Psychologica*, *185*, 65–71. https://doi.org/10.1016/j.actpsy.2018.01.010

Suchotzki, K., Kakavand, A., & Gamer, M. (2018). Validity of the reaction time concealed information test in a prison sample. *Frontiers in Psychiatry*, *9*, 745. https://doi.org/10.3389/fpsyt.2018.00745

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*, 991–1002.

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

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*, 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

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. https://doi.org/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*, 342–351. https://doi.org/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. https://doi.org/10.3389/fpsyg.2013.00146

Received 26 June 2020;
Received in revised form 16 October 2020;
Accepted 16 October 2020;
Available online 19 January 2021