[WORKING PAPER. Last Update: Jan. 09, 2020]

(For Tables & Figures, see end of this preprint)

What are you hiding? Initial validation of the reaction time-based searching

Concealed Information Test

Dave Koller^{1,2*}, Franziska Hofer^{3,4}, Tuule Grolig¹, Signe Ghelfi³, Bruno Verschuere²

¹Department of Psychology, University of Zurich, Zurich, Switzerland

²Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

³Zurich State Police, Airport Division, Research and Development

⁴brainability, Developing Human & Organizational Potentials, Zurich, Switzerland

* Corresponding author

Corresponding author

Dave Koller
Binzmühlestrasse 14 / Box 22
8050 Zurich
Switzerland
E-mail: d.koller@psychologie.uzh.ch

Abstract

The reaction time-based Concealed Information Test (RT-CIT) has been used to judge the veracity of an examinee's claim to be naïve about an incident by testing his/her memory of relevant details. Here, we explore the validity of the RT-CIT to generate new knowledge about the incident – the searching CIT. In a mock terrorism study ($n = 60$) the searching RT-CIT allowed to detect crime information well above chance, with sampling error possibly underestimating its true potential. A simulation study confirms the potential of the searching RT-CIT, and identifies conditions under which it best performs. We used an archival data-set that met these conditions (high CIT effect, large number of item repetitions), and found better item classification performance than in the mock terrorism study. Although the classification is far from perfect, the searching RT-CIT could be a new, promising investigative tool to reveal new crime details to the investigative party.

Keywords: memory detection, searching Concealed Information Test (CIT), deception, external validity, application

Introduction

By testing a suspect on crime information that only a perpetrator, a witness or a victim could have, the Concealed Information Test (CIT) also known as Guilty Knowledge Test (Lykken, 1959) can connect an examinee to knowledge about the crime.

To illustrate how the CIT can be used, imagine the following scenario: Two burglars broke into a storage hall of Pravay (a chemical plant) with the use of a crowbar. They stole large quantities of concentrated sulfuric acid that can be used to synthesize explosives. Based on low quality CCTV footage and a terror watch list, the police bring in a suspect for questioning. He denies all knowledge about the break in. With the information the police officers have about the crime, they can construct a *known solution* CIT by taking the true crime information (Pravay, crowbar, sulfuric acid; so-called *probes*) and adding plausible alternatives (company names, other tools often used to break in, different chemicals; so-called *irrelevants*). When asked about the crime, a naïve person cannot distinguish between the

Validation of the searching RT-CIT

probes and the irrelevant items and therefore does not show a systematic difference regarding the response to the stimuli. On the other hand, a knowledgeable person shows recognition of the probes and may attempt to hide that (Klein Selle, Verschuere, Kindt, Meijer, & Ben-Shakhar, 2017). These processes are typically accompanied by an increase in skin conductance response (SCR), in P300 amplitude, and in increased response times; all of which can be used to classify individuals into knowledgeable/naïve well above chance (e.g. Meijer, Klein Selle, Elber, & Ben-Shakhar, 2014; Seymour, Seifert, Shafto, & Mosmann, 2000). The CIT paradigm can therefore be used to detect past actions.

Detecting past actions is often not enough. In a real life scenario similar to the described break in, the police are not just interested in assessing whether the suspect may be involved in the burglary, but they are also, perhaps primarily, eager to prevent the attack. For that purpose, it would be helpful to get an answer to such questions as: Who is the second burglar? Are there more people involved? Where is the explosive synthesized and stored? Where and when do they intend to execute the attack? The police might have a list of critical infrastructure and possible targets with many casualties, but to act effectively with limited time and resources, the police needs to know the target of the upcoming attack. The approach to tackle this challenge is called the *searching CIT* as the police is searching for the probe amongst a set of probable alternatives. Contrary to the known solution CIT, the police do not know the crime information in the searching CIT. However, if the actual crime information is included in the CIT, a knowledgeable person does still recognize this information. The person tries to hide the knowledge which leads to the aforementioned effects (e.g. increased RTs). Based on the observed data, the searching CIT tries to classify each item as either being crime irrelevant or crime relevant (i.e. an irrelevant item or a probe item). The difference to the known solution CIT arises in the way the data is analyzed. One of the few single subject searching CIT studies was conducted by Meixner and Rosenfeld (2011) using EEG. Participants were assigned either to the guilty condition in which they were asked to plan a

Validation of the searching RT-CIT

terror attack with three testable crime information (type of attack: bomb; location: Houston; time: July) or the innocent condition in which they planned a vacation. These items were tested against five irrelevant items in each information category. In order to find the probes, for each participant, they compared the two items with the largest mean P300 amplitude within each category. If the difference (measured by comparing bootstrapped means) between these items was sufficiently big, it was concluded that the item with the largest mean P300 is crime relevant (probe), otherwise it was concluded that there is no probe item in this category and participant. This algorithm achieved a probe classification accuracy of 67% (chance performance was 20%). The technical requirements of EEG-systems and the well trained personal needed to administer an EEG are barriers to apply it in practice. A cheaper (no specialized equipment or personal needed), easier (automated data collection), but still unbiased (automated data analysis) alternative could be reaction time (RT) measures.

As far as we know, the present study is the first to explore the validity and applicability of the Reaction-Time based searching CIT (searching RT-CIT). We evaluate two searching algorithms for the scenario where the investigators themselves do not know what the critical details are (i.e. searching RT-CIT). In contrast to the common known-solution CIT where the examinee is tested for critical details that the examiner knows are related to the crime, we pretend to be ignorant about the crime and aim to classify the items as crime relevant/irrelevant and use this to classify participants as guilty/innocent in an airport setting using a mock crime paradigm (Study 1). Because Study 1 showed an effect that is likely to impact the performance of the searching RT-CIT algorithms negatively, an effect that can only be explained by chance, we used simulated data to examine the potential of the algorithms in the absence of any bias (Study 2). Finally, the algorithms were cross validated on an unbiased data-set (Study 3).

The two algorithms we evaluate are inspired by Meixner and Rosenfeld (2011), and Noordraven and Verschuere (2013). We expect above-chance classification performance for

the items (Hypothesis 1a) and in a second step for participants (Hypothesis 1b), on the basis of the item classification for both algorithms.

Study1: Applying the Searching RT-CIT in an Airport Setting

Study 1 used a mock crime paradigm at an international airport, with a guilty group that planned and partially executed a mock terror attack, and an innocent control group. Two searching RT-CIT algorithms were used to detect crime relevant information and to classify participants. A priori, we expected both algorithms to show above chance classification performance for items and participants, but we had no predictions when it came to comparing the two algorithms.

In order to draw conclusions about the searching RT-CIT in an airport setting, we first need to validate the known solution RT-CIT in that setting; an environment with high security standards (enforced by the police) that, in addition, is relatively unfamiliar to participants and therefore likely to cause higher agitation levels in all participants than a laboratory setting at a university does. Thus, we predict a larger standardized probe-irrelevant difference in RTs ($\frac{M(\text{probe}) - M(\text{irrelevant})}{SD(\text{irrelevant})}$) as introduced by Noordraven and Verschuere (2013) and here forth called *CIT-effect*) in the guilty vs. innocent group (Hypothesis 2a). In a similar vein, we expect the guilty and innocent classification accuracy based on the CIT-effect to be greater than 50% (Hypothesis 2b).

As a secondary aim, Study1 also investigated potential effects of richer memory traces of past actions compared to intentions (e.g. Cohen, 1981) on the CIT-effect. Although it has been shown that reaction times (Noordraven & Verschuere, 2013) can be used to detect intentions with the CIT, it is unknown if there is a difference in how well past actions and intentions can be detected using RTs. The insight we gain is of high practical relevance as it will show if the RT-CIT is suitable for exposing planned criminal actions before the crime is committed which is especially important in the context of terrorism.

Method

The experiment was approved by the ethical committee of the Faculty of Arts and Social Sciences of the University of Zurich (Approval number: 2018.2.11). The study is exploratory,¹ data and code can be found on osf.io/69yrj.

Participants

Participants were sixty undergraduate students from University of Zurich (M age = 22.5 years; SD = 3.1 years, range 19-32 years, 47 female). To end up with a balanced design, we recruited until we had 60 participants after applying the preregistered exclusion criteria. Of all the tested participants (n = 68), 8 were excluded (1 due to poor performance in the task [more than 50% errors in at least one item category], 6 exceeded the two-error-limit in the post-CIT recognition task, and 1 participant failed both criteria). Participants were recruited via participants' mailing-list, postings on bulletin boards at the university and advertisements in lectures. The participants were enrolled to the study when the following inclusion criteria were met: age between 18 and 35, high school degree or higher, and fluent in German. Before the experiment started all participants were asked to read and sign the informed consent. It was clearly stated that the participation is voluntary, and withdrawal is possible at any time during the course of experiment with full compensation. All participants received 20 CHF ($\approx$ 20.40 USD) or course credits for 1.5 hours of study participation (participants' choice). All participants were told that they have a chance to earn an additional 5 CHF ($\approx$ 5.10 USD) by appearing innocent (both groups) until the instructed task (described in the next chapter) is completed. However, independent of their performance, all participants received the additional 5 CHF.

Half of the participants were asked to plan for a mock-crime (guilty suspects; n =30; M age = 22.10 (SD = 2.80); 24 female; 23 right-handed). The other half was asked to plan for a

¹The study was preregistered (osf.io/69yrj/) but that preregistration was premature and the authors decided to analyse the classification based on adaptations of already existing algorithms and refrained from calculating the preregistered Bayesian index I . Since this is an integral part of the study, it should be considered exploratory.

Validation of the searching RT-CIT

non-criminal act (innocent suspects; $n=30$; M age = 22.83 ($SD = 3.34$); 23 female, 27 right-handed).

Procedure

The experimental procedure consisted of four phases - *planning phase*, *execution and interception*, *RT-CIT*, and *target and probe recognition*.

Planning Phase

All participants were contacted by e-mail and requested to bring a self-packed cabin bag for a day trip with an airplane. It was explicitly stated that the bag must not contain any forbidden items. The participants were informed that new security measures and communication protocols between different divisions of the airport police had been introduced. Participants were told that this study is part of an airport security check to test these measures. Upon arrival to the airport, an experiment leader welcomed the participant and brought him/her to an office room where they read the instructions on a sheet of paper. The participants in the guilty group were told to take part in a mock-terror attack to test the newly implemented security measures. The innocent group was given no additional information beside that they are to test the efficacy of the new security protocol. All participants received mock flight documents issued to their name and a map of the airport. They were given up to seven minutes to plan their tasks.

The participants in the guilty condition were instructed to go to a location marked on the map where they will find an envelope with a code word on it. The envelope contained a numbered key to a safe deposit box at the airport in about 5 min walking distance. In the safe deposit box, they found two items which they should take with in their hand luggage. Item 1 was to be smuggled through the security check and handed over to a confederate (airside), whereas Item 2 was to be used as a sign to be recognized by the confederate. The purpose of each item was explained in the instructions (see Appendix A). The guilty participants were instructed to carefully hide the items in their luggage in order to avoid detection. After

Validation of the searching RT-CIT

crossing the security check, they were asked to a specific gate, take out Item 2, and wear it or hold it visibly until they are approached by the confederate which they will recognize by a certain feature (e.g. red shoes) and hand over Item 1. The guilty participants were also instructed that in case nobody approached them within five minutes, they should put the item at a certain place and then meet the experiment leader at a location described in the instructions (see Table 1 for all the items). To control for possible item effects, the information used in the instructions was balanced over the course of the experiment in the sense that within each category, every item was used exactly five times.

The innocent participants were instructed to go to four specific stores at the airport that were indicated by numbers on a map they received. In each store they were to select an item and write down its name and price on a sheet of paper received from the experiment leader. After completing their task, the participants were asked to use the provided flight documents and go through the security check. Thereafter, they were told to have 15 minutes waiting time to be filled as they wanted. Next, the participants were requested to meet the experiment leader at the baggage check. This procedure ensured that innocent participants had to complete a task that took a similar amount of time as the mock crime, that lead them to the same areas of the airport, and that also included orienting and planning based on a map and navigating through the airport preferably relying on their memory.

All participants (innocent and guilty) executed their tasks as planned up until they were about to use their flight documents (enacted tasks). The activities that should have followed after the security check were considered planned future acts (intentions).

Execution & Interception

All participants carried out the tasks up to the point when they should use the flight documents to proceed to the security check. Both guilty and innocent participants were intercepted by a confederate posing as an undercover police officer and brought to an office room where they were asked to complete a security test (RT-CIT). The items the guilty

Validation of the searching RT-CIT

participants encountered before this interception (code word, number on the key, item 1 and item 2) were considered to be part of the enacted tasks whereas gate, recognition feature, item placement, and meeting spot were considered as intention-items.

RT-CIT

The RT-CIT was programmed with MATLAB version 9.4.0 (The MathWorks, 2018) with the Psychtoolbox extension version 3.0.14 (Brainard, 1997) on a Dell Latitude E6530 with a 15.6" screen running on Windows 7 (Service Pack 1). Participants were seated approximately 50cm from the screen.

In the beginning, participants were asked to learn eight target items to which they should respond with “YES, this is connected to the crime” in the subsequent task. Participants could end the learning phase on their own as soon as they felt well prepared. They were then presented with all 48 items of the test and were asked to click on the target items with the mouse to ensure that the items have been memorized. The selected items were highlighted to help the participants to keep track of their choices. Participants with more than one error in this recognition test received feedback that they have made more than one error and that they should learn the target items again. This procedure was repeated until no more than one error was made.

After passing the target recognition test, the reaction time trials started. On each trial, a single item was presented in the middle of the screen, either a probe, a target or an irrelevant item. Participants were asked to indicate as fast and as accurately as possible whether the item has a connection to the crime by pressing either “e” or “i” on the keyboard. The answers were assigned in a way that NO was always pressed with the participants’ dominant hand. The response-stimulus-interval varied randomly from 500-1000ms.

There were three practice blocks of 32 trials each to familiarize the participants with the task before the actual test trials started. Every item was presented twice in this practice phase. The first practice block had a response time limit of 10 seconds. The items remained

Validation of the searching RT-CIT

on screen until the participant pressed a button or the time limit was reached. For the second and third practice phases, participants were instructed to respond within 1.2 seconds and 0.8 seconds respectively. A red “TOO SLOW” message appeared above the stimulus if the participants did not respond within those instructed time intervals. However, every response given up until 1.5 seconds after stimulus presentation was recorded. Participants also received feedback on whether they responded correctly (i.e. YES to targets, NO to probes and irrelevant) by indicating errors with the display of a red “X” below the stimulus. The actual test phase consisted of twenty blocks. In each block, every item is presented once, resulting in 960 trials in total which took approximately 23 minutes to complete. Participants were given the opportunity for two short, self-paced breaks after blocks eight and fifteen.

Target and probe recognition

After the RT-CIT ended, there was another target recognition test to ensure that participants did not forget the targets during the test. Additionally, participants in the guilty group had to complete a probe recognition test. They were told that only the guilty participants see this test, that it is needed to evaluate the study properly, and that they should therefore answer truthfully.²

Searching CIT Algorithms

In contrast to the known solution CIT, we assume to be ignorant of the guilt of the participant and about the items that were involved in the crime. We applied two searching-CIT algorithms aimed to classify each item as relevant/irrelevant and each participant as guilty/innocent.

Standardization algorithm

² The innocent group did not complete this recognition test because the probes and targets were counterbalanced. However, innocent participants were presented with all the items and were asked to indicate the most plausible one for each category. The only category that showed a significant after correcting for multiple comparison (Bonferroni).

Conceptually, the first algorithm that we tried is one where every item is treated as the possible probe and compared to all the other items in its category. Items with CIT-scores above a certain cut-off are classified as probes, the others as irrelevant. Based on the idea of standardizing the probe-irrelevant difference within a participant (Noordraven & Verschuere, 2013), this algorithm uses within category standardized CIT-scores. For each participant and every item i in item category j , the CIT-scores are calculated as $dCIT_{i,j} = (M(RT_{i,j}) - M(RT_{k \neq i,j})) / SD(RT_{k \neq i,j})$. For participant-classification, we use the mean of the largest $dCIT_{i,j}$ scores of each category $dCIT_p = M(\max(dCIT_{i,j}))$ and classify a participant as ‘guilty’ if $dCIT_p$ is above a certain threshold and as ‘innocent’ otherwise.

1st to 2nd bootstrap algorithm

The second algorithm we applied on our RT data has been successfully used in the P300 CIT (Meixner & Rosenfeld, 2011). The rationale behind this algorithm is that the probes should have the largest RT and that they should only be classified as probe if the difference to the item with the second largest RT is sufficiently big. In a first step, the item with the largest mean RT in each category is identified for every participant. These items are considered possible probes. The item with the second largest RT is considered irrelevant and will be used for comparison. All other items are also considered irrelevant, but they are ignored for the rest of the algorithm.

In a second step, 2000 bootstrap sample means are calculated for the possible probe and the presumed irrelevant for each category. A sample is created by drawing (with replacement) as many RTs from the responses to a given item as there are valid trials for that item. In each of these 2000 iterations, the mean RT for the possible probe item is compared to the mean RT of the irrelevant item. The possible probe is then classified by the percentage of iterations where its mean RT was larger than the mean RT for the irrelevant item. Participants are classified based on the average of those percentages over all item categories.

Results

Validation of the searching RT-CIT

Target trials, trials with response errors, with unusually slow (i.e., 1500ms or more) or unusually fast (i.e., 150ms or faster) response times were excluded from the analysis.

Known-solution group analysis

Before we tried the searching CIT algorithms, we verified that there was a larger RT-CIT effect in the guilty condition than in the innocent condition. A one-sided Bayesian independent samples t-test on the CIT-effect between the guilty and the innocent group (using a weakly informative Cauchy prior; scale = .707) was used to compare the hypothesis H_1 (larger CIT-effects for the guilty group compared to the innocent group) to H_0 (no difference in the CIT-effect between the two groups). The test revealed very strong evidence for H_1 ($BF_{10} = 2.82 \times 10^6$) with a between group effect size $dCIT_{\text{between}} = 1.76$ (95% credible interval [1.09, 2.28]) showing that the guilty group ($M dCIT_{\text{within}} = .36$; $SD = .22$) has larger within group CIT-effects than the innocent group ($M dCIT_{\text{within}} = .04$; $SD = .14$).³

Known solution participant classification

We plotted the receiver operating characteristics (ROC) curve to assess if $dCIT$ can be used to discriminate between guilty and innocent participants. The area under the ROC curve (AUC) is an often-used index of diagnostic power (Fawcett, 2006). The AUC was 0.91 (95% CI: [0.84, 0.98]) and well above chance level. We used leave-one-out cross-validation (LOO CV) and applied the cut-off that maximizes the Youden's J statistic ($J = \text{sensitivity} + \text{specificity} - 1$; Youden, 1950) in the model building sample for individual classification. This procedure achieved a cross-validated classification accuracy of 85% (25 of 30 or 83% of guilty participants and 26 of 30 or 87% of innocent participants were classified correctly).

The reason Youden's J was used is that it is not biased by the base rate of guilty people in the population under investigation because it weights an increase of 1% in

³ CIT-effects were calculated as: $dCIT_{\text{between}} = \frac{M(dCIT_{\text{guilty}}) - M(dCIT_{\text{innocent}})}{\sqrt{\frac{var(dCIT_{\text{guilty}}) + var(dCIT_{\text{innocent}})}{2}}}$; $dCIT_{\text{within}} = \frac{M(RT_{\text{probe}}) - M(RT_{\text{irrelevant}})}{SD(RT_{\text{irrelevant}})}$

Validation of the searching RT-CIT

sensitivity and a 1% increase in specificity equally. To illustrate this, let us assume 10'000 people with a guilty base rate of 1% (i.e. 100 guilty, 9900 innocent people) should be classified. A 5% increase in sensitivity will classify five additional guilty people as such which is an increase in accuracy of .05% whereas a 5% increase in specificity will classify 495 additional innocent people correctly which is an increase of 4.95%. However, Youden's J will in both cases increase by .05. Why this is a desirable property becomes evident when comparing classifiers from different scenarios: A naïve classifier that classifies everyone as innocent would reach 99% accuracy (but $J = 0$) in this example – an almost perfect classifier with 100% true positive and 2% false positive would reach 98% accuracy with $J = .98$. The same two classifiers with a guilty base rate of 50% would achieve accuracies of 50% and 99% while J remains unchanged.

Task enactment: Past versus future behaviour

A two-tailed Bayesian paired samples t-test with a weakly informative Cauchy prior (scale $r = .707$) was conducted to compare the CIT-effects of guilty participants in enacted and intent items using JASP (JASP Team, 2019). We found moderate evidence ($B_{01} = 4.91$; 95%-credible interval $[-.52, .39]$) that the data are more likely under the null-hypothesis stating that the CIT-effects between enacted and intent items do not differ than under the alternative hypothesis that CIT-effects do differ.

Searching CIT

Item classification performance of the searching algorithms was assessed by the Youden's J at the optimal cut-off and at a sensitivity of .9. As explained above, accuracy is not a suitable measure when the base rates are very different from .5. A naïve classifier (e.g., 'all items are irrelevant') would achieve 90% accuracy because 90% of the to-be-classified items belong to the irrelevant category. Also, the AUC used in the known solution participant classification cannot be used with the 1st to 2nd bootstrap algorithm because only one item in each category is classified based on a criterion. This means that no matter how liberal or

Validation of the searching RT-CIT

conservative the criterion is set; the algorithm will never reach a sensitivity or specificity of 1; which is a prerequisite to interpret the AUC.

The standardization algorithm achieved a Youden's J of .40 (sensitivity = .69; specificity = .71) in classifying the items as probes. 166 of 240 probes and 1537 of 2160 irrelevant items were classified correctly. At a pre-set sensitivity of .9, the algorithms specificity was .33 resulting in J = .23.

Participant classification was above chance with AUC = .68 (95% CI: [.54, .82]) but significantly worse than the known solution CIT (DeLong's test for two ROC curves: $D = -2.98$; $p < .01$; Robin et al., 2011). LOO CV resulted in a classification accuracy of 65% (19 of 30 guilty participants and 20 of 30 innocent participants were classified correctly).

The 1st to 2nd bootstrap algorithm achieved a Youden's J of .34 (sensitivity = .50; specificity = .84) in the probe classification task. 119 of 240 probes and 1818 of 2160 irrelevant items were classified correctly. This algorithm cannot achieve any arbitrary sensitivity or specificity; it strongly depends on the number of the probes that are selected as possible probes in the first step of the algorithm (in the current study, 120 of 240 probes). The algorithm could therefore not be evaluated at a sensitivity of .9.

Participant classification was above chance level with an AUC of .74 (95% CI: [.61, .87]) but significantly worse than the known solution CIT (DeLong's test for two ROC curves: $D = -2.33$; $p = .02$; Robin et al., 2011). LOO CV resulted in a classification accuracy of 68% (18 of 30 guilty participants and 23 of 30 innocent participants were classified correctly).

Discussion

In Study 1, we conducted a CIT in an airport setting. The known solution CIT - to investigate if items related to an enacted task show larger CIT-effects than items related to intentions - showed larger CIT-effects for guilty than for innocent participants with a classification performance (85% accuracy) well above chance. We found no evidence for an

Validation of the searching RT-CIT

effect of enactment (i.e. differences in the CIT-effect for items related to enacted tasks and items related to future intentions).

For the first time, we showed that RTs can be used to find crime relevant information and distinguish knowledgeable from naïve participants using the searching RT-CIT. Both searching algorithms achieved item and participant classification on a similar and above chance level. While encouraging and providing initial evidence for the validity of the searching RT-CIT, and therefore new applications, it was also evident that the known solution CIT is substantially more accurate in classifying participants.

This study demonstrated that item classification based on RTs is possible, but the level of classification performance should be interpreted cautiously. The small CIT-effect that the innocent group showed for the intent items (see Table 2.) indicates that, by chance, some items were particularly relevant to some participants - over and above item effects (e.g. word length, saliency), as these effect would be cancelled out by the balancing and matching we used. It is therefore likely that the searching CIT classification performance found in study 1 is an underestimation of how these algorithms could perform in the absence of this chance effect.

Study 2: Simulation Study

The CIT-effect in innocent participants for the intent-items in study 1 can only be explained by chance. If the CIT contains items that are specifically relevant to some innocent participants, the searching CIT's classification performance is likely to be influenced negatively. To explore the searching RT-CIT's potential without an innocent bias and under various conditions, we ran a simulation study. We simulated a wide array of data-sets that varied along the dimensions of CIT-effect size (eight levels: $dCIT_{within} = .2, .3, .36, .4, .45, .5, .6, 1$), number of item categories (i.e. information that could be tested; eight levels: 1-8), and number of trials per item (five levels: 5, 10, 20, 50, 100). This resulted in a total of 320 ($8*8*5$) data-sets with 1000 simulated participants each (500 guilty, 500 innocent). The effect

Validation of the searching RT-CIT

sizes $dCIT_{within} = .36$ and $dCIT_{within} = .45$ were simulated to compare the algorithms' performance on the simulated data to their performance on empirical data with the same effect sizes. The effect size of $dCIT_{within} = .36$ was the CIT-effect found in Study 1, $dCIT_{within} = .46$ ⁴ corresponds to the CIT-effect found in an independent study (Verschuere & Kleinberg, 2016) whose data will be used to validate the simulations in Study 3.

Data Generation

The data was generated on the trial level with the following assumptions: (1) People differ in their baseline reaction times. As an estimate of baseline RTs, we used the reaction time of innocent participants to irrelevant items. (2) Knowledgeable participants differ in their response to the probe (e.g., due to different perceived salience of the probes, or different ability to suppress the initial YES response to probes) which results in different CIT-effects in knowledgeable participants. (3) Innocent participants do not recognize the probe and therefore do not show a CIT-effect. (4) Reaction times on every trial are influenced by unsystematic noise. For now, we did not model item-effects or effects of item category.

Following these assumptions, the response time for participant i on trial j is generated by adding the different components. For innocent participants and for irrelevant items of guilty participants this results in $RT_{i,j} = \text{baseline } RT_i + \text{noise}_j$ and for probe trials of guilty participants it is $RT_{i,j} = \text{baseline } RT_i + \text{CIT-effect}_i + \text{noise}_j$. Reaction times of targets were not simulated as they do not influence the analysis.

Results

Figure 1 presents the searching CIT algorithms' performance on simulated data with the effect size found in Study 1 ($d_{within} = 0.36$) but without a bias for innocents or item effects (e.g. word length, numbers versus words). The simulations show a mean maximized Youden's J of .47 (sensitivity = .68, specificity = .79) for the standardization algorithm when

⁴ $dCIT_{within} = .45$ is based on initial calculations that contained a mistake in data aggregation. The correct effect size is $dCIT_{within} = .46$ but we refrained from running new simulations because the difference is minimal.

Validation of the searching RT-CIT

each item is presented twenty times compared to the empirically observed J of .40 in Study 1. The 1st to 2nd bootstrap algorithm achieved a mean maximized J of .45 compared to the empirically found J of .34 (sensitivity = .60, specificity = .85). This implies that with the same effect size and under better conditions (i.e. optimal items and no bias for innocents) the algorithms are expected to perform better than what we found in Study 1.

At least four conclusions can be drawn from the simulation outcomes. First, like observed in Study 1, the simulations indicate that, at the maximized Youden's J, the standardization algorithm is more liberal (i.e. is more likely to categorize items as probes). This results in a somewhat higher sensitivity but also a somewhat lower specificity than the 1st to 2nd bootstrap algorithm. Second, across all the simulations (see Appendix B) the standardization algorithm tends to outperform the 1st to 2nd bootstrap algorithm. Third, evidently, the algorithms perform better when the dCIT_{within} is larger (Figure 2). Fourth, the simulations indicate that both algorithms would profit from increasing the number of repetitions per item.

Discussion

The goal of Study 2 was to explore the searching algorithms' performance in the absence of bias and item effects under various conditions. Both algorithms show very similar benefits from more repetitions per item and from larger CIT-effects.

The main differences in item classification performance between the two algorithms root in the sensitivity limitations of the 1st to 2nd bootstrap algorithm. Its sensitivity cannot exceed the proportion of probes that has been marked as possible probe in the first step of the algorithm, no matter how liberal the criterion in the second step.

Direct comparison between the empirical data and the simulated data with the same effect size showed that the innocent bias and item effects influenced the searching algorithms negatively. The performances based on the simulated data should therefore be considered estimates of the ceiling performance. However, because the empirical data was confounded

by an innocent bias, it cannot be concluded that the simulated ceiling performance is a realistic estimate in the absence of this bias. The simulations need validation on an unbiased data-set.

Study 3: Validation on Archival Data

The simulation study showed that both algorithms have the potential to perform better than what we found in Study 1 when the innocent bias is eliminated, the number of item repetitions is increased, and the CIT effect is larger. To validate the algorithms on a second data-set and to show that the simulations yield realistic results, we applied the analysis from Study 1 on the autobiographical RT-CIT data from Verschuere and Kleinberg (2016) that fits two of the three optimized conditions identified in Study 2 (i.e., no bias, larger CIT effect but the same number of repetitions). These results were compared to those of the simulated data with the same effect size.

Method

For a detailed description of the method we refer the reader to Verschuere and Kleinberg (2016). The data can be found on osf.io/cg5es. In brief, an autobiographical RT-CIT was used with five item categories (first name, last name, university course, birthday, country of origin) and twenty trials per item for a total of 600 trials. Participants that were instructed to hide their identity showed a mean CIT-effect of $M(dCIT_{within}) = .46$ ($SD = .23$), whereas unknowledgeable participants had a CIT-effect of $M(dCIT_{within}) = -.01$ ($SD = .13$).

Results

Maximizing the Youden's J of the searching algorithms for the empirical data of Verschuere and Kleinberg (2016) resulted in a Youden's J of .53 (sensitivity = .83; specificity = .69) for the standardization and a J of .48 (sensitivity = .61; specificity = .87) for the 1st to 2nd bootstrap algorithm. Therefore, both algorithms showed above chance classification performance. We also obtained further indications that the standardization algorithm is more liberal (at the cost of lower specificity) and that it shows slightly better overall

discriminability. Figure 3 visualizes the performance of the algorithms in comparison with the expected performance based on simulated data.⁵

Participant classification based on the searching algorithms was with $AUC = .68$ (95% CI: [.56, .81]) for the standardization and $AUC = .69$ (95% CI: [.56, .81]) for the 1st to 2nd bootstrap algorithm above chance for both algorithms.

Discussion

The aim of Study 3 was to validate the searching CIT algorithms on an independent, unbiased data-set, and to show that the simulations yield realistic results. As predicted by the simulations, both algorithms showed better item classification than in Study1. Furthermore, the finding from Study 1 that the standardization algorithm is more liberal when the optimal cut-off is used was replicated.

Study 3 therefore showed the validity of the searching algorithms on an independent data-set and presented additional evidence that our data simulation can be used to estimate the optimal performance those algorithms can realistically achieve given a certain effect size.

General Discussion

The present study used the RT-CIT in a mock-terror attack scenario at an international airport and explored the potential of two searching algorithms to reveal critical information about the attack and to classify participants.

We first showed that the known solution RT-CIT can be applied in an airport setting with a high classification accuracy of 85%. The between group effect size of $d_{\text{between}} = 1.76$ is in line with the effect sizes reported in a recent meta-analysis for RT-CITs using relevant stimuli (Suchotzki et al., 2017). High security standards, police presence and the unfamiliar environment of the airport, that possibly lead to higher agitation levels than studies conducted in a university setting, did not seem to influence the validity of the RT-CIT negatively.

⁵ Applying the searching algorithms on simulated data with the same effect size ($M(d_{\text{CIT}_{\text{within}}}) = .45$) resulted in $J = .56$ (sensitivity = .74; specificity = .81) and $J = .54$ (sensitivity = .68; specificity = .86) for the standardization and 1st to 2nd bootstrap algorithm respectively.

Especially in the terror context, the police are interested in detecting malicious intent to prevent the attack. To investigate if intentions can be detected to the same degree as past actions, we compared the CIT-effect of items that the participants physically interacted with to items related to their intentions. We found moderate evidence that the CIT-effect is not influenced by enactment. These results could be explained in different ways: It could be that the richness of the memory trace (if the memory of enacted and intent items is sufficiently strong) indeed does not influence the CIT-effect. An alternative explanation could be that the effect was masked by increased focus on the intent items as they were still relevant to execute the mock crime successfully. Theoretically, because the mock crime scenario did not allow us to balance the items between the past action and intent condition, this finding could also be a result of the item selection. Although we cannot definitely conclude that enactment does not influence the CIT-effect, our results provide further evidence that the CIT is well suited to detect memory of past actions and intentions.

Finally, we set out to investigate whether response times can be used to reveal new crime details to the investigative party. Study 1 showed that searching CIT algorithms can be used to identify crime relevant information above chance level. The standardization algorithm showed slightly higher discriminability, but the main difference was that its sensitivity was higher at the cost of lower specificity compared to the 1st to 2nd bootstrap algorithm. However, this only applies when the algorithms are evaluated at their maximized Youden's J. When the criterion in the standardization procedure is set to match a certain sensitivity or specificity of the bootstrap procedure, they both achieve the same performance. Furthermore, the searching CIT achieved above chance classification performance of participants into guilty/innocent but both algorithms were considerably worse than the known solution CIT. Gathering useful information before testing a suspect is therefore still needed to get the most accurate guilty/innocent classification.

Validation of the searching RT-CIT

In order to explore the algorithms' potential without biases and item effects, we turned to simulated data. Whereas our simulations shed light on what would happen with different numbers of trials and different effect sizes, it is limited in the number of factors it takes into account and currently disregards known moderators of the CIT effect (e.g., motivation, saliency, countermeasures; Suchotzki et al. (2017)). Both algorithms show very similar benefits from more repetitions per item and from larger CIT-effects. The simulations further indicated that the searching RT-CIT could achieve substantially better classification performance given the right conditions. Optimization of the paradigm to increase the CIT-effect is very challenging and will take time, but testing the validity of the RT-CIT with large numbers of trials and investigating the effects of fatigue and habituation might give valuable insight that could be incorporated quickly. In addition, this knowledge could be used to refine the simulations. Especially in the exploration phase of this new field of searching algorithms in the RT-CIT, data simulation could be a valuable tool to explore the properties of algorithms. Using simulated data to explore the behaviour of a system (e.g. algorithm, computational model) in a wide array of conditions is well established in cognitive psychology (Sun, 2008) and could be a promising direction for CIT research in general.

The validation of the results from our simulations using independent data is further evidence that our data simulation can be used to estimate the optimal performance those algorithms can realistically achieve given a certain effect size. The remaining discrepancy between the performance on the simulated and empirical data is most likely due to item effects in dimensions that are likely to influence the CIT-effect or response times in general, such as saliency (Kleinberg & Verschuere, 2015; Verschuere, Kleinberg, & Theodoridou, 2015), word length (Barton, Hanif, Eklinder Björnström, & Hills, 2014), and word frequency (Rayner & Duffy, 1986).

In general, both algorithms show very similar classification performance with a slight advantage for the standardization algorithm, especially with small number of repetitions

Validation of the searching RT-CIT

per item. Possibly the most relevant difference between the algorithms is that the classification criteria of the standardization algorithm can be set freely to achieve any desired sensitivity/specificity, whereas the 1st to 2nd bootstrap algorithm's sensitivity is limited to the proportion of probes that are considered "possible probe" in the first step. How the criterion should be set in practice is determined by the circumstances. If a high sensitivity is needed (e.g. terror prevention) the criterion is set lower than in scenarios with very limited resources that must not be spent on false alarms – a flexibility that is not achieved by the 1st to 2nd bootstrap algorithm.

Another important difference between the algorithms is the susceptibility to an irrelevant item showing a CIT-effect. This could be due to an involuntary reaction of a participant to an item or it could be part of a countermeasure used by guilty participants. While both algorithms are expected to be affected in a similar way for innocent participants, the effect for guilty participants can be different. Let us assume a CIT-effect of that irrelevant item is of same size as the actual probe. For the standardization algorithm, this would result in the same $dCIT_{i,j}$ scores, but they would still be larger than zero and therefore diagnostic (although smaller than without the irrelevant signal because the difference of the means decreases and the SD increases; see Study 1 for the formula). Using the optimal cut-off, both items would be classified as probes while the other irrelevant items of guilty and innocent participant would still be classified correctly. Furthermore, $M(\max(dCIT_{i,j}))$ could still be used to classify participants. The 1st to 2nd bootstrap algorithm would, in the first step, only treat half of the real probes as "possible probes" limiting the sensitivity to a maximum of .5. The bootstrap comparison in step two would take place between the real probe and the irrelevant item that showed the same CIT-effect, yielding a mean of 50% - the same as when two irrelevant items of an innocent participant are compared. In this case, the second step will not improve the classification. The best performance would be reached when every possible probe is classified as the probe. Participant classification, however, would not be possible

Validation of the searching RT-CIT

since this guilty participant would show no bootstrap difference, just like innocent participants.

From a practical point of view, the standardization algorithm has the advantage that it takes less computational resources and therefore less time, since it does not rely on bootstrapping. For those reasons, we conclude that the standardization algorithm has more desirable properties and should currently be favoured over the 1st to 2nd bootstrap algorithm.

Applicability

The results from the empirical data and the simulations suggest that the RT-CIT is suitable not only to test if someone possesses specific crime knowledge (known solution CIT) but also to find unknown crime information among plausible but crime unrelated alternatives (searching CIT). The known solution RT-CIT could already be applied in specific situations such as testing a suspect for crime knowledge as part of a police investigation, at the border when the police have suspect that the country of origin person provided is wrong and the police have a specific suspicion, or possibly testing the knowledge about a substance found in a passengers' luggage which the passenger claims not to have packed. Our finding now suggests, that it can also be used to test for intentions such as plans for a journey or a terror attack.

The searching algorithms open up an additional spectrum of scenarios in which the RT-CIT can be applied. In the context of an investigation, for example when the police caught someone carrying illegal substances, they could use the searching RT-CIT to narrow down or prioritize where to look for the seller; or in a situation where the police have some information about a planned terror attack but do not know where it will take place, but they have a suspect that they believe to have knowledge about the attack. It could be used to get hints on where the attack will take place, what kind of bomb to look for, the day of the attack and alike. Although not addressed in this study, the searching RT-CITs performance can most likely be increased if multiple people sharing the same crime knowledge, as it has been done

with the physiological CIT (e.g. Breska, Ben-Shakhar, & Gronau, 2012; Breska, Zaidenberg, Gronau, & Ben-Shakhar, 2014; Elaad, 2016), can be tested. This reduces the impact of one person showing a distinct reaction to an irrelevant item for any reason.

Limitations

Although we used a highly realistic scenario in Study 1 by getting participants to the airport, making them execute the mock attack in a high security environment with real police present, and a believable cover story, three important aspects are very different from a real life scenario. 1) Apart from not getting the monetary bonus of 5 CHF, there were no negative consequences being classified as guilty. In reality, this would be an extremely high stakes crime and the suspects would be very motivated not to be classified as guilty. Although high stakes crimes need more investigation, there is evidence that motivation does not affect the validity of the RT-CIT (Kleinberg & Verschuere, 2016). 2) The participants were given instructions about the mock crime, learned them and then planned the execution. Planning the attack from scratch and considering possible alternatives might impact the CIT-effect of alternatives that were considered but not chosen, which might influence the classification performance of both the known solution and the searching CIT. 3) We used a student population, which is not representative of the general population.

The sample size of $N = 60$ of Study 1 was not enough to find conclusive evidence for the null hypothesis that stated that there is no effect of enactment, even though the difference in the CIT-effect was minimal. Although the results are promising, studies with larger sample sizes are needed to reach conclusive evidence on the matter.

Study 1 & 3 showed that the searching algorithms' performance on the simulated data was above what was observed on empirical data. This was expected, considering the way the data was simulated (e.g. no item effects). The simulations give estimates of the ceiling performance of the algorithms. Allowing for item effects and possibly fatigue or habituation with large number of repetitions should give more accurate estimates. More detailed

simulations and validations on other independent data-sets should be performed to validate the results of this exploratory study.

Future Studies

Searching algorithms on RT-CIT data is a research field that remains to be explored. We encourage other researchers to develop new algorithms and use the simulated data to explore boundary conditions. Promising research directions include non-binary classifications (i.e. providing a measure of certainty that the classification is correct), machine learning approaches and using converging evidence of multiple algorithms.

References

- Barton, J. J. S., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology*, 31(5–6), 378–412.
<https://doi.org/10.1080/02643294.2014.895314>
- Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision*, 10, 433–436.
- Breska, A., Ben-Shakhar, G., & Gronau, N. (2012). Algorithms for detecting concealed knowledge among groups when the critical information is unavailable. *Journal of Experimental Psychology: Applied*, 18(3), 292–300. <https://doi.org/10.1037/a0028798>
- Breska, A., Zaidenberg, D., Gronau, N., & Ben-Shakhar, G. (2014). Psychophysiological detection of concealed information shared by groups: An empirical study of the searching CIT. *Journal of Experimental Psychology: Applied*, 20(2), 136–146.
<https://doi.org/10.1037/xap0000015>
- Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, 22(1), 267–281. <https://doi.org/https://doi.org/10.1111/j.1467-9450.1981.tb00402.x>
- Elaad, E. (2016). Extracting critical information from group members' partial knowledge using the searching concealed information test. *Journal of Experimental Psychology: Applied*, 22(4), 500–509. <https://doi.org/10.1037/xap0000101>

- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- JASP Team. (2019). JASP (Version 0.9.1.0) [Computer software].
- klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the Concealed Information Test. *Psychophysiology*, 54(4), 628–639. <https://doi.org/10.1111/psyp.12825>
- Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test memory. *PLoS ONE*, 10(4), 1–17. <https://doi.org/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. <https://doi.org/10.1016/j.jarmac.2015.11.004>
- Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43(6), 385–388. <https://doi.org/10.1037/h0046060>
- Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, 51(9), 879–904. <https://doi.org/10.1111/psyp.12239>
- Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology*, 48(2), 149–154. <https://doi.org/10.1111/j.1469-8986.2010.01050.x>
- Noordraven, E., & Verschuere, B. (2013). Predicting the Sensitivity of the Reaction Time-based Concealed Information Test. *Applied Cognitive Psychology*, 27(3), 328–335. <https://doi.org/10.1002/acp.2910>
- Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3),

191–201. <https://doi.org/10.3758/BF03197692>

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011).

pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1). <https://doi.org/10.1186/1471-2105-12-77>

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.

<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(4), 428–453. <https://doi.org/10.1037/bul0000087>

Sun, R. (2008). *The Cambridge handbook of computational psychology*. (R. Sun, Ed.).

Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772>

The MathWorks. (2018). MATLAB. Natick, Massachusetts.

Verschuere, B., & Kleinberg, B. (2016). ID-Check: Online Concealed Information Test

Reveals True Identity. *Journal of Forensic Sciences*, 61(January), S237–S240.

<https://doi.org/10.1111/1556-4029.12960>

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.

<https://doi.org/10.1016/j.jarmac.2015.01.001>

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.

[https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)

Validation of the searching RT-CIT

Table 1. *Items used in the RT-CIT sorted by enactment (enacted task/intention) and item category.*

Enactment	Item category	Item
Enacted	Code word	Gondel [gondola], Magnet [magnet], Brett [board], Anker [anchor], Stempel [stamp], Klammer [clip]
Enacted	Number of safe deposit box	Six unique random numbers from 1 to 98 (newly generated for each participant)
Enacted	Item 1, to be smuggled	USB-stick, watch, unknown liquid, phone, unknown powder, wallet containing a foreign identity card
Enacted	Item 2, used for recognition	thermos, map, book, headphones, pen, bag
Intended	Gate	A81, B34, C06, D55, E27, F19
Intended	Confederate feature	red shoes, grey pants, brown jacket, blue scarf, white hat, golden sunglasses
Intended	Item placement	in a trash can, in a PET container, in an aluminium recycling bin, under a seat behind a pillar, behind a door
Intended	Meeting spot	information desk, Hour Passion, Center Bar, Amavita, Marché, NZZ Café

Validation of the searching RT-CIT

Table 2. Mean reaction times (in ms; SDs in parentheses), CIT-effect of innocent and guilty participants by item type and enactment (enacted, intention, collapsed), and between group effect sizes by enactment.

	Innocent				Guilty				dCIT _{between}
	RT			dCIT _{within}	RT			dCIT _{within}	
	Irrelevant	Probe	Target		Irrelevant	Probe	Target		
Enacted	482 (154)	481 (152)	600 (132)	-.00 (.19)	486 (146)	532 (181)	627 (135)	.36 (.30)	1.45
Intention	491 (151)	501 (162)	618 (139)	.08 (.21)	502 (151)	551 (182)	646 (148)	.39 (.34)	1.07
Collapsed	486 (153)	491 (158)	609 (136)	.04 (.14)	494 (148)	541 (182)	637 (142)	.36 (.22)	1.76

Validation of the searching RT-CIT

Figure 1. Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the 1st to 2nd bootstrap algorithm (right) on simulated data with CIT-effect size $d_{\text{CIT}_{\text{within}}} = .36$ (lines) and on the empirical data of Study 1 (cross).

Validation of the searching RT-CIT

Figure 2. Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the 1st to 2nd bootstrap algorithm (right) on the simulated data-set with 8 item categories.

Validation of the searching RT-CIT

Figure 3. Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the 1st to 2nd bootstrap algorithm (right) on simulated data with CIT-effect size $d_{\text{CIT}_{\text{within}}} = .45$ (lines) and on the empirical data of Study 3 (cross).

Appendix A. *Instructed purpose of Item 1 in the mock terror attack*

USB stick: “Der USB-Stick wird dazu verwendet einen Virus in das System des Flugzeugs einzuschleusen, um die Kontrolle darüber zu erlangen und das Flugzeug zu entführen.”

[English: The USB-stick is used to infiltrate the aircrafts computer system with malware in order to gain control and abduct the aircraft.]

Watch: “Die Uhr dient als Zeitzündler einer Bombe.” [English: The watch serves as a time fuse of a bomb.]

Unknown liquid: “Die an sich harmlose Flüssigkeit wird mit einer zweiten vermischt. Zusammen ergeben sie den Sprengstoff.” [The inherently harmless liquid will be mixed with a second one. This results in an explosive.]

Phone: “Das Telefon dient als Fernzündler und Energiequelle einer Bombe.” [English: The phone serves as a remote detonator and energy source for a bomb.]

Unknown powder: “Das an sich harmlose Pulver wird mit einem zweiten vermischt. Zusammen ergeben sie den Sprengstoff.” [English: The inherently harmless powder will be mixed with a second one. This results in an explosive.]

Wallet containing a foreign identity card: “Das Portemonnaie enthält gefälschte Ausweispapiere, die es dem Terroristen erlauben in abgesperrte Bereiche des Flughafens zu gelangen.” [English: The wallet contains forged identification documents which give the terrorist access to restricted areas of the airport.]

Appendix B.

Appendix B, Figure B1. Standardization Algorithm

Validation of the searching RT-CIT

Appendix B, Figure B2. 1st to 2nd Bootstrap Algorithm