



UvA-DARE (Digital Academic Repository)

Novel applications of response time-based memory detection

Koller, D.C.

Publication date

2022

Document Version

Final published version

[Link to publication](#)

Citation for published version (APA):

Koller, D. C. (2022). *Novel applications of response time-based memory detection*. [Thesis, fully internal, Universiteit van Amsterdam, University of Zurich].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

NOVEL APPLICATIONS OF RESPONSE TIME-BASED MEMORY DETECTION

DAVE COLLIN KOLLER

NOVEL APPLICATIONS OF RESPONSE TIME-BASED MEMORY DETECTION

DAVE COLLIN KOLLER





**Universität
Zürich** ^{UZH}

NOVEL APPLICATIONS OF RESPONSE TIME- BASED MEMORY DETECTION

Thesis (cumulative thesis)

presented to the Faculty of Arts and Social Sciences

of the University of Zurich

for the degree of Doctor in Philosophy

by

Dave Collin Koller

Accepted in the spring semester 2022

on the recommendation of the doctoral committee composed of:

Prof. Dr. Klaus Oberauer (main supervisor)

Dr. Bruno Verschuere

Zurich, 2022

© Dave C. Koller

ISBN:

This work was carried out under the supervision of prof. dr. Klaus Oberauer and dr. Bruno Verschuer.

All rights reserved. No part of this publication may be reproduced or transmitted in any form by any means, without permission of the author.

NOVEL APPLICATIONS OF RESPONSE TIME-BASED MEMORY DETECTION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. K. I. J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op dinsdag 28 juni 2022, te 12.00 uur

door Dave Collin Koller

geboren te Wettingen AG

Doctorate Committee for the University of Amsterdam / Promotiecommissie UvA:

Supervisors / Promotores:

prof. dr. K. Oberauer	Universität Zürich
dr. B.J. Verschuere	Universiteit van Amsterdam

Other members / Overige leden:

prof. dr. R.W.H.J. Wiers	Universiteit van Amsterdam
prof dr. B. Orobio de Castro	Universiteit van Amsterdam
dr. J. Masip	Universidad de Salamanca
prof. dr. L. Jäncke	Universität Zürich
prof. dr. J. Ullrich	Universität Zürich

The research reported in this thesis was supported by the Swiss Federal Office of Civil Aviation and the Zurich State Police.

This thesis was prepared within the partnership between the University of Amsterdam and the University of Zurich with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Social and Behavioural Sciences at the University of Amsterdam and in the Faculty of Arts and Social Sciences at the University of Zurich.

Het hier beschreven onderzoek werd mede mogelijk gemaakt door steun van de Zwitserse Federale Dienst voor de Burgerluchtvaart en de Staatspolitie van Zürich.

Dit proefschrift is tot stand gekomen binnen een samenwerkingsverband tussen de Universiteit van Amsterdam en de Universität Zürich met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid in de Faculteit der Maatschappij- en Gedragwetenschappen van de Universiteit van Amsterdam en de Philosophische Fakultät van de Universität Zürich.

ENGLISH ABSTRACT

Distinguishing truth from lie is integral for police investigations, the judiciary, and high-risk employers. But even professionals are only slightly better than chance without technological aids. The Concealed Information Test, a questioning protocol designed to reveal memory, has been shown to validly detect feigned ignorance. Moreover, such ‘memory detection’ can be done with a simple behavior measure – response time. But the applied possibilities are restricted.

This dissertation addresses restrictions limiting the applied viability of the response time-based Concealed Information Test (RT-CIT) by investigating possible remedies. First, previous research has shown that the RT-CIT’s validity is reduced when only one testable piece of information is available, which might be the case in practice. The findings suggest that examinees approach the task differently when there is only one testable piece of information (and that this reduces the test validity), but also that the presentation in different modalities increases the validity. Second, the RT-CIT detects concealed knowledge but not the (perhaps innocent) source of that knowledge. Therefore, critical information available to uninvolved people leads to false positive classifications. Contrary to the inventor’s initial publication, the modified Inducer RT-CIT protocol – a proposed remedy for this information contamination – was not immune to it, but an alternative RT-based test, the autobiographical Implicit Association Test, was (though that test is likely to be more vulnerable to faking). Third, until now, the evaluation of RT-CIT data requires knowledge of critical information on the examiner’s side. I tested two novel analytic methods that allow the RT-CIT to be used as an information gathering tool (i.e., to discover new information, that the examiner does not have). The two searching algorithms detected unknown crime information with above chance accuracy. Participant classification performance was substantially below what is typically observed when the information is known to the examiner, but above chance level. Fourth, a technological innovation to test RT-CIT theory and to explore new RT-CIT measures was introduced. The analog keyboard can detect minimal finger movements and, therefore, partial errors that would otherwise go unnoticed. Concealed information was marked by more partial errors – though they were quite rare. As an indicator for response conflict, they could be included in the classification procedure and possibly be a theoretically interesting measure not only for the CIT but also

for other conflict tasks. Taken together, the current dissertation demonstrates new applications of response time-based memory detection and thereby increases their appeal for practice.

NEDERLANDSE SAMENVATTING (DUTCH ABSTRACT)

Het onderscheiden van waarheid van leugen is van essentieel belang in politieonderzoeken, maar ook voor hoog risico werkgevers. Maar zonder technologische hulpmiddelen doen zelfs professionals het nauwelijks beter dan kans. De Verborgene Informatie Test is een ondervragingsprotocol dat herkenning meet en geveinsde onwetendheid kan detecteren. Bovendien kan deze vorm van "geheugendetectie" gebeuren met een eenvoudige gedragsmaat: Reactiesnelheid. De toegepaste mogelijkheden van geheugendetectie zijn evenwel beperkt.

Dit proefschrift richt zich op belemmeringen in de toepasbaarheid van de op reactiesnelheid gebaseerde Verborgene Informatie Test (RT-VIT), en onderzoekt mogelijke oplossingen. Ten eerste heeft eerder onderzoek aangetoond dat de validiteit van de RT-VIT beperkt is wanneer er slechts één stuk informatie beschikbaar is - wat in de praktijk nogal eens voorkomt. Mijn onderzoek toont dat examinandi de taak inderdaad anders benaderen wanneer er slechts één stuk informatie getest wordt (en dat dit de validiteit van de test vermindert), maar ook dat het presenteren van dat stuk informatie in verschillende modaliteiten de validiteit verhoogt. Ten tweede detecteert de RT-VIT weliswaar de herkenning van verborgen (bijvoorbeeld misdrijf-gerelateerde) informatie, maar niet de (misschien wel onschuldige) bron van die kennis. Onschuldigen die beschikken over kritische informatie (bijvoorbeeld via de media) kunnen zodoende verkeerd beoordeeld worden (vals positief testresultaat). Als mogelijke oplossing hiervoor werd recent een aangepast RT-VIT testprotocol aangedragen. Ik vond evenwel dat dit protocol niet de verhoopte immuniteit bood tegen het lekken van informatie. De autobiografische Impliciete Associatie Test (aIAT) deed dat wel, maar van die test is bekend dat die eenvoudiger te omzeilen is. Ten derde, vereist de inzet van de RT-VIT kennis van kritische informatie aan de kant van de onderzoeker. De onderzoekers beschikken zelf evenwel niet altijd over die informatie (bijvoorbeeld, waar is het lijk gebleven?). Ik testte twee nieuwe data-analytische methoden die het mogelijk maken de RT-VIT te gebruiken om nieuwe informatie te ontdekken. De twee zoekalgoritmes konden herkenning van tot dan toe onbekende misdaadinformatie aantonen. De accuraatheid was wel aanzienlijk lager dan wat typisch behaald wordt wanneer de informatie reeds bekend is bij de onderzoeker. Ten vierde introduceerde ik een technologische innovatie om het mechanisme van de RT-VIT te toetsen en om nieuwe maten voor

de RT-VIT te verkennen. Een analoog toetsenbord (ontwikkeld voor digitale spelletjes) kan minimale vingerbewegingen detecteren en zo gedeeltelijke fouten opsporen die met een klassiek toetsenbord onopgemerkt zouden blijven. Ik vond dat herkenning van verborgen informatie gekenmerkt wordt door gedeeltelijke fouten - hoewel die wel zeldzaam waren. Als maat van responsconflict zouden gedeeltelijke fouten niet alleen kunnen bijdragen aan het aantonen van verborgen informatie maar ook bijdragen aan theorievorming (bij de RT-VIT, maar ook andere taken). Samengevat toont dit proefschrift een aantal nieuwe toepassingen van geheugendetectorie gebaseerd op reactiesnelheid en vergroot het daarmee wellicht ook de aantrekkelijkheid ervan voor de praktijk.

CONTENTS

Chapter 1	General introduction	13
Chapter 2	Different target modalities improve the single probe protocol of the response time-based concealed information test	27
Chapter 3	Nationality check in the face of information contamination: Testing the inducer-CIT and the autobiographical IAT	43
Chapter 4	What are you hiding? Initial validation of the reaction time-based searching concealed information test	67
Chapter 5	Assessing partial errors via analog keyboards in response conflict tasks: A preregistered pilot with the concealed information test	99
Chapter 6	General discussion	123

CHAPTER 1

General introduction

On the 3rd of September 2020, the police found five murdered children in their beds in the German town Solingen (WDR, 2021). Forensics showed that they were first sedated with a mix of pharmaceutical drugs and then drowned in the bathtub or suffocated. The mother claimed that a stranger entered the apartment and made her give the drugs to her children, then tied her up and killed her children. Later in court, the mother's lawyer pleaded not guilty due to the lack of evidence, but the judges decided otherwise and found her guilty of first-degree murder (Euronews, 2021).

Without any witnesses, except for the alleged stranger, the statement of the mother was the only direct account on what had happened that morning in the apartment. While the investigation did not yield evidence of a stranger in the apartment (the mother claimed that he wore gloves), this is not strong evidence to conclude that he was not there (Tuzet, 2015). It is cases like these that illustrate how valuable reliable veracity assessments of statements could be. With the expert's reports of psychologists and psychiatrists, motives (e.g., revenge on the husband) and other sources of evidence, the judges deemed they had enough evidence to convict her. But there is a substantial number of cases that were dismissed due to the lack of evidence (Factly, 2020), as well as hundreds of wrongful convictions (<https://innocenceproject.org/>).

HUMAN'S INNATE ABILITY TO DETECT DECEPTION

The importance of assessing the veracity of statements extends beyond police interviews and court proceedings. Insurance companies worldwide face fraudulent claims resulting in an estimated cost of forty billion dollars (Roy & George, 2017), migration agencies need to verify people's identities to ensure fair proceedings, government agencies and private companies want to assess the personnel's eligibility for classified information or high security jobs, and airport security is concerned with detecting forbidden goods and passenger's terroristic intentions. This non-exhaustive list illustrates that many people are affected by veracity assessments of agencies and companies, maybe unknowingly.

This is at the stark contrast to people's innate inability to tell lies from truths. An extensive meta-analysis showed that we, the general public and experts alike, are only slightly above chance performance when asked to judge the veracity of statements (i.e., 54% with 50% representing the chance

level; Bond & DePaulo, 2006; Vrij, 2008) and the interindividual differences are negligible (Bond & DePaulo, 2008). Yet, police investigators estimated their own accuracy to be 77% (Kassin et al., 2007) which can lead to inappropriate interrogation techniques, false confessions, and wrongful convictions (Drizin & Leo, 2004).

But why is it so difficult to tell lies from truths? A qualitative literature review suggests that our beliefs about deceptive cues plays an important role. For example, many people believe that behavioral cues such as gaze aversion, fidgeting, posture shifts, and self-manipulations are indicators for deception (Hartwig & Granhag, 2014). However, a comprehensive meta-analysis showed that these cues are either non-diagnostic or, in the case of fidgeting with an object, show an effect opposite to the widespread believe (DePaulo et al., 2003). Also, some often mentioned paralinguistic cues (e.g., pauses, speech disturbances) are empirically unsubstantiated. But not all our beliefs are wrong. The meta-analysis showed that lies are less plausible, less consistent, less immediate, and that liars' speech has a higher pitch. The integration of information from invalid cues (mostly behavioral) and valid cues (mostly verbal) results in the overall classification accuracy of 54%. The superiority of verbal cues is also illustrated by the finding that audio statements were classified more accurately than audio-visual statements (Bond & DePaulo, 2006). However, our partially incorrect beliefs are not the only reason for our poor performance. Even if a cue is diagnostic, the effect sizes are generally small. Out of 88 cues, only two, namely immediacy and cooperativeness, showed medium sized effects ($|d| > .5$). Twenty-three cues showed at least small effect sizes ($|d| > .2$; Bond & DePaulo, 2003). Although this makes it difficult to achieve high accuracies if only few cues are considered, there are validated content-based methods that incorporate a wide array of cues in the analysis of transcripts (Oberlader et al., 2021).

Since content-based methods are intricate and naturally processed behavioral and verbal cues are insufficient to assess the veracity of the mother's account accurately, could technological approaches provide reliable measures and classification? Several technological deception detection systems have been developed over the years to help distinguish truths from lies. More recent inventions include the Silent Talker (Orshea et al., 2018), an automated video-based system developed to increase security at the Schengen land borders as one part of the Intelligent Portable Control System (<https://www.iborderctrl.eu/>), and the commercial eye-tracking-based EyeDetect (Cook et al., 2012;

<https://converus.com/>). The most famous technology is the Control Question Technique (CQT) polygraph (Reid, 1947), which is based on differential physiological responses between control questions and relevant questions. Despite its lack of theory, its dependence on the examiner, its susceptibility to countermeasures, and its high false positive rate (Iacono & Ben-Shakhar, 2018; Meijer & Verschuere, 2015; National Research Council, 2003), the CQT polygraph is still widely used, for example for personnel selection and in investigations in the US, to monitor convicted sex offenders in the UK, and as evidence in Belgian courts. Interestingly, it is not the technological part (the polygraph) that is in question, but the CQT interview protocol. It has been argued that the questions play a more central role than the cues itself, both in technological and non-technological approaches (Meijer et al., 2016; Vrij & Granhag, 2012). Therefore, this thesis focusses on the Concealed Information Test (CIT; Lykken, 1959), a scientifically sound interview protocol. The detailed discussion of the aforementioned technological approaches is beyond the scope of this work.

THE CONCEALED INFORMATION TEST

The CIT (Lykken, 1959) is a questioning protocol designed to detect knowledge instead of lies. The rationale of the CIT is that if an examinee sees items that could be related to the incident under investigation (e.g., different drugs), they show a distinct response to the truly incident related item if and only if they recognize it as such. Naïve examinees should not be able to distinguish the incident related item (probe) from other plausible but non-related items (irrelevants). This important assumption can be tested beforehand with non-suspects (Doob & Kirschenbaum, 1973). Decades of research demonstrated the validity of the CIT (for reviews, see Meijer et al., 2014; Verschuere et al., 2011) in combination with physiological measures (Gamer, 2011a; Lykken, 1959), EEG (Rosenfeld, 2011), fMRI (Gamer, 2011b), eye-tracking (Gamer & Yoni, 2018), and response times (RTs; Farwell & Donchin, 1991; Seymour et al., 2000; Suchotzki et al., 2017). Table 1 illustrates exemplary items and the expected RT pattern.

Table 1. *Exemplary items and expected data pattern for the RT-CIT*

Item type	Correct response	Items (example)	Data pattern (expected)
Probe	“No”	Palexia	<p>Response Time</p> <p>Knowledgeable Naïve</p> <p>■ Probe ■ Irrelevant □ Target</p>
	“No”	Pethidin	
Irrelevant	“No”	Nalbuphin	
	“No”	Zalviso	
	“No”	Dipidolor	
	“No”	Durogesic	
Target	“Yes”	Codein	

While the basic rationale holds for all measures, some capture different processes than others. Recent studies using guilty and knowledgeable innocent (i.e., witnesses) participants showed a dissociation between the skin conductance response (SCR) on the one hand and respiration measures and heart rate on the other hand (Klein Selle et al., 2016, 2017). Both groups showed larger SCRs to probes than to irrelevants but only the guilty group also showed shorter respiration line lengths and slower heart rates. Because both groups recognized the probes but only the guilty group tried to conceal it, they argued that SCR reflects the recognition based orienting response (Sokolov, 1963) while respiration line length and heart rate reflect arousal inhibition. Using the same paradigm, Rosenfeld et al. (2017) found that the P300 amplitude, the most common EEG-based measure in the CIT literature, was larger for the probes than for irrelevants in both groups but the effect was significantly larger for the guilty than for the witness group. This indicated that the P300 amplitude was affected by the orienting response and by arousal inhibition. Additionally, the guilty but not the witness group showed delayed N200/N300 responses. For the RT-based CIT (RT-CIT), research suggests a third process, response conflict or response inhibition, to play an essential role (Seymour & Schumacher, 2009; Verschuere & De Houwer, 2011). Response conflict arises from the introduction of target items (Seymour et al., 2000) which share a feature with the probe (e.g., familiarity) but require different responses. Therefore, the faster familiarity-based response needs to be overridden (or inhibited) in favor

of the slower recollection-based response (Yonelinas, 2002). For a thorough discussion on the theories underlying the CIT, see Klein Selle et al. (2018).

Considering the large effect sizes ($d = 1.56$ for P300, $d = 1.57$ for SCR, and $d = 1.30$ for RTs; Meijer et al., 2014; Suchotzki et al., 2017) it might come as a surprise that Japan is the only country using the (physiological) CIT on a large scale with approximately 5000 examinations per year and that the CIT is rarely allowed as evidence in court, even in Japan (Osugi, 2011). Interestingly, there are two important differences between how the CIT is used in the field and how it is studied in the laboratory (Ogawa et al., 2015). First, laboratory research evaluates the data on an examinee level (i.e., aggregating the data of multiple questions) while the Japanese police evaluates the knowledge of the examinee for every question individually. Second, the police uses the CIT mainly to find new evidence (searching CIT) and not to provide evidence of knowledge (known solution CIT; Osugi, 2014). This means that instead of only one comparison per question (probe vs irrelevants) in the known solution CIT, every item could be the probe and needs to be compared to all other items in the searching CIT, resulting in as many comparisons as there are items. Albeit these differences limit the generalization of laboratory findings to the Japanese practice, they are primarily differences in how the data is analyzed. If police case data and laboratory data turn out to be comparable, like first results of Osugi (2010) and Zaitzu (2016) suggest, we might learn a lot about the applied performance from re-analyses of existing experimental data with ground truth. Simultaneously, suggestions to develop more sophisticated classification algorithms could be tested (Matsuda et al., 2012).

Additional barriers preventing extensive field application of the physiological CIT might be the need for trained personnel, specialized equipment, as well as the similarity to the CQT polygraph for people unfamiliar with the questioning protocols. Substituting the polygraph with EEG or fMRI would require even more training and expensive equipment with little effect on the classification performance (Meijer et al., 2016). However, the RT-CIT, showing similar performance (Meijer et al., 2016; Suchotzki et al., 2017), could be a viable alternative. Data collection and analysis can be done automatically on any computer, and large numbers of participants can be tested simultaneously without additional resources through remote (online) testing (Kleinberg & Verschuere, 2015; Verschuere & Kleinberg, 2016). Large scale testing could lead to new applications (e.g., screening of military

applicants for connections to extremist groups). Albeit these beneficial properties, there are also limitations and challenges to the RT-CIT.

APPLIED CHALLENGES FOR THE RT-CIT

The RT-CIT faces several applied challenges, some of which are also valid concerns for the physiological CIT (Ben-Shakhar, 2012). This chapter discusses the state-of-the-art RT-CIT prior to our empirical work and its challenges for field application.

Limited scope

Research has shown several conditions that need to be met for the RT-CIT to perform well. While these conditions might influence the experimental design in laboratory research, they pose limitations to the scope of the RT-CIT in real-life situations. Addressing the limitations to known information, to scenarios with multiple testable items, and to scenarios without information contamination is the primary focus of this thesis. The willingness of subjects to take an RT-CIT in high stakes situations was not investigated.

One limitation is that the examiner needs to know the relevant information to construct the RT-CIT. Returning to the example case, the police (most likely) did not know which child was killed first and could therefore not construct a CIT item from this information. However, this restriction does not apply to the searching CIT. The searching CIT includes all possible answers to question at hand, in this case the names of the five children, and the examiner tries to infer what happened by analyzing the data for each item in comparison to all others regarding this question. (Because most questions have a much larger set of possible answers, examiners need to identify the most likely ones to construct the CIT and add an “other” item for all non-included possibilities.) The searching CIT has only been used with physiological (Breska et al., 2014; Elaad, 2016; Meijer et al., 2013; Osugi, 2011) and EEG measures (Meixner & Rosenfeld, 2011) but not RTs. As for now, the RT-CIT is limited to scenarios in which the relevant information is known to the examiner.

Unfortunately, a single testable piece of information (e.g., how the victims were suffocated: by hand, with a bag, with a pillow, with a rope etc.) might not be enough for the RT-CIT. Eom et al. (2016) found that the probe-irrelevant difference increased with the number of information categories (e.g., the first victim, the way of killing the youngest child, where in the apartment was the oldest child murdered) while keeping the number of trials fixed. Furthermore, Verschuere et al. (2015) showed that randomly presenting items of multiple information categories in the same test block (so-called multiple probe protocol) increased the validity of the RT-CIT compared to the single probe protocol in which categories were tested in separate blocks. In an attempt to increase the validity of the single probe protocol, Lukács et al. (2017) added familiarity-related filler items (e.g., “RECOGNIZED”, “UNFAMILIAR”) that participants also needed to classify to the RT-CIT. This modification increased the probe-irrelevant difference drastically and even showed a larger effect than the multiple probe protocol (also see Olson et al., 2020). Although the filler protocol also led to more exclusions due to high error rates, it might alleviate the restriction to scenarios with multiple testable information.

The CIT detects recognition and with that comes a susceptibility to information contamination. That is, the possibility of innocent people learning about the critical information e.g., due to media coverage or information disclosed during police interviews. Because the RT-CIT cannot discriminate the ways the examinee learned the information, using possibly contaminated information should be avoided (Bradley et al., 2011). As for our example case, items about the looks of the crime scene or the drugs used to sedate the children could not be used since the mother had the opportunity to look at the crime scene and was forced to hand the drugs to the perpetrator. There are two ways to address this problem: First, the examinee could be asked to disclose any information they know about the incident and exclude those items from the RT-CIT. Second, proposed by Lukács and Ansorge (2019), the RT-CIT could be modified in a way that response conflict is induced by a feature other than familiarity (e.g., self-relatedness), leading to conflict only for the perpetrator. Lukács and Ansorge (2019) replaced target items with self- or other-referring inducer items that needed to be classified along with the probe and irrelevant items of the RT-CIT. While the contaminated group did not show a CIT-effect, the CIT-effect of the knowledgeable group was also reduced, resulting in similar classification performance as the standard RT-CIT.

Lastly, because participants have an active role in the RT-CIT, the willingness to engage with the test is a prerequisite. It is difficult to foresee under what circumstances people would be willing to participate in an RT-CIT.

While these are considerable limitations to the scope of the RT-CIT in its current state, it should be considered that research on the RT-CIT as a stand-alone task has only picked up speed in the last decade (cf. Suchotzki et al., 2017) - future adaptations might widen its scope. But even if this is not the case, the RT-CIT might be a valuable tool in some applied situations, provided that the classification performance from the laboratory studies persists in the field.

Box 1. <i>Applied limitations to the scope of the RT-CIT</i>	
Known critical information:	The examiner needs to know the critical information to construct and evaluate the RT-CIT.
Multiple pieces of information:	The validity of the RT-CIT is reduced when only one piece of information is tested.
No information contamination:	The critical information should not be known by people uninvolved in the crime to reduce false positive results.
Willingness to participate:	Because the RT-CIT measures RTs of voluntarily controlled actions, the willingness to engage is a prerequisite.

Transferability from lab to field

When a situation within the scope of the RT-CIT has been identified, the next important question is if the RT-CIT is valid in this scenario or if the results from the laboratory do not generalize to real-life situations. The main concerns in this regard are the tested population, the test circumstances, the item encoding, and the classification model. The following paragraphs address these issues one by one.

The (psychology) student population commonly used in laboratory studies differs from the general population on many dimensions (e.g., age, education, experience in study participation). Some

dimensions may impact the validity of the RT-CIT significantly, which could impair the generalizability laboratory findings to applied scenarios. First studies with prisoners (Suchotzki et al., 2019) and children (7-10 years; Visu-Petra et al., 2016) confirmed the test's validity in these samples. The CIT-effect of prisoners was comparable to the age and education matched group. Although further research is needed to draw conclusions on the generalizability, the initial results are promising.

While different populations can be tested in the laboratory, it is unlikely that test circumstances (e.g., high stakes, arousal level, involvement level, actual crime committed) will ever match real-life scenarios. Nevertheless, researchers used incentives to pass the RT-CIT and motivational instructions to increase participant's motivation level. A meta-analysis showed smaller CIT-effects in higher motivation conditions but the CIT-effects remained large (Suchotzki et al., 2017). Addressing effects of instructed vs. self-initiated wrongdoings, Geven et al. (2018) gave participants the opportunity to cheat on a trivia quiz with monetary incentive. Self-initiated cheaters (mimicking real-life) showed the same RT pattern as the instructed cheaters (typical laboratory setting). Although these results, too, are promising, the effects of high stakes (e.g., receiving insurance money or not, getting a job or not) with all its consequences (e.g., arousal, preparation or even training to beat the CIT) are widely unknown.

Another important difference between laboratory and field studies relates to the encoding and memory of items. In laboratory studies, it is made sure that participants encode (and learn) the critical information, often with the same stimuli later used in the RT-CIT (e.g., Eom et al., 2016; Suchotzki et al., 2021; Visu-Petra et al., 2014). Examiners in applied settings, especially forensic settings, cannot be certain what parts of the crime scene the perpetrator encoded. Imagine a burglar getting surprised by the homeowner. In a panic reaction, the burglar grabs a small sculpture that is in reach and throws it at the owner before he flees, fatally wounding the owner. Despite the statue being the murder weapon, the burglar might not have encoded the statue well enough (or from a different angle) to recognize it in a CIT. He might not even have realized that it was a statue. The validity of the RT-CIT depends on choosing well encoded, salient items (Seymour & Fraynt, 2009; Verschuere et al., 2015) but it is unknown to what degree we can identify suitable items in a given scenario. A recent study further substantiated the importance of encoding and item selection, showing that if an item is encoded on an

exemplar level (e.g., “necklace”) but tested on a categorical level (e.g., “jewelry”), or vice versa, the CIT-effect disappeared (Geven et al., 2019).

There is also a statistical reason why classification performance in the field might be lower than in laboratory studies. Laboratory studies with the aim to test the classification accuracy usually have a knowledgeable group and a naïve control group (real or simulated). The classifier is trained on a subset of data from both groups and applied to the remaining data (i.e., cross-validation). In many applied settings, however, it is impossible to build a model both on naïve and knowledgeable people because there is only the one knowledgeable person that we aim to identify. Instead, one could use generic decision rules, such as the cutoff $dCIT > .2$ proposed by Noordraven and Verschuere (2013), use a classifier based on the deviance of the suspect’s probe RT distribution from their irrelevant RT distribution (Seymour et al., 2000, 2013), or use a classifier based on the deviance from a naïve control group.

Researchers put in a lot of effort to design realistic experiments, but only few studies addressed the other threats to the generalizability of laboratory studies. It is up for debate if these issues should be investigated more thoroughly in laboratory studies or if the evidence to justify field testing is sufficient. Ultimately, however, only field studies will show with certainty if the RT-CIT can retain the performance found in laboratory experiments.

Misclassifications

Misclassifications (false positives and false negatives) can never be avoided completely but there are at least two ways how they can be influenced. First, by adjusting the criterion at which a person is classified as knowledgeable. Shifting towards a more liberal criterion (i.e., classifying the examinee as knowledgeable requires less strong evidence) leads to less false negative but also more false positive classifications (Tanner & Swets, 1954). While experimental studies can set the criterion to maximize the accuracy, Youden’s J (Youden, 1950), or another benchmark, choosing a good criterion in practice is more complex and highly dependent on the situation at hand. In a pre-employment screening for a high security position of a critical infrastructure, for example, a false negative classification could lead

to sabotage and very high costs whereas a false positive classification would only result in the dismissal of the applicant and another applicant would take their place. From the employer's perspective, a liberal criterion seems desirable. In a criminal trial, however, a false positive result could lead to wrongful convictions (Drizin & Leo, 2004). Therefore, a more conservative criterion might be appropriate.

Second, misclassifications can be reduced by improvements to the paradigm, the classifier, or the data quality. Most research focused on improvements to the RT-CIT: using the multiple-probe protocol (Verschuere et al., 2015), familiar targets (Suchotzki et al., 2018), familiarity related filler items (Lukács et al., 2017; Olson et al., 2020), secondary tasks (Hu et al., 2013; Visu-Petra et al., 2013), or short response deadlines (Suchotzki et al., 2021). But even a refined paradigm needs to be coupled with the appropriate classifier to minimize misclassifications. While there is some diversity of classification procedures used for the RT-CIT (e.g., see Noordraven & Verschuere, 2013; Seymour et al., 2013; Seymour & Fraynt, 2009), a systematic comparison to identify the gold standard has not been conducted yet. Lastly, technological improvements could provide the classifier with better or additional data, thereby increasing its performance. The only study exploring this approach combined ocular measures with RTs (Seymour et al., 2013). Although the combined measure failed to outperform the classification based only on RTs, this is most likely due to the ceiling performance of the RT-based classification (98% accuracy) and not a sign of futility of the approach itself.

Misclassification rates of 10% or more, which are not uncommon in the RT-CIT literature, might seem too high for applied purposes. However, established methods like expert fingerprint classification showed an error rate of 9% (Kellman et al., 2014) and even automated fingerprint systems showed 5%-10% misclassifications until very recently (Mohamend, 2021).

Countermeasures

Countermeasures are actions taken by the examinee to manipulate the test outcome in their favor. They are an important obstacle to overcome for every deception detection procedure. While various countermeasures have been studied for the physiological CIT (for a review, see Ben-Shakhar, 2011) and a promising countermeasure resistant CIT protocol has been proposed for EEG-based CIT

(Rosenfeld et al., 2008), research on countermeasures in the RT-CIT is scarce. This is especially surprising as pressing a button is under more voluntary control than effecting (neuro)physiological measures. The only study on RT-CIT countermeasures prevention to this day showed that a strict response deadline rendered the strategy to slow down responses to irrelevant items ineffective (Suchotzki et al., 2021). An initial experiment on physical countermeasures (i.e., pressing down the toes when an irrelevant item is presented) found that they were not effective (Norman et al., 2020; Experiment 3). Countermeasure strategies, their prevention, and their detection remain an under-researched field and were also not addressed in this thesis.

THIS THESIS

The goal of the current thesis is to bring the RT-CIT closer towards real life application. Three empirical studies addressed challenges regarding the scope of the RT-CIT, the fourth study explored a possible technological improvement.

The first study addressed the problem that the CIT-effect, and in consequence the validity, is reduced when only a single testable piece of information is available (Verschuere et al., 2015), limiting the scenarios in which the RT-CIT can show high classification performance. We investigated if the CIT-effect of this so-called single probe protocol could be increased by presenting the information in different modalities. In this online study, we asked participants to conceal their nationality. Nationality information were presented as words (e.g., “Sweden), flags or maps. Independently manipulating the number of target modalities and probe/irrelevant modalities helped us to a better understanding of why the single probe protocol shows reduced validity and revealed a viable way to increase it.

The second study was concerned with the CIT’s susceptibility to information contamination (e.g., innocent people also know details of a crime) which results in more frequent false positive classifications. Because of this limitation to the scope of the RT-CIT, Lukács and Ansorge (2019) proposed a modified paradigm, the Inducer-CIT, which should be immune to information contamination. A second, theoretically viable, alternative could be the autobiographical Implicit Association Test (aIAT; Sartori et al., 2008), but it has never been tested in an information contamination

scenario before. In this online study, Dutch and British participants were asked to pretend to be British while completing either the Inducer-CIT or the aIAT. The contamination group consisted of British citizens with good knowledge about the Netherlands, which was assessed in a pretest. The value of this study is three-fold: first, to independently replicate the results of Lukács and Ansorge (2019); second, to test the assumed viability of the aIAT in a contamination scenario; and third, to find the recommended procedure by directly comparing both tests.

The third study addressed the restriction of the standard RT-CIT that the examiner needs to know the to-be-tested information. While polygraph- and EEG-based CIT studies showed that this is not necessarily the case and that the CIT can be used to reveal new information (i.e., searching CIT; (Breska et al., 2012, 2014; Elaad, 2016; Meijer et al., 2013; Meixner & Rosenfeld, 2011; Osugi, 2011), this has not been investigated with RTs. In an elaborate study at an international airport, we asked participants to carry out the first part of a mock terror attack or a control activity, followed by the RT-CIT. We contrasted two searching CIT algorithms on their performance to reveal new information and to identify the mock terrorists. We additionally ran a simulation study to better understand how the algorithms are expected to perform under different circumstances and applied the algorithms to an independent data set. This study not only explored an additional applied use of the RT-CIT, but its realistic scenario also provided further evidence on the validity of the RT-CIT in real life settings.

The fourth and last study of this thesis introduced a technological modification, an analog keyboard, that can give insights about the response dynamics beyond RTs. In the long run, the additional data could increase classification performance or help to detect countermeasures. But as this was the first experiment to use analog keyboards, the goal was to measure response conflict directly by recording partial button presses corresponding to the conflicting answer - similar to EMG measures (Seymour & Schumacher, 2009) but without its drawbacks (specialized equipment, adaptations to the paradigm, training of the examiner). In this laboratory study, participants were asked to complete two conflict tasks, an autobiographical RT-CIT and a modified Sternberg task (Oberauer, 2001). To foreshadow the results, partial errors were more frequent in conflict trials than in non-conflict trials. While we focused on response conflict due to its relevance to the RT-CIT, the rich data provided by analog keyboards could be valuable for experimental psychologists in general.

CHAPTER 2

Different target modalities improve the single probe protocol of the response time-based concealed information test

This chapter is published as:

Koller, D., Hofer, F., & Verschuere, B. (2021). Different target modalities improve the single probe protocol of the response time-based concealed information test. *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1016/j.jarmac.2021.08.003>

ABSTRACT

To detect if someone hides specific knowledge (called ‘probes’), the reaction time-based Concealed Information Test (RT-CIT) asks the examinee to classify items into two categories (targets/non-targets). Within the non-targets, slower RTs to the probes reveals recognition of concealed information. The preferred protocol examines one piece of information per test block (single probe protocol), but its validity is suboptimal. The aim of this study was to improve the validity of the single probe protocol by presenting the information in multiple modalities. In a preregistered study ($n = 388$) participants were instructed to try to hide their nationality. The items referring to the nationality were presented as words, flags, and maps. Increasing the number of modalities of the targets ($BF_{10} = 37$), but not of the probes and irrelevants ($BF_{01} = 6$), increased the CIT-effect. This broadens the range of the RT-CIT’s applicability, which is an important step towards application in practice.

Keywords: memory detection, Concealed Information Test, CIT, deception, single probe protocol, lie detection

GENERAL AUDIENCE SUMMARY

Law enforcement agencies, intelligence service, or even companies can encounter situations in which they would like to find out if a person knows something about an incident even though he or she claims not to. In situations like these, the reaction time-based Concealed Information Test (RT-CIT) could be applied. In this automated test, examinees first learn a few so-called target items (words and/or pictures). Then they will see different items one by one. These items are either the learned targets, items related to the incident (so-called probes) or unrelated items (so-called irrelevant items). The examinees' task is to indicate with two keys on the keyboard if he or she recognizes this item or not (i.e., YES-button for the targets and NO-button for the other items). Slower responding to items related to the incident than to irrelevant items is an indication of concealed recognition. The RT-CIT works best if multiple pieces of information about the incident (e.g., location, stolen jewelry, tool used for the break-in) are tested. In reality, however, an examiner might have only one information that can be tested. We explored two possibilities to improve the performance of the RT-CIT in such a scenario. In an online study with 388 participants, they were asked to hide their true nationality and claim to be from another country which was used as the target information. We investigated if the performance of the RT-CIT could be increased if we present the items not only as words (e.g., "United Kingdom") but also as flags and maps of countries. Presenting the target information in different ways increased the difference between probes and irrelevant items which implies more correct detections. This makes the RT-CIT applicable in a wider range of situations.

INTRODUCTION

The concealed information test (CIT; Lykken, 1959), can be used to reveal if a person has specific knowledge he/she claims not to possess and is frequently used by Japan's police (Osugi, 2018). The rationale of the CIT is that a person shows a different reaction to an item whose recognition he/she tries to conceal, compared to similar yet irrelevant items. When for example crime related items (e.g., the location where the victim was found: the woods) are presented among other plausible items (e.g., the river, the sewer, the shed), only a person with crime knowledge is expected to show a distinct reaction to the crime related items (so-called *probes*) compared to the other items (so-called *irrelevants*). The typically observed reactions for concealed recognition in the CIT include increased response times, increased skin conductance response, and increased P300 amplitude (see Verschuere & Meijer, 2014 for a review). With its potential to easily test large groups of people remotely (Verschuere & Kleinberg, 2016), there has been renewed interest in response times as a CIT index.

The response time-based CIT (RT-CIT) includes a third item type, the so-called targets (Farwell & Donchin, 1991; Rosenfeld et al., 1988; Seymour et al., 2000). Examinees are instructed to press the YES button only for targets, and the NO button for all other items (including the probes). Meijer et al. (2016) found a weighted average of the area under the receiver operating characteristic curve (AUC) of .82 based on 981 participants across the nine analysed experiments, which is well above chance. The classification performance is known to vary with several factors, one of which is the testing protocol (Lukács et al., 2017; Verschuere et al., 2015).

There are two main RT-CIT testing protocols: the *single probe protocol* and the *multiple probe protocol*. In the single probe protocol, each block contains items of one item category (Lykken, 1959). For instance, a first block could test the examinee for stolen goods, the next block for crime locations, etc. In the multiple probe protocol, the items of the different item categories are all presented intermixed in each block (e.g., Seymour et al., 2000). Research showed superior classification performance for the multiple probe protocol compared to the single probe protocol in the RT-CIT (Lukács et al., 2017; Verschuere et al., 2015). Experiment 2 of Verschuere et al. (2015), for example, showed larger effect sizes for the multiple probe protocol compared to the single probe protocol ($d_{within,MP} = 1.52$; $d_{within,SP} =$

0.59) as well better classification ($AUC_{MP} = .86$; $AUC_{SP} = .69$). The application of the multiple probe protocol is limited to situations in which more than one critical piece of information is known. Furthermore, when the RT-CIT is used as an investigative tool in the form of a searching RT-CIT (Koller et al., 2020), the single probe protocol may be the only option. For example, to reveal the exact location of a planned terror attack, the examiner would first need to test for the city, then for a specific location within that city. Without this serial testing, the number of items that would need to be included in the CIT would increase rapidly.

Given the need for a more accurate single probe protocol, Lukács et al. (2017) introduced the addition of familiarity related filler items that needed to be classified as familiar (e.g., the filler word “RECOGNIZED”) or unfamiliar (e.g., the filler word “UNKNOWN”). This modification led to larger CIT-effects as it assures semantic processing of the stimuli and/or may enhance response conflict for the probes. While familiarity related fillers seem to be a good way to improve the single probe protocol, exploring alternative solutions still has its merits (e.g., to find combinations of effective techniques or to overcome potential shortcomings of one solution).

In the present study, we examined whether presenting items in different modalities (e.g., a country as name, flag, or map) is sufficient to increase the validity of the single probe protocol.¹ Further, we explored whether it is the probe or the target modalities that contribute to the effect. The ultimate goal of this study was to make the RT-CIT applicable to a wider range of scenarios.

METHOD

The experiment was approved by the ethical committee of the Faculty of Social and Behavioral Sciences of the University of Amsterdam (approval number: 2014-CP-3389). Preregistration, material, data, and scripts can be found on <https://osf.io/d536j/>.

¹ We infer increased validity from larger within-participant CIT-effects. Lukács and Specker (2020) showed that this inference might not be valid if the standard deviations of the within-participant CIT-effect increases substantially. This was not the case in our study (see Table 3 and supplementary materials on OSF; <https://osf.io/d536j/>).

Deviations from preregistration

We had 3 exclusion criteria, one stated: “Participants with mean RT of irrelevant items deviating more than $\pm 3 SE$ of their respective group means of irrelevant items (only correctly answered trials are considered in this analysis) will be excluded”. However, this is an unfortunate typing error and was meant to state “ $\pm 3 SD$ ”.

Participants

Participants were eligible to enroll if they were between 18 and 45 years old and of one of the following nationalities: British, Portuguese, Spanish, German, Italian, Austrian, or Swedish (see Procedure). Completion of this study took participants about 14 minutes and was reimbursed with 1.4 GBP (≈ 1.8 USD).

Following the preregistered recruitment procedure, four hundred participants were recruited using the online platform Prolific (M age = 27.88, SD = 7.36, 51% female). Twelve participants (3%) were excluded based on the three preregistered criteria. Eight indicated that they provided wrong information about their nationality in the pre-CIT questionnaire, three due to poor performance in the task (more than 50% errors in at least one item type), and one due to slow RTs to irrelevant items (M $RT_{\text{irrel}} > \text{group mean} + 3*SD$), resulting in a final sample of $N = 388$ (M age = 27.74, SD = 7.35, 51% female). Per inclusion criteria, participants were British (42%), Portuguese (33%), Spanish (11%), German (2%), Italian (10%), Austrian (0%), and Swedish (2%).

Procedure

At the beginning of the experiment, participants were asked to indicate their nationality from a list of the seven nationalities that were eligible for the study and “other”.² Participants that chose

² Due to a programming mistake, participants did not see the informed consent form and therefore did not give us explicit consent before the study. However, in agreement with the ethics review board of the University of Amsterdam, we are convinced that it is ethical to use the collected data for the following reasons. 1) Data collection was done on Prolific, an online platform for running scientific studies to which participants signed up. 2)

“another country” were directed to the end of the study explaining that they are not eligible to participate in this study. The chosen country was the critical probe item in the CIT. We used Prolific’s built-in nationality filter to invite eligible participants only. To further improve data quality, we asked participants at the end of the study whether they indicated their true nationality in the beginning. It was made clear that this does not have any influence on their payment or on in-/exclusions for future studies. These measures give us confidence in the truthfulness of the reported nationalities.

If participants chose one of the seven countries as their nationality, they were subsequently asked to indicate up to one other country from the remaining six that is of significance to them. This country was removed from the item pool of irrelevant items in the CIT to assure that only the country of origin stands out amongst the other countries in the CIT. Not removing other significant countries would lead to a lower sensitivity. If no other country was indicated as significant, one was removed at random.

Next, participants were asked to imagine a scenario in which an online service they need to use is not available for people of certain nationalities and theirs is one of them (e.g., Sweden). Therefore, they had to pretend to be from another country (randomly assigned from the remaining five countries, e.g., Spain). As they were suspected of cheating (the online service provider was said to have detected a mismatch between the location of the computers IP-address and the claimed nationality) they were asked for additional verification: the RT-CIT. Participants were told that this verification tries to detect their true nationality and that their goal is to hide that information and to convince the service provider to be from their indicated country (Spain). Then, the RT-CIT started. After the RT-CIT, participants were asked if they indicated their true nationality in the beginning of the study, thanked and redirected to the Prolific website. Payment was processed through Prolific within a few days.

Participants were informed about the nature of the study before they chose to participate. More specifically they were informed that they need overcome an information verification test, that the test is based on reaction times, that the test cannot be paused, that they need to install a plugin but instructions to uninstall the plugin will be provided, that only participants of certain nationalities are eligible to participate (UK, Germany, Spain, Portugal, Italy, Sweden, Austria), and they were informed about the time the study will take and reimbursement they will receive. 3) Participants had the possibility to revoke their consent at any point by “returning” their submission, as Prolific calls it. 4) Participants could revoke their consent by contacting the first author using Prolific’s built-in messaging system.

RT-CIT

The RT-CIT was programmed with Inquisit 5 (2016) and ran on the participants' computer using the Inquisit Web plugin, which they downloaded just before starting the test.

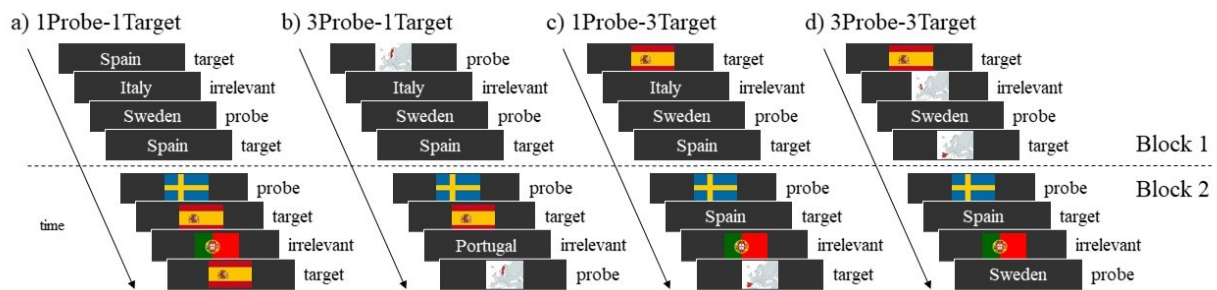
Participants were instructed to answer the question "Is this your nationality?" as fast and accurately as possible by responding YES (A-key) only for items that belong to their fake nationality (*targets*) and NO (L-key) to all other items (*probes* and *irrelevants*) and therefore denying their true nationality. So, a person from Sweden pretending to be from Spain would answer YES only to Spain, and NO to Austria, *Sweden*, Portugal, and Germany. Three modalities (word, flag, map), with six items (1 probe, 1 target, 4 irrelevant) per modality were used as items in the RT-CIT. The three target modalities (i.e., word, flag, map corresponding to their fake identity) were presented on the screen for 10 seconds together with a reminder that participants should respond to these items with YES, followed by a repeatable practice block in which each of the 18 items was presented once. Then followed three test blocks, each consisting of 6 burn-in trials at the beginning of the block that were discarded to avoid possible artefacts (e.g., due to finger placement, accustoming to the pace) and 126 test trials (totaling 378 test trials).

Every trial started with the stimulus presentation in the center of the screen which stayed until a response was given or the response deadline of 1.5s was reached. The question "Is this your nationality?" was displayed on top of the screen as well as the answer labels (YES, NO) and key labels (A-key, L-key) to the left and right of the stimuli, see Figure 1. Feedback (red "WRONG" displayed below the stimulus for 250ms for incorrect responses; red "TOO SLOW" above the stimulus 800ms after stimulus onset) was given only in the practice trials. The practice block was repeated up to three times or until the accuracy of each item type (probe, target, irrelevant) was above 50% and the mean RT of irrelevant items was below 800ms. Reminder instructions were given before each repetition. The response-stimulus-interval varied randomly from 400-800ms. Until the end of the practice phase, the task was identical for all participants.

For the test phase, we manipulated the number of modalities of targets (one vs. three) and the number of modalities of probes and irrelevant items (one vs. three) that participants saw in any given













block. Participants had been randomly divided into four groups (1Probe-1Target with $n = 89$, 3Probe-1Target with $n = 103$, 1Probe-3Target with $n = 96$, 3Probe-3Target with $n = 100$) based on the automatically generated participant number. The 1Probe-1Target modality condition (see Figure 1a) used one probe and one target of the same modality in each block (e.g., flags in the first, words in the second, and maps in the third block). This condition mimics the traditional single probe protocol. The 3Probe-1Target modality condition (see Figure 1b) used all three modalities of the probe but only one modality of the target per block. This means that in one block, the target information was only presented as a flag, for example, while the probes and irrelevants were presented in all three modalities. Similarly, the 1Probe-3Target modality condition (see Figure 1c) used one modality for the probes and irrelevants, but three modalities for targets per block. The 3Probe-3Target modality condition (see Figure 1d) used all modalities for probes, irrelevants, and targets which results in three identical blocks. Table 1 shows the word-block for all conditions as an example. Each participant saw every item 21 times over the course of the 3 blocks. The blocks were presented in random order.

Figure 1. *Exemplary segments of the four conditions*



Note. This illustration shows exemplary segments from two of the three blocks for each condition.

Table 1. *Experimental conditions illustrated for the word-block*

	Item type		
	Probe	Target	Irrelevant
1P1T	Sweden	Spain	Germany, Portugal, Austria, Italy
3P1T	Sweden,  	Spain	Germany,  
1P3T	Sweden	Spain,  	Germany, Portugal, Austria, Italy
3P3T	Sweden,  	Spain,  	Germany,  

Note. Cells that only contain words would be changed to flags or maps for the flag-block and map-block, respectively. The mixed cells remain unchanged.

RESULTS

We only included correctly answered non-target test trials with RTs between 200ms and 1500ms in the analysis. A total of 1209 non-target trials (0.99%) were excluded.³ We obtained the typical CIT effect in response times (i.e., larger RTs for probes than for irrelevant items) in each of the four conditions, see Table 2. Consequently, the main dependent variable, the within participant CIT-effect ($dCIT = (M_{RT_{Probe}} - M_{RT_{Irrel}}) / SD_{RT_{Irrel}}$; Kleinberg & Verschuere, 2015), was credibly greater than zero for all groups (i.e., the lower bound of the 95% CI greater than zero; see Table 2).

Table 2. Response times, within-participant CIT-effect, and probe-irrelevant effect size

Condition	RT			$dCIT$	δ
	Probe	Irrelevant	Target		
1P1T ($n = 89$)	418 (64)	405 (52)	491 (51)	0.15 (0.30) [0.10, 0.23]	0.61 [0.38, 0.85]
3P1T ($n = 103$)	458 (58)	441 (51)	520 (51)	0.17 (0.24) [0.13, 0.22]	0.82 [0.59, 1.04]
1P3T ($n = 96$)	458 (57)	436 (51)	545 (54)	0.24 (0.28) [0.19, 0.30]	1.08 [0.75, 1.39]
3P3T ($n = 100$)	509 (59)	482 (55)	575 (53)	0.28 (0.32) [0.23, 0.35]	1.03 [0.75, 1.27]

Note. Mean response times (in ms; SD in parentheses), mean within-participant CIT-effect (SD in parentheses; 95% credible interval in brackets), and probe-irrelevant effect size δ (95% credible interval in brackets) estimated with a Bayesian Wilcoxon signed-rank test.

We used the BayesFactor package (Morey & Rouder, 2018) extension for R (R Core Team, 2020) to conduct a Bayesian fixed effects ANOVA with a JZS prior (Cauchy prior with scale = .5) to test for effects of number of probe modalities and number of target modalities on the within-participant CIT-effect ($dCIT$). The data was most likely under M_{Tar} , the model containing only the main effect of number of target modalities (Table 3). We found very strong evidence for this model compared to the null-model M_0 ($BF_{Tar,0} = 37.1$), supporting the hypothesis that there is an effect of number of target modalities on the within participant CIT-effect ($dCIT$). Both model comparison that can be used to assess the effect of number of probe modalities showed moderate evidence against such an effect ($BF_{Pro,0} = .17$; $BF_{Tar,Main} = 5.3$). Those comparisons show that the data is about 5 times more likely under the model

³ Due to the low error rate and following the preregistration, errors rates were not analysed. Error rates per condition are presented in the supplementary materials on OSF (<https://osf.io/d536j/>).

without an effect of number of probe modalities. We also found evidence against an interaction effect ($BF_{Main,Full} = 6.3$; $BF_{Tar,Full} = 33.5$).⁴ Therefore, our results showed a benefit of using multiple target modalities but no effect of multiple probe modalities.

The magnitude of the effect of target modality was assessed by the parameter's (β_{Tar}) posterior distribution of M_{Tar} ($M \beta_{Tar} = .10$; 95% HDI = [.04, .15]). Because the 95% HDI does not include 0, it can be concluded that presenting the targets in three different modalities instead of one lead to a credible increase of the within participant CIT-effect ($dCIT$) of .1, on average. This effect is independent of the number of probe modalities, since the model comparison showed evidence against an interaction effect.

Table 3. Bayesian model comparison

Models	Compared to the intercept only model (i.e., $BF_{Model, 0}$)	Compared to the best model (i.e., $BF_{Tar, Model}$)
M_{Pro} $dCIT \sim nProbes$	0.17 ($\pm 0\%$)	213.38 ($\pm 0\%$)
M_{Tar} $dCIT \sim nTargets$	37.12 ($\pm 0\%$)	1.0 ($\pm 0\%$)
M_{Main} $dCIT \sim nProbes + nTargets$	7.06 ($\pm 0.97\%$)	5.26 ($\pm 0.97\%$)
M_{Full} $dCIT \sim nProbes + nTargets + nProbes \times nTargets$	1.11 ($\pm 1.06\%$)	33.46 ($\pm 1.06\%$)

Note. In the first column, the models are compared to the null-model. Bayes factors > 1 indicate that the data is more likely under this model than the null-model. In the second column, the best model (M_{Tar}) is compared to the other models. Bayes factors indicate how much more likely the data is under M_{Tar} compared to the other models.

DISCUSSION

Some situations in practice require the single probe protocol of the RT-CIT that tests for one item per block, but it has lower validity than the multiple probe protocol that tests several items per block. The present study aimed at investigating how the validity of the single probe protocol could be improved to make it applicable in a wider range of situations. We tested for one piece of information (nationality) and presented it either in one or in three different modalities (word, flag, map). We independently manipulated the number of target modalities and probe modalities. We found moderate evidence against an effect of number of probe modalities on the within-participant CIT-effect ($dCIT$), but strong evidence

⁴ A reviewer proposed an alternative model comparison (using Baws factors), which essentially lead to the same results. This analysis can be found in the supplementary materials on OSF (<https://osf.io/d536j/>).

for an effect of number of target modalities. This study suggests that the validity of the single probe protocol can be increased not by presenting the probes in different modalities, but by presenting the targets in different modalities.

The impact of the number of target modalities is especially interesting, as researchers initially thought targets would not matter at all and could even be discarded (e.g., Matsuda et al., 2009). Nevertheless, it has become clear that the target items influence the CIT-effect. Gamer et al. (2007) argued that perceptual similarity between target items and test items (probes and irrelevant) influences the encoding of test items. Dissimilar target items can be identified easily without deep encoding, leading to smaller probe-irrelevant differences. However, not only perceptual similarity seems to impact the CIT-effect. Suchotzki et al. (2018) manipulated the familiarity of the target items. They argued that this increased the feature overlap (in the familiarity feature) between targets and probes, and therefore the response conflict for probes, which lead to larger probe-irrelevant differences. Suchotzki et al. (2018) also observed a small increase in the probe-irrelevant difference when four targets were used as compared to two targets. A crucial difference to our study is that they added targets (e.g., ‘Spain’ and ‘Greece’) whereas we presented the same target in different modalities (e.g., flag and name of Spain). This increased the number of semantic objects participants had to keep in mind.

The reason why we found larger CIT-effects in the three target modality conditions could have been because it might have altered the way examinees approached the task. With a single target modality, examinees can perform the task by focussing on a unique perceptual feature of the target (while attempting to ignore other features). Such perceptual processing reduces the influence of other features needed for the CIT effect (i.e., familiarity, saliency). Targets share those features with the probes but not the irrelevant, which leads to slower RTs for probes than for irrelevant (Gamer et al., 2007; Suchotzki et al., 2018). With multiple target modalities, there is not a single perceptual feature that allows to do the task and therefore requires semantic encoding. This leads to the incorporation of a wider array of features in the decision-making process, including features that lead to response conflict for probes. We call this explanation *target focus hypothesis*. A large body of research on the Stroop effect (Stroop, 1935) and the Garner interference (Garner, 1974) showed that it is possible to primarily focus the attention on the feature of interest but other, irrelevant, features are often not ignored

completely and thus influence the decision also (for reviews, see Algom & Fitousi, 2016; MacLeod, 1991). Therefore, two crucial processes for the target focus hypothesis, focused attention on specific features and integration of information from multiple feature dimensions into a binary decision, have been shown in other paradigms. Note that a connection between Stroop-like interference and the CIT-effect has been suggested before (e.g., Seymour & Schumacher, 2009). The addition of other ways to increase the reliance on these conflict inducing features, as it was done by using familiar targets (Suchotzki et al., 2018) or familiarity-related filler items (Lukács et al., 2017) could improve our multiple modality single probe protocol even further.

It is fortunate, especially for applied purposes, that it seems to be sufficient to increase the target modalities rather than the probe modalities. It would be an additional restriction if only probes could be used for which different presentations not only exist, but for which the examinee also has a strong internal representation – an important factor to obtain strong CIT-effects (Geven et al., 2019). For example, if the police want to test someone only on the pseudonym of a cybercriminal, it would be very challenging to find different visual modalities for that pseudonym.

Limitations & Future Studies

This study exclusively looked at nationality information and limited the number of countries that were included in the RT-CIT. Of those seven countries, 75% of participants were British or Portuguese. While we expect that our results generalize to other countries, a replication study with a more diverse population and a balanced design should be conducted.

Due to the randomization and the nature of an online study, we cannot rule out some degree of selective attrition. However, it seems unlikely that it is a major issue since the group sizes were very similar.

Furthermore, we cannot exclude that our results are unique to the modalities (word, flag, map) used in this study. While words are commonly used when testing for participants' nationality, flags and maps have not been studied often with the RT-CIT. Exploring the possibilities of switching between

visual and auditory presentation within a block or other means that prevent participants to focus on a perceptual feature (e.g., using synonyms or very closely related stimuli as targets [e.g., gun, pistol, firearm]) seems highly relevant for applied purposes.

An obvious but important limitation to consider from an applied perspective is that the increased validity for multiple target modalities might not generalize from the RT-CIT to the physiological CIT, the only version currently applied in the field (Osugi, 2018). It is likely that different mechanisms are involved in the RT-CIT and the physiological CIT (see e.g., Klein Selle et al., 2018; Seymour & Schumacher, 2009).

With the introduction of familiarity related fillers (Lukács et al., 2017) and our multiple modality approach, we now know of two ways to increase the validity of the single probe protocol. We want to encourage future research to explore the combination of both approaches.

Conclusion

Our findings show that in situations in which only one piece of information is available for testing, the validity of the RT-CIT can be increased by using multiple modalities for the target item. Presenting the target in several modalities may prevent a purely perceptual way of processing and assure semantic processing of the stimuli in the RT-CIT. This brings about the processing of feature dimensions that induce response conflict and therefore the CIT-effect.

Author Contributions

Dave Koller proposed the initial study design which was refined in collaboration with Bruno Verschuere. Programming, data collection, and analysis was done by Dave Koller. The manuscript was mainly written by Dave Koller but in close collaboration with Bruno Verschuere and in consultation of Franziska Hofer. Franziska Hofer and Bruno Verschuere supervised the project.

Acknowledgments

We thank the Swiss Federal Office of Civil Aviation (project number: 2016-106) and the Zurich State Police, Airport Division for their financial support.

CHAPTER 3

Nationality check in the face of information contamination: Testing the inducer CIT and the autobiographical aIAT

This chapter is currently under revision for:

Psychology, Crime & Law

Initial submission date: 20.5.2021

ABSTRACT

Government agencies including border control have an interest to detect if someone provides false information about their nationality. While response time tasks have been proposed to be able to detect someone's true nationality, there is the risk that they will often err, particularly in the face of information contamination (i.e., someone having thorough knowledge of the country). We screened 2,200 participants to create three groups: Dutch participants (n=118), and British participants with (n=99) and without knowledge of the Netherlands (n=118). They were tested with either the autobiographical Implicit Association Test (aIAT) or the Inducer-Concealed Information Test (I-CIT). While both tests could discriminate Dutch participants from British participants without knowledge of the Netherlands ($AUC_{I-CIT}=.65$; $AUC_{aIAT}=.88$), only the aIAT could also discriminate Dutch participants from British participants with knowledge ($AUC_{I-CIT}=.52$; $AUC_{aIAT}=.86$). Therefore, the aIAT, but not the I-CIT, could be a helpful tool to detect false nationality claims, even when information contamination is suspected.

Keywords: Concealed Information Test, autobiographical Implicit Association Test, Memory Detection, Deception Detection, Reaction time

INTRODUCTION

The Covid-19 pandemic forced the world to drastic measures. In Europe, the pandemic challenges the core principles of the European Union: solidarity, policy coordination, and especially the free movement across national borders (Biancotti et al., 2020). Being able to travel between countries and continents both due to available means of transportation and international policy agreements is one major achievement of modern human society. However, the motivation to travel is not always due to innocent or voluntary nature. For example, human trafficking, international terrorism, and structural criminality are some of the negative concomitants of globalization. According to the report of the United Nations, the number of international migrants worldwide reached nearly 272 million in 2019, of which about 24 million are considered refugees and asylum seekers (United Nations, 2019). Between 2010 and 2017, the number of refugees and asylum seekers increased by about 13 million (World Bank, 2019). The process for asylum is clearly regulated in the law. In the case of Switzerland, the Asylum Act states the possibilities to get asylum. Furthermore, in most cases the application for asylum must be submitted directly at the border. Usually, according to the Swiss State Secretary of Migration an interrogation is conducted about the travel itinerary, family background, and other aspects. However, the time for the final decision can take years (Roos et al., 2018; Swiss State Secretary of Migration, 2020). To this end, the person seeking asylum has most likely endured an exhausting, maybe even dangerous journey, and is often not allowed to take a job. In most cases, if the application for asylum is rejected, the person has to go back to the home country independent of the time spent abroad (Federal Act on Foreign Nationals and Integration, 2005). Consequently, whether due to criminal intention or despair, one might be motivated to provide false information about the own nationality and other autobiographical aspects (e.g., travel itinerary, already applied asylums, family background) to increase the chance of getting the application for asylum approved. In other words, people might lie about their true nationality to be permitted to stay in the country or to be allowed to take a job. Let us take a look at the following example: Due to the humanitarian catastrophe in Eritrea (European Asylum Support Office, 2019), asylum applications were very likely to be approved (European Asylum Support Office, 2015). This was not the case for the neighbouring country Ethiopia. However, Ethiopia is also facing political instability (e.g., Pilling & Schipani, 2020). Therefore, for someone with Ethiopian origin suffering from hunger and

having lost hope of a better future, it could be attractive to seek asylum by falsely claiming to be from Eritrea. As the person might be able to tell a lot about Eritrea – perhaps because of thorough research, family history, or because of having lived in Eritrea – it could be feasible to pretend to have an Eritrean origin and hide their true nationality. On the other hand, a person from Eritrea might mistakenly be accused to be from Ethiopia and rejected for asylum. So, the question is: is it possible to detect the true nationalities?

In general, it is a well replicated fact that humans are not very good at detecting whether someone is telling the truth or not (on average 54% accuracy; e.g., Bond & DePaulo, 2006, 2008; Hartwig & Bond, 2011). Specific techniques, as for example those arising from the cognitive approach of lie detection (e.g., imposing cognitive load, encouraging to say more, or asking unanticipated questions; see Vrij, 2015 for a review) can be used in order to amplify verbal and nonverbal differences between liars and truth tellers. In a meta-analysis, Vrij and colleagues assessed the effectiveness of this approach to be around 71% correct classification (Vrij et al., 2017). While that seems to be an improvement compared to the 54% base rate, the authors' analysis has been criticized by Levine et al. (2017) stating that the true detection rate of the cognitive load approach is much lower, and the benefit was overestimated. Moreover, the most recent meta-analysis points out that the benefit of the cognitive approach is reduced remarkably when considering publication bias (Mac Giolla & Luke, 2020).

Apart from novel interviewing techniques, researchers also worked on technology-assisted ways to detect deception. A well-investigated test is the reaction time (RT)-based Concealed Information Test (CIT) originally named as Guilty Knowledge Test (Lykken, 1959; Seymour et al., 2000; for a review see Verschuere & De Houwer, 2011). After Kleinberg and Verschuere (2015) showed the RT-CIT can be run reliably and validly via the internet, this test has also gained attention by practitioners. The test allows to identify concealed knowledge, i.e., it can detect if somebody has specific knowledge about a topic of interest (e.g., a country, crime, name, etc). In the RT-CIT, a series of stimuli is presented on the computer screen. There are three types of stimuli: the item of interest (so-called probes), irrelevant items that have no specific relevance to the participants nor are connected to the case, and targets. Targets are used to make sure that participants are paying attention and processing the stimuli. For the Eritrea-Ethiopia asylum case, an RT-CIT could be operationalized in the following way: the probes

could be names of towns, locations, mountains, or products of Ethiopia that are mainly known by indigenous people and rather not known by non-indigenous. Target items could be names of towns, locations, mountains, or products of Eritrea. Irrelevants could be names of rather unknown towns, locations, mountains, etc. of other eastern African countries. The question used in the RT-CIT would then be “Is this connected to your home country?”. Participants would then be instructed to respond with “Yes” to the targets (names of towns, locations, mountains, or products of Eritrea) and with “No” to all other items. People with an Ethiopian origin recognize the probes and therefore show longer RTs for the probes than for the irrelevant due to response conflict (Seymour & Schumacher, 2009; Verschuere & De Houwer, 2011). Thus, recognition is inferred based on systematically longer RTs to Ethiopian items (i.e., the probes) as compared to the irrelevant stimuli (i.e., towns in other African countries).

Research shows that the RT-CIT constitutes a reliable and valid method of detecting concealed recognition and the potential applications are manifold. The meta-analysis of 114 studies on 3307 participants by Suchotzki et al. 2017 showed a large effect ($d = 1.049$; 95% CI [0.930; 1.169]). One major concern, the RT-CITs vulnerability to faking, was recently addressed by Suchotzki et al. (2021). In a series of studies, they showed that faking was ineffective when participants only had a short response window. However, another major restriction is that the RT-CIT and the physiological CIT are not immune against information contamination, meaning that validity is threatened if the critical information (probe) could also be recognized by innocent examinees (Lukács & Ansorge, 2019; see Bradley et al., 2011 for a review). For example, a person from Eritrea seeking asylum could know a lot about Ethiopia, e.g., because he/she has worked or lived in the country, has heard news about Ethiopia in the press, or just because of personal interest. In this case, the person would show longer RTs to the probe although not being Ethiopian.

Depending on the context, this so-called information contamination could happen quite often. Consequently, lack of immunity of the CIT against information contamination is one of the most often mentioned concerns of practitioners (see Podlesny, 2003; and interviews conducted by the authors with Swiss law enforcement offices of different branches). Theoretically, there are at least two ways to tackle this challenge. In the context of a crime, information contamination could be mitigated by a very concise information management by the investigative authority. However, for cases such as the mentioned

asylum example, this possibility seems less feasible as due to globalization, digitalization, and social media information are distributed worldwide. For example, a person stating to be from a specific country has many possibilities to gain as much knowledge as possible about this country. He or she could learn online about the politics, could walk on the streets in the capital city with Google maps, and could follow people on social media twittering about daily events. Keeping information within a restricted group of people poses an increasing challenge in the globally networked world. Therefore, the second possibility to reduce the negative effect of information contamination could be achieved by further developing the RT-CIT method to increase the method's robustness against information contamination.

In this regard, Lukács and Ansoerge (2019) recently proposed an adjusted RT-CIT. This Inducer-CIT (I-CIT) is closely related to the Association-based CIT (Lukács et al., 2017). The I-CIT utilizes a shared feature (e.g., home-relatedness) between the so-called inducer items and the probes to induce response conflict. To illustrate the principle, we again come back to our Eritrea-Ethiopia example. Imagine that you are an asylum seeker from Eritrea that has relatives in Ethiopia which you visit regularly. Therefore, you know the name of different lesser-known Ethiopian mountains, towns, and products. In the I-CIT you are now asked to categorize home-referring inducers (e.g., "HOME", "NATIVE"), foreign-referring inducers (e.g., "FOREIGN", "OTHERS"), and various names of mountains, products, and towns. The name of the Ethiopian items and the foreign-referring inducers share the same response key. While you recognize the names of Ethiopian mountains, towns, and products, this is not a problem for you. Residents of Ethiopia, however, would experience response conflict. For them, Ethiopian items refer to home, but they must not press the response key associated with the home-referring inducers. Lukács and Ansoerge (2019) used autobiographical information to show that informed innocent participants did not show an I-CIT-effect but the group that has been instructed to hide their identity did. The I-CIT was therefore immune to information contamination.

The I-CIT relies on the similar assumptions as the autobiographical Implicit Association Test (aIAT; Sartori, et al., 2008), a test for assessing autobiographical memory. Autobiographical memory refers to events that constitute part of one's life and are part of the long-term memory (Tulving, 1983). The aIAT is a variant of the Implicit Association Test - a reaction-time based test to measure associations between two different categories or concepts that one is not willing or not able to reveal (Greenwald et

al., 1998). With the aIAT it is possible to evaluate which one of two autobiographical events is true (for a review see Agosta & Sartori, 2013). True events are identified by shorter RTs when sharing the same response key as the category ‘true’ compared to when the event shares the response key with the category ‘false’. Furthermore, the aIAT can also be used to identify intentions and motives of past and future actions (Agosta et al., 2013; Zangrossi et al., 2015). Verschuere and Kleinberg (2017) showed that the aIAT can also be used successfully in web-experiments for assessing autobiographical memory. Besides the applied potential of the aIAT, it must be taken into account that the meta-analysis by Suchotzki et al. (2017) reports lower average effect size compared to the CIT. One explanation could lie in the different design of the aIAT itself, as trials with and without response conflict are presented in two different blocks, whereas this is not the case in RT-CIT studies. Participants can therefore try to control their behavior block-wise in the aIAT, e.g., try to slow down their responses in the block without response conflict.

The goal of this study was to examine the impact of information contamination on the I-CIT and the aIAT. To our knowledge, this is the first study that investigates information contamination on the aIAT. This was done with an online experiment using a scenario of identifying the true nationality. For the study, Dutch and British participants were recruited. Within the British participants there were two groups: The information-contamination group consisted of British people which all had good knowledge about factual details of the Netherlands. The other group consisted of British people without any specific knowledge about the Netherlands. Therefore, these three groups resulted: Dutch group, UK naïve group, and UK contamination group. Our main predictions were that the I-CIT and the aIAT show above chance classification performance both when discriminating the Dutch from the UK naïve as well as when discriminating the Dutch from the UK contamination group. We further test the hypotheses that the predictors of the I-CIT ($dCIT$) and the aIAT (D_i) are larger in the UK contamination group than in the UK naïve group which would be a sign against immunity to information contamination. Finally, we investigated if there is a difference in the Dutch/UK classification performance between the I-CIT and the aIAT.

METHOD

The experiment was approved by the ethical committee of the Faculty Social and Behavioral Sciences of the University of Amsterdam (approval number: 2020-CP-2352). Data, code, and preregistration are publicly available on <https://osf.io/8wyf6/>.

Deviations from Preregistration

After data collection, but before testing the hypotheses, we realized that our post-test recognition test was suboptimal and would have led to the exclusion of most participants. We therefore did not apply the strict participants exclusion criteria based on the post-test recognition (i.e., 1) that participants in the Dutch group and in the UK contamination group need to indicate all the correct and only the correct items with confidence ≥ 5 and 2) that participants in the naïve group must not indicate any of the correct items with confidence ≥ 5). Because participants could indicate as many items as they wanted, most participants in the knowledgeable groups indicated more than three items, which would exclude them automatically. These strict criteria would have led to the exclusion of 64% and 76% of the participants in the aIAT and I-CIT respectively.

Procedure

Stage 1 - Screening

Dutch and British participants that were between 18 and 50 years old were eligible to participate in the screening. This took participants about 3 minutes to complete, and participants were compensated with 0.35 GBP (≈ 0.50 USD). Participants were recruited over the online platform Prolific.

Participants provided informed consent. Participants were asked to indicate the country that they would call their ‘home’ and their ‘native’ country, to estimate the amount of time they spent in the Netherlands in their entire life and to rate their knowledge about the Netherlands on a 7-point Likert scale (1= very poor, 7 = very good). Then, participants were presented with ten single-choice knowledge

questions about the Netherlands in which participants needed to indicate which of the five options (e.g., Zeldonk, Ipelo, Utrecht, Winddicht, Omert) is connected to the Netherlands. For the complete list of items, see <https://osf.io/8wyf6/>. To discourage participants to look up the answers, we did not use incentives for correct answers and each question needed to be answered within 10 seconds. After every question, participants were asked to indicate their confidence that the answer is correct on a 7-point Likert scale (1 = not confident at all; I guessed, 4 = relatively confident, 7 = extremely confident).

Because we were interested in the knowledge of participants, we required a question to be answered correctly and with a confidence of 5 or higher to be scored as ‘known’. Based on the answers and confidence ratings, we looked for the three questions that divide the screened participants most evenly into the following three groups while including as many of the screened participants as possible. Group 1 (Dutch): Participants who indicated the Netherlands as their home and native country, and who knew the answers to those questions about the Netherlands whose items will be used in the I-CIT. Group 2 (UK contamination): Participants who indicated the United Kingdoms as their home and native country, and who also knew the answers to those questions about the Netherlands. Group 3 (UK naïve): Participants who indicated the United Kingdoms as their home and native country, and who answered the questions about The Netherlands incorrectly or with a confidence of 1 (i.e., guessing).

Stage 2 – RT-Task

In a second stage, one to three days after the screening, participants were invited to the RT-task. The possibility to participate in the second stage ended eight days after the invitation. Participants again provided informed consent.

Participants were asked to imagine a scenario in which they need to use an online service that is not available for Dutch citizens. During the registration for this service, they indicated to be from the United Kingdoms. Participants were told that the service provider is required to use precautions against people providing wrong nationality information and that the following task is used to verify if they truly are from the United Kingdoms. Participants should try to convince this automated test to be from the United Kingdoms. This was the truth for the UK naïve and UK contamination group, and a lie for the Dutch group. Then, the task specific instructions followed, which differed depending on the assignment

to either the I-CIT or the aIAT task. Task assignment was done based on the participant number which was randomly assigned by Inquisit (version 6.2.1; Inquisit, 2020) – the software used for the RT-tasks.

Inducer-Concealed Information Test

The I-CIT was mimicked after Lukács and Ansorge (2019). Participants were asked to answer the question ‘Is this connected to my home?’. They were presented with the three YES-inducer items (‘Home’, ‘Native’, ‘Local’) and six NO-inducer items (e.g., ‘Foreign’, ‘Abroad’, ‘Others’). Participants were told that, to be convincing, they need to press YES (‘A’-key) only when one of the YES-inducer items appeared. In all other cases they need to press NO (‘L’-key). Figure 1 shows an exemplary segment of the I-CIT. The complete list of items used in the I-CIT is displayed in Table 1.⁵ They were also instructed to respond as fast and accurately as possible. Feedback was given throughout the task by a ‘TOO SLOW’ appearing on top of the stimulus after 800ms until a response was given or the response deadline of 1500ms was reached, and by a ‘WRONG’ message appearing below the stimulus for 400ms in case of an error. Participants completed a practice block with 24 trials in which each item was presented once in random order. If participants had less than 50% correct for any item type, the practice block was repeated (up to two times). The main task contained three blocks, one for each information (Location, City, Club) with a short, self-paced break between blocks. Each block started with 8 burn-trials that were excluded from the analysis to get participants accustomed to the task again after the break. Then followed the 135 test trials that were divided into three sub-blocks of 45 trials (5 YES-inducer, 10 NO-inducer, 6 probes, 24 irrelevant). Item order within a sub-block was randomized.⁶ Every inducer was presented 5 times, and every irrelevant or probe item 18 times per block. This results in a total of 405 test trials (54 probe, 216 irrelevant, 45 YES-inducer, 90 NO-inducer).

⁵ Due to changes in the scenario, we could not use the same inducer items as Lukács and Ansorge (2019). However, we contacted the developer of the I-CIT, Gáspár Lukács, to discuss our adjustment before data collection. He agreed that the inducers should be adjusted to the scenario and suggested a minor change to one of the inducer items which we incorporated.

⁶ Note that this is more random than in the original study by Lukács and Ansorge (2019) in which an inducer was never followed by another inducer.

Figure 1. *Exemplary Segments of the I-CIT*

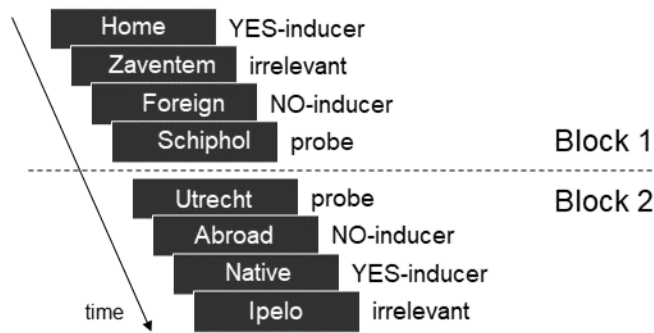


Table 1. *Items used in the I-CIT*

Item Category	Items		
	Location	City	Club
Irrelevant	Zaventem, Kortrijk,	Winddicht, Omert,	Cres, Teucer, Eetion,
	Ondermeer, Oostende	Ipelo, Zeldonk	Priam
Probe	Schiphol	Utrecht	Ajax
YES-Inducer	Home, Native, Local		
NO-Inducer	Foreign, Abroad, Others, Theirs, Another, Alien		

Autobiographical Implicit Association Test

The scenario in the aIAT condition was the same. Participants were asked to use the ‘A’ and ‘L’-keys to categorize statements into response categories shown on the top left and top right of the screen. Statements that were presented in green (logical statements) needed to be categorized into TRUE or FALSE; statement presented in white (autobiographical statements) into ‘I am from the Netherlands’ and ‘I am from the United Kingdoms’ (see Table 2 for the complete list of statements).⁷ Participants were instructed to answer as fast and accurately as possible. Feedback was given throughout the task by a ‘WRONG’ message appearing below the statement until the correct response was given (following the D_1 scoring scheme of Greenwald et al., 2003). The aIAT is divided into seven blocks

⁷ We consulted the developer of the aIAT, Giuseppe Sartori, about the labels and items (logical and autobiographical) and adapted two logical items according to his suggestions.

with brief instructions at the beginning of each block. The logical discrimination block (block 1, 20 trials) consisted of logical statements only. The initial autobiographical discrimination block (block 2, 20 trials) consisted of autobiographical statements only. In the initial double categorization practice block (block 3, 20 trials), and in the initial double categorization block (block 4, 60 trials) participants alternately saw logical and autobiographical statements and needed to classify them both according to their respective, color coded, categories. The reversed autobiographical discrimination block (block 5, 20 trials) consisted of autobiographical statements only, but with reversed category-key bindings compared to block 2. Finally, the reversed double categorization practice block (block 6, 20 trials) and the reversed double categorization block (block 7, 60 trials) combined the category-key bindings from block 1 and block 5. The category-key bindings of block 2 were assigned based on the subject number which was randomly generated by Inquisit. Data from blocks 3, 4, 6, and 7 are used in the analysis (see D_t measure in Greenwald et al., 2003).

Table 2. *Statements used in the aIAT*

Category		Statements
Logical	True	I am in front of a screen
		I am participating in a study
		I am reading words
		I am classifying sentences
	False	I am using a keyboard
		I am climbing a mountain
		I am in the sea
		I am looking at birds
Autobiographical	The Netherlands	I am watching a movie
		I am using a pan
		The Netherlands is my home country
		I am native to the Netherlands
		My nationality is Dutch
	United Kingdoms	I own a Dutch Passport
		I come from the Netherlands
		The United Kingdom is my home country
		I am native to the United Kingdom
		My nationality is British
I own a British Passport		
I come from the United Kingdom		

Post-test recognition test

After the RT-task, participants were asked to select the items from the list of all probes and irrelevant items for which they are ‘fairly certain to very certain’ that they are connected to the Netherlands.⁸ Finally, participants were thanked, debriefed, and given the code to be entered on Prolific.

Participants

After screening the preregistered maximum of 2200 participants (300 Dutch, 1900 British), we had 209 participants in the Dutch group, 155 in the UK contamination group, and 260 in the UK naïve group, for a total of 624 participants eligible for the RT task. Of those, 355 (57%) completed the second stage. Twenty participants (6%) were excluded from the analysis. One participant revoked the consent, 19 participants (8 Dutch, 2 contamination, 9 naïve) did not reach the pre-registered minimal requirement of at least 40% correct on each item type in the I-CIT.⁹ No participant needed to be excluded in the aIAT condition. The final sample consisted of $N = 335$ participants (54% of the invited participants). It took participants about 10 minutes to complete the RT-task which was reimbursed with 1.10 GBP (≈ 1.50 USD). Demographic information per condition is shown in Table 3.

⁸ On average, participants selected 4.7 items even though only 3 items were connected to the Netherlands. While this is a sign that we chose realistic items, it also indicates that participants applied a low threshold of certainty to select an item. However, the post-test recognition test can only serve its purpose as a knowledge check when a high certainty threshold is applied. Therefore, we did not use data of the post-test recognition test.

⁹ While this seems like a low benchmark, there is a strong response tendency towards the ‘no’ button in the I-CIT because participants need to press ‘yes’ only for home-referring inducers which make up only 11% of test trials.

Table 3. *Demographic information*

		Dutch	Contamination	Naïve	Collapsed
I-CIT	N	45	44	64	153
	Sex	33% female	20% female	81% female	49% female
	Age	26.8 (7.4)	34.5 (7.7)	27.7 (8.4)	29.4 (8.5)
aIAT	N	73	55	54	182
	Sex	34% female	36% female	70% female	46% female
	Age	25.9 (7.2)	34.2 (8.0)	28.1 (7.5)	29.1 (8.3)
Collapsed	N	118	99	118	335
	Sex	34% female	29% female	76% female	47% female
	Age	26.3 (7.3)	34.3 (7.8)	27.9 (8.0)	29.2 (8.4)

Note. Standard deviations are reported in the parentheses.

RESULTS

Following the preregistration, we excluded incorrectly answered trials and trials with latencies smaller than 200ms or larger than 1499ms from the analysis in the I-CIT. A total of 3480 trials (5.6%; 2695 inducer trials, 785 non-inducer trials) were excluded. In the aIAT, we only excluded trials with latencies larger than 10,000ms as preregistered and suggested by Greenwald et al. (2003). A total of 6 trials (0.02%) were excluded.

To classify participants in the I-CIT condition, we used the normalized within participant probe-irrelevant difference $dCIT = \frac{M_{RT(\text{probe})} - M_{RT(\text{irrelevant})}}{SD_{RT(\text{irrelevant})}}$ (Kleinberg & Verschuere, 2015). Dutch participants are expected to show a positive $dCIT$ score indicating recognition of the probes and a link between the probe and the participants' home, according to Lukács and Ansorge (2019). Naïve participants are expected to show a $dCIT$ score around zero indicating that the probe could not be distinguished from irrelevant items. If the I-CIT indeed is immune to information contamination (Lukács & Ansorge, 2019), then we expect $dCIT$ scores of the contamination group to be around zero (because

the probe is not linked to the participants' home); if not, we expect positive *dCIT* scores (due to probe recognition).

Individual classification in the aIAT was done using the D_I -measure proposed by Greenwald et al. (2003). D_I is an equal weight average between $D_{Practice}$ and D_{Test} . ($D_I = \frac{D_{Practice} + D_{Test}}{2}$). $D_{Practice}$ and D_{Test} are the differences in mean RT between the congruent versions (the labels 'TRUE' and 'I am from the United Kingdoms' share the response key) and incongruent versions (the labels 'TRUE' and 'I am from the Netherlands' share the response key) of the respective block divided by their pooled standard deviation (e.g., $D_{Test} = \frac{M_{RT}(\text{congruent testblock}) - M_{RT}(\text{incongruent testblock})}{SD_{Pooled}}$, with $SD_{Pooled} = \sqrt{\frac{(n_{congruent} - 1) * var(RT_{congruent}) + (n_{incongruent} - 1) * var(RT_{incongruent})}{n_{congruent} + n_{incongruent} - 2}}$). D_I values greater than zero result from slower responding in congruent blocks than in the incongruent blocks (i.e., a response conflict in the congruent block). Therefore, positive D_I values are expected for Dutch participants. Negative D_I values on the other hand indicate British nationality.

While we rely, per our preregistration, solely on the *dCIT* and D_I scores for hypothesis testing, we report RTs for both tasks in Table 4 to provide a comprehensive picture of how participants of the three conditions responded in the two tasks. Inspection of Table 4 shows larger probe-irrelevant differences for knowledgeable participants (Dutch, UK contamination) than for naïve participants in the I-CIT. Table 4 further shows that the Dutch group had longer RTs for the TRUE/ 'I am from the United Kingdoms' block than for the TRUE/ 'I am from the Netherlands' block. Both UK groups showed the reversed pattern.

To help assess the evidential value of the seemingly small between group differences in the I-CIT, we conducted a Bayesian analysis of variance with a Cauchy prior (scale = .5) of group (Dutch, UK contamination, UK naïve) on the probe-irrelevant difference. This analysis was not preregistered. The data were 7 times more likely under the hypothesis that there is an effect of group on the probe-irrelevant difference than under the null-hypothesis without a group effect ($BF_{10} = 7.0$). We also conducted post-hoc group comparisons using a two-sided Bayesian t-test with a Cauchy prior (scale = .707). The uncorrected Bayes factors indicating how many times more likely the data is under the

hypothesis that there is a difference between the respective groups relative to the null-hypothesis were: $BF_{\text{Dutch}/\text{Naïve},0} = 7.2$, $BF_{\text{Dutch}/\text{Contamination},0} = .23$, and $BF_{\text{Naïve}/\text{Contamination},0} = 7.3$. The same Bayesian analysis for the aIAT showed the data was 4×10^{11} times more likely under the hypothesis that there is an effect of group relative to the null-hypothesis ($BF_{10} = 4.0 \times 10^{11}$). The results of the individual post-hoc comparisons show that the data were more likely under the alternative hypothesis (i.e., that there is a between group difference) relative to the null-hypothesis for Dutch vs. UK naïve and Dutch vs UK contamination comparison but not for the UK naïve vs UK contamination comparison ($BF_{\text{Dutch}/\text{Naïve},0} = 3.9 \times 10^9$; $BF_{\text{Dutch}/\text{Contamination},0} = 1.4 \times 10^8$; $BF_{\text{Naïve}/\text{Contamination},0} = .33$).

Table 4. Mean reaction times

Group	I-CIT			aIAT		Block Difference
	Probe	Irrelevant	Probe-Irrelevant Difference	True/ 'I am from the UK' and False/ 'I am from the Netherlands'	False/ 'I am from the UK' and True/ 'I am from the Netherlands'	
Dutch	442 (50)	438 (50)	4.2 (15.8)	1026 (183)	962 (216)	64 (166)
UK contamination	426 (45)	423 (43)	3.2 (11.9)	823 (159)	983 (224)	-159 (185)
UK naïve	406 (46)	409 (45)	-3.4 (11.7)	887 (213)	1087 (213)	-199 (211)

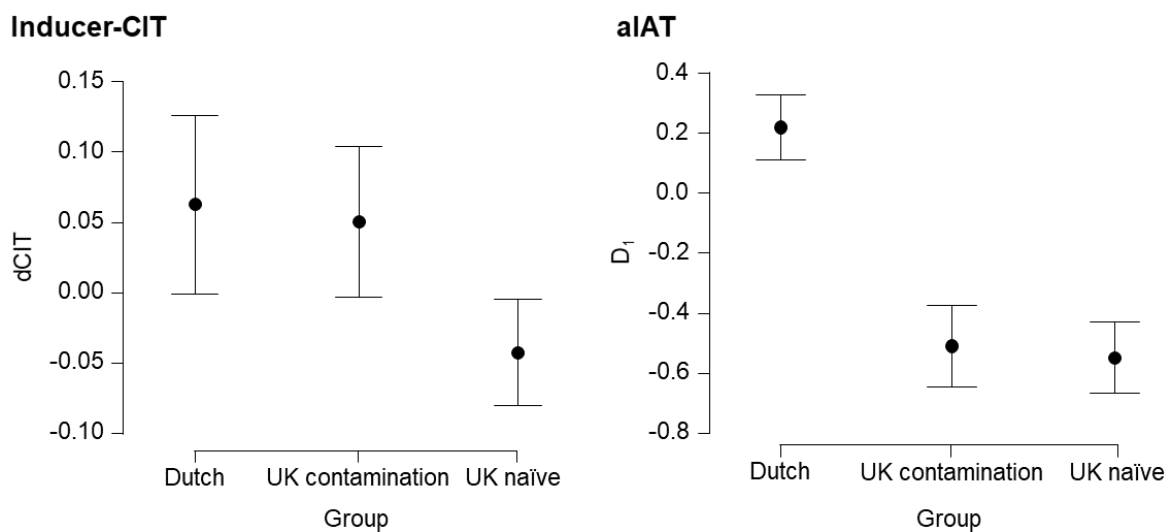
Note. RTs are reported in ms, standard deviations are reported in parentheses.

Without information contamination (i.e., discrimination between Dutch and UK naïve participants), the classification performance of the I-CIT was slightly above chance with an area under the receiver operator characteristics curve (AUC) of .65 (95% CI: [.55, .76]). In other words, the I-CIT can distinguish between Dutch and naïve British participants above chance, but only slightly. The critical classification in a scenario with information contamination, however, is between the Dutch and the UK contamination group. The I-CIT could not distinguish between the two groups (AUC = .52; 95% CI: [.40, .64]) which is not significantly different from chance level performance.

Using D_I for the same classifications in the aIAT conditions showed above chance performance, not only for the Dutch versus UK naïve classification (AUC = .88; 95% CI: [.82, .94]), but also for the critical Dutch versus UK contamination classification (AUC = .86; 95% CI: [.80, .93]). These results show that not only can the aIAT also successfully distinguish between Dutch and knowledgeable British participants, it can do so with a performance comparable to the Dutch-naïve classification.

While successful Dutch versus UK contamination classification is crucial for applied purposes, this is not sufficient to show that a test protocol is immune to information contamination. If a test protocol is truly immune to information contamination, then the index used to classify a person should not be influenced by information contamination. Two separate Bayesian one-sided independent samples t-tests on $dCIT$ and D_I with Cauchy priors (scale = .707) showed strong evidence that $dCIT$ for the contamination group is larger than for the naïve group ($BF_{10} = 17.8$) and moderate evidence against the larger D_I scores for the contamination group ($BF_{10} = .29$). The I-CIT was therefore not immune to information contamination while the aIAT showed little to no influence of information contamination (see Figure 2).

Figure 2. Mean I-CIT Scores ($dCIT$) and aIAT Scores (D_I)



Note. The 95% credible interval indicates the range in which the true parameter falls with a probability of 95%.

Finally, we directly compared the classification performance of the I-CIT and the aIAT in the task that resembles the problem practitioners face the most (i.e., to identify the Dutch participants among all three groups). We calculated DeLong's test for differences in ROC curves. The classification performances differed significantly, $D(226.6) = 4.65$, $p < .001$, with the aIAT showing more accurate classification (AUC = .87; 95% CI: [.82, .92]) than the I-CIT (AUC = .60; 95% CI: [.50, .70]).

Exploratory Analyses

To make sure that our deviation from the preregistration did not drastically change the results and conclusions, we also tested the main hypotheses on the reduced data set (i.e., after exclusions based on the post-test recognition; leaving $n = 103$ divided into $n_{I-CITdutch} = 10$, $n_{I-CITcontamination} = 6$, $n_{I-CITnaïve} = 21$, $n_{aIATdutch} = 20$, $n_{aIATcontamination} = 23$, and $n_{aIATnaïve} = 23$). The only result that qualitatively changed was that the I-CITs performance of discriminating Dutch and naïve UK participants was not above chance in the reduced data set (AUC = .58; 95% CI: [.33, .82]). Note that this result was based on only 31 participants (10 Dutch, 21 naïve UK). The aIAT showed high classification performance also in the reduced data set (AUC = .87; 95% CI: [.77, .98]). Regarding the Dutch versus UK contamination classification, the I-CIT did not perform above chance (AUC = .53; 95% CI: [.21, .85]) but the aIAT did (AUC = .82; 95% CI: [.70, .95]), confirming the results of the main analysis.

DISCUSSION

Computerized paradigms such as the RT-CIT have been shown to be able to detect that someone provides false information about their nationality. However, assessing recognition of details related to nationality, there is a risk the test will err for truth tellers who are knowledgeable about the tested nationality. The present study addressed this information contamination scenario, which is problematic for the traditional RT-CIT (Lukács & Ansorge, 2019). However, two other RT-based paradigms might not be affected by information contamination: the novel I-CIT (Lukács & Ansorge, 2019) and the aIAT (Sartori et al., 2008). We tested Dutch participants, and British participants with and without knowledge

about the Netherlands. We were primarily interested to see how well the two tests would perform in the most challenging situation: to discriminate Dutch participants from British participants with knowledge about the Netherlands (i.e., contamination group). Our results showed that the aIAT was able to discriminate between the Dutch and the contamination group, but the I-CIT was not.

Contrary to the results of Lukács and Ansoerge (2019), we found a larger I-CIT-effect for the UK contamination group than for the UK naïve group, suggesting the I-CIT is vulnerable to contamination. In their original study introducing the I-CIT, Lukács and Ansoerge (2019) used autobiographical information (country of origin, date of birth, and favorite animal) and familiarity related inducers (e.g., ‘Familiar’, ‘Recognized’, ‘Mine’, ‘Unfamiliar’, ‘Unknown’, ‘Foreign’ etc.). The inducers indicating familiarity required a YES response, all other stimuli including the probes a NO response. For guilty participants, they argued that because familiar inducers and the probes share the self-relatedness feature, but require a different response, there is a response conflict. This leads to longer response times for probes as compared to the irrelevant items (which are not self-related, hence no conflict). And indeed, they found the guilty group, but not the contaminated innocent group, to show an I-CIT effect. In contrast, we found I-CIT effects (of similar magnitude) both in the Dutch and the UK contamination group which is a strong indication that home-relatedness did not cause the effect. We see at least two possible reasons for this discrepancy: our implementation to test for home-relatedness was suboptimal or home-relatedness is not a suitable feature for the I-CIT. First, we tested for the nationality whereas Lukács and Ansoerge (2019) tested for the identity of participants. As a result, we used home-relatedness as the conflict inducing feature and adapted the inducer items accordingly (incorporating the feedback from the original paper’s first author). It could be, however, that the probes (Schiphol, Ajax, and Utrecht) were not strongly connected to “home” for all participants. For instance, for someone typically relying on the local airport instead of Schiphol. If this is true for a substantial part of Dutch participants, this could explain why did not find larger I-CIT effects for the Dutch than the UK contamination group. Second, home-relatedness might simply not be a suitable feature to induce response conflict in the I-CIT. Due to the lack of research on the I-CIT, we can only speculate about what properties a feature needs to reliably lead to an I-CIT effect (e.g., high saliency).

The I-CIT did not differentiate between the Dutch and UK contaminated group, but we did find a small I-CIT effect in both groups. If the response conflict was not induced by home-relatedness, what mechanism might explain this small I-CIT effect? We see at least two possibilities. First, our choice of irrelevant items could be argued to have caused an I-CIT-effect. We used fictional and lesser-known names as irrelevant items. This led to a feature (word familiarity) shared between the probes and the inducers. While this feature is not only shared with the YES- but also with the NO-inducers (since these are also words), some degree of response conflict or at least response uncertainty could have influenced the Dutch and the UK contamination group. Second, the probes were recognized as task relevant by the Dutch and the UK contamination group. This made the probes stand out among a majority of irrelevant items which could lead to an orienting response (Lykken, 1974; Sokolov, 1963) disrupting the decision-making and increasing the RTs (Verschuere et al., 2004; for a review see Verschuere & Ben-Shakhar, 2011). The slowing due to the orienting response is relatively small but comparable to the RT differences reported in our study. The two explanations are not mutually exclusive. But in either case response conflict due to home-relatedness cannot explain the observed I-CIT effects.

Verschuere and De Houwer (2011) argued that stimulus-response incompatibility and the resulting response conflict is crucial to find robust probe-irrelevant differences. In the classic RT-CIT, target items are used to manipulate the stimulus-response compatibility of probes depending on the participant's knowledge. Naïve participants can perform the task solely by judging the familiarity of items since only the targets are familiar to them. For knowledgeable participants, probes are also familiar, but they are instructed to press the key related to unfamiliarity leading to stimulus-response incompatibility. Suchotzki et al. (2018) showed that increased target familiarity and more targets lead to larger probe-irrelevant differences. They argued that participants relied more on familiarity to do the RT-CIT. Matsuda et al. (2009) omitted target items entirely and did not find a probe-irrelevant differences in RTs (while differences in the event related potentials persisted). That is not to say that the familiarity of targets is the only way to induce response conflict. Lukács and Ansorge (2019) argued that other features (e.g., self-relatedness) also lead to response conflict if the feature is shared by the YES-inducers and the probe. However, of the three YES-inducers that they used only "MINE" was self-referring - "FAMILIAR" and "RECOGNIZED" both referred to familiarity. It is therefore unclear why

these should not have induced response conflict in the contamination group as well. This, in addition to our findings, warrants caution and further investigation about the assumptions and boundary conditions of the I-CIT before applied use should be considered.

The aIAT accurately discriminated between Dutch and British participants and did not show reduced classification performance for British participants with knowledge about the Netherlands. This is plausible because the aIAT is not based on recognition of information but on the associations between the response labels (Sartori et al., 2008). In blocks in which the autobiographic label ‘I am from the Netherlands’ is paired with the logical label ‘True’, response times for participants for which the statement ‘I am from the Netherlands’ is true (the labels are associated; i.e., Dutch participants) will respond faster than in blocks in which ‘I am from the Netherlands’ is paired with ‘False’. Those associations are independent of the participant’s knowledge. Our results are in line with previous aIAT studies without a contamination group (for reviews see Agosta & Sartori, 2013; Suchotzki et al., 2017) and show that the aIAT can be a valid tool to assess someone’s nationality even if that person has knowledge about the country of his fake nationality.

The good classification accuracy of the aIAT, even in the face of information contamination, does not imply the aIAT is flawless. If participants are instructed on how to fake the aIAT, e.g., by slowing down responses in one block, they are able to do so (Agosta et al., 2011a; Hu et al., 2012; Suchotzki et al., 2017; Verschuere et al., 2009). Although faking could not be prevented so far, Agosta et al. (2011a) developed algorithms to detect faking. However, it seems likely that this algorithm could be tricked if the participants get instructed to also slow their responses in the single categorization blocks (blocks 1, 2, and 4). Another limitation of the aIAT regards the labels and statements that can be used. Agosta et al. (2011b) showed that negative statements (e.g., ‘I do not own a Dutch passport’), counter-affirmative statements (e.g., ‘I own a passport from another country than the Netherlands’), and negative labels (e.g., ‘I am not from the Netherlands’) reduced the aIATs classification performance. Therefore, the aIAT needs to contrast two specific autobiographical facts. In a forensic setting, this could be the crime and the alibi. In the context of nationalities, the examiner needs to have a strong suspicion about the true country of origin for the aIAT to perform optimally. Lastly, even though information contamination does not seem to be a problem for the aIAT, association contamination might be.

Dhammapeera et al. (2020) showed that imagining a false alibi reduces the detection performance of the aIAT. An equivalent in the nationality confirmation setting could be a British citizen that was born and raised in the Netherlands, that views the Netherlands as her/his home. Formally, this person is British but emotionally she/he is Dutch. Research on how to overcome this challenge might be valuable for practical applications.

Any imperfect classification system in the context of deception detection raises ethical issues about when it should be applied and how the results should be incorporated in the decision-making process. It could be tempting to strongly rely on such tests, since they do not pose any risk to the examinee, they are easily applicable, and they give the examiner an objective result. However, after carefully validating the test outside of the laboratory, we urge practitioners to use the information provided by tests like these as an addition to other sources of information and not as a replacement.

Limitations

This study is not without its limitations. We see at least three limitations worth discussing. First, we chose a quasi-experimental design to increase the ecologic validity by using pre-existing knowledge about the Netherlands. In terms of internal validity, however, this is inferior compared to a truly experimental design and introduces the possibility of between group differences (e.g., gender, age, level of education, motivation) that could impact the response times other than the nationality and knowledge. For instance, gender happened to covary with group. But an exploratory analysis showed there was no impact on the I-CIT-effect, see supplementary materials (<https://osf.io/8wyf6/>).

Second, the applied goal of this study also called for a scenario in which such a test might be applied, such as nationality confirmation. While we did our best to optimize both tests for this scenario, it might not be ideal for the I-CIT. Strictly speaking, we did not test nationality directly, but we tested the home-relatedness of the probes. It seems reasonable to assume that the probes are not home related for the two British groups, but it is less clear for the Dutch group. For instance, not all participants may strongly relate the specific probes used (Schiphol, Ajax, and Utrecht) to their home. If and by how much this diminished the I-CIT-effect for the Dutch group cannot be estimated in this study. However, the

comparison between the UK naïve and the UK contamination group is not affected by this. A second point connected to the items used in the I-CIT also needs to be addressed. Unexpectedly, we found a slightly negative I-CIT effect for naïve participants, which means that naïve participants responded quicker to the probes than to the irrelevant items. A core assumption of the CIT is that probes and irrelevant items should be indistinguishable by naïve participants and lead to similar RTs. We can only assume that this is due to item properties. However, such item effects reduce the *dCIT* scores for all groups, hence cannot explain that the contaminated groups showed an I-CIT effect.

Third, it is yet to be investigated to what extent our results generalize to other situations. Again, we started out with a problem practitioners face and applied both test to the best of our ability and in consultation of the tests' inventors. For the aIAT, this meant using unambiguous, simple labels and sentences that needed to be classified. One could argue that we did not use the same information in both tests (e.g., to use 'Schiphol' in both tests, we could have used sentences like 'Schiphol is in my home country') to make the aIAT and the I-CIT more comparable in this regard. However, this would be an unnecessarily indirect way of assessing the citizenship with the aIAT and could lead to worse performance. If our results still hold in other scenarios (e.g., a mock crime with 'wallet' as a probe for the I-CIT and the corresponding sentence 'I stole the wallet' for the aIAT) remains to be investigated. It could also be that the home-relatedness feature is not strong enough to induce response conflict (and therefore not suitable to assess the nationality) but the I-CIT would show better performance with identity information (as used by Lukács & Ansorge, 2019) or specific crime knowledge.

Conclusion

Recognition-based paradigms such as the RT-CIT are error-prone when truth tellers know the to-be tested information. The present study suggests the I-CIT is no exception. At the same time, our results show that the aIAT was not affected by knowledge about the falsely claimed nationality. Therefore, the aIAT could be a valuable instrument to identify people claiming to have a false nationality – a problem that border control and immigration agencies face daily.

Acknowledgments

We thank the Swiss Federal Office of Civil Aviation (project number: 2016-106) and the Zurich State Police, Airport Division for their financial support. We also thank Gáspár Lukács and Giuseppe Sartori for their advice on the construction of the Inducer-CIT and the aIAT, respectively.

Conflict of interest statement

The authors declare no conflict of interest.

Data availability statement

The material, data, and scripts that support the findings of this study are openly available on OSF (<https://osf.io/8wyf6/>).

Funding details

This work was supported by the Federal Office of Civil Aviation and the Zurich State Police under Grant 2016-106.

CHAPTER 4

What are you hiding?

Initial validation of the reaction time-based searching concealed information test

This chapter is published as:

Koller, D., Hofer, F., Grolig, T., Ghelfi, S., & Verschuere, B. (2020). What are you hiding? Initial validation of the reaction time-based searching concealed information test. *Applied Cognitive Psychology*, 34, 1406-1418. <https://doi.org/10.1002/acp.3717>

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 by using RTs to test for recognition 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 RT-CIT not only allowed to link suspects to known crime details, but also allowed to reveal new crime details well above chance. A simulation study confirms the potential of the searching RT-CIT and identifies conditions under which it performs best. We used an archival dataset that met these conditions (high CIT effect, large number of item repetitions), and found better item classification performance than in the mock terrorism study. The searching RT-CIT could be a new, promising investigative tool to reveal new (e.g., 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 utilized to synthesize explosives. Based on low quality closed-circuit television 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 probes and the irrelevants 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), response times, and P300 amplitude as well as a decrease of the heart rate and respiration line length; 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; Suchotzki, Verschuere, van Bockstaele, Ben-Shakhar, & Crombez, 2017). Classification performances range from an area under the curve (AUC) of $AUC = .74$ for heart rate to $AUC = .88$ for P300 amplitude (Meijer et al., 2014) with response times achieving $AUC = .82$ (Meijer et al., 2016).¹⁰

Detecting if a suspect is involved in the crime of interest 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?

¹⁰ AUC = .5 represents chance performance; AUC = 1 is perfect classification performance.

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, for the searching CIT to work, it is crucial to have a set of items that includes the true crime information with a very high probability. 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 are analyzed.

The idea of the searching CIT is not new. Autonomic measures have been used to extract information from groups of participants with shared complete (e.g., Breska, Zaidenberg, Gronau, & Ben-Shakhar, 2014; Meijer, Bente, Ben-Shakhar, & Schumacher, 2013) or partial crime knowledge (e.g., Elaad, 2016) in experiments, and Japanese law enforcements regularly use autonomic measures based searching CIT (Osugi, 2011).

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 terror attack with three testable crime information (the 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. Therefore, the probes were always present among the tested items. 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 trained personal needed to

administer an EEG are barriers to applying it in practice and limiting factors when it comes to scalability. The physiological CIT, although cheaper, cannot be scaled up easily for the same reasons. The RT-CIT, however, requiring only a computer and data collection and analysis possibly being fully automated, allows for remote and parallel testing with little additional resources needed.

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). Based on these results, the performance of the algorithms under different conditions was explored using a simulation study (Study 2). Finally, the algorithms were cross validated on independent data (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), based on the item classification for both algorithms.

STUDY 1: 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 a mock terror attack and partially executed it, and an innocent control group (see Procedure section). 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.

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*) for participants 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, Study 1 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 can be used to detect intentions with the CIT (Noordraven & Verschuere, 2013), 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,¹¹ data and code can be found on osf.io/69yrj.

Participants

Participants were sixty students from the 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

¹¹The study was preregistered (osf.io/69yrj) but that preregistration was premature and the authors decided to analyze 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.

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 (≈ 20.40 USD) or course credits for 1.5 hours of study participation (participants' choice). All participants were told that they will earn an additional 5 CHF (≈ 5.10 USD) if they can complete their task without being accused of anything (indicated by a search of their hand luggage). They were specifically instructed that simply being suspected is not enough to lose this bonus. 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 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 he/she 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 within 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 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.

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

Enactment	Item category	Item
Enacted	Code word	Gondola, magnet, board, anchor, stamp, 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, liquid, phone, powder, wallet
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 trash can, in PET container, in aluminium recycling bin, under seat, behind pillar, behind door
Intended	Meeting spot	Information desk, Hour Passion, Center Bar, Amavita, Marché, NZZ Café

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

Immediately before participants used their flight documents, they 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 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

Upon entering the CIT-room, participants were orally instructed to turn off/mute their phone and to put it aside together with their bag (and possibly other items). They were asked to sit down in

front of the computer and the confederate posing as a police officer informed them that the following test was designed to determine if he/she has knowledge about a crime. They were told that all the instructions will be presented on the screen but if they had a question at any point, he/she could ask the police officer who stayed in the same room in hearing distance but not visible.

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.

The RT-CIT consisted of eight item categories (4 past action categories, 4 intention categories), with six items per category (1 probe, 1 target, 4 irrelevant). In the beginning, participants were asked to learn the 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 on the 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 the 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.¹²

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

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

¹² 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. χ^2 -tests for each item category showed no significant effects after correcting for multiple comparisons (Bonferroni).

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 presumed 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 m^{th} bootstrap sample of category j is denoted $Pboot_{j,m}$ and $Iboot_{j,m}$ for the possible probe and the presumed irrelevant respectively. The possible probe of category j is then classified by $boot_j$, the percentage of iterations in which its mean RT was larger than the mean RT for the irrelevant item ($boot_j = \frac{\text{count}_m(M(Pboot_{j,m}) > M(Iboot_{j,m}))}{2000}$). Participants are classified based on the mean of those percentages over all item categories ($\frac{\sum boot_j}{8}$).

Results

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. 1.58% of irrelevant and probe trials were excluded from the analysis.

Post-CIT recognition

The final sample of 60 participants, target recognition accuracy after the RT-CIT was 95.2%. The probe recognition accuracy of the 30 guilty participants was 84.2%. Note that participants with more than two errors in either test were excluded (and replaced by another participant) and therefore not included in the final sample.

Known-solution group analysis

Before we tried the searching CIT algorithms, we verified that participants in the guilty condition showed larger RT-CIT effects than participants 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 ($BF_{10} = 2.82 \cdot 10^6$) 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-participant CIT-effects than the innocent group ($M dCIT_{\text{within}} = .04$; $SD = .14$; see Table 2).¹³

Table 2. Reaction times and effect sizes by group

	Innocent				Guilty				
	RT				RT				
	Irrelevant	Probe	Target	$dCIT_{\text{within}}$	Irrelevant	Probe	Target	$dCIT_{\text{within}}$	$dCIT_{\text{between}}$
Enacted	482 (154)	481 (152)	600 (132)	-0.00 (0.19)	486 (146)	532 (181)	627 (135)	0.36 (0.30)	1.45
Intention	491 (151)	501 (162)	618 (139)	0.08 (0.21)	502 (151)	551 (182)	646 (148)	0.39 (0.34)	1.07
Collapsed	486 (153)	491 (158)	609 (136)	0.04 (0.14)	494 (148)	541 (182)	637 (142)	0.36 (0.22)	1.76

Note. Mean reaction time (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.

¹³ CIT-effects were calculated as: $dCIT_{\text{between}} = \frac{M(dCIT_{\text{guilty}}) - M(dCIT_{\text{innocent}})}{2\sqrt{\text{var}(dCIT_{\text{guilty}}) + \text{var}(dCIT_{\text{innocent}})}}$; $dCIT_{\text{within}} = \frac{M(RT_{\text{probe}}) - M(RT_{\text{irrelevant}})}{SD(RT_{\text{irrelevant}})}$

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). For sake of comparison, we note that the commonly used cutoff $d = .2$ (see Noordraven & Verschuere, 2013) led to a specificity of 86.7% and a sensitivity of 80% (overall accuracy 83.3%).

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 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 behavior

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 null-hypothesis stating that the CIT-effects between enacted and intent items do not differ is better supported by the data than 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. 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 if at least one irrelevant item shows a larger mean RT than the probe in this category (the probe is therefore automatically classified as irrelevant), the algorithm will never reach a sensitivity of 1 no matter how liberal the criterion is set; which is a prerequisite to interpret the AUC.

Using LOO CV procedure for item classification, the standardization algorithm achieved a Youden's J of .37 (sensitivity = .68; specificity = .69). 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 for item classification achieved a cross-validated Youden's J of .33 (sensitivity = .50; specificity = .83). It should be noted that 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). 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 test crime knowledge and 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 effect of enactment (i.e., differences in the CIT-effect for items related to enacted tasks and items related to future intentions). Although we tried to make the enacted and intent items and categories comparable (e.g., each contained one alphanumeric non-word and neither contained emotionally loaded items), the mock crime scenario did not allow for counterbalancing between enacted and intent items. The possibility that this null effect is due to item selection can therefore not be discarded completely.

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.

STUDY 2: SIMULATION STUDY

The results of Study 1 warrant further exploration on how the searching algorithms are influenced by different factors. We ran a simulation study to do so. We simulated a wide array of datasets 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 ($8*8*5$) datasets with 1000 simulated participants each (500 guilty, 500 innocent). The effect 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, $dCIT_{within} = .46$ ¹⁴ 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

Data were 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 include 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 $RT_{i,j} = \text{baseline } RT_i + \text{CIT-effect}_i + \text{noise}_j$. For both irrelevant and probe items, the noise was drawn from a right skewed distribution with a mean of 0 (exponentially modified gaussian distributions with a $\mu = 0$, $\sigma = 56$, and $\beta = 120$ for irrelevant items and $\mu = 0$, $\sigma = 72$, and $\beta = 154$ for probes). These values were derived from fitting an exponentially modified gaussian model to the data of Study 1 (see osf.io/69yrj/ for further information). Reaction times of targets were not simulated as they do not influence the analysis.

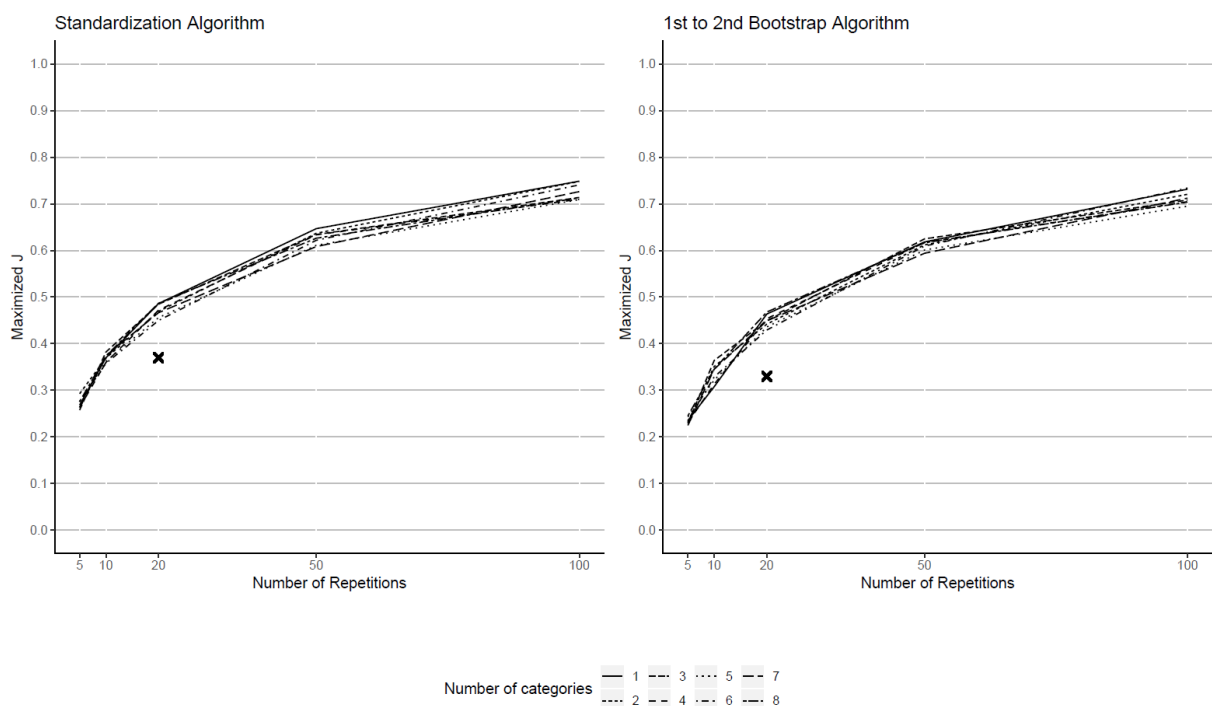
Results

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 any item effects (e.g., word length, numbers

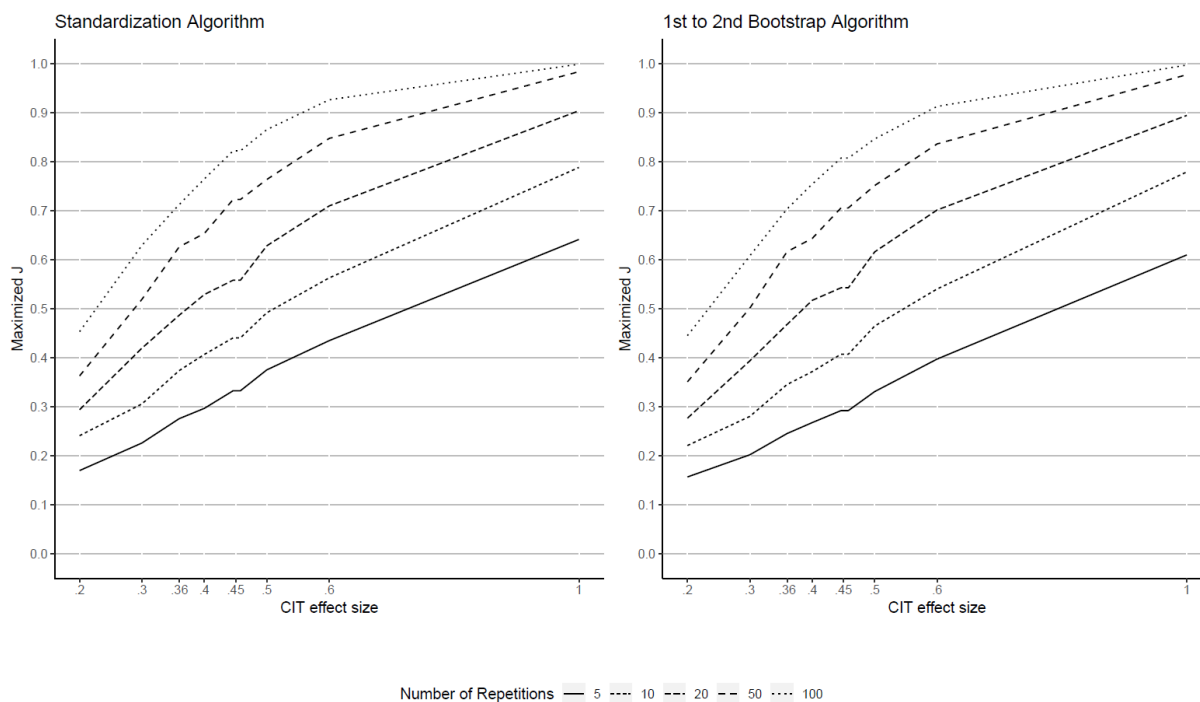
¹⁴ $dCIT_{within} = .45$ is based on initial calculations that contained a mistake in data aggregation. The correct effect size is $dCIT_{within} = .46$ but we refrained from running new simulations because the difference is minimal.

versus words, salience). The simulations show a mean maximized Youden's J of .47 (sensitivity = .68, specificity = .79) for the standardization algorithm when each item is presented twenty times compared to the empirically observed J of .37 in Study 1. The 1st to 2nd bootstrap algorithm achieved a mean maximized J of .45 compared to the empirically found J of .33 (sensitivity = .60, specificity = .85). This implies that with the same effect size and under optimal conditions (i.e., no item effects) the algorithms could perform better than what we found in Study 1.

Figure 1. Item classification performance for the CIT-effect size of Study 1



Note. Youden's J with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on simulated data with CIT-effect size $d_{\text{CIT}_{\text{within}}} = 0.36$ (lines) and on the empirical data of Study 1 (cross).

Figure 2. Simulated item classification performance by CIT-effect size and number of repetitions

Note. Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on the simulated dataset with eight item categories.

At least four conclusions can be drawn from the simulation outcomes. First, as 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 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 item effects seem to influence the searching algorithms negatively. The performances based on the simulated data should therefore be considered estimates of the theoretical ceiling performance.

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 if the number of item repetitions is increased or if the CIT effect is larger. Although this has not been manipulated in the simulation, the comparison between simulated and empirical data suggests that reducing item effects could have a considerable impact on the searching CIT performance also. To validate the algorithms on a second dataset and to show that the simulations yield useful ceiling estimates, we applied the analysis from Study 1 on the autobiographical RT-CIT data from Verschuere and Kleinberg (2016). This study had the same number of repetitions as Study 1, larger CIT-effects (which is expected due to the high relevance of the autobiographical information), and possibly smaller items effect. Smaller item effects can be expected because there are no differences in how well the different information was learned (since the information is autobiographic and does not need to be learned) and because participants could indicate if an irrelevant item stood out to them which lead to the exclusion of that item, reducing saliency effects among the irrelevant items.

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

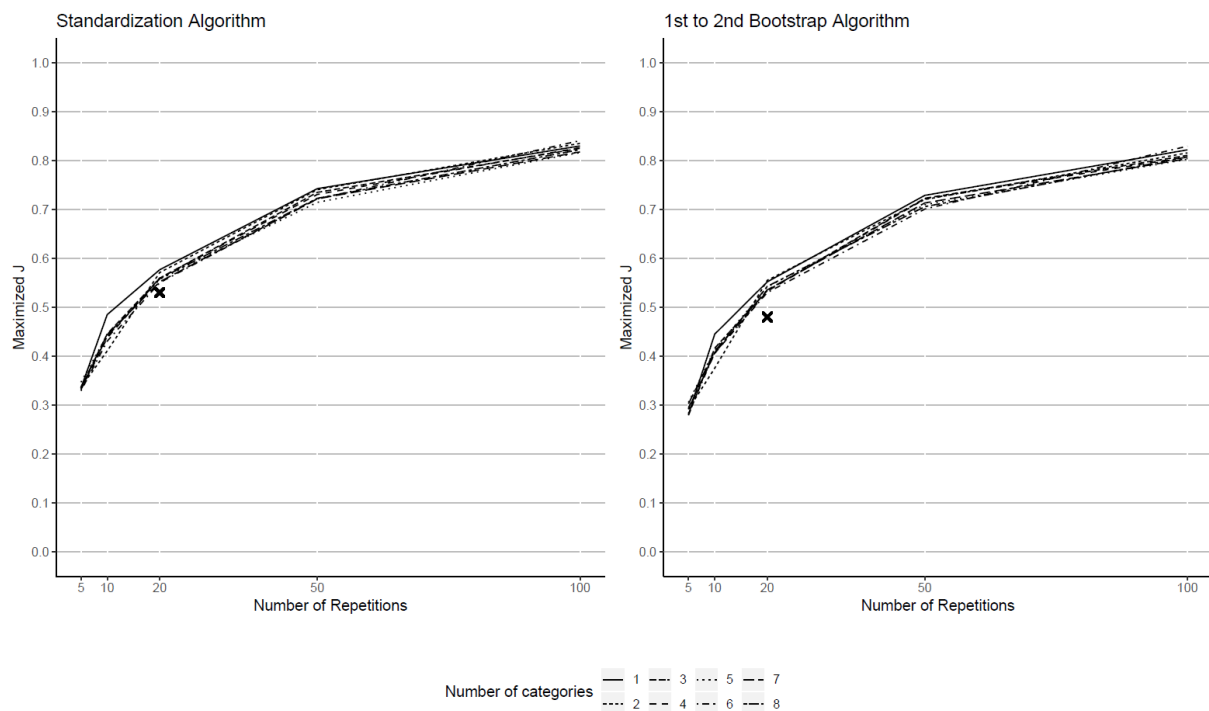
trials per item for a total of 600 trials. Participants that were instructed to hide their identity showed a mean CIT-effect of $M(dCIT_{within}) = .46$ ($SD = .23$), whereas unknowledgeable participants had a CIT-effect of $M(dCIT_{within}) = -.01$ ($SD = .13$).

Results

Maximizing the Youden's J of the searching algorithms for the empirical data of Verschuere and Kleinberg (2016) using LOO CV resulted in a Youden's J of .53 (sensitivity = .83; specificity = .69) for the standardization and a J of .48 (sensitivity = .62; specificity = .86) 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 theoretical ceiling performance based on simulated data.¹⁵

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.

¹⁵ Applying the searching algorithms on simulated data with $M(dCIT_{within}) = .45$ 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.

Figure 3. Item classification performance for the CIT-effect size of Verschuere and Kleinberg (2016)

Note. Youden's J achieved with the optimal cut-off using the standardization algorithm (left) and the first to second bootstrap algorithm (right) on simulated data with CIT-effect size $d_{\text{CIT}_{\text{within}}} = 0.45$ (lines) and on the empirical data of Study 3 (cross).

Discussion

The aim of Study 3 was to validate the searching CIT algorithms on an independent dataset and to show that the simulations yield realistic results. As predicted by the simulations, both algorithms showed better item classification than in Study 1. 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 dataset and presented additional evidence that our data simulation can be used to estimate the ceiling performance those algorithms can theoretically 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 (AUC = .91; using a commonly used cutoff: 85%). This shows that high accuracies can also be achieved in situations with possibly higher agitation levels (due to high security standards, police presence, and the unfamiliar airport environment) than studies conducted in university settings.

Especially in the terror context, the police are interested in detecting malicious intent to prevent an 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 an 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 with the CIT is therefore still needed to get the most accurate guilty/innocent classification.

To explore the algorithms' potential under different conditions and without 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 (i.e., increased CIT-effects and more repetitions per item, see Figure 2). Note that possible effects of habituation and fatigue could not be taken into account due to the lack of research in this area of the RT-CIT. 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 and could be done 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 behavior 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 maximal performance those algorithms can theoretically 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 et al., 2015), word length (Barton et al., 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 numbers of repetitions 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 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 the same size as the actual probe. For the standardization algorithm, this would result in the same $dCIT_{ij}$ score for this irrelevant item as for the probe, but they would still be larger than zero and therefore diagnostic. Note that the $dCIT_{ij}$ score of the probe would be smaller than without an irrelevant signal because the difference of the means decreases and the SD of irrelevants 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 participants would still be classified correctly. Furthermore, $M(\max(dCIT_{ij}))$ 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 since this guilty participant would show no bootstrap difference, just like innocent participants.

The possibility of stronger countermeasures that result in a CIT-effect of the irrelevant item larger than the CIT-effect of the probe or of applying countermeasures to multiple irrelevant items must be considered also. These possibilities lead to a further decrease in classification performance for both algorithms but to a lesser extent in the standardization than in the 1st to 2nd bootstrap algorithm, following the same rationale as in the presented example.

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 (Table 3).

Finally, a general limitation of the searching CIT needs to be considered. In this study, as in most others, the true probe was always present in the searching CIT which does not need to be the case in practice. Real-life situations rarely have a closed set possible probes in which the investigator knows that the real probe is included. This either means that the searching CIT (irrespective of the measures used) should only be applied in very rare situations or that an item that covers all other possibilities should be included. The latter is done in Japan, the only country that uses the searching CIT on a large scale (Osugi, 2018). The effects of this practice have yet to be thoroughly investigated.

Table 3. *Overview of the searching algorithm's evaluation*

	Standardization algorithm	First to second bootstrap algorithm
Classification performance	Fair	Fair
Sensitivity space	Unrestricted	Restricted
Vulnerability to countermeasures	Vulnerable	Very vulnerable
Computational resources needed	Low	Medium

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 suspect that the country of origin provided by a person is wrong and they have a specific suspicion where the person might be from (e.g., by testing knowledge about lesser known towns of that country), or possibly testing the knowledge about a substance found in a passengers' luggage which the passenger claims not to have packed. Our finding 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 can be tested, as it has been done with the physiological CIT (e.g., Breska, Ben-Shakhar, & Gronau, 2012; Breska, Zaidenberg, Gronau, & Ben-Shakhar, 2014; Elaad, 2016). 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, the meta-analysis of Suchotzki et al. (2017) did not find an effect of motivation on the size of the CIT-effect. 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.

As with any deception detection tool that might get used in practice, its susceptibility to countermeasures is an important concern that needs to be addressed. Suchotzki et al. (2017) found RT-based deception detection measures to be vulnerable to countermeasures but further research is warranted as different RT-based paradigms had to be analyzed together due to the few countermeasure studies that were conducted. Furthermore, susceptibility to countermeasures does not necessarily mean that detection methods cannot be used. If countermeasures can be detected, this can also be a valuable piece of information to the examiner by itself.

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.

Acknowledgments

We thank the Swiss Federal Office of Civil Aviation for the financial support (project number: 2016-106). We thank the Zurich State Police, Airport Division for their financial support and the possibility to use their infrastructure to conduct Study 1. We thank Zoé Dolder for her help in data collection.

APPENDIX**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 Zeitzünder einer Bombe.” [English: The watch serves as a time fuse of a bomb.]

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.]

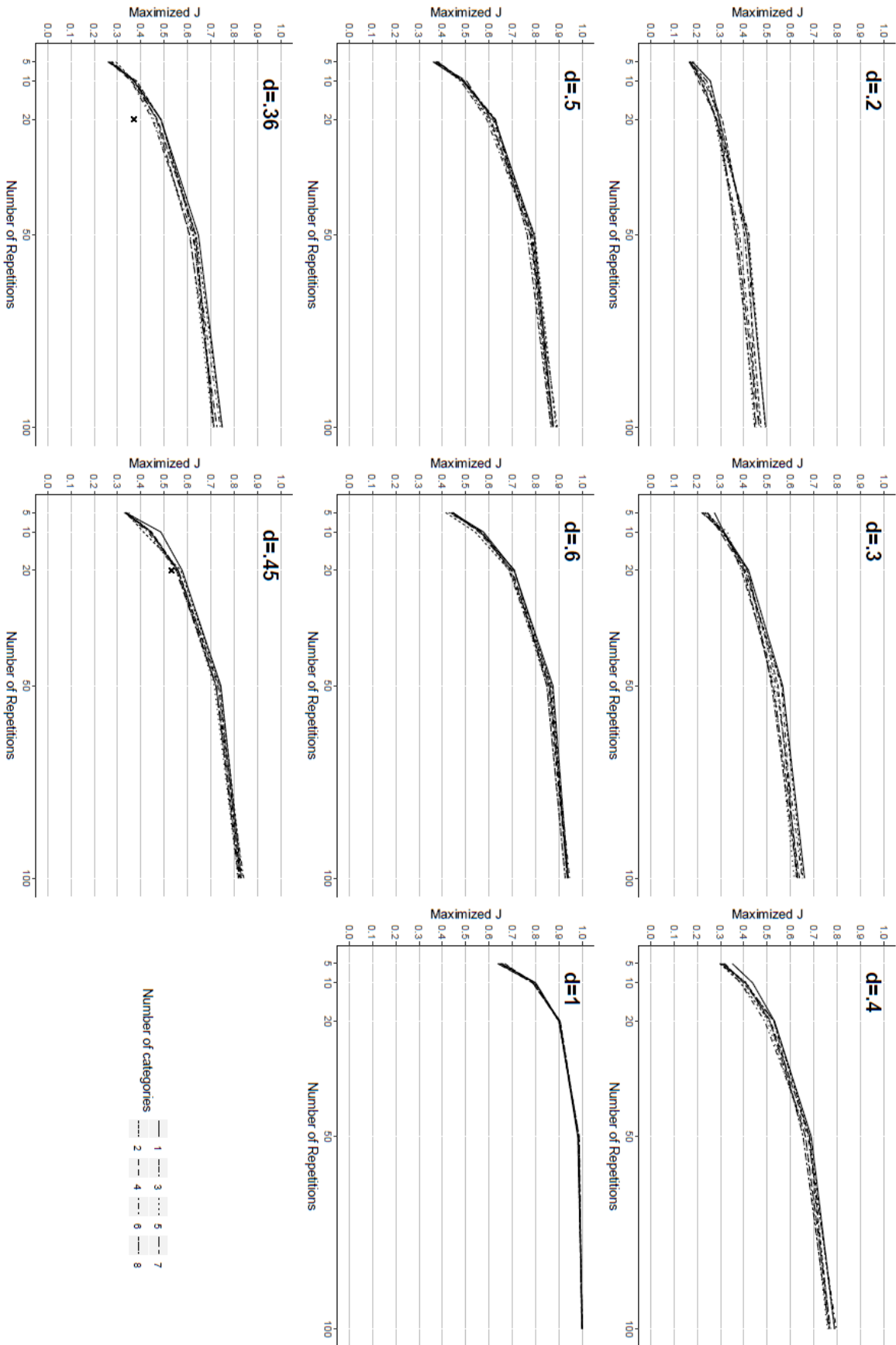
Phone: “Das Telefon dient als Fernzünder und Energiequelle einer Bombe.” [English: The phone serves as a remote detonator and energy source for a bomb.]

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.]

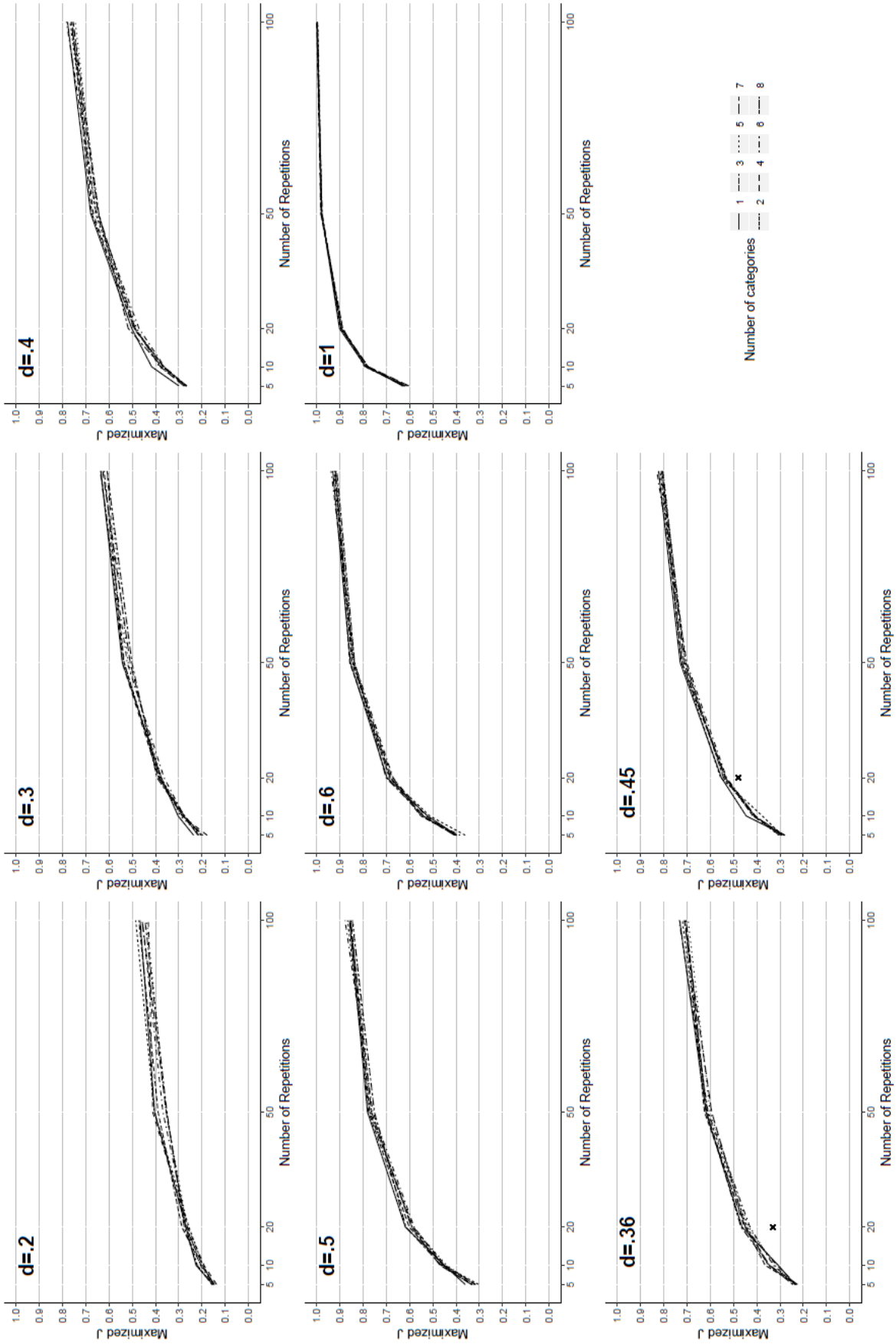
Wallet: “Das Portemonnaie enthält gefälschte Ausweispapiere, die es dem Terroristen erlauben in abgesperrte Bereiche des Flughafens zu gelangen.” [English: The wallet contains forged identification documents which give the terrorist access to restricted areas of the airport.]

Appendix B

Appendix B, Figure B1. Standardization Algorithm



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



CHAPTER 5

Assessing partial errors via analog keyboards in response conflict tasks: A preregistered pilot with the concealed information test

This chapter is in preparation for the initial submission.

ABSTRACT

The response time-based Concealed Information Test (RT-CIT) is an established memory detection paradigm. Slower RTs to critical information (called ‘probes’) compared to control items (called ‘irrelevants’) reveal recognition. Different lines of research indicate that response conflict is a strong contributor to this RT-difference. Previous studies used electromyography to measure response conflict, but this requires special equipment and trained examiners. The aim of this study was to explore if response conflict can also be measured with an analog keyboard that is sensitive to minimal finger movements. In a preregistered study ($n = 35$), participants completed an autobiographical RT-CIT and a cued recognition task (modified Sternberg task). Partial errors, partial button presses of the incorrect response key, were more frequent in trials with response conflict than in trials without conflict ($BF_{\text{CIT}} = 275$; $BF_{\text{Sternberg}} = 102$) but still rare (CIT: 2.9%; Sternberg: 1.7% of conflict trials). This is the first evidence that analog keyboards can measure partial errors. Practical and theoretical implications are discussed.

Keywords: memory detection, Concealed Information Test, CIT, deception, Sternberg task, response tendency, analog keyboard

INTRODUCTION

The Concealed Information Test (CIT) aims to detect if someone has specific knowledge that they cannot or do not want to reveal (Lykken, 1959). Examinees are presented with several, equally plausible pieces of information (e.g., names) and they are asked to indicate whether they recognize the information. The examinee's name (e.g., Alex) amongst a series of other names (e.g., Frank, David, Mark) typically shows a distinct behavioral (Seymour et al., 2000), physiological (Lykken, 1959), and neurophysiological response (Langleben et al., 2002; Rosenfeld et al., 1988, 2008) that can be used to infer recognition of the presented information (for a review, see Verschuere and Meijer, 2014).

Several theories have been formulated to explain the results of decades of CIT research (for a review, see Klein Selle et al., 2018). While most are unitary approaches, Klein Selle et al. (2017) introduced a response fractionation model for the physiological CIT that holds that not all measures are driven by the same mechanism (also see Barry, 1982). More specifically, they propose that the orienting response (Sokolov, 1963) drives the skin conductance response, while heart rate and respiration are linked to arousal inhibition. The response time-based CIT (RT-CIT) effect – the slower responding to concealed information than to control items – might be linked to response conflict and response inhibition (Seymour & Schumacher, 2009; Suchotzki et al., 2015). Other than the classical physiological CIT that only consists of the concealed items (also called *probe* items) and the control items (also called *irrelevant* items), the RT-CIT additionally has so-called *target* items. Targets are items to which examinees are instructed to respond differently than to all other items (i.e., press YES when you recognize the target; Farwell & Donchin, 1991). Targets are typically learned before the test and are therefore familiar to the participant. Because familiarity is a valid cue that is in line with recollection for irrelevant and target items (which make up five out of six trials) and because the RT-CIT is a speeded paradigm, participants might strongly rely on the fast familiarity-based responding (Ratcliff & McKoon, 2008; Yonelinas, 2002). For probes, however, the familiarity-based response (YES, because it is familiar) contradicts the recollection-based response (NO, because recognition should be concealed) which is expected to lead to response conflict and therefore slower RTs.

Different lines of research showed converging evidence for the importance of response conflict in the RT-CIT. One line of research aimed to manipulate response conflict in the RT-CIT experimentally. Lukács et al. (2017) added familiarity related “filler” items to the RT-CIT which needed to be classified as familiar or unfamiliar. They argued that these filler items could increase the reliance on familiarity and therefore should increase response conflict. While they found larger probe-irrelevant RT differences in the filler condition (replicated by Olson et al., 2020), they note that this could also be due to deeper semantic encoding or disruption of a target focused response strategy (also see Koller et al., 2021). A more direct approach that did not modify the RT-CIT paradigm, and also succeeded in increasing the RT difference, is using personally familiar instead of learned targets (Suchotzki et al., 2018). The reasoning behind this manipulation is similar as for the fillers: Familiarity-based responding becomes a more viable strategy to do the CIT, since targets and irrelevant items can be classified correctly and quickly based on familiarity alone. For probes, however, familiarity is an invalid cue and familiarity-based responding needs to be inhibited. Increasing target familiarity probably also increased target saliency and therefore the response conflict due to overlap in the saliency dimension between targets and probes. Since we are interested in response conflict in general, this is not problematic. But the manipulation also introduced task difficulty as a possible confound. The familiar target condition might be easier because targets did not need to be learned and retained.

Another line of research investigated the mechanisms involved in the CIT using neurophysiological measures linked to response conflict detection and resolution. fMRI studies showed increased activation in the ventral fronto-parietal network (for a meta-analysis, see Gamer, 2011). This network is connected to multiple potentially important mechanisms for the CIT like response inhibition (Zhang et al., 2017), but also to orienting response (Strange et al., 2000), and memory processes (Nyberg et al., 2003) which complicates isolated inferences about one of those mechanisms. Furthermore, the insights from fMRI-based CIT studies – that typically have a slower pace and no targets - might not be directly transferable to the RT-CIT. A neurophysiological measure that can be combined with the RT-CIT paradigm is EEG. The most consistent finding is that recognized probes show larger P300 amplitudes than irrelevant items (e.g., Allen et al., 1992; Rosenfeld et al., 1988; Rosenfeld, 2011) which is closely related to the orienting response (Donchin et al., 1984; Nieuwenhuis et al., 2011). Rosenfeld

et al. (2017) additionally found increased N200/N300 latencies at F3 in guilty participants that tried to hide their crime knowledge but not in the witness group that should reveal their knowledge. Although some research links N200/N300 latency to inhibition processes (e.g., Falkenstein et al., 1999), Rosenfeld et al. (2017) acknowledge that “A great deal more research with N200/N300 is required to elucidate in detail [...] the specific kind of inhibition possibly required from CIT suspects” (Rosenfeld et al., 2017, p. 646). Interestingly, this study used the complex trial protocol (Rosenfeld et al., 2008) in which participants only need to acknowledge that they saw the stimulus which deems response conflict or response inhibition unlikely to be the mechanisms driving the N200/N300 latency. Also, attempts to link measures of executive control to probe-irrelevant differences in RTs did not provide evidence for a connection (Suchotzki et al., 2015; Visu-Petra et al., 2012, 2014).

A more direct approach to measure response conflict in the RT-CIT used electromyography (Seymour & Schumacher, 2009; for a related approach see Hadar et al., 2012). Electrodes were placed on the triceps brachii of each arm to measure muscle activity. Participants held two cylinders with electric switches and responded by exerting a “moderate downward force” (Seymour & Schumacher, 2009, p. 76) to those cylinders. This study found that probes elicited subthreshold muscle activity in the arm indicating recognition more frequently than irrelevant items. These so-called partial errors were used as evidence for response conflict in other conflict tasks before (e.g., Burle et al., 2002) and are considered small corrected errors (e.g., Allain et al., 2009). By measuring response related muscle activity, electromyography can provide strong evidence for response tendencies and response conflict, but it comes with its drawbacks. It requires specialized equipment, trained personnel to place the electrodes correctly, and often requires adaptations of well established experimental tasks that typically use a keyboard.

Could partial errors also be assessed with an analog keyboard which not only registers if a key is pressed or not but how far a key is pressed at any given time? Such would provide us with a relatively simple tool to detect response conflict in individual trials for a wide array of RT-tasks without the need to modify the experimental paradigm. For the RT-CIT, partial button presses could also increase classification performance or help detect countermeasures. Just like the partial errors picked up by the electromyogram, we expect that response conflict leads to partial errors in the form of partial button

presses (the precise definition is provided in the Method section). We manipulated the amount of response conflict by using either learned or familiar targets (Suchotzki et al., 2018). From this, we derived the following four main hypotheses. The first two hypotheses pertain to the benchmark probe-irrelevant difference in RTs and the replication of Suchotzki et al. (2018) on the effect of familiar targets on RTs: 1) Probes show larger RTs than irrelevant items and 2) the probe-irrelevant difference in RTs is larger in the high familiarity condition (i.e., familiar targets) compared to the low familiarity condition (i.e., learned targets). Since we expect partial button presses to measure response conflict, we predicted the same effects for partial button presses: 3) partial button presses occur more frequently for probes than for irrelevant items and 4) we expect a larger probe-irrelevant difference in the frequency of partial button presses in the high familiarity condition compared to the low familiarity condition.

While our focus is on the RT-CIT, partial button presses should also occur in other, non-deceptive, conflict tasks. To ensure that partial button presses are not unique to the RT-CIT and that potential differences between the familiarity conditions are not due to task difficulty, we employed the modified Sternberg task (Oberauer, 2001), a cued recognition task, as a secondary response conflict task. Conflict was manipulated by the proportion of trials for which familiarity is a valid cue (match and new trials; see Method section) compared to intrusion trials for which familiarity induces response conflict. For this additional task, we had the following hypotheses: 5) RTs for intrusion items are larger than for new items¹⁶, so-called intrusion costs. 6) Intrusion costs in the high conflict condition are larger than in the low conflict condition. Concerning partial button presses, we expected that 7) partial button presses occur more frequently in intrusion trials compared to new trials and that 8) the difference in the frequency of partial button presses between intrusions and new trials as well as between intrusions and matches is larger in the high conflict condition than in the low conflict condition.

¹⁶ In the preregistration, they were called non-presented lures.

METHOD

The experiment was approved by the ethics committee of the Faculty of Social and Behavioural Sciences of the University of Amsterdam (approval number: 2020-CP-12001). Preregistration, material, data, and scripts can be found on <https://osf.io/x8ecn/>.

Deviations from preregistration

One Swiss participant was tested at the University of Amsterdam although only German and Dutch participants were preregistered as eligible. However, this criterion was based on the demographics of students at the University of Amsterdam and not on the study design. Because the inclusion of this participant does diminish the validity of this study in any way, we decided to not exclude this participant.

Participants

Participants were eligible to enroll if they were at least 18 years old and if they have moved at least once in the past five years. Data was collected simultaneously at the University of Zurich and the University of Amsterdam. Participating at the University of Zurich required proficiency in German and one of the following nationalities: Swiss, German, Austrian. Participants at the University of Amsterdam were required to be proficient in English and either Dutch or German. Completion of this study took participants about 75 minutes and was reimbursed according to the standard rates of the respective universities (19 CHF at the University of Zurich, 12.50 EUR at the University of Amsterdam). Participants were recruited via a participant mailing list and via the research study platform of the University of Amsterdam.

Following the preregistered recruitment procedure, we concluded data collection based on our time deadline. A total of forty-three participants were recruited but two participants were excluded prior to data analysis due to illegibility or technical errors. Of the forty-one participants that entered the data analysis, five participants (12%) were excluded based on the preregistered language proficiency criteria (LexTALE score > 70; Lemhöfer & Broersma, 2012). One participant had to be excluded from the RT-

CIT because the RT-CIT could not be constructed due to item familiarity (see below) resulting in a sample of $n = 35$ (M age = 25.89, $SD = 5.14$, 80% female) for the RT-CIT. Of the thirty-six participants, three had to be excluded from the modified Sternberg task, due to poor task performance (less than 60% correct in at least one item category) resulting in a final sample $n = 33$ (M age = 25.33, $SD = 4.59$, 78.8% female) for the modified Sternberg task. Of the thirty-six participants, twenty-six (72.2%) participated at the University of Zurich (22 Swiss, 3 German, 1 Austrian) and 10 at the University of Amsterdam (7 Dutch, 2 German, 1 Swiss; see deviations from preregistration).

Procedure

The experimenter welcomed the participants and asked them to read and sign the informed consent. It was clearly stated that participation is voluntary and that participants can withdraw their consent at any time without giving reasons or disadvantages. They were further informed that data containing their personal information will be treated confidentially and that an anonymized version of the data will be made publicly accessible on a data repository. After providing consent, participants then completed the RT-CIT and the modified Sternberg task. The task order was balanced between participants (before exclusions). After the two response time tasks, participants completed the LexTALE language proficiency task (Lemhöfer & Broersma, 2012). Finally, participants were debriefed, reimbursed, and thanked for their participation.

RT-CIT

Before the RT-CIT started, we asked participants for autobiographical information (name, surname, date of birth as well as the street and city they currently live in). We also asked them to provide their former address (street and city) as well as the name, surname, and date of birth of a good friend of the same sex. The information were entered by the participant but under supervision of the experimenter to ensure that the format is consistent with the other items used in the RT-CIT (e.g., no abbreviations).

Next, we presented participants with lists of seven items, one list per information category (i.e., seven names, seven surnames, etc.), and asked them to indicate up to two items that were of personal relevance to them by clicking on them. Erroneous clicks could be corrected by clicking on the same item again. The indicated items were removed from the item pool that we used to construct the upcoming RT-CIT. Participants were instructed to contact the experimenter if more than two items in a list were of personal relevance because in that case, the RT-CIT could not be constructed.

We then asked the participants to imagine that they want to flee a country, but the police and border control are looking for them that is why they carry a fake ID with them. They get stopped by the border control at the airport and tested for their identity. Participants were instructed to hide their true identity and to pretend to be the person on the fake ID whose information (i.e., name, surname, date of birth, street, and city) was shown on the screen (for similar scenarios see e.g., Verschuere & Kleinberg, 2016). To do so, they should press YES when presented with any information of the fake ID (targets) and NO for all other information (irrelevant items and probes). We asked participants to learn the information of their fake identity and tested their memory using free recall. Only participants without errors in the free recall could proceed to the RT-CIT. Participants were redirected back to the learning phase if they made an error.

The RT-CIT consisted of the five information categories (name, surname, date of birth, street, and city), with six items per category (1 probe, 1 target, 4 irrelevant items; within-subjects factor). The true autobiographical information were used as probes. The irrelevant items were randomly selected from a pre-selected pool of potential irrelevant/target items (see <https://osf.io/x8ecn/>). Target items were either all randomly selected from the item pool (low familiarity condition) or the friend's information and the participant's previous address were used as targets (high familiarity condition; between-subjects factor).¹⁷ On each trial, a single item was presented in the middle of the screen. Participants were instructed to answer the question "Is this you?" as quickly and accurately as possible by pressing either "i" or "e" on the keyboard. The NO response was mapped to the participant's dominant hand. Participants should keep their index fingers on the response keys throughout the RT-CIT. The items

¹⁷ Items were adapted depending on the test site (German vs. Dutch cities and street names) and nationality (German vs. Dutch names).

were displayed until a response was given or the response deadline was reached. The response-stimulus interval varied randomly between 500 ms and 1000 ms. However, if participants were pressing a response key when the next trial was supposed to start, a message to fully release all keys was displayed. The next trial started between 500 ms and 1000 ms after the keys were released.

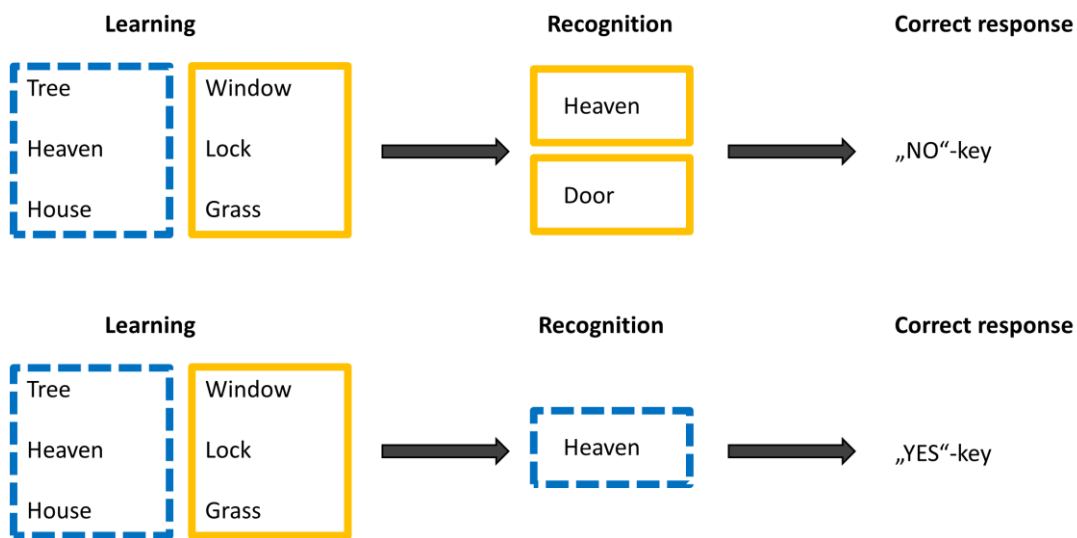
The RT-CIT started with three practice blocks of 30 trials each, in which every item was presented once. A red “X” (in case of an error) or a red “TOO SLOW” message displayed for 200 ms below the item provided feedback in the practice phase. The “TOO SLOW” message was shown if the response time was larger than 10 s in the first practice block, larger than 1.2 s in the second, or larger than 0.8 s in the third practice block. Response deadlines for were 10 s, 1.5 s, and 1.5 s respectively. Participants had to repeat the third practice phase if they had less than 50% correct for any item type (probe, target, irrelevant) or a mean response time larger than 800 ms. The test phase consisted of 20 blocks of 30 trials each, resulting in 600 test trials in total (100 probes, 100 targets, and 400 irrelevant items). Every item was presented once per block and the response deadline was set to 1.5 s. Participants could take a short self-paced break after 10 blocks. The RT-CIT was followed by a free recall of target items to ensure that participants did not forget the targets during the test.

Modified Sternberg task

The modified Sternberg task (Oberauer, 2001) is a cued recognition task (Figure 1). In the learning phase consisted of two lists of three nouns each that were presented side by side in colored rectangles (blue and yellow). The six items were presented simultaneously for 4.8 s followed by a blank screen of 800 ms. In the recognition test, one word was shown in either a blue or yellow rectangle. Participants’ task was to indicate as quickly and accurately as possible if the presented word was in the list of the cued color. There are three possible trial types (within-subjects factor: match, intrusion, new) depending on the word-color combinations. In a match trial, the word was in the list of the cued color. If the word was part of one list but is presented with the color of the other list, this is a so-called intrusion trial. Finally, if a word is presented that was not in either list, it is called a new trial. Match trials require a YES response while intrusion and new trials require a NO response. Like in the RT-CIT, “e” and “i” were

the response keys and the NO response was mapped to the participant's dominant hand. Participants were also instructed to keep their index fingers on the response keys throughout the task. The items were displayed until a response was given or the response deadline was reached. The response-stimulus interval varied randomly between 500 ms and 1000 ms. However, if participants were pressing a response key when the next trial was supposed to start, a message to fully release all keys was displayed. The next trial started between 500 ms and 1000 ms after the keys were released.

Figure 1. Illustration of the different trial types of the Modified Sternberg Tasks



Note. We used solid lines in the experiment. The dashed lines are for visibility for grayscale printouts.

We manipulated the validity of familiarity as a cue to solve this task (low validity, high validity; between-subjects) by changing the proportion of new and intrusion trials. The low validity condition consisted of 40% intrusion trials and 10% new trials, the high validity condition used 15% intrusion trials and 35% new trials. The task consisted of 50% match trials in both conditions to ensure that there is no dominant response key. Consequently, purely familiarity-based responding would lead to 60% and 85% correct responses in the low and high familiarity condition respectively.

The modified Sternberg task started with two practice blocks of 10 trials each. A red "X" (in case of an error) or a red "TOO SLOW" message displayed for 500 ms below the item provided feedback in the practice phase. The "TOO SLOW" message was shown if the response time was larger than 6 s in the first practice block or larger than 1.5 s in the second practice block. Response deadlines for were

6 s and 2.5 s respectively. The test phase consisted of 120 trials with a response deadline of 2.5 s. Participants could take a short self-paced break after 40 and 80 trials. Cue color and word position within the list for match and intrusion trials was balanced across test trials. No word was presented more than once.

LexTALE

We used the MATLAB (The Math Works, 2018) based LexTALE versions provided on LexTALE's website (www.lextale.com). The language tested by the LexTALE corresponded to the language of the RT-CIT and the modified Sternberg task (i.e., German for participants at the University of Zurich; English for participants at the University of Amsterdam). In this test, participants were presented with 60 strings of letters – 40 real words (e.g., scornful, ablaze), 20 pseudowords (e.g., mensible, pulsh) and their task was to indicate whether this string is a word of the tested language or not. If they recognized a word but did not know its meaning, they should still indicate “yes”. However, if they are unsure, they should indicate “no”. The LexTALE score is calculated as $\% correct_{av} = ((2.5 * \text{number of words correct}) + (5 * \text{number of nonwords correct})) / 2$. This score highly correlates with other language proficiency measures such as the Quick Placement Test (2001) ($r = .63$) and translational scores ($r = .75$; Lemhöfer & Broersma, 2012). For more detailed information about the LexTALE, see Lemhöfer and Broersma (2012).

After participant exclusions due to low scores in the LexTALE ($\% correct_{av} \leq 70$), participants had a mean score of $M \% correct_{av} = 85.3$ ($SD = 6.61$; range: 71.25 – 96.25). This corresponds to a high level of language proficiency (cf. Frank et al., 2019; Lemhöfer & Broersma, 2012).

Partial button presses

We used the Wooting Two Lekker edition keyboard to measure partial button presses (see https://wooting.io/wooting_two_lekker). This keyboard uses hall effect switches to translate the position of any key into an analog value ranging from zero to one. Keys that are not pressed down have an analog

value of zero, fully pressed keys have an analog value of one. However, our testing showed that if a key is pressed at an angle, the value might not quite reach one. Therefore, we decided to set the threshold of when we consider a key to be fully pressed to analog values $> .95$. The analog values were retrieved at a rate of 1000 Hz. To reduce the size of the data files, we only recorded the analog values and the corresponding timestamp when the analog value changed since the last retrieval. We speak of a partial button press if both response keys showed analog values > 0 before the response threshold (analog value $> .95$) was reached.

RESULTS

Analyses were conducted in R (version 4.0.3; R Core Team, 2020) with the BayesFactor (Morey & Rouder, 2018) and brms (Bürkner, 2017) package.

RT-CIT

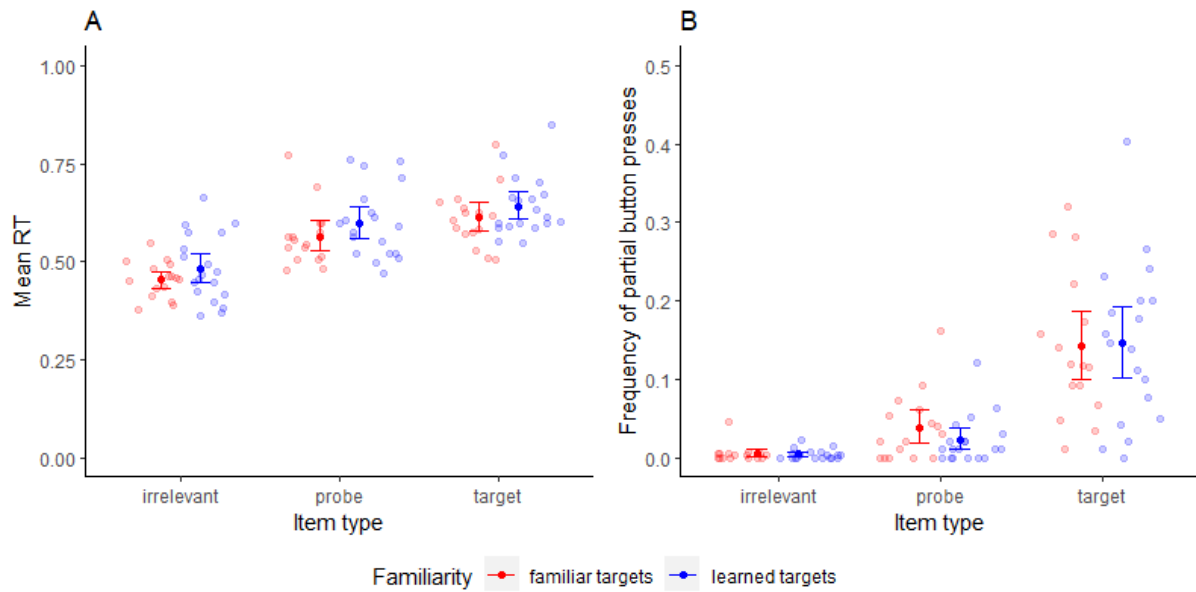
Preregistered analyses

Following (Koller et al., 2021), we excluded target trials, trials with response times smaller than 200 ms or larger than 1500 ms, and trials with response errors. We also excluded trials that start with a partially pressed key (analog value > 0 in the first 5 ms of a trial) to avoid accidental key presses. In total, 1.82% of probe and irrelevant trials were excluded.

RTs. To test for the CIT effect in RTs (hypothesis 1) and for the effect of target familiarity on the CIT effect in RTs (hypothesis 2), we conducted a two (Item type: probe vs. irrelevant; within-subjects) by two (Target familiarity: learned targets vs. familiar targets; between-subjects) Bayesian mixed effects ANOVA with JZS priors (Cauchy priors with scale = .5) on the participant mean RTs (Figure 2: A). Comparing the main effects model M_{Main} , the model with both main effects, to the model with only the main effect of familiarity (M_{Fam}) showed that the data is much more likely under M_{Main} ($BF_{\text{Main,Fam}} = 2.0 \times 10^9$), providing strong evidence for the predicted probe-irrelevant difference in RTs ($M RT_{\text{probe}} = 583$ ms, $SD = 75$ ms versus $M RT_{\text{irrelevant}} = 469$ ms, $SD = 72$ ms). Comparison of the model

with both main effects and the interaction (M_{Full}) and M_{Main} showed anecdotal evidence *against* an interaction ($BF_{Full,Main} = .33$). In other words, the data is more likely under the model without the interaction than under the full model. Hypothesis 2, the increased probe-irrelevant difference in the familiar target condition (Suchotzki et al., 2018), was therefore not supported by the data. The results were robust to changes in the width of the cauchy prior.

Partial button presses. We also predicted a CIT effect (hypothesis 3), moderated by target familiarity (hypothesis 4), for partial errors. Therefore, we tested these hypotheses in an analogous manner to the RT analyses. We conducted a two (Item type: probe vs. irrelevant; within-subjects) by two (Target familiarity: learned targets vs. familiar targets; between-subjects) Bayesian mixed effects ANOVA with JZS priors (Cauchy priors with scale = .5) on the frequency of partial button presses (Figure 1: B). The data was more likely under the main effects model than under the model with only a main effect of familiarity ($BF_{Main,Fam} = 275$), providing strong evidence for an effect of item type (hypothesis 3). This means that the CIT effect was also apparent in the frequency of partial button presses (M Proportion partial presses_{probes} = 2.93% , $SD = 3.73\%$ versus M Proportion partial presses_{irrelevants} = .46%, $SD = .88\%$). Comparing the full model to the main effects model showed anecdotal evidence against an interaction effect ($BF_{Full,Main} = .59$) and therefore against hypothesis 4. The results did not qualitatively change when we used the arcsine transformed data and the results were robust to changes in the width of the cauchy prior.

Figure 2. Participant mean RTs and frequency of partial button presses by item type in the RT-CIT

Note. The error bars indicate the bootstrapped 95% confidence intervals. The CIT effect is calculated as the difference between the probe and the irrelevant items.

Non-preregistered analyses

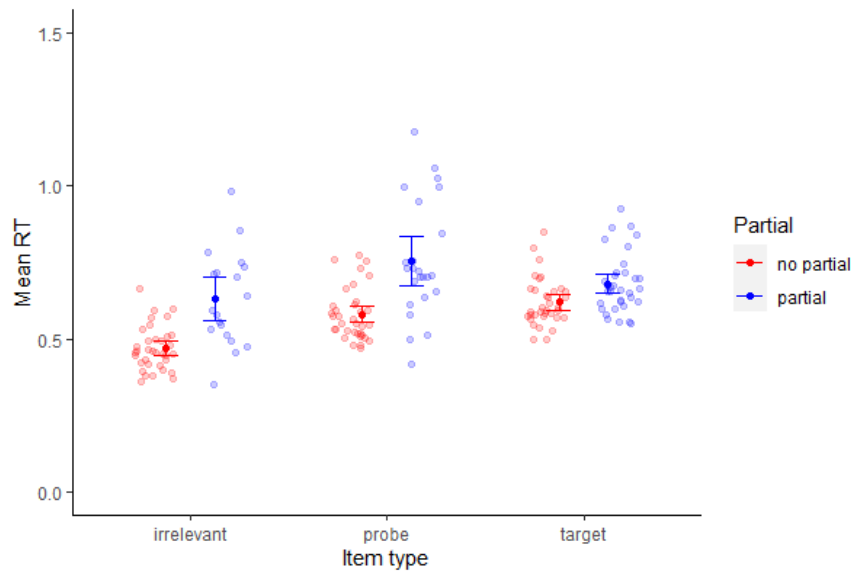
Because we did not find evidence for an effect of target familiarity, we do not distinguish between the two groups in the exploratory analysis. We calculated the mean RTs for trials with and without partial button presses per participant and item type and found larger RTs for trials in which a partial button press occurred. This effect was present for all item types (Figure 3) but less pronounced for targets. However, since partial button presses are more frequent in target trials, aggregation gives more weight to partial button presses of irrelevant and probe trials than to target trials. (One person's mean RT of probes with partial button presses might rely on very few trials while the mean RT of targets with partial button presses relies on more trials, but aggregation results in two data points with equal weight.) Therefore, we fitted an exponentially modified gaussian distribution model to the individual trial data using *brms* (Bürkner, 2017). The model included the main effects of item type and partial button press, their interaction, and random intercepts of participants and information category (e.g., name, surname, date of birth).

$$\text{RT} \sim 1 + \text{item type} * \text{partial} + (1|\text{participant}) + (1|\text{information})$$

$$\text{sigma} \sim \text{item type} + \text{partial}$$

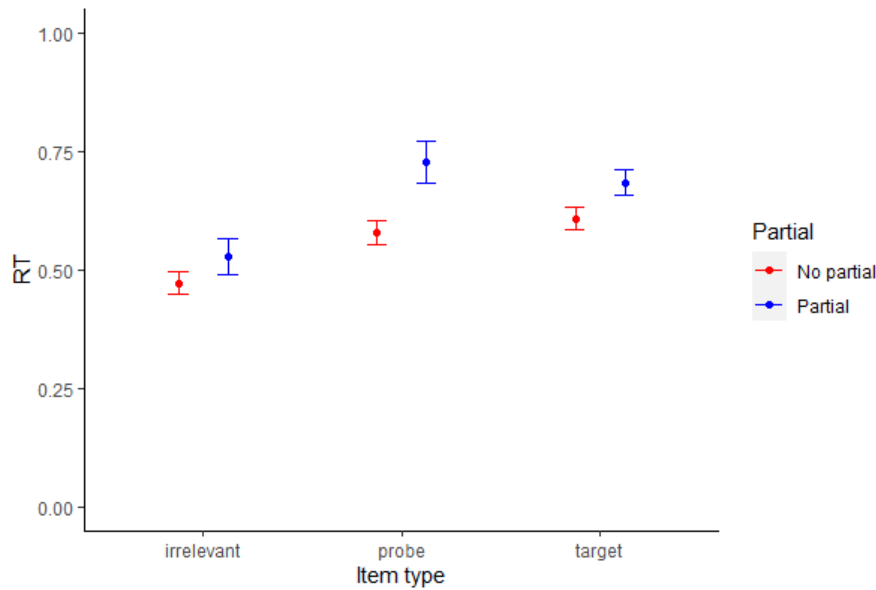
$$\text{beta} \sim \text{item type} + \text{partial}$$

Figure 3. Comparison of participant mean RTs for trials with and without partial button presses



Note. The error bars indicate the bootstrapped 95% confidence intervals.

The posterior means (Figure 4) showed larger RTs for trials with partial button presses compared to trials without partial button presses. The mean RT costs of partial errors varied with item type (irrelevant: $M = 56$ ms, probe: $M = 149$ ms, target: $M = 77$ ms), which could reflect the different stages at which the conflict occurs. For probes, we expected conflict when recollection provides the information that the correct response is "no", contrary to the familiarity based information. The expected conflict for targets is based on the predominant "no"-response in the CIT (five out of six items require a "no"-response) that conflicts with the familiarity based "yes"-response. Therefore, conflict occurs before recollection information is available. For irrelevants, we did not expect any response conflict.

Figure 4. *Posterior means of the exponentially modified gaussian distribution model*

Note. The error bars indicate the 95% credible intervals.

Modified Sternberg task

Preregistered analyses

Trials with response times smaller than 200 ms or larger than 2500 ms, trials that start with a partially pressed key (analog value > 0 in the first 5 ms of a trial), and trials with response errors were excluded from the analysis. 617 out of 3960 trials (15.58%) were excluded (18.03% of match trials, 17.55% of intrusion trials, 7.21% of new trials). Out of the 617 excluded trials, 543 (88%) were excluded due to response error.

RTs. We conducted a two (Item type: intrusion vs. new; within-subjects) by two (Validity of familiarity: low vs. high; between-subjects) Bayesian mixed effects ANOVA with JZS priors (Cauchy priors with scale = .5) on the participant mean RTs (Figure 5: A). Comparing the main effects model (M_{Main}) to the model with only the main effect of familiarity ($M_{\text{Familiarity}}$) showed that the data is much more likely under M_{Main} ($BF_{\text{Main,Familiarity}} = 1.7 \times 10^9$). Therefore, we found strong evidence for intrusion costs in RTs (hypothesis 5; $M RT_{\text{intrusion}} = 1216$ ms, $SD = 231$ ms versus $M RT_{\text{new}} = 941$ ms, $SD = 237$ ms). The comparison between the full model (M_{Full}) and M_{Main} showed the data were about equally likely under the model with vs without the familiarity \times item type interaction ($BF_{\text{Full,Main}} = 1.46$). With the

$BF_{Full,Main}$ being close to 1, the current data does not allow to reach a conclusion on the presence (or absence) of the interaction predicted by hypothesis 6. The results were robust to changes in the width of the cauchy prior.

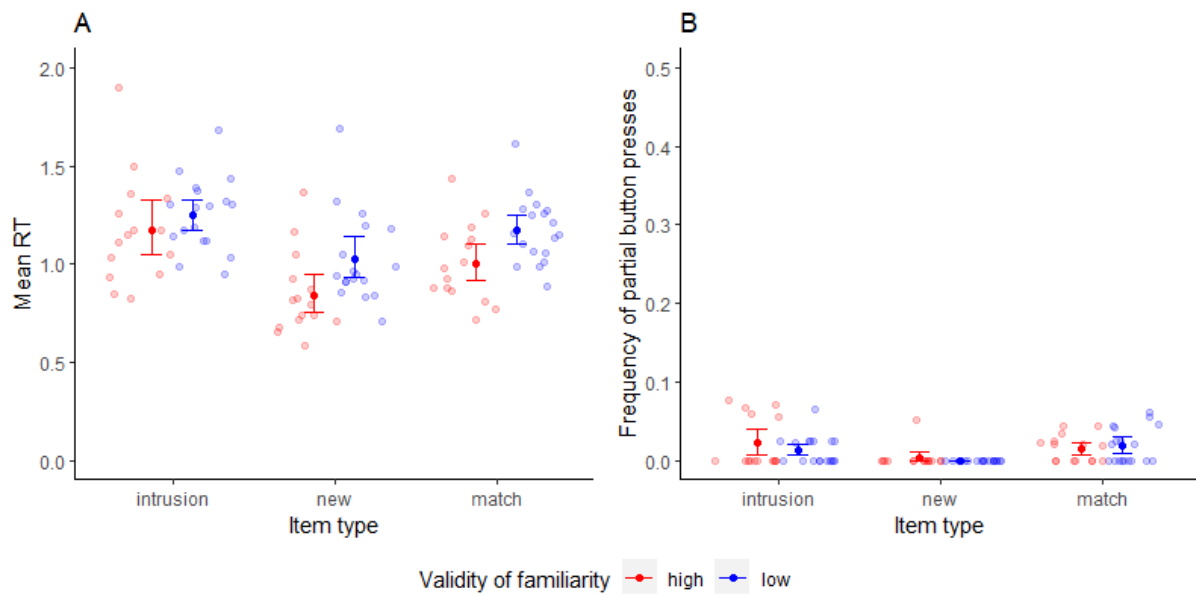
Partial button presses. We conducted a three (Item type: intrusion vs. new vs. match; within-subjects) by two (Validity of familiarity: low vs. high; between-subjects) Bayesian mixed effects ANOVA with JZS priors (Cauchy priors with scale = .5) on the frequency of partial button presses (Figure 5: B). As predicted by hypothesis 7, we found strong evidence for a main effect of item type ($BF_{Main,Familiarity} = 127$) but anecdotal evidence against an interaction effect ($BF_{Full,Main} = .34$), contrary to hypothesis 8. Pairwise group comparisons were conducted using a paired one-sided Bayesian Wilcoxon signed-rank test between intrusion and new trials ($BF_{Itemtype,0} = 102$)¹⁸ and a one-sided Bayesian t-test (Cauchy prior with scale = .707) between intrusion and match trials ($BF_{Itemtype,0} = .18$). The proportion of trials with partial button presses was low (1.3% of valid trials; $M_{partial_{intrusion}} = 1.71\%$, $SD = 2.54\%$; $M_{partial_{new}} = .16\%$, $SD = .89\%$; $M_{partial_{match}} = 1.72\%$, $SD = 1.95\%$). The results did not qualitatively change when we used the arcsine transformed data and the results were robust to changes in the width of the cauchy prior. The results of the of partial button presses should be interpreted cautiously as they are based on very few trials and the majority of participants did not show any partial errors in new and intrusion trials.

Non-preregistered analyses

The preregistered comparison of intrusion trials and new trials might not be the best comparison to assess the cost of response conflict. New trials can be resolved without using recollection altogether. Therefore, we also compared intrusion trials to match trials. Both require recollection but only the intrusion trials involve response conflict. We conducted a two (Item type: intrusion vs. match; within-subjects) by two (Validity of familiarity: low vs. high; between-subjects) Bayesian mixed effects ANOVA with JZS priors (Cauchy priors with scale = .5) on the participant mean RTs (Figure 5: A).

¹⁸ Normality assumption of the preregistered t-test was violated, we therefore report the results of the Wilcoxon signed rank test. Regardless, we also conducted the pairwise group comparison using paired one-sided Bayesian t-tests (Cauchy prior with scale = .707) between intrusion and new trials ($BF_{Itemtype,0} = 115$). The results do not differ qualitatively.

Figure 5. Participant mean RTs and frequency of partial button presses by item type in the Modified Sternberg Task



Note. The error bars indicate the bootstrapped 95% confidence intervals.

Comparing the main effects model (M_{Main}) to the model with only the main effect of familiarity ($M_{\text{Familiarity}}$) showed that the data is much more likely under M_{Main} ($BF_{\text{Main},\text{Familiarity}} = 2.1 * 10^5$). Therefore, we found strong evidence for intrusion costs in RTs (hypothesis 5; $MRT_{\text{intrusion}} = 1216$ ms, $SD = 231$ ms versus $MRT_{\text{match}} = 1094$ ms, $SD = 200$ ms). The comparison between the full model (M_{Full}) and M_{Main} showed the data were slightly more likely under the model with vs without the familiarity \times item type interaction ($BF_{\text{Full},\text{Main}} = 2.65$) providing anecdotal evidence for the interaction. We did not further analyse partial button presses in the Modified Sternberg task due to their very rare occurrence.

DISCUSSION

Response conflict is an integral part of various psychological tasks. An established direct measure of response conflict are partial errors in electromyographic data - increased muscular activity associated with the conflicting response option. Electromyograms require specialized equipment and often also require adaptations to the way participants respond in a task (e.g., pressing down a cylinder instead of

pressing a key on a keyboard; see Seymour & Schumacher, 2009). These complications lead us to explore if analog keyboards could be an alternative to assess partial errors. Our results show that partial errors are rare but they occur more frequently in conflict trials than in control trials in the RT-CIT. While we found the typical probe-irrelevant difference in RTs, we could not replicate the target familiarity effect (Suchotzki et al., 2018) despite sufficient statistical power (> 95%). We therefore consider this response conflict manipulation unsuccessful. Similarly, we found intrusion costs in RTs and increased frequency of partial errors in the Modified Sternberg task but ambiguous evidence regarding the response conflict manipulation.

While the response conflict manipulations would have helped to investigate the role of familiarity-based responding in more detail, we can still contrast conflict (probes; intrusions) to non-conflict (irrelevant; new, match) trials and compare the RT-CIT results to EMG findings.

Comparison to EMG data

The comparison of our results to the EMG results of Seymour et al. (2009) shows qualitative similarities between keyboard and EMG partial errors (i.e., higher relative frequency of partial errors for probes than for irrelevant items) but also quantitative differences (probes: 28% EMG vs 3% keyboard; irrelevant: 2% EMG vs. 0.5% keyboard). We see three possible reasons for this discrepancy.

First, Seymour et al. (2009) had an uncharacteristic data pattern compared to most RT-CIT studies. They found RTs for probes to be larger than for targets, and accuracies for probes to be lower than for targets (cf. Lukács et al., 2019, 2020; Meijer et al., 2007; Noordraven & Verschuere, 2013; Varga et al., 2015). The exceptionally low probe accuracy could indicate that participants forgot about the source of the probe over the course of the experiment. Participants were asked to learn and recall the probes before continuing to a 10 minutes long distractor task followed by the RT-CIT. Unfortunately, Seymour et al. (2009) did not include a second probe recall task after the RT-CIT to ensure that the probes were still remembered. If participants recognize the probe as familiar but cannot retrieve its context information, there would be no response conflict between familiarity and recollection based responding. Yet, they would experience response ambiguity because 50% of familiar items are targets

that require a YES-response (Jacoby, 1991; Seymour, 2001). Therefore, the quantitative difference between Seymour et al. (2009) and our study could be due to probe recollection performance. Because we used autobiographical probes, it is reasonable to assume perfect probe recollection in the RT-CIT and an unconfounded measure of partial errors in this regard. Further tentative support for the influence of response ambiguity on partial errors comes from the Modified Sternberg task: Match trials (without response conflict but response ambiguity if recollection failed) showed more partial errors than new trials.

The second explanation for the quantitative difference could be that the analog keyboard is inherently less sensitive to detect partial errors than EMG. While this cannot be ruled out, there are some factors that might have negatively influenced the analog keyboard's sensitivity to partial errors. For example, while we instructed participants to not remove the fingers from the keyboard, we do not know if they followed the instruction. But this is obviously a prerequisite if small movements should be measured. This possibility could be addressed by filming the participant's finger positions and excluding trials in which the fingers were not on the response keys, or by requiring that both response keys are minimally pressed for the next trial to start (although this might slow down the task and give participants too much control over the task without additional precautions). Furthermore, even though we stressed both speed and accuracy, participants had a long response window and, therefore, might have relied less on familiarity (the presumably conflict inducing feature; e.g., Ratcliff & McKoon, 2008). But even if it turns out that the analog keyboard is moderately less sensitive to detect partial errors, the compatibility with all tasks that use a standard keyboard and the ease of data collection compared to EMG might outweigh the reduced sensitivity.

Third, based on the parallel task set model (Seymour, 2001), partial button presses would be expected to occur at a lower rate than EMG partial errors. According to this model, partial errors that can be detected by the analog keyboard occur only when response conflict is detected during the response execution step of the familiarity based response. The recording of sub-threshold muscular activity by the EMG, however, should also be sensitive to response conflict that is detected during the response preparation phase of the familiarity based response.

Implications

For the RT-CIT, this method of detecting response conflict directly, especially the increased frequency of partial errors for probes compared to irrelevants, provides researchers with a new measure that could be used to detect knowledge in the RT-CIT. However, its incremental predictive value beyond RTs remains to be tested. Partial errors might also help detecting countermeasures such as intentionally slower responding (Norman et al., 2020; Suchotzki et al., 2021). We would expect that slower responding reduces the impact of familiarity and of the predominant "no"-response which, in consequence, decreases the frequency of partial errors for both probes and targets.

On a more general note, the relatively large number of partial errors in target trials indicates that partial errors might have been significantly influenced by the tendency towards the predominant "no"-response, given that five out of six trials required this response (e.g., Ratcliff & McKoon, 2008). It could be that this response bias made it more difficult to evoke familiarity-recollection-based partial errors. This suggests that the analog keyboard might be better suited for speeded conflict tasks with balanced responses (e.g., Eriksen Flanker task, Simon task; Eriksen & Eriksen, 1974; Simon & Wolf, 1963).

The analog keyboard allows us to identify individual trials in which the decision making process was already at the motor stage (initiated movement of the error response) but then managed to stop and correct their response – something that has gone unnoticed using regular keyboards. Our exploratory analysis suggests that response corrections at the motor stage comes at a greater cost than when the conflict was detected in the pre-motor stage (as evidenced in the slower RTs of probe trials with partial error compared to probe trials without partial error). Additionally, first results hint that the stopping cost might be dependent on when the conflict occurs. The mean stopping cost in target trials (for which familiarity information is sufficient to realize that the pre-dominant "no"-response needs to be stopped) seems to be smaller than in probe trials (for which recollection information is needed for the conflict to occur). Note that these are post-hoc explanations of exploratory results and need to be tested thoroughly before any conclusions should be drawn.

The more detailed view on the response behavior provided by the analog keyboard and the occurrence of partial errors might call for extensions of contemporary response models. A widely used family of models, sequential sampling models (for a review, see Forstmann et al., 2016), generally assume that evidence accumulates over time until a decision threshold is reached upon which the motor response is initiated. These models successfully capture many characteristics of RT data but do not have mechanisms that could account for behavioral partial errors. Another model, the Parallel Task Set model (Seymour, 2001), predicts both pre-motor partial errors (e.g., measured with EMG) and behavioral partial errors due to conflicting response preparation of familiarity-based and recollection-based response. However, a discussion on how the models could be extended is out of the scope of this manuscript and would be premature given that the current study only provides a first glimpse at the pattern of partial errors.

Future Studies

This was the very first study to explore analog keyboards as an alternative to EMG to measure partial errors. Considering our results but also the quantitative difference to EMG partial errors (Seymour et al., 2009), follow up studies should combine both measures to allow for a direct comparison and investigate if our results generalize to other speeded response conflict tasks (e.g., Erikson Flanker task, Simon task; Eriksen & Eriksen, 1974; Simon & Wolf, 1963).

To assess the applied value of partial errors the RT-CIT, future studies should include a naïve control group to assess the incremental value of partial errors for classification purposes both for the known-solution participant classification and the item classification of the searching RT-CIT (Koller et al. 2020). Different ways on how to include partial errors information in the classification procedure should be explored (e.g., partial error rate as an additional independent variable or the addition of partial errors to regular errors and use this as a predictor).

We also urge researchers to independently replicate the target familiarity effect (Suchotzki et al., 2018) that has only been studied in two, although well-powered, online experiments ($n = 357$, $n = 499$) before and we failed to replicate. It would be valuable for researchers to know if this is a robust

manipulation that can be used to manipulate the reliance on familiarity and therefore response conflict, and for practitioners have a way to improve the classification performance using familiar targets.

Conclusion

Our study showed that analog keyboards can detect partial errors although they occurred in a small minority of conflict trials. The frequency of partial errors was influenced by familiarity-recollection-based response conflict and by a predominant response option. Albeit its limited sensitivity in this study, analog keyboards could be a valuable tool to further our understanding of response conflict.

Author Contributions

Dave Koller proposed the initial study design which was refined in collaboration with Bruno Verschuere. Programming and data collection was done by research assistants under the supervision of Dave Koller. The analysis was done by Dave Koller. The manuscript was mainly written by Dave Koller but in close collaboration with Bruno Verschuere and in consultation of Franziska Hofer. Franziska Hofer and Bruno Verschuere supervised the project.

Acknowledgments

We thank the Swiss Federal Office of Civil Aviation (project number: 2016-106) and the Zurich State Police, Airport Division for their financial support. We also thank Moritz Truninger for programming this experiment.

CHAPTER 6

General discussion

Cues to deception, verbal and behavioral, are faint (DePaulo et al., 2003) and even experts cannot distinguish lies from truths accurately without suitable questioning protocols and measures (Bond & DePaulo, 2006). The CIT (Lykken, 1959) is a well validated questioning protocol to detect knowledge with possible applications in various fields (e.g., criminal investigations and court trials, background checks of personnel for high security positions, validating insurance claims). Albeit its potential, only the Japanese police apply the (physiological) CIT on a large scale (Osugi, 2011). More recently, behavioral measures (RTs) have been shown to be a valid alternative to physiological measures (Seymour et al., 2000; Suchotzki et al., 2017; Verschuere & De Houwer, 2011). In contrast to the physiological CIT, the RT-CIT does not require specialized equipment, and many people can be tested simultaneously and remotely. These desirable properties might even lead to new applications (e.g., screening of new recruits, remote passenger screening as part of the flight check-in procedure). But the RT-CIT also faces applied challenges that limit its scope. The goal of this thesis was to increase the applied viability of the RT-CIT by exploring ways to alleviate some of those restrictions.

SUMMARY OF RESULTS

Chapter 2 addressed the limitation that the validity of the RT-CIT is reduced when only one testable piece of information is known to the examiner (e.g., due to the lack of available information or because information was leaked to the public and should, therefore, not be used anymore; Lukács et al., 2017; Verschuere et al., 2015). To better understand the reason for the reduced validity and to improve the suboptimal single-probe protocol that tests for a single item, participants were asked to conceal their nationality. We presented the nationality information in different modalities (flag, word, map) and independently manipulated the number of target modalities and number of probe/irrelevant modalities per block. Our results show that the validity of the single-probe RT-CIT can be increased when the target information is presented in different modalities, while the number of probe/irrelevant modalities did not have an effect. On one hand, we found a viable modification to the RT-CIT for applied purposes since the selection of target items is under the examiners control (allowing the examiner to choose targets that can be presented in multiple modalities), and on the other hand, our results indicate that the difference

in validity between the single probe protocol and the multiple probe protocol might originate in how participants approach the task.

Chapter 3 was concerned with the RT-CIT's susceptibility to information contamination – probably the biggest threat to the CITs validity (Bradley et al., 2011). The RT-CIT can detect knowledge but cannot determine how the knowledge was acquired, which becomes problematic when probe information was disseminated. We set out to replicate the finding that the newly developed I-CIT (Lukács & Ansorge, 2019) is immune to information contamination and to investigate the aIAT's (Sartori et al., 2008) susceptibility to information contamination. Three groups of participants (Dutch, British with and without thorough knowledge about the Netherlands) were instructed to convince in either the aIAT or the I-CIT to be from the United Kingdoms. Contrary to Lukács and Ansorge (2019), we found the I-CIT to be susceptible to information contamination, but the aIAT was not. The aIAT showed high classification performance between the Dutch and both British groups individually ($AUC \geq .86$) and could, therefore, be a valuable tool to detect false nationality claims, even when information contamination is suspected.

Chapter 4 was the first study to explore searching algorithms in the RT-CIT which could expand its scope to situations in which the critical information is unknown to the examiner (e.g., the police arrested a thief and want to find out where the stolen goods are hidden). In a realistic mock-crime scenario at an international airport, participants were instructed to commit a mock-terror attack (or a control activity). They were intercepted by an experimenter posing as an undercover police officer and brought to a room for an additional security test (RT-CIT). We evaluated two searching algorithms inspired by Meixner and Rosenfeld (2011) and Noordraven and Verschuere (2013). Both algorithms classified the participants and the specific crime information with above-chance performance but with a considerable number of misclassifications. Simulations suggest that the performance could be increased by increasing the number of trials per item. Chapter 4, therefore, showed that the RT-CIT can be applied when the crime information is unknown, but its efficacy to classify the examinee is reduced compared to the known-solution RT-CIT in which the crime information is known. Additionally, the realistic mock-crime setting used in this study provided further evidence for the validity of the RT-CIT in the field, but actual field studies testing the validity have not been conducted yet.

Chapter 5 introduced a technological modification, an analog keyboard, to test the RT-CIT theory and explore new RT-CIT measures. With the information of how far a key is pressed down at any time, the goal was to explore if response conflict leads to partial response errors, similar to partial errors in EMG measures (Seymour & Schumacher, 2009). In two response conflict tasks, the autobiographical RT-CIT and the modified Sternberg task (Oberauer, 2001), we found that partial errors were more frequent in conflict than in non-conflict trials, suggesting that the analog keyboard could be used to measure response conflict. However, partial errors were rare, even in the conflict trials (CIT: 2.9%; Sternberg: 1.7%). The additional information from the analog keyboard could provide valuable insights on decision making and responding processes on a trial-by-trial level for a multitude of tasks. For the RT-CIT specifically, partial errors might help to identify countermeasures or potentially increase classification performance.

IMPLICATIONS AND FUTURE RESEARCH

Taken together the findings of this thesis, the promising results regarding the RT-CIT's validity in non-student populations (Suchotzki et al., 2019; Visu-Petra et al., 2016), its validity in self-initiated cheating scenarios (Geven et al., 2018), and the prevention of countermeasures (Suchotzki et al., 2021), field testing is more called for than ever. This is not to say that less laboratory research should be conducted, quite the contrary. With the application as a possible next step after field testing come real-life consequences and well-controlled research is necessary to provide practitioners with the optimal paradigm, classifiers, and boundary conditions.

Laboratory research

While most research focused on understanding the psychological processes involved in the RT-CIT (Gamer et al., 2007; Seymour & Schumacher, 2009; Suchotzki et al., 2015; Verschuere & De Houwer, 2011; Visu-Petra et al., 2012, 2014; Visu-Petra et al., 2016), on improving the paradigm (Lukács, Gula, et al., 2017; Lukács et al., 2017; Lukács & Ansorge, 2019; Seymour et al., 2000, 2013;

Suchotzki et al., 2018), or on identifying boundary conditions and expanding its scope (Georgiadou et al., 2019; Geven et al., 2019; Kleinberg & Verschuere, 2015, 2016; Lukács et al., 2020; Noordraven & Verschuere, 2013; Norman et al., 2020; Seymour & Fraynt, 2009; Seymour & Kerlin, 2008; Suchotzki et al., 2019, 2021, 2017; Varga et al., 2015; Verschuere et al., 2015; Verschuere & Kleinberg, 2016), little is known about the optimal test specifications and data processing – research areas that might improve the classification performance and, important for field application, might give practitioners guidelines on how to set up the RT-CIT and analyze the data.

With regard to test specification, I consider the number and frequency of target items, the item ratio, the number of probes, timing, and the number of trials crucial aspects that need more research. In two experiments, Suchotzki et al., (2018) contrasted a two-target and a four-target condition in an RT-CIT with two probes. The four-target condition showed larger probe-irrelevant differences, but this effect might not be based on the number of targets but on the frequency of each target item (which was halved in the four-target condition to keep the ratio between probes, irrelevants, and targets constant). It is further unknown, if the often used 1:1:4 ratio for target, probe, and irrelevants indeed achieves the best results. Systematic research on the optimal number of probes is also scarce. Although six or more are considered ideal (Seymour & Fraynt, 2009), Eom et al. (2016) found that using more than three probes was not beneficial. However, the validity of this study is questionable due to the confounded experimental design. Regarding timings, Suchotzki et al. (2021) showed that a response deadline prevents participants from faking (by intentionally slowing down the responses to irrelevant items) but it came at the cost of higher exclusion rates and reduced validity for non-faking participants. Additional studies are needed to understand the effects of response deadlines and explore possible effects of inter-trial-interval specifications (e.g., little variation might lead to rhythmic responding that could influence the task's validity; Nobre et al., 2007). A last important test specification is the number of trials. Lukács (2021) split the data of twelve experiments and analyzed the probe-irrelevant differences between the first half and second half of the trials. He found that although the difference slightly decreased, the cumulative AUC increased over the number of trials suggesting that more trials lead to better results. Supplementary analyses (Lukács, 2021, fig. A2) suggest that there is an optimal number of trials that maximizes the AUC for most experiments but this prediction and the reasons for it (e.g., habituation,

strategy development, fatigue) have yet to be tested. Although the RT-CIT, with the commonly used specification, is a valid memory detection paradigm (e.g., Suchotzki et al., 2017), exploring alternative specifications might reveal potential for further improvements.

With regard to data processing, possibly the biggest advancement towards application might come from a systematic analysis of scoring systems (e.g., in-/exclusion of error trials or accuracy information; similar to work on the IAT see Greenwald et al., 2003) and classification procedures (i.e., the integration of information to reach a classification; see Matsuda et al., 2012) to improve classification performance and develop sound guidelines for practitioners. While different combinations of scoring systems and classification procedures (e.g., see Noordraven & Verschuere, 2013; Seymour et al., 2013; Seymour & Fraynt, 2009) proved viable, they have never been compared directly. Their evaluation on the same data sets could identify the most suitable data processing and analysis pipeline. The classification performance might improve further by using more sophisticated analyses that could incorporate a wider range of predictive data patterns (e.g., from an analog keyboard) or provide confidence estimates for the classification (e.g., in the form of likelihood ratios) instead of binary classifications.

Considering possible field testing and application, research on countermeasures (e.g., ascribing relevance to irrelevant items, randomly delay responding by sometimes taking the hands off the keyboard, or slowing of responses to irrelevant items; Norman et al., 2020; Suchotzki et al., 2021) should be intensified. This pertains the effectiveness of different countermeasures, their detection (e.g., intentional slowing or taking the hands off the keyboard might reduce partial errors for probes and targets or it might lead to exceptionally wide RT distributions), and their prevention (e.g., response deadlines (Suchotzki et al., 2021) or trials only start when both response keys are slightly pressed, to ensure the fingers are not lifted).

Field research

The empirical studies of this thesis provided further evidence that the RT-CIT could be useful to practitioners and a natural next step would be to test the RT-CIT in the field. The first step in field

research should purely focus on the validation of the RT-CIT and neither replace nor influence current proceedings. This recommendation is not only to avoid consequences for the examinee based on a possibly invalid test, but also to reduce the potential for nonscientific criticism that could impede discussions regarding implementation (e.g., the media confusing the CIT with the CQT polygraph “lie detector” and transferring justified criticism of the CQT to the CIT).

The main problem of field validation is the lack of ground truth. Every possible criterion (e.g., court decision, confession, forensic evidence) is error prone (see Drizin & Leo, 2004; Kellman et al., 2014) and attempts to minimize those errors by only including clear-cut cases might introduce a bias. In scenarios with least one valid, albeit not perfect, alternative procedure (e.g., evidence collected during an investigation) that measures the same construct (e.g., possession of crime knowledge), this procedure could be used to assess the comparative validity. In this case, the RT-CIT should be conducted as soon as sufficient testable information is available and before any information is disclosed to avoid information contamination. To obtain accurate false-positive estimates, suspects and non-suspects should take the RT-CIT. Test results should be withheld until the regular proceedings were concluded and a verdict was reached to prevent biases (cf. Iacono & Ben-Shakhar, 2018).

A highly relevant example of a scenario (Hofer et al., 2021) for which no valid alternative procedure exists (except for, possibly, extensive background checks by police and intelligence agencies which might not be feasible if a large number of people need to be screened or due to time pressure) is the identification of connections to a violent fundamentalism. One attempt to address this problem could be to test former members or supporters of the violent fundamentalist groups (e.g., ISIS) and different control groups (e.g., moderate Muslims, other (non-)religious groups) on in-group knowledge (Figure 1) to demonstrate validity of this specific RT-CIT in the laboratory. Additionally, large scale non-anonymous testing of high-risk groups (e.g., testing people immediately before travelling to Syria at the time when ISIS was on the rise) could provide valuable information on the test’s sensitivity. If there is evidence that a tested person joined the fundamentalist group (e.g., pictures, video footage of the person, records of the ISIS administration), researchers could go back to the person’s test results and identify it as either a true positive or a false negative result. Therefore, sensitivity could be estimated with more and more confirmed cases over time, under the premise that that the RT-CIT data of confirmed

fundamentalists do not differ from unconfirmed fundamentalists. Specificity estimates are more difficult, as they require evidence that someone does not have a connection to the fundamentalist group. Experts might find criteria that indicate that a connection is unlikely (e.g., short stay, family in Syria) or one could assume that a person is not connected if no connection could be confirmed after several years, but this is most likely more error prone than the confirmation of cases and it might lead to a biased sample. Despite these uncertainties, an estimation of the RT-CIT’s specificity is needed to evaluate the test’s validity, since even undiagnostic tests can be highly sensitive at the cost of low specificity.

Figure 1. Exemplary items to detect knowledge of violent Islamic fundamentalism (Hofer et al., 2021, p.16)

Probe	Target	Irrelevant 1	Irrelevant 2	Irrelevant 3	Irrelevant 4	Irrelevant 5	Irrelevant 6	Irrelevant 7
								
DABIC	ECOS	avis	monop	DIAGEN	Esquire	PROG	MONOCLE	Tattva Viveka

Note. Items to the question “Do you recognize this item in connection to Islamism?”

Given the complexity of the second scenario (organizationally, legally, and design-wise), it might be advisable for first field studies to focus on situations in which a valid alternative classification procedure is available. Possible scenarios include detection of crime information (police investigations serves as ground truth) and verification of origin in migration, using information restricted to the examinee’s suspected origin (e.g., idiosyncratic expressions; linguistic analyses could serve as ground truth; Verrips, 2010). But irrespective of the scenario and exact methodology, the current stage of research, despite the remaining open questions, justifies taking this next important step towards application – towards a scientifically valid tool to test claims of nescience.

CONCLUSION

The criterion validity of the RT-CIT has been shown in many laboratory studies (Suchotzki et al., 2017) but there are conditions for the RT-CIT to perform optimally that might not always be met in practice.

By addressing insufficient testable information, information contamination, and unknown critical information, our studies showed that the RT-CIT's scope may be wider than previously thought. The explored technological innovation might be a seed to further increase the validity of the RT-CIT. Combined with its ease of use, automatic data collection and analysis, and scalability due to remote testing, the RT-CIT might be a viable tool for practitioners in various fields. However, its validity in real-life situation remains to be tested and research on other essential aspects (e.g., countermeasures, classification algorithms) is still scarce.

REFERENCES

- Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011a). Detecting Fakers of the autobiographical IAT. *Applied Cognitive Psychology*, 25, 299–306.
<http://dx.doi.org/10.1002/acp.1691>
- Agosta, S., Mega, A., & Sartori, G. (2011b). Detrimental effects of using negative sentences in the autobiographical aIAT. *Acta Psychologica*, 136(3), 296–306.
<https://doi.org/10.1016/j.actpsy.2010.05.011>
- Agosta, S., Pezzoli, P., & Sartori, G. (2013). How to detect deception in everyday life and the reasons underlying it. *Applied Cognitive Psychology*, 27(2), 256–262. <https://doi.org/10.1002/acp.2902>
- Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology*, 4, 1–12. <https://doi.org/10.3389/fpsyg.2013.00519>
- Algom, D., & Fitousi, D. (2016). Half a century of research on Garner interference and the separability-integrality distinction. *Psychological Bulletin*, 142(12), 1352–1383.
<https://doi.org/10.1037/bul0000072>
- Allain, S., Burle, B., Hasbroucq, T., & Vidal, F. (2009). Sequential adjustments before and after partial errors. *Psychonomic Bulletin and Review*, 16(2), 356–362.
<https://doi.org/10.3758/PBR.16.2.356>
- Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, 29(5), 504–522. <https://doi.org/10.1111/j.1469-8986.1992.tb02024.x>
- Barry, R. J. (1982). Novelty and significance effects in the fractionation of phasic OR measures: A synthesis with traditional OR theory. *Psychophysiology*, 19(1), 28–35.
<https://doi.org/10.1111/j.1469-8986.1982.tb02595.x>

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>
- Ben-Shakhar, G. (2011). Countermeasures. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory Detection: Theory and Application of the Concealed Information Test* (pp. 200–214). Cambridge University Press. <https://doi.org/10.1017/CBO9780511975196.012>
- Ben-Shakhar, G. (2012). Current research and potential applications of the concealed information test: An overview. *Frontiers in Psychology*, *3*, 1–11. <https://doi.org/10.3389/fpsyg.2012.00342>
- Biancotti, C., Borin, A., Cingano, F., Tommasino, P., & Veronese, G. (2020, March 18). *The case for a coordinated COVID-19 response: No country is an island*. VoxEU.
<https://voxeu.org/article/case-coordinated-covid-19-response-no-country-island>
- Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2
- Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. <https://doi.org/10.1037/0033-2909.134.4.477>
- Bradley, M., Barefoot, C., & Arsenault, A. (2011). Leakage of information to innocent suspects. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory Detection: Theory and Application of the Concealed Information Test* (pp. 187–199). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511975196.011>
- 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>

REFERENCES

- 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>
- Bürkner, P. C. (2017). brms : An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1). <https://doi.org/10.18637/jss.v080.i01>
- Burle, B., Possamaï, C. A., Vidal, F., Bonnet, M., & Hasbroucq, T. (2002). Executive control in the Simon effect: An electromyographic and distributional analysis. *Psychological Research*, 66(4), 324–336. <https://doi.org/10.1007/s00426-002-0105-6>
- Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, 22(1), 267–281. <https://doi.org/10.1111/j.1467-9450.1981.tb00402.x>
- Cook, A. E., Hacker, D. J., Webb, A. K., Osher, D., Kristjansson, S. D., Woltz, D. J., & Kircher, J. C. (2012). Lyin’ eyes: Ocular-motor measures of reading reveal deception. *Journal of Experimental Psychology: Applied*, 18(3), 301–313. <https://doi.org/10.1037/a0028307>
- DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118. <https://doi.org/10.1037/0033-2909.129.1.74>
- 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(2), 266–282. <https://doi.org/10.1037/xap0000250>
- Donchin, E., Heffley, E., Hillyard, S. A., Loveless, N., Maltzman, I., Öhman, A., Rösler, F., Ruchkin, D., & Siddle, D. (1984). Cognition and event-related potentials II. The orienting reflex and P300. *Annals of the New York Academy of Sciences*, 425(1), 39–57. <https://doi.org/10.1111/j.1749-6632.1984.tb23522.x>
- Doob, A. N., & Kirschenbaum, H. M. (1973). Bias in police lineups - Partial remembering. *Journal of Police Science and Administration*, 1, 187–293.

REFERENCES

- Drizin, S. A., & Leo, R. A. (2004). The problem of false confessions in the post-DNA world. *North Carolina Law Review*, *21*(2), 891–1007.
- 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>
- Eom, J., Sohn, S., Park, K., Eum, Y., & National, C. (2016). Effects of varying numbers of probes on RT-based CIT accuracy. *International Journal of Multimedia and Ubiquitous Engineering*, *11*(2), 229–238. <https://doi.org/10.14257/ijmue.2016.11.2.23>
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. <https://doi.org/10.3758/BF03203267>
- Euronews. (2021). Mutter (28) wegen Mord ihrer 5 Kinder zu lebenslanger Haft verurteilt. Retrieved January 04, 2022, from <https://de.euronews.com/2021/11/04/mutter-28-wegen-mord-ihrer-5-kinder-zu-lebenslanger-haft-verurteilt-solingen>
- European Asylum Support Office (2015). *EASO-Bericht über Herkunftsländer-Informationen: Länderfokus Eritrea*. <https://www.easo.europa.eu/sites/default/files/public/BZ0415327DEN.pdf>
- European Asylum Support Office (2019). *Eritrea - National service, exit, and return*. https://coi.easo.europa.eu/administration/easo/PLib/2019_EASO_COI_Eritrea_National_service_exit_and_return.pdf
- Factly. (2020). About 23% of all IPC cases are disposed of by the Police for 'Lack of Evidence'. Retrieved January 26, 2020, from <https://factly.in/about-23-of-all-ipc-cases-are-disposed-of-by-the-police-for-lack-of-evidence/>
- Falkenstein, M., Hoormann, J., & Hohnsbein, J. (1999). ERP components in Go/Nogo tasks and their relation to inhibition. *Acta Psychologica*, *101*(2–3), 267–291. [https://doi.org/10.1016/s0001-6918\(99\)00008-6](https://doi.org/10.1016/s0001-6918(99)00008-6)

REFERENCES

- Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy (“lie detection”) with event-related brain potentials. *Psychophysiology*, 28, 531–547.
<https://doi.org/10.1111/j.1469-8986.1991.tb01990.x>
- 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>
- Federal Act on Foreign Nationals and Integration of 16 December 2005, SR 142.20 (2005).
https://sherloc.unodc.org/cld/uploads/res/document/federal-act-on-foreign-nationals_html/Federal_Act_on_Foreign_National_EN.pdf
- Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, 67, 641–666. <https://doi.org/10.1146/annurev-psych-122414-033645>
- Frank, A., Biberici, S., & Verschuere, B. (2019). The language of lies: a preregistered direct replication of Suchotzki and Gamer (2018; Experiment 2). *Cognition and Emotion*, 33(6), 1310–1315.
<https://doi.org/10.1080/02699931.2018.1553148>
- Gamer, M. (2011a). Detecting concealed information using autonomic measures. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 27–45). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511975196.003>
- Gamer, M. (2011b). Detecting of deception and concealed information using neuroimaging techniques. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 90–113). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511975196.006>
- Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human Brain Mapping*, 28(12), 1287–1301. <https://doi.org/10.1002/hbm.20343>

REFERENCES

- Gamer, M., & Yoni, P. (2018). Detecting concealed knowledge from ocular responses. In J. P. Rosenfeld (Ed.), *Detecting Concealed Information and Deception: Recent Developments* (pp. 169–186). Academic Press. <https://doi.org/10.1016/B978-0-12-812729-2.00008-2>
- Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.
- Georgiadou, K., Chronos, A., Verschuere, B., & Sauerland, M. (2019). Reaction time-based Concealed Information Test in eyewitness identification is moderated by picture similarity but not eyewitness cooperation. *Psychological Research*, 1–11. <https://doi.org/10.1007/s00426-018-1139-8>
- Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in Cognitive Science*, 1–24. <https://doi.org/10.1111/tops.12353>
- Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2019). It's a match!?! Appropriate item selection in the Concealed Information Test. *Cognitive Research: Principles and Implications*, 4(1), 11. <https://doi.org/10.1186/s41235-019-0161-8>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197–216. <https://doi.org/10.1037/0022-3514.85.2.197>
- Hadar, A. A., Makris, S., & Yarrow, K. (2012). The truth-telling motor cortex: Response competition in M1 discloses deceptive behaviour. *Biological Psychology*, 89(2), 495–502. <https://doi.org/10.1016/j.biopsycho.2011.12.019>
- Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4), 643–659. <https://doi.org/10.1037/a0023589>

REFERENCES

- Hartwig, M., & Granhag, P. A. (2014). Exploring the nature and origin of beliefs about deception. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting Deception: Current Challenges and Cognitive Approaches* (pp. 125–154). John Wiley & Sons.
<https://doi.org/10.1002/9781118510001.ch6>
- Hofer, F., Dolder, Z., & Koller, D. (2021). *Abschlussbericht studie 6: Einsatzszenarien*. [Unpublished manuscript].
- Hu, X., Evans, A., Wu, H., Lee, K., & Fu, G. (2013). An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychologica*, *142*(2), 278–285. <https://doi.org/10.1016/j.actpsy.2012.12.006>
- Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating automatic autobiographical associations: The effect of instruction and training in strategically concealing information in the autobiographical Implicit Association Test. *Psychological Science*, *23*(10), 1079–1085.
<https://doi.org/10.1177/0956797612443834>
- Iacono, W. G., & Ben-Shakhar, G. (2018). Current status of forensic lie detection with the comparison question technique: An update of the 2003 National Academy of Sciences report on polygraph testing. *Law and Human Behavior*. <https://doi.org/10.1037/lhb0000307>
- Inquisit (2016). Inquisit (Version 5) [Computer software]. <https://www.millisecond.com>
- Inquisit (2020). Inquisit (Version 6) [Computer software]. <https://www.millisecond.com>
- Jacoby, L. L. (1991). A process dissociation framework : Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 513–541.
- JASP Team. (2020). JASP (Version 0.14.0.0) [Computer software]. <https://jasp-stats.org>
- Kassin, S. M., Leo, R. A., Meissner, C. A., Richman, K. D., Colwell, L. H., Leach, A. M., & Fon, D. La. (2007). Police interviewing and interrogation: A self-report survey of police practices and beliefs. *Law and Human Behavior*, *31*(4), 381–400. <https://doi.org/10.1007/s10979-006-9073-5>

REFERENCES

- Kellman, P. J., Mnookin, J. L., Erlikhman, G., Garrigan, P., Ghose, T., Mettler, E., Charlton, D., & Dror, I. E. (2014). Forensic comparison and matching of fingerprints: Using quantitative image measures for estimating error rates through understanding and predicting difficulty. *PLoS ONE*, *9*(5), e94617. <https://doi.org/10.1371/journal.pone.0094617>
- klein Selle, N., Verschuere, B., & Ben-Shakhar, G. (2018). Concealed Information Test: Theoretical background. In J. P. Rosenfeld (Ed.), *Detecting Concealed Information and Deception* (pp. 35-57). Academic Press. <https://doi.org/10.1016/B978-0-12-812729-2.00002-1>
- klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2016). Orienting versus inhibition in the Concealed Information Test: Different cognitive processes drive different physiological measures. *Psychophysiology*, *53*(4), 579–590. <https://doi.org/10.1111/psyp.12583>
- 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>
- Koller, D., Hofer, F., Grolig, T., Ghelfi, S., & Verschuere, B. (2020). What are you hiding? Initial validation of the reaction time-based searching concealed information test. *Applied Cognitive Psychology*, *34*, 1406–1418. <https://doi.org/10.1002/acp.3717>
- Koller, D., Hofer, F., & Verschuere, B. (2021). Different target modalities improve the single probe protocol of the response time-based Concealed Information Test. *Journal of Applied Research in Memory and Cognition*. <https://doi.org/10.1016/j.jarmac.2021.08.003>

REFERENCES

- Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O'Brien, C. P., & Childress, A. R. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeuroImage*, *15*(3), 727–732.
<https://doi.org/10.1006/nimg.2001.1003>
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343.
<https://doi.org/10.3758/s13428-011-0146-0>
- Levine, T. R., Blair, J. P., & Carpenter, C. J. (2017). A critical look at meta-analytic evidence for the cognitive approach to lie detection: A re-examination of Vrij, Fisher, and Blank (2017). *Legal and Criminological Psychology*, *23*, 7–19. <https://doi.org/10.1111/lcrp.12115>
- Lukács, G. (2021). Prolonged response time Concealed Information Test decreases probe-control differences but increases classification accuracy. *Journal of Applied Research in Memory and Cognition*.. <https://doi.org/10.1016/j.jarmac.2021.08.008>
- Lukács, G., & Ansorge, U. (2019). Information leakage in the response time-based Concealed Information Test. *Applied Cognitive Psychology*, *33*(6), 1178–1196.
<https://doi.org/10.1002/acp.3565>
- Lukács, G., Grządziel, A., Kempkes, M., & Ansorge, U. (2019). Item roles explored in a modified P300-Based CTP Concealed Information Test. *Applied Psychophysiology Biofeedback*, *44*(3), 195–209. <https://doi.org/10.1007/s10484-019-09430-6>
- Lukács, G., Gula, B., Szegedi-Hallgató, E., & Csifcsák, G. (2017). Association-based Concealed Information Test: A novel reaction time-based deception detection method. *Journal of Applied Research in Memory and Cognition*, *6*(3), 283–294. <https://doi.org/10.1016/j.jarmac.2017.06.001>
- Lukács, G., Kleinberg, B., Kunzi, M., & Ansorge, U. (2020). Response time Concealed Information Test on smartphones. *Collabra: Psychology*, *6*(1), 4. <https://doi.org/10.1525/collabra.255>

REFERENCES

- Lukács, G., Kleinberg, B., & Verschuere, B. (2017). Familiarity-related fillers improve the validity of reaction time-based memory detection. *Journal of Applied Research in Memory and Cognition*, 6(3), 295–305. <https://doi.org/10.1016/j.jarmac.2017.01.013>
- Lukács, G., & Specker, E. (2020). Dispersion matters: Diagnostics and control data computer simulation in Concealed Information Test studies. *PLoS ONE*, 15(10), 1–22. <https://doi.org/10.1371/journal.pone.0240259>
- 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>
- Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist*, 29, 725–739. <https://doi.org/10.1037/h0037441>
- Mac Giolla, E., & Luke, T. J. (2020). Does the cognitive approach to lie detection improve the accuracy of human observers? *Applied Cognitive Psychology*, 35, 1–8. <https://doi.org/10.1002/acp.3777>
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Review*, 109, 163–203. <https://doi.org/10.1037/0033-2909.109.2.163>
- Matsuda, I., Nittono, H., & Allen, J. J. B. (2012). The current and future status of the concealed information test for field use. *Frontiers in Psychology*, 3, 1–11. <https://doi.org/10.3389/fpsyg.2012.00532>
- Matsuda, I., Nittono, H., Hirota, A., Ogawa, T., & Takasawa, N. (2009). Event-related brain potentials during the standard autonomic-based concealed information test. *International Journal of Psychophysiology*, 74(1), 58–68. <https://doi.org/10.1016/j.ijpsycho.2009.07.004>
- Meijer, E. H., Bente, G., Ben-Shakhar, G., & Schumacher, A. (2013). Detecting concealed information from groups using a dynamic questioning approach: Simultaneous skin conductance measurement and immediate feedback. *Frontiers in Psychology*, 4, 1–6. <https://doi.org/10.3389/fpsyg.2013.00068>

REFERENCES

- 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>
- Meijer, E. H., Smulders, F. T. Y., & Merckelbach, H. L. G. J. (2010). Extracting concealed information from groups. *Journal of Forensic Sciences*, *55*(6), 1607–1609. <https://doi.org/10.1111/j.1556-4029.2010.01474.x>
- Meijer, E. H., Smulders, F. T. Y., Merckelbach, H. L. G. J., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, *66*(3), 231–237. <https://doi.org/10.1016/j.ijpsycho.2007.08.001>
- Meijer, E. H., & Verschuere, B. (2015). The polygraph: Current practice and new approaches. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting Deception: Current Challenges and Cognitive Approaches* (pp. 59–80). John Wiley & Sons. <https://doi.org/10.1002/9781118510001.ch3>
- 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*(5), 593–604. <https://doi.org/10.1111/psyp.12609>
- 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>
- Mohamend, H. M. (2021). Fingerprint classification using a deep convolutional neural network. *Journal of Electrical and Electronic Engineering*, *2*, 9(5), 147–152. <https://doi.org/10.11648/j.jeee.20210905.11>
- Morey, R. D., & Rouder, J. N. (2018). Computation of Bayes Factors for common designs (Version 4.2) [Computer software]. <https://richarddmorey.github.io/BayesFactor/>

REFERENCES

- National Research Council. (2003). *The Polygraph and lie detection. Committee to review the scientific evidence on the polygraph. Division of behavioral and social sciences and education.* National Academies Press. <https://doi.org/10.17226/10420>
- Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology*, 48(2), 162–175. <https://doi.org/10.1111/j.1469-8986.2010.01057.x>
- Nobre, A., Correa, A., & Coull, J. (2007). The hazards of time. *Current Opinion in Neurobiology*, 17(4), 465–470. <https://doi.org/10.1016/j.conb.2007.07.006>
- 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>
- Norman, D. G., Gunnell, D. A., Mrowiec, A. J., & Watson, D. G. (2020). Seen this scene? Scene recognition in the reaction-time Concealed Information Test. *Memory and Cognition*. <https://doi.org/10.3758/s13421-020-01063-z>
- Nyberg, L., Marklund, P., Persson, J., Cabeza, R., Forkstam, C., Petersson, K. M., & Ingvar, M. (2003). Common prefrontal activations during working memory, episodic memory, and semantic memory. *Neuropsychologia*, 41(3), 371–377. [https://doi.org/10.1016/S0028-3932\(02\)00168-9](https://doi.org/10.1016/S0028-3932(02)00168-9)
- Oberauer, K. (2001). Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27(4), 948–957. <https://doi.org/10.1037/0278-7393.27.4.948>
- Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2021). Validity of content-based techniques for credibility assessment—How telling is an extended meta-analysis taking research bias into account? *Applied Cognitive Psychology*, 35(2), 393–410. <https://doi.org/10.1002/acp.3776>

REFERENCES

- Ogawa, T., Matsuda, I., Tsuneoka, M., & Verschuere, B. (2015). The Concealed Information Test in the laboratory versus Japanese field practice: Bridging the scientist – practitioner gap. *Archives of Forensic Psychology, 1*(2), 16–27. <https://doi.org/10.16927/afp.2015.1.2>
- Olson, J. M., Rosenfeld, P. J., & Perrault, E. (2020). Familiarity-related filler items enhance the RT CIT (but not the P300 CIT) with differential effects on episodic compared to semantic protocols. *International Journal of Psychophysiology, 158*, 370–379. <https://doi.org/10.1016/j.ijpsycho.2020.10.001>
- Orshea, J., Crockett, K., Khan, W., Kindynis, P., Antoniadis, A., & Boultadakis, G. (2018). Intelligent deception detection through machine based interviewing. *Proceedings of the International Joint Conference on Neural Networks, July*. <https://doi.org/10.1109/IJCNN.2018.8489392>
- Osugi, A. (2010). Gap and connection between laboratory research and field application of the CIT in Japan. *International Journal of Psychophysiology, 77*(3), 238. <https://doi.org/10.1016/j.ijpsycho.2010.06.356>
- Osugi, A. (2011). Daily application of the concealed information test: Japan. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory Detection: Theory and Application of the Concealed Information Test* (pp. 253–275). Cambridge University Press. <https://doi.org/10.1017/CBO9780511975196.015>
- Osugi, A. (2014). Review and analysis of the practical data conducted in Japanese criminal investigation. *International Journal of Psychophysiology, 94*(2), 131. <https://doi.org/10.1016/j.ijpsycho.2014.08.617>
- Osugi, A. (2018). Field findings from the Concealed Information Test in Japan. In J. P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 97-121). Academic Press. <https://doi.org/10.1016/B978-0-12-812729-2.00005-7>
- Pilling, D., & Schipani, A. (2020, November 18). *Ethiopia crisis: 'a political mess that makes fathers fight sons'*. Financial Times. <https://www.ft.com/content/b888c23a-45ed-4937-9154-3117cc23e202>

REFERENCES

- Podlesny, J. A. (2003). A paucity of operable case facts restricts the applicability of the guilty knowledge technique in FPI criminal polygraph examinations. *Forensic Science Communications*, 5(3), 20-37.
- Quick Placement Test. (2001). Oxford: Oxford University Press.
- R Core Team (2020). R: A language and environment for statistical computing. [Computer software]. <https://www.R-project.org/>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- 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>
- Reid, R. E. (1947). A revised questioning technique in lie-detection tests. *The Journal of Criminal Law and Criminology, Including the American Journal of Police Science*, 37(6), 542–547. <https://doi.org/10.2307/1138979>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77. <https://doi.org/10.1186/1471-2105-12-77>
- Roos, K., Wenger, I., Sowe, R., & Indermühle, Y. (2018). Addressing barriers to work for asylum seekers: Report from Switzerland. *World Federation of Occupational Therapists Bulletin*, 74 (2), 123–127. <https://doi.org/10.1080/14473828.2018.1540100>
- Rosenfeld, J. P., Cantwell, G., Nasman, V. T., Wojdac, V., Ivanov, S., & Mazzeri, L. (1988). A modified, event-related potential-based Guilty Knowledge Test. *International Journal of Neuroscience*, 42(1–2), 157–161. <https://doi.org/10.3109/00207458808985770>

REFERENCES

- Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, *45*(6), 906–919. <https://doi.org/10.1111/j.1469-8986.2008.00708.x>
- Rosenfeld, J. P., Ozsan, I., & Ward, A. C. (2017). P300 amplitude at Pz and N200/N300 latency at F3 differ between participants simulating suspect versus witness roles in a mock crime. *Psychophysiology*, *54*(4), 640–648. <https://doi.org/10.1111/psyp.12823>
- Rosenfeld, P. J. (2011). P300 in detecting concealed information. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 63–89). Cambridge University Press. <https://doi.org/10.1017/CBO9780511975196.005>
- Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 1-6. <https://doi.org/10.1109/ICCPCT.2017.8074258>
- Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772–780. <https://doi.org/10.1111/j.1467-9280.2008.02156.x>
- Seymour, T. L. (2001). A EPIC model of the “guilty knowledge effect”: *Strategic and automatic processes in recognition*. *Dissertation Abstracts International: Section B. The Sciences & Engineering*, *61*, 5591.
- Seymour, T. L., Baker, C. A., & Gaunt, J. T. (2013). Combining blink, pupil, and response time measures in a concealed knowledge test. *Frontiers in Psychology*, *3*, 1–15. <https://doi.org/10.3389/fpsyg.2012.00614>
- Seymour, T. L., & Fraynt, B. R. (2009). Time and encoding effects in the concealed knowledge test. *Applied Psychophysiology Biofeedback*, *34*(3), 177–187. <https://doi.org/10.1007/s10484-009-9092-3>

REFERENCES

- Seymour, T. L., & Kerlin, J. R. (2008). Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Applied Cognitive Psychology*, 22, 475–490.
<https://doi.org/10.1002/acp.1375>
- Seymour, T. L., & Schumacher, E. H. (2009). Electromyographic evidence for response conflict in the exclude recognition task. *Cognitive, Affective and Behavioral Neuroscience*, 9(1), 71–82.
<https://doi.org/10.3758/CABN.9.1.71>
- 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>
- Simon, J., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics*, 6(1), 99–105. <https://doi.org/10.1080/00140136308930679>
- Sokolov, E. N. (1963). *Perception and the conditioned reflex*. Macmillan.
- Strange, B. A., Henson, R. N. A., Friston, K. J., & Dolan, R. J. (2000). Brain mechanisms for detecting perceptual, semantic, and emotional deviance. *NeuroImage*, 12(4), 425–433.
<https://doi.org/10.1006/nimg.2000.0637>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662. <https://doi.org/10.1037/h0054651>
- 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. (2019). Validity of the reaction time concealed information test in a prison sample. *Frontiers in Psychiatry*, 10, 1–8.
<https://doi.org/10.3389/fpsy.2018.00745>

REFERENCES

- Suchotzki, K., Verschuere, B., & Gamer, M. (2021). How vulnerable is the reaction time Concealed Information Test to faking? *Journal of Applied Research in Memory and Cognition*, January. <https://doi.org/10.1016/j.jarmac.2020.10.003>
- Suchotzki, K., Verschuere, B., Peth, J., Crombez, G., & Gamer, M. (2015). Manipulating item proportion and deception reveals crucial dissociation between behavioral, autonomic, and neural indices of concealed information. *Human Brain Mapping*, 36(2), 427–439. <https://doi.org/10.1002/hbm.22637>
- 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*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511816772>
- Swiss State Secretary of Migration (2020). *Monitoring Asylsystem: Bericht 2019*. <https://www.sem.admin.ch/dam/sem/de/data/publiservice/berichte/monitoring-asyl/monitoring-asylsystem-2019.pdf>
- Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401–409. <https://doi.org/10.1037/h0058700>
- The MathWorks. (2018). MATLAB (version R2018a) [Computer software]. <https://www.mathworks.com>
- Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.
- Tuzet, G. (2015). On the absence of evidence. In T. Bustamante & C. Dahlman (Eds.), *Argument Types and Fallacies in Legal Argumentation* (112th ed., pp. 37–51). Springer. <https://doi.org/10.1007/978-3-319-16148-8>

REFERENCES

- United Nations (2019). *International Migration 2019: Report*.
https://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/InternationalMigration2019_Report.pdf
- Varga, M., Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2015). The “good cop, bad cop” effect in the rt-based concealed information test: Exploring the effect of emotional expressions displayed by a virtual investigator. *PLoS ONE*, *10*(2), 1–27. <https://doi.org/10.1371/journal.pone.0116087>
- Verrips, M. (2010). Language analysis and contra-expertise in the Dutch asylum procedure. *International Journal of Speech, Language and the Law*, *17*(2), 279–294.
<https://doi.org/10.1558/ijssl.v17i2.279>
- Verschuere, B., & Ben-Shakhar, G. (2011). Theory of the Concealed Information Test. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory Detection. Theory and Application of the Concealed Information Test*, (pp. 128–148). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511975196.008>
- Verschuere, B., Ben-Shakhar, G., & Meijer, E. H. (2011). *Memory Detection: Theory and Application of the Concealed Information Test*. Cambridge University Press.
- Verschuere, B., Crombez, G., & Koster, E. H. W. (2004). Orienting to guilty knowledge. *Cognition and Emotion*, *18*(2), 265–279. <https://doi.org/10.1080/02699930341000095>
- Verschuere, B., Crombez, G., Koster, E. H. W., Van Bockstaele, B., & De Clercq, A. (2007). Startling secrets: Startle eye blink modulation by concealed crime information. *Biological Psychology*, *76*(1–2), 52–60. <https://doi.org/10.1016/j.biopsycho.2007.06.001>
- Verschuere, B., & De Houwer, J. (2011). Detecting concealed information in less than a second: Response latency-based measures. In B. Verschuere, G. Ben-Shakhar, & E. H. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 128–148). Cambridge University Press. <https://doi.org/10.1017/CBO9780511975196.004>
- Verschuere, B., & Kleinberg, B. (2016). ID-check: Online Concealed Information Test reveals true identity. *Journal of Forensic Sciences*, *61*, 237–240. <https://doi.org/10.1111/1556-4029.12960>

REFERENCES

- Verschuere, B., & Kleinberg, B. (2017) Assessing autobiographical memory: The web-based autobiographical Implicit Association Test. *Memory*, 25(4), 520-530.
<https://doi.org/10.1080/09658211.2016.1189941>
- 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>
- Verschuere, B., & Meijer, E. H. (2014). What's on your mind? Recent advances in memory detection using the concealed information test. *European Psychologist*, 19(3), 162–171.
<https://doi.org/10.1027/1016-9040/a000194>
- Verschuere, B., Prati, V., & Houwer, J. De. (2009). Cheating the lie detector: Faking in the autobiographical Implicit Association Test: Research report. *Psychological Science*, 20(4), 410–413. <https://doi.org/10.1111/j.1467-9280.2009.02308.x>
- Visu-Petra, G., Miclea, M., Buş, I., & Visu-Petra, L. (2014). Detecting concealed information: The role of individual differences in executive functions and social desirability. *Psychology, Crime & Law*, 20(1), 20–36. <https://doi.org/10.1080/1068316X.2012.736509>
- Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Applied Cognitive Psychology*, 26(3), 342–351. <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, 1–11. <https://doi.org/10.3389/fpsyg.2013.00146>
- Visu-Petra, L., Jurje, O., Ciornei, O., & Visu-Petra, G. (2016). Can you keep a secret? Introducing the RT-based Concealed Information Test to children. *Psychology, Crime and Law*, 22(3), 276–301.
<https://doi.org/10.1080/1068316X.2015.1109085>

REFERENCES

- Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). John Wiley & Sons.
- Vrij, A. (2015). A cognitive approach to lie detection. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Wiley series in the psychology of crime, policing and law. Detecting deception: Current challenges and cognitive approaches* (pp. 205–229). Wiley-Blackwell.
- Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22, 1–21. <https://doi.org/10.1111/lcrp.12088>
- Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, 1(2), 110–117. <https://doi.org/10.1016/j.jarmac.2012.02.004>
- WDR. (2021). Fünf Kinder ermordet: «Es sah aus, als würden sie schlafen.». Retrieved January 04, 2022, from <https://www1.wdr.de/nachrichten/rheinland/prozess-solingen-mutter-angeklagt-wegen-mordes-100.html>
- World Bank (2019). *Refugee population by country or territory of asylum*. <https://data.worldbank.org/indicator/SM.POP.REFG>
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3), 441–517. <https://doi.org/10.1006/jmla.2002.2864>
- 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)
- Zaitso, W. (2016). External validity of Concealed Information Test experiment: Comparison of respiration, skin conductance, and heart rate between experimental and field card tests. *Psychophysiology*, 53(7), 1100–1107. <https://doi.org/10.1111/psyp.12650>
- Zangrossi, A., Agosta, S., Cervesato, G., Tessarotto, F., & Sartori, G. (2015). "I didn't want to do it!" The detection of past intentions. *Frontiers in Human Neuroscience*, 9, 608. <https://doi.org/10.3389/fnhum.2015.00608>

REFERENCES

- Zhang, R., Geng, X., & Lee, T. M. C. (2017). Large-scale functional neural network correlates of response inhibition: an fMRI meta-analysis. *Brain Structure and Function*, 222(9), 3973–3990.
<https://doi.org/10.1007/s00429-017-1443-x>

ACKNOWLEDGEMENTS

A word of thanks.

Bruno, thank you for mentorship and support. Your can-do attitude and encouragement for innovation allowed me to explore new approaches despite the uncertain outcome and our informal meetings often were a source of inspiration. I also want to express my gratitude for helping me to settle in in Amsterdam and for suggesting to pursue a joint PhD in the first place. It was a great experience that I will never forget. I am looking forward to reading more of the interesting research coming from the Lielab and, hopefully, future collaborations.

Klaus, I am grateful that you gave me the opportunity to do this PhD in the extraordinary research group you put together, even though my topic was quite distinct from the rest of the group's research. The possibility to discuss any problem on short notice or even during our communal lunches coupled with the freedom to pursue my interests provided a great research environment.

Franziska, thank you for helping me to keep an eye on the applied aspects and challenges. The conversations with practitioners from various fields that you enabled were very insightful and gave me a different perspective on the topic.

I also want to thank the members of the Doctorate Committee, Prof. Dr. Reinout Wiers, Prof. Dr. Bram Orobio de Castro, Prof. Dr. Lutz Jäncke, Prof. Dr. Johannes Ullrich, and Dr. Jaume Masip for taking the time to assess this thesis.

Further, I want to thank the Federal Office of Civil Aviation and the Kantonspolizei Zürich that financed this project.

ACKNOWLEDGEMENTS

To all the members of the cognition lab in Zurich and the Lielab in Amsterdam, thank you for the welcoming atmosphere in your team and the great time at work and after work. To my long-time office mates Clara, Samuel, Julia, and Hannah: Thank you all for the great time we had. No matter what was going on, your company brightened the day.

A special word of thanks also goes out to my parents, my brother, and his family. Thank you for your support, for never doubting me, and for giving me a home when COVID-19 forced me to go back to Switzerland before I intended. Knowing you all behind me made this journey a lot easier. To my nieces, Jarina and Elina, I cannot wait to see you growing up and to create many beautiful memories together.

I want to thank all my long-time friends that supported me from the beginning and all the newly made friends during my PhD. Kyle and Lane, stumbling across your post on Reddit was the best thing that could have happened to me. Our weekly meetups lead to friendships with so many wonderful people that I will cherish forever. You all were a huge part of what made my stays in Amsterdam the great experience that they were.

Ninna, I can't express how grateful I am to have you in my life. You are the most amazing person and the way we met still feels like a fairy tale to me. You supported me more than I could have ever hoped for. Your believe in me, your kind words in difficult times, the shared joy of even the smallest accomplishments, and the knowledge that you are there for me anytime and no matter what gave me the energy to carry on. Your friendship and love mean the world to me.