DECEPTION IN COURT: OPEN ISSUES AND DETECTION TECHNIQUES

EDITED BY: Cristina Scarpazza and Giuseppe Sartori
PUBLISHED IN: Frontiers in Psychiatry

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-830-7
DOI 10.3389/978-2-88963-830-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

DECEPTION IN COURT: OPEN ISSUES AND DETECTION TECHNIQUES

Topic Editors:
Cristina Scarpazza, University of Padova, Italy
Giuseppe Sartori, University of Padova, Italy

Citation: Scarpazza, C., Sartori, G., eds. (2020). Deception in Court: Open Issues and Detection Techniques. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-830-7

Table of Contents

frontiers in Psychiatry

EDITORIAL
published: 25 May 2020
doi: 10.3389/fpsyt.2020.00476

Editorial: Deception in Court—Open Issues and Detection Techniques

Cristina Scarpazza and Giuseppe Sartori*

Department of General Psychology, University of Padova, Padova, Italy

Keywords: lie detection, memory detection, forensic psychiatry, criminal cases, translational application

Editorial on the Research Topic

Deception in Court—Open Issues and Detection Techniques

"Deceiving others. That is what the world calls a romance." – Oscar Wilde

Forensic psychiatric assessment is an extremely difficult task that is even more complicated by the risk of deception and malingering. Due to the high prevalence of the latter (around 40%) (1, 2), an accurate and thorough evaluation is a cornerstone issue in forensic practice. This is especially true in the case of insanity evaluation, where the assessment of psychiatric and cognitive symptoms is further complicated by the fact that these symptoms can be easily faked or exaggerated (3) for defensive purposes, although the majority of offenders found not guilty by reason of insanity have had previous contacts with psychiatric services. Taking this problem into consideration is even more important during evaluation of defendants who do not have had a previous psychiatric history.

The importance of assessing malingering is unfortunately still underestimated by clinicians, who usually are overconfident on their clinical ability to detect deceptive behavior (4). However, scientific research suggests that experienced individuals (i.e. judges, psychiatrists, etc) performance in detecting deception is only slightly better than chance (5). For these reasons, in the last few years, there has been increasing interest in the application of cutting edge methods for the detection of deception to enhance its accuracy in the legal setting.

The aim of this Research Topic is twofold: first, it aims to provide an updated overview of the techniques currently used to detect deception and malingering in court. Second, it aims to provide new perspectives, emerging concepts, and novel deception detection techniques that could potentially expand the future role of neuro-scientific evidences in court.

The Research Topic opens with a comprehensive review of approaches to detect malingering in forensic context (Walczyk et al.), where the authors summarize the strategies currently applied to detect malingering of psychiatric symptoms and cognitive impairment. Critically, the shortcomings of each method are described. The review also analyzes in detail behavioral, reaction based memory detection techniques, such as the Concealed Information Test (CIT), and the Autobiographic Implicit Association Test (aIAT). A final emphasis is placed on new methods, grounded on theoretical accounts of deception, attempting to induce greater cognitive load on liars than truth tellers.

Two original studies deepen our knowledge on classical lie detection techniques. First, the interesting work from Curci et al. investigates the accuracy of relying on experiential criteria as paraverbal aspects and cognitive complexity to identify liars from videotaped interview. The results confirm previous literature that the accuracy in discriminating liars from truth-tellers is far from accurate and that the identification of truth is more accurate than the one of lie. Critically, the study

OPEN ACCESS

Edited and reviewed by:
Thomas Nilsson,
University of Gothenburg, Sweden

****Correspondence:***
Cristina Scarpazza
cristina.scarpazza@gmail.com

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *20 March 2020*
Accepted: *11 May 2020*
Published: *25 May 2020*

Citation:
Scarpazza C and Sartori G (2020)
Editorial: Deception in Court—Open
Issues and Detection Techniques.
Front. Psychiatry 11:476.
doi: 10.3389/fpsyt.2020.00476

also highlights that judges' accuracy is poorly related to their confidence in their detection and this should be taken into account in real life settings to avoid wrongful decisions. Second, Mazza, Burla et al. used the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) in a very big sample (n=400) of post-divorce child custody court controversies, revealing that these individuals showed higher scores in underreporting and lower scores in overreporting validity scales. These results are critical as they highlight the necessity to interpret the MMPI-2 profile of these clients in light of normative data collected specifically in a forensic setting and the urgent need to identify alternative/complementary methods.

Five papers expanded the topic of CIT. Matsuda et al. shared their expert knowledge on the use of CIT in real criminal investigations in Japan, where the CIT is widely used in association with the polygraph to detect deception. Interestingly, they underline the difference between laboratory and field CIT and discussed some practical problems in its use and interpretation, such as the determination of statistical methods to be applied, the selection of a discriminative threshold to identify cheaters and the need to add additional measures to reduce the inconclusive cases. An original study by Ambach et al. focused on the impact of evaluative observation on psysiological responding in CIT. In a between-subjects manipulation, participants were divided into two groups, based on whether or not they were observed through a camera and were faced with the real-time video of the experimenter watching them while completing the CIT. Physiological measures were recorded. Results revealed that the expected enhanced CIT effect under evaluative observation was not present. A second study on CIT, by Rosenfeld et al., aims to investigate the influence of instruction and motivation on the P300 CIT effect and found that the financial motivation does not impact the P300 CIT effect and that financial incentives has no incremental effect after participants are instructed to defeat the test. The third original study of this section aims to differentiate between innocent suspects who have knowledge of crime information and guilty suspects. To this aim, Kim et al. used eye tracking to study eye movement of participants while viewing crime-relevant, crime-irrelevant and neutral stimuli. The interesting results revealed that guilty individuals show attentional avoidance as they focused their attention on crime-relevant and irrelevant stimuli for a shorter period of time compared with innocent individuals who have knowledge of crime. The potential translational application of these results is worth further investigations. The translational application of reaction-time (RT) CIT has been expanded by Suchotzki et al., where it has been used in a forensic sample submitted to an imaginary mock crime task. The data revealed that the RT CIT produced medium to large effects in both error rate and RTs, supporting the hypothesis that RT CIT is a promising techniques also in real life contexts. Second, the CIT effect was stronger in the inmate group compared to the control group, when error rates are analyzed. Third, the CIT effect does not correlate with impulsivity, rejecting the hypothesis that CIT effect in forensic samples can be attributable to differences in response inhibition.

Two interesting papers cover the topic of verbal lie detection and underline the need of more strategy-based research in the field of verbal lie detection. In particular, Vrij et al., besides providing a comprehensive review of the verbal lie detection techniques, focuses on the Model Statement, a technique where interviewers elicit participants to provide additional information on a specific topic. Based on prior knowledge of the different cognitive abilities used by individuals during truth telling or during lie, the method relies on the quality (rather than the quantity) of information that is reported to classify a narration as truthful or deceitful. Importantly, the authors describe how to use the model statement in real life, providing important and practical suggestions for scientists. The critical need to dig deeper into the language of liars, looking for traces of deception in the quality of the details provided during the narration is further expanded by Nahari and Nisin, who, in their opinion paper, wisely suggest to proceed with an in depth analysis of the narration, that, qualitatively, will greatly differ between truth tellers and liars.

Two papers cover the topic of detection of malingered amnesia for the crime. An interesting review by Jelicic summarizes the scientific evidences on crime related amnesia and described the methods actually used to evaluate its genuineness. Of particular relevance, the author also describe the approach to use the symptom validity testing strategy created on details from crime scenes and explains in which cases and why to adopt this approach can be considered more reliable than relying on other techniques to determine the authenticity of memory loss. The topic is further expanded by the original investigation from Zago et al., where the efficacy of three techniques, namely the aforementioned symptom validity testing, the facial thermography and the kinematic analysis of mouse movements, is compared with regard to the detection of feigned amnesia for crime. Besides confirming the efficacy of symptom validity testing in detecting feigned amnesia, the results also support the usefulness of kinematic analysis of mouse movements in differentiating truth tellers from liars in the case of amnesia malingering.

In the current Research Topic, new detecting deception techniques have also been proposed and their real potential translational application in court has been discussed. In particular, the fascinating idea of using the mouse trajectory dynamics as a tool for lie detection has been proposed in Monaro et al., where this technique has been used, during a two alternative forced choice task on symptoms of depression, to detect the simulation of depressive symptoms. The authors stressed that this tool has the key advantage that the kinematic movement is not consciously controllable by the individuals, and thus it is almost impossible to deceive. A complex data analysis performed using machine-learning models trained on mouse dynamics features, reached a classification up to 96% in distinguishing liars from depressed patients and truth-tellers. The usefulness of machine learning algorithms to enhance the accuracy detection of malingerers of psychopathology is also a key topic of Mazza, Monaro et al., where these algorithms applied to the Minnesota Multiphasic Personality Inventory-2

Restructured Form data, collected through a computerized form, revealed 95% of accuracy in detecting malingerers when subjects were instructed to respond under time pressure.

The Research Topic concludes with an important paper from Burgoon that, besides providing a summary of verbal and non verbal signals on which people rely to formulate gut judgments on authenticity, suggests to adopt an holistic approach based on the convergence of evidences principle, where multiple indicators of lie from different techniques are applied together to improve detection deception accuracy.

Lie and memory detection techniques are enormously promising as they have high potential translational value. As each method is characterized by specific drawbacks, scientists, and forensic experts should be well aware of them to select the most appropriate technique depending on their research or real life question, in order to enhance their future application into real world forensic practice. The innovative techniques discussed in this special issue are of interest both at the fact finding investigative stage (e.g. verbal lie detection) as well as in the verification stage (e.g. CIT). We hope that the readers will find this Research Topic a useful reference reflecting the current state of art in this emerging field of neuroscience based detecting deception tools.

AUTHOR CONTRIBUTIONS

CS and GS wrote the manuscript.

REFERENCES

1. Greve KW, Ord JS, Bianchini KJ, Curtis KL. Prevalence of malingering in patients with chronic pain referred for psychologic evaluation in a medico-legal context. *Arch Phys Med Rehabil* (2009) 90(7):1117–26. doi: 10.1016/j.apmr.2009.01.018
2. Mittenberg W, Patton C, Canyock EM, Condit DC. Base rates of malingering and symptom exaggeration. *J Clin Exp Neuropsychol* (2002) 24(8):1094–102. doi: 10.1076/jcen.24.8.1094.8379
3. Rosenhan DL. On being sane in insane places. *Science* (1973) 179(4070):250–8. doi: 10.1126/science.179.4070.250
4. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol* (1999) 77(6):1121–34. doi: 10.1037/0022-3514.77.6.1121
5. Kassin SM, Gudjonsson GH. The Psychology of Confessions: A Review of the Literature and Issues. *Psychol Sci Public Interest* (2004) 5(2):33–67. doi: 10.1111/j.1529-1006.2004.00016.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychiatry

ORIGINAL RESEARCH
published: 08 June 2018
doi: 10.3389/fpsyt.2018.00249

The Detection of Malingering: A New Tool to Identify Made-Up Depression

Merylin Monaro[1], Andrea Toncini[1], Stefano Ferracuti[2], Gianmarco Tessari[2], Maria G. Vaccaro[3], Pasquale De Fazio[4], Giorgio Pigato[5], Tiziano Meneghel[6], Cristina Scarpazza[1†] and Giuseppe Sartori[1†]*

[1] Department of General Psychology, University of Padova, Padova, Italy, [2] Department of Human Neurosciences, University of Roma "La Sapienza," Rome, Italy, [3] Neuroscience Center, Department of Medical and Surgical Science, University "Magna Graecia," Catanzaro, Italy, [4] Department of Psychiatry, University "Magna Graecia," Catanzaro, Italy, [5] Psychiatry Unit, Azienda Ospedaliera di Padova, Padova Hospital, Padova, Italy, [6] Dipartimento di Salute Mentale, Azienda Unità Locale Socio Sanitaria 9, Treviso, Italy

OPEN ACCESS

Edited by:
Thomas Masterman,
Karolinska Institutet (KI), Sweden

Reviewed by:
Axel Haglund,
Swedish National Board of Forensic Medicine, Sweden
Nubia G. Lluberes,
Baylor College of Medicine, United States

***Correspondence:**
Cristina Scarpazza
cristina.scarpazza@gmail.com;
cristina.scarpazza@unipd.it

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Forensic Psychiatry, a section of the journal Frontiers in Psychiatry

Received: 26 February 2018
Accepted: 23 May 2018
Published: 08 June 2018

Citation:
Monaro M, Toncini A, Ferracuti S, Tessari G, Vaccaro MG, De Fazio P, Pigato G, Meneghel T, Scarpazza C and Sartori G (2018) The Detection of Malingering: A New Tool to Identify Made-Up Depression. Front. Psychiatry 9:249. doi: 10.3389/fpsyt.2018.00249

Major depression is a high-prevalence mental disease with major socio-economic impact, for both the direct and the indirect costs. Major depression symptoms can be faked or exaggerated in order to obtain economic compensation from insurance companies. Critically, depression is potentially easily malingered, as the symptoms that characterize this psychiatric disorder are not difficult to emulate. Although some tools to assess malingering of psychiatric conditions are already available, they are principally based on self-reporting and are thus easily faked. In this paper, we propose a new method to automatically detect the simulation of depression, which is based on the analysis of mouse movements while the patient is engaged in a double-choice computerized task, responding to simple and complex questions about depressive symptoms. This tool clearly has a key advantage over the other tools: the kinematic movement is not consciously controllable by the subjects, and thus it is almost impossible to deceive. Two groups of subjects were recruited for the study. The first one, which was used to train different machine-learning algorithms, comprises 60 subjects (20 depressed patients and 40 healthy volunteers); the second one, which was used to test the machine-learning models, comprises 27 subjects (9 depressed patients and 18 healthy volunteers). In both groups, the healthy volunteers were randomly assigned to the liars and truth-tellers group. Machine-learning models were trained on mouse dynamics features, which were collected during the subject response, and on the number of symptoms reported by participants. Statistical results demonstrated that individuals that malingered depression reported a higher number of depressive and non-depressive symptoms than depressed participants, whereas individuals suffering from depression took more time to perform the mouse-based tasks compared to both truth-tellers and liars. Machine-learning models reached a classification accuracy up to 96% in distinguishing liars from depressed patients and truth-tellers. Despite this, the data are not conclusive, as the accuracy of the algorithm has not been compared with the accuracy of the clinicians; this study presents a possible useful method that is worth further investigation.

Keywords: depression, malingering, decetpion, machine learning, automatic

INTRODUCTION

Major depression is a high-prevalence [7%; (1)] mental disease with major socio-economic impact for both direct (medications and hospitalization) and indirect (mortality, work absence and turnover, disability compensation) costs (2). Strikingly, the length of absence from work due to depressive disorder is significantly longer than that due to organic serious illnesses such as heart disease, back pain, diabetes mellitus and hypertension (3). Although important, absenteeism is not the only cost whereby depression burdens the public health, as a critical percentage of the national health system income is devolved for the provision of invalidity pensions. The Italian government, for instance, recognizes an invalidity of up to 80% for people suffering with endogenous depression, with the consequent allocation of monthly disability checks amounting from 270 to 500 euro per person (4). In addition, in Italy, insurance companies spend weighty annual sums for the compensation of psychic damage, including depression, which could result, for example, from road accidents, stalking and mobbing (5).

Due to the undeniable economic advantages of being clinically depressed, major depression symptoms can be faked or exaggerated in order to obtain economic compensation. The literature on this topic is still at its infancy (6). In Italy, the problem of people feigning a wide range of symptoms to obtain disability pensions is of critical relevance. Indeed, in some regions of Italy, people feigned many conditions, from inability to walk to blindness to obtain economic advantages (https://www.ilfattoquotidiano.it/2013/02/09/falsi-invalidi-meccanico-in-svizzera-da-2500-prendeva-pensione-da-1300-in-italia/494390/; http://www.iltempo.it/economia/2016/03/31/news/tre-milioni-di-invalidi-100-mila-falsi-1005788/). Critically, the malingered conditions are simple to be feigned. The same might be true for depression; its symptoms are very intuitive for naïve people (7), as everyone has experienced low mood during life. Importantly, experienced clinicians are trained not to only rely on the self-reported symptoms provided by patients. Indeed, according with the DSM-5 guidelines, outstandingly important pieces of information also came from direct observation of signs of depression, observation of the non-verbal behavior of the patients, and convergence of the information reported by patients and relatives. However, the behavioral observations rely heavily on information that could be consciously controlled by the patient. This is because, as depression is a very common disorder, individuals who want to feign a depressive disorder do not require any particular knowledge or specific training to produce clinically reliable depressive symptoms and signs. Furthermore, a large majority of both symptoms and signs easy to fake: lack of concentration, restlessness, lack of interest for daily life activities, feelings of guilt, and so on are easy to fake if one wanted and planned to. For this reason, depression is one of the mental disorders that are more frequently and easily faked to achieve financial or other advantages, and this underlines, in the forensic setting, the necessity to couple the psychiatric examination with a different methodology, which is less influenced by the individuals' overt decisions to malinger a psychiatric disorder.

Malingering is defined as the voluntary fabrication or exaggeration of mental or physical symptoms to gain secondary benefits, which could include financial compensations or other advantages, such as leniency, drugs, avoiding obligations (school, work, army), or just getting the attention of other people (8).

Although malingering is not considered to be a mental disorder, recent scientific knowledge suggest that it should be the focus of clinical attention, so much so that it has been introduced in the Diagnostic and Statistical Manual of Mental Disorders [DSM-5; (1)]. This has been an important step forward in the scientific community, as the fabrication/accentuation of symptoms or the concealment of a disorder are very frequent, especially when the evaluation takes place in forensic contexts (9). Although it is hard to define it reliably, literature reports an estimate of the prevalence of malingering in a forensic setting as ranging from 20 to 40%. (9–11). In regards to depression, Mittenberg et al. reported that 16.08% of depressive syndromes which are diagnosed in litigation or compensation cases are feigned (10).

Currently, a diagnosis of a psychiatric disorder, including depression, is formulated according to the subjective experience that is reported by the patient and to the observation of signs and non-verbal behavior that are easily manipulated by the individual will to deceive (12). In other words, the psychic symptom exists because the patient refers to it, and the assessment of malingering is mostly based on clinical judgment (13). While this aspect is less (or none) than a problem in the clinical setting, in which patients are seeking help for their sufferance and in which malingering itself become a symptom (as in the case of factious disorder or Munchausen syndrome), the forensic context is a quite different situation. Indeed, as already introduced, in the psychiatric setting, symptoms can be exaggerated or faked to obtain a secondary advantage. Thus, classical psychiatric or clinical evaluation itself is not reliable when dealing with forensic-relevant topic. The limitations of classic psychiatric evaluation alone have been provocatively investigated in a well-known experiment conducted by Rosenhan (14), in which "pseudo-patients" feigning hallucinations were all admitted to the psychiatric department of 12 different highly specialized hospitals: all but one (who was diagnosed as having a bipolar disorder) received a clinical diagnosis of full-blown schizophrenia. In another study (15), the authors reported that experienced psychiatrists distinguished actors and depressed patients during a clinical interview with an accuracy close to the chance level. Furthermore, the clinicians rated their confidence in their diagnoses as 6.5 out of 10 in the case of patients and 7.1 in the case of actors, denoting that they were equally certain of their right and wrong diagnoses. Considered together, the results of these studies highlighted the urgent need to have complementary and integrative tools that may strengthen the process of achieving a correct psychiatric diagnosis.

This low reliability of the classical clinical evaluation used alone in detecting malingering in forensic setting led to an exponential growth of the research in this topic over the last 15 years (16). Different strategies to detect malingering in clinical and forensic setting have been proposed, and *ad hoc* tests have been developed. Strategies are varied, but until recently, they were mainly based on self-report questionnaires such as the M-Test and the Structured Inventory of Malingered Symptomatology (SIMS). The former was conceived to be specifically applied to the feigning of schizophrenic symptoms (17), while the latter was conceived to detect malingering of both psychiatric syndromes, including depression, and cognitive deficits (18). Although these instruments have been undoubtedly useful, they share the important limitation of the psychiatric assessment as described above: they are based on the "patients"' report of symptoms, and they can be easily deceived by coaching (19). As psychiatric symptoms are easy to be feigned, unmasking the simulation of psychiatric syndromes is much more challenging than unmasking other pathologies (e.g., cognitive disorders), and thus the instrument to detect them should be more sophisticated.

An important advance over the self-report questionnaires has been achieved in the last few years, as behavioral-based lie detection techniques to spot the simulation of psychiatric symptoms have been introduced. Contrary to the self-report questionnaire, which took into consideration the explicit answers of subjects, the behavioral methods mainly rely on implicit measures not fully under the explicit and conscious control of the evaluated subject. For example, the autobiographical Implicit Association Test [aIAT; (20)] and the Concealed Information Test [CIT; (21)] are able to identify liars based on their response time (TR). Concerning malingering detection, the aIAT has been successfully applied to detect whiplash malingering, confirming an accuracy of around 90% (22), as well as to unveil phantom-limb pain (13) and psychogenic amnesia (16). The CIT has been principally used to assess the simulation of amnesia (21). The main limitation of these implicit tools is that they can only investigate one symptom at a time. In other words, more than one aIAT or CIT would be necessary to establish whether the subject is feigning a psychiatric syndrome or not, checking all of the symptoms, one by one, with a specific test.

Interestingly, different studies in literature have shown that deception can be captured through analysis of hand-motor responses while the subject is engaged in a double-choice task (23–26). More in general, the kinematic analysis can be used as an implicit online measure of the mental operations that are put in place by the subject during a task (27). A simple hand movement, such as the movement of the mouse on the computer screen, reflects, in real time, the evolution of the cognitive processes underlying the action. Because lying requires great cognitive resources (28), the motor response to a stimulus is altered in terms of spatial and temporal features compared to the truth telling (23, 26).

The aim of the present study is to present a new tool specifically developed to detect malingering, which has the important advantages of: (i) being conceived specifically to evaluate the truthfulness of depressive syndrome (but that might also be adapted to other psychopathologies); (ii) relying on an implicit measure, i.e. the kinematic of movement, which is not consciously controllable; (iii) being not easy to cheat by knowing the symptomatology or by coaching; (iv) being based on machine learning algorithm that will allow the identification of liars at individual level; and (v) considering at the same time both implicit variables and clinical symptoms.

METHODS

Participants

As machine-learning algorithms require to be built and tested using two independent samples, two independent groups of participants have been selected for this research.

Group 1

Seventy-two Italian-speaking participants were recruited, with the aim to build machine-learning classification models. In detail, 26 patients suffering from depression were recruited (see below for details), as well as 46 age- and gender-matched healthy volunteers.

Before the experiment, the Beck Depression Inventory [BDI, (29)] was administered to all participants with the aim to exclude possible sub-clinical participants (undiagnosed depressed participants) within the healthy controls and to exclude responder participants (defined as clinically diagnosed participants under medications who did not manifest depressive symptoms) from the clinical group. Six sub-clinical participants and six responders were identified and excluded from the experiment. The final sample consisted of 60 participants (39 females, 21 males). The average age was 38.60 years ($SD = 14.74$), and the average education level was 15.15 years ($SD = 2.98$).

Twenty participants were depressed patients, and the remaining 40 individuals were healthy subjects randomly assigned to the truth-teller (e.g., non-depressed participants who were instructed to respond truthfully to the test; $n = 20$) or liar (e.g., non-depressed participants who were instructed to respond deceitfully to the test; $n = 20$) condition. An ANOVA confirmed that the three groups were similar in terms of age and schooling ($p > 0.01$ for both age and schooling), whereas a Chi-squared test (χ^2) confirms that they were similar also for gender (all $ps > 0.01$). On the contrary, the groups differed in the BDI score [$F_{(2,\,57)} = 83.41$, $p < 0.01$]: the post hoc test highlighted that, tautologically, the truth-tellers' BDI average score of 6.1 ($SD = 3.97$) was similar to the average liars' score of 6.2 ($SD = 3.62$), while the BDI score of the depressed patients clearly differed from the one of the healthy volunteers score of 29.5 ($SD = 10.09$).

The patients suffering from depression were recruited from Azienda Ospedaliera Sant'Andrea di Roma ($n = 4$), Ospedale Ca' Foncello di Treviso ($n = 10$), Unità Operativa di Psichiatria, Dipartimento di Scienze della Salute dell'Università Magna Grecia (Catanzaro; $n = 5$), and Azienda Ospedaliera di Padova ($n = 1$). These patients were diagnosed according to the DSM-IV criteria by an expert psychiatrist at each site. At the time of the study, all of the depressed participants were under pharmacological medications, and seven of them were attending psychotherapy.

Group 2

A second group comprising 27 Italian-speaking participants was also enrolled with the aim to test the model and its generalization (30). This second group consisted of 8 males and 19 females with an average age of 35.37 years ($SD = 21.42$) and an average education level of 13.15 years ($SD = 3.59$) and did not statistically differ from Group 1 in any demographic data (age: $p > 0.01$, education: $p > 0.01$). No sub-clinical or responder participants have been identified in this second group. As for Group 1, the participants enrolled in group 2 comprised 9 depressed patients (which were recruited from the same Institutions and with the same modalities of Group 1) and 18 healthy participants. Again, the healthy participants were randomly assigned to the truth-teller ($n = 9$) or liar ($n = 9$) condition. The three groups did not differ in age ($p > 0.01$), schooling ($p > 0.01$), or gender (all $ps > 0.01$), while they differ in BDI score [$F_{(2, 24)} = 34.65$, $p < 0.01$], with the depressed patients scoring higher than the healthy participants (truth-tellers: $M = 5.8$, $SD = 4.02$; liars: $M = 5.8$, $SD = 3.67$; depressed: $M = 27.9$, $SD = 9.87$). All of the depressed participants were under pharmacological medications, and two were in psychotherapy treatment.

All of the participants provided informed consent before the experiment. The experimental procedures were approved by the ethics committee for psychological research of the University of Padova and were in accordance with the declaration of Helsinki and its later amendments.

Stimuli

The stimuli adopted in the current study consisted of simple and complex questions about symptoms of depression and concerning the experimental condition. Please see Monaro et al. (31) for a description. The typical symptoms of depression were extracted from the Depression Questionnaire (QD) of the Cognitive Behavioural Assessment 2.0 [CBA 2.0; (32)] and from the Structured Clinical Interview for Mood Spectrum [SCI MOODS; (33)].

Simple questions referred to only one piece of information related to the experimental condition (e.g., "Are you carrying out a questionnaire?") or one piece of information related to a single symptom of depression (e.g., "Do you feel tired very easily?"). Each simple question required a "yes" or "no" response. Contrarily, complex questions are questions which comprised two (or more) pieces of information. A complex question required a "yes" response when both pieces of information were true, whereas it requires a "no" response when at least one of the two pieces of information was false. Asking complex questions is a method used to overcharge the cognitive load of liars (34). In fact, literature showed that the increment of the liar's cognitive load is an effective strategy to spot deceptive responses (35). While a truth-teller can easily decide whether each information is true or false, the liar has firstly to match each piece of information with his lie and then decide about it. In other words, the greater the number of pieces of information, the greater the liar's cognitive effort to monitor its plausibility (36).

More in depth, the experimental task included nine different types of questions that could be categorized as follows:

Simple Questions ($n = 30$):

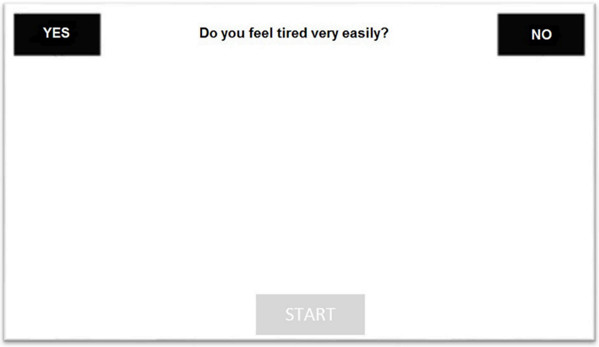
- 5 items referred to the experimental condition (EX; e.g., "Are you wearing shoes?"). These are control questions to which all participants are required to respond truthfully.
- 10 items referred to depressive symptoms (DS; e.g., "Do you think more slowly than usual?").
- 15 items referred to very atypical symptoms (VAS). These questions were taken from the Affective Disorders (AF) scale of the Structured Inventory of Malingered Symptomatology [SIMS; (18); e.g., "Do you rarely laugh?"]. The SIMS is a questionnaire designed to detect malingering through a number of bizarre experiences and highly atypical psychiatric symptoms reported by each participant. The AF scale consists of 15 items about very atypical symptoms of anxiety and depression. An individual is classified as malingering if he/she reports more than five atypical symptoms.
- Complex Questions ($n = 46$):
- 15 items consisted of two *discordant symptoms* (2DS-d): a typical symptom of depression and an atypical symptom of depression (e.g., "Do you face difficulties to concentrate at work, and **are you full of energy?**").
- 15 items consisted of two *concordant symptoms* (2DS-c); both of them were typical symptoms of depression (e.g., "Do you feel abandoned from the others, and **is your mood sad all day?**").
- 5 items consisted of two *discordant pieces of information*: a typical depression symptom and an information about the experimental condition (DS&EX-d). One piece of information was correct (in other words, it required a "yes" response), while the other one was not correct (it required a "no" response; e.g., "Do you have difficulties in concentrating, and **are you in Paris?**").
- 5 items consisted of two *concordant pieces of information*: a typical depression symptom and a piece of information about the experimental condition (DS&EX-c). Both of them are correct (both of them required a "yes" response; e.g., "Are you often sad, and **are you sitting on a chair?**").
- 3 items consisted of two discordant pieces of information, both of them concerning the experimental condition (2EX-d). One information was correct (in other words, it required a "yes" response), while the other one was not correct (it required a "no" response; e.g., "Are the questions written in red, and **are you wearing shoes?**").
- 3 items consisted of two concordant pieces of information, both of them concerning the experimental condition (2EX-c). Both of them are correct (both of them required a "yes" response; e.g., "Are you responding with the mouse, and **are you in a room?**").

The complete list of questions are reported in the Online Supplementary Information. Questions required responding "yes" or "no." All participants were expected to respond in the same way to the questions concerning the experimental condition (EX) and the very atypical depressive symptoms (VAS). On the contrary, truth-tellers, liars and depressed participants were expected to respond in different ways to the questions, including the depressive symptoms (DS). Indeed, truth-tellers were expected to give 19 "yes" responses and 76 "no" responses in cases in which they denied all depressive symptoms. Depressed participants were expected to give 44 "yes" responses and 57 "no" responses in cases in which they manifested all of the typical depressive symptoms. However, we contemplated that some healthy participants could express few depressive symptoms and, conversely, some depressed patients could deny any of the typical symptoms. For this reason, no feedback was presented in the case of participants who gave an unexpected response (e.g., a healthy participant who responded "yes" to the question "Are you in trouble falling asleep without drugs?"). Finally, liars were expected to give some "yes" responses to depressive symptoms similar to the ones provided by depressed participants. In other words, liars and depressed subjects were expected to declare an equal number of depressive symptoms.

Experimental Procedure

Just before the experimental task, participants assigned to the liar group were instructed to lie about their mood. In particular, they were asked to simulate a depressive status. To increase the compliance, participants were given a little scenario: "*Now, imagine being examined by an insurance policy commission to receive compensation for psychological damage. You have to make them believe that the damage has caused severe depression. So, you have to respond questions simulating a depression, trying to be credible and avoiding being unmasked.*" Conversely, truth-tellers and depressed subjects were asked to answer all the questions truthfully.

The task was programmed and run using *MouseTracker* software (37). Each participant was presented with 76 randomized questions displayed in the upper part of the computer screen. The squares containing YES and NO response labels were located in the upper left and upper right parts of the screen. Participants were instructed to press the START button (located in the lower part of the screen) to let the questions appear and to then respond to questions by clicking with the mouse on the correct label (YES or NO). **Figure 1** shows an example of the computer screen as it appeared to the subjects during the task. The experimental procedure was preceded by 10 training questions, to allow participants to familiarize themselves with the task.



Do you feel tired very easily?

FIGURE 1 | The figure reports an example of the computer screen as appeared to the subjects during the task.

Data Collection

For each answer, motor response was tracked using *MouseTracker* software (37). To permit averaging and comparison across multiple trials, the software performs a time normalization. Specifically, each trajectory is normalized in 101 time frames through linear interpolation. This resulted in each time frame corresponding to specific x and y coordinates in a binary space. In other words, the software derived the position of the mouse along the axis over the 101 time frames (X_n,Y_n). The software also describes the motor response in terms of spatial and temporal features, such as onset, duration, shape, stability and direction of the trajectory. The space–time features recorded by *MouseTracker* are described in detail in **Table 1**. For each of these features, the average value of the responses in the different types of questions (EX, DS, VAS, 2DS-d, 2DS-c, DS&EX-d, DS&EX-c, 2EX-d, 2EX-c) were computed. In addition, the average velocity (v) and acceleration (a) of the mouse movement between two time frames, respectively, on the x-axis ($v_x = X_n - X_{n-1}$ and $a_x = v_{xn} - v_{xn-1}$) and y-axis ($v_y = Y_n - Y_{n-1}$ and $a_y = v_{yn} - v_{yn-1}$) were calculated. The number of symptoms reported by the participants (DS, 2DS-d, 2DS-c, DS&EX-d, DS&EX-c, and VAS) and the number of errors in the control questions related to the experimental condition (EX, 2EX-d, 2EX-c) were also computed. This procedure led to a total of 83 variables that were entered as predictors in machine-learning models (please see the online supplements for a detailed description of the 83 features).

TABLE 1 | The table reports the description of the space-time features recorded by *MouseTracker* software.

	Feature	Description
Temporal features	Initiation time (IT)	Time between the appearance of the question and the beginning of the mouse movement
	Reaction time (RT)	Time from the appearance of the question to the click on the response box
	Maximum deviation time (MD-time)	Time to reach the point of maximum deviation
Spatial features	Maximum deviation (MD)	The largest perpendicular distance between the actual trajectory and the ideal trajectory
	Area under the curve (AUC)	The geometric area between the actual trajectory and the ideal trajectory
	x-flip	Number changes in direction along the x-axis
	y-flip	Number changes in of direction along the y-axis

RESULTS

Visual Analysis of Mouse Trajectories

A preliminary visual analysis was carried out comparing the trajectories of the three experimental groups. **Figure 2** compares

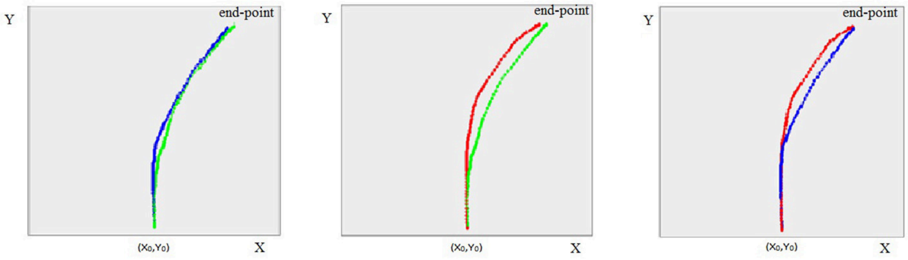


FIGURE 2 | The figure represents the average trajectories between the participants, respectively for liars (in red), truth-tellers (in green) and depressed subjects (in blue), to all questions (EX, DS, VAS, 2DS-d, 2DS-c, DS&EX -d, DS&EX-c, 2 EX-d, 2EX-c).

the average trajectories of liars, truth-tellers and depressed subjects, considering their responses to all the 76 questions. The visual pattern is similar to the one observed in other studies that spot liars through mouse dynamics (26). The trajectories of liars and truth-tellers seem to differ in both AUC and MD parameters. Indeed, both healthy and depressed truth-tellers outlined a more direct trajectory connecting the starting point with the correct response. By contrast, in the initial phase of the response, the liars spent more time moving on the y-axis, and they then deviated toward the response with a delay compared to truth-tellers.

Univariate Statistical Analysis

In **Table 2**, the descriptive statistics of the three experimental groups were reported for the space–time features collected by the software considering all the 76 items of the task.

A univariate one-way ANOVA was performed on each of the 83 collected features with the aim of identifying the variables that statistically differed between groups. **Table 3** reports the variables that differed the three groups.

Finally, a Tukey test was run as a post hoc test to verify which groups accounted for the significant differences found by ANOVA. The results are reported in **Table 4**.

TABLE 2 | The table reports means (M) and standard deviations (SD) for each feature collected by the software.

Feature	Truth-tellers		Depressed		Liars	
	M	***SD***	***M***	***SD***	***M***	***SD***
IT	620.35	491.94	408.67	332.99	559.57	399.84
RT	4018.79	1466.98	6641.81	3204.22	4030.60	1203.67
MD-time	2392.37	1001.23	4199.85	1818.62	2297.64	620.98
MD	0.44	0.31	0.51	0.30	0.57	0.24
AUC	1.01	0.90	1.09	0.76	1.25	0.67
x-flip	7.64	2.24	8.75	2.94	9.23	2.72
y-flip	7.74	2.63	8.05	2.88	9.20	2.67
v_x	0.00627	0.00060	0.00582	0.00076	0.00573	0.00059
v_y	0.01326	0.00014	0.01315	0.00010	0.01326	0.00017
a_x	−0.00001	0.00004	0.00001	0.00002	0.00000	0.00002
a_y	−0.00004	0.00008	−0.00006	0.00004	−0.00001	0.00009

IT, initiation time; RT, reaction time; MD-time, maximum deviation time; MD, maximum deviation; AUC, area under the curve; x-flip, y-flip, average velocity and acceleration on x and y axis = v_x, v_y, a_x, a_y, respectively for liars, truth-tellers and depressed subjects responding to all questions.

TABLE 3 | The table reports *F*-value, degrees of freedom (*gdl*), *p*-value and effect-size (Omega-squared, ω^2) resulting from the comparison of the three experimental groups for the features that reached the statistical significance.

Feature	One-way ANOVA (*gdl*, *F-value*, *p-value*, *effect-size*)
DS	$F_{(2,57)} = 93.59$, $p < 0.01$, $\omega = 0.87$
EX	$F_{(2,57)} = 3.84$, $p < 0.05$, $\omega = 0.29$
2DS-d	$F_{(2,57)} = 22.94$, $p < 0.01$, $\omega = 0.64$
2DS-c	$F_{(2,57)} = 91.42$, $p < 0.01$, $\omega = 0.86$
DS&EX-c	$F_{(2,57)} = 23.49$, $p < 0.01$, $\omega = 0.65$
VAS	$F_{(2,57)} = 85.6$, $p < 0.01$, $\omega = 0.85$
RT	$F_{(2,57)} = 9.87$, $p < 0.01$, $\omega = 0.47$
RT DS	$F_{(2,57)} = 15.22$, $p < 0.01$, $\omega = 0.56$
RT EX	$F_{(2,57)} = 4.52$, $p < 0.05$, $\omega = 0.32$
RT 2DS-d	$F_{(2,57)} = 9.24$, $p < 0.01$, $\omega = 0.46$
RT 2DS-c	$F_{(2,57)} = 11.3$, $p < 0.01$, $\omega = 0.50$
RT DS&EX-d	$F_{(2,57)} = 7.06$, $p < 0.01$, $\omega = 0.41$
RT DS&EX-c	$F_{(2,57)} = 7.50$, $p < 0.01$, $\omega = 0.42$
RT 2EX-d	$F_{(2,57)} = 4.29$, $p < 0.05$, $\omega = 0.31$
RT 2EX-c	$F_{(2,57)} = 6.06$, $p < 0.01$, $\omega = 0.38$
RT VAS	$F_{(2,57)} = 7.35$, $p < 0.01$, $\omega = 0.41$
MD-time	$F_{(2,57)} = 14.68$, $p < 0.01$, $\omega = 0.55$
MD-time DS	$F_{(2,57)} = 18.25$, $p < 0.01$, $\omega = 0.60$
MD-time EX	$F_{(2,57)} = 3.25$, $p < 0.05$, $\omega = 0.26$
MD-time 2DS-d	$F_{(2,57)} = 11.68$, $p < 0.01$, $\omega = 0.51$
MD-time 2DS-c	$F_{(2,57)} = 14.47$, $p < 0.01$, $\omega = 0.55$
MD-time DS&EX-d	$F_{(2,57)} = 9.04$, $p < 0.01$, $\omega = 0.45$
MD-time DS&EX-c	$F_{(2,57)} = 9.2$, $p < 0.01$, $\omega = 0.46$
MD-time 2EX-d	$F_{(2,57)} = 4.27$, $p < 0.05$, $\omega = 0.31$
MD-time 2EX-c	$F_{(2,57)} = 12.61$, $p < 0.01$, $\omega = 0.52$
MD-time VAS	$F_{(2,57)} = 12.62$, $p < 0.01$, $\omega = 0.52$
v_x	$F_{(2,57)} = 3.85$, $p < 0.05$, $\omega = 0.29$
v_y	$F_{(2,57)} = 4.06$, $p < 0.05$, $\omega = 0.30$

TABLE 4 | Differences between truth-tellers and liars, liars and depressed, truth-tellers and depressed.

Feature	Difference between groups	Tukey test, *t*-value, *p*-value
TRUTH-TELLERS vs. LIARS		
DS	−7.75	$t = -13.38$, $p <01$
2DS-d	5.25	$t = 6.73$, $p < 0.01$
2DS-c	−11.85	$t = -13.46$, $p < 0.01$
DS&EX-c	−2.40	$t = -6.17$, $p < 0.01$
VAS	−6.80	$t = -13.02$, $p < 0.01$
v_x	0.00053	$t = 2.59$, $p < 0.05$
LIARS vs. DEPRESSED		
DS	2.45	$t = 4.23$, $p < 0.01$
2DS-d	−2.15	$t = -2.75$, $p < 0.05$
2DS-c	5.00	$t = 5.68$, $p < 0.01$
VAS	2.85	$t = 5.46$, $p < 0.01$
RT	−2611.21	$t = -3.84$, $p < 0.01$
RT DS	−2036.49	$t = -4.79$, $p < 0.01$
RT 2DS-d	−3507.4	$t = -3.89$, $p < 0.01$
RT 2DS-c	−2880.60	$t = -4.14$, $p < 0.01$
RT DS&EX-c	−5171.2	$t = -3.60$, $p < 0.01$
RT 2EX-d	−2708.4	$t = -2.79$, $p < 0.05$
RT 2EX-c	−1403.2	$t = -2.77$, $p < 0.05$
RT VAS	−2012.3	$t = -3.14$, $p < 0.01$
MD-time	−1902.21	$t = -4.80$, $p < 0.01$
MD-time DS	−1182.22	$t = -5.40$, $p < 0.01$
MD-time 2DS-d	−2494.9	$t = -4.39$, $p < 0.01$
MD-time 2DS-c	−2076.4	$t = -4.81$, $p < 0.01$
MD-time DS&EX-d	−1741.8	$t = -2.97$, $p < 0.05$
MD-time DS&EX-c	−4187.4	$t = -3.92$, $p < 0.01$
MD-TIME 2EX-d	−1968.1	$t = -2.88$, $p < 0.05$
MD-TIME 2EX-c	−13.45.5	$t = -4.54$, $p < 0.01$
MD-TIME VAS	−1514.67	$t = -4.35$, $p < 0.01$
v_y	0.00013	$t = 2.56$, $p < 0.05$
TRUTH-TELLERS vs. DEPRESSED		
DS	−5.30	$t = -9.15$, $p < 0.01$
EX	0.45	$t = 2.72$, $p < 0.05$
2DS-d	3.10	$t = 3.97$, $p < 0.01$
2DS-c	−6.85	$t = -7.78$, $p < 0.01$
DS&EX-c	−2.20	$t = -5.66$, $p < 0.01$
VAS	−3.95	$t = -7.56$, $p < 0.01$
RT	−2623.02	$t = -3.85$, $p < 0.01$
RT DS	−2019.08	$t = -4.75$, $p < 0.01$
RT EX	−1027.1	$t = -3.01$, $p < 0.05$
RT 2DS-d	−3183.4	$t = -3.53$, $p < 0.01$
RT 2DS-c	−2836.75	$t = -4.08$, $p < 0.01$
RT DS&EX-d	−3480	$t = -3.70$, $p < 0.01$
RT DS&EX-c	−4350.8	$t = -3.30$, $p < 0.05$
RT 2EX-c	−1621.3	$t = -3.20$, $p < 0.01$
RT VAS	−2225.7	$t = -3.47$, $p < 0.01$
MD-time	−1807.48	$t = -4.56$, $p < 0.01$
MD-time DS	−1100.85	$t = -5.03$, $p < 0.01$
MD-time EX	−560	$t = -2.50$, $p < 0.05$
MD-time 2DS–d	−2240.2	$t = -3.94$, $p < 0.01$
MD-time 2DS-c	−1937.7	$t = -4.491$, $p < 0.01$
MD-time DS&EX-d	−2407.9	$t = -4.11$, $p < 0.01$
MD-time DS&EX-c	−3742.3	$t = -3.50$, $p < 0.01$
MD-time 2EX-c	−1219.7	$t = -4.12$, $p < 0.01$
MD-time VAS	−1510.78	$t = -4.34$, $p < 0.01$

The table reports t-value, p-value and effect-size resulting from Tukey test and the value of the difference between the compared groups. Only the results that reached the statistical significance are reported.

TABLE 5 | The table reports the 6 features resulted from the features selection.

Feature	Ranked attributes
DS	0.55
2DS-c	0.52
2DS-d	0.41
VAS	0.52
MD-time 2DS-d	0.36
MD-time VAS	0.37

The second column reports the value of the correlation between the feature and the dependent variable (truth-teller vs. liar vs. depressed).

Multivariate Analysis: Features Selection

In order to select the variables to be entered in machine-learning models, a features selection was performed using WEKA 3.9 (38). Features selection is a widely used procedure in machine learning that allows the removal of redundant and irrelevant features and an increase of model generalization by reducing overfitting (39). In the current paper, a correlation-based feature selector (CFS) was used to reduce the number of features (40). This algorithm selects the independent variables with the maximum correlation with the dependent variable (truth-teller vs liar vs depressed) and the minimum correlation across independent variables (the 83 features), using greedy stepwise as search method. The features selected by the CFS are the following: the number of very atypical symptoms (VAS) reported by each participant, the number of depressive symptoms reported by each participant in simple questions (DS), the number of symptoms reported by the participants when they responds to 2DS-c and 2DS-d questions (i.e., complex questions, concordant or discordant, about depressive symptoms), the time needed to reach the point of maximum deviation in 2DS-d questions (MD-time 2DS-d) and the time needed to reach the point of maximum deviation in questions about very atypical symptoms (MD-time VAS). The selected features are reported in **Table 5**.

Multivariate Analysis: Machine-Learning Models

The six features mentioned above were entered in different machine-learning (ML) classifiers. Particularly, we selected four different classifiers that differ for the classification strategy (41–44): Naive Bayes, Sequential Minimal Optimization (SMO), Logistic Model Tree (LMT) and Random Forest (RF). For each classifier, a three-class classification (as the model is required to classify depressed patients, liars and truth-tellers) was run using a 10-fold cross-validation procedure, as implemented in WEKA 3.9 (38). In 10-fold cross-validation, the sample of 60 participants is randomly partitioned into 10 equal size subsamples ($n = 6$). Of the 10 subsamples, a single subsample is retained as the validation set for testing the model, and the remaining 9 subsamples are used as training sets. The cross-validation process is recursively repeated 10 times, each time with one of the 10 subsamples used as a validation set. The 10 results from the folds are finally averaged to produce a single classification accuracy estimation. The classification accuracies

TABLE 6 | Accuracies obtained by four different ML classifiers in 10-fold cross-validation and in test set.

Classifier	Accuracy in 10-fold-cross validation (n = 60) (%)	Accuracy in test set (n = 27) (%)
Naïve Bayes	90	96.3
SMO	83.3	92.6
LMT	81.6	96.3
Random Forest	80	92.6

Classifiers are Naïve Bayes, Sequential Minimal Optimization (SMO), Logistic Model Tree (LMT) and Random Forest.

TABLE 7 | Classification accuracies of liars and depressed participants by four different ML classifiers in 10-fold cross-validation and in test set.

Classifier	Accuracy in 10-fold-cross validation (n = 60) (%)	Accuracy in test set (n = 28) (%)
Naïve Bayes	80	94.4
SMO	82.5	88.9
LMT	80	88.9
Random Forest	87.5	94.4

Classifiers are Naïve Bayes, Sequential Minimal Optimization (SMO), Logistic Model Tree (LMT) and Random Forest.

obtained by the four classifiers in 10-fold cross-validation are reported in **Table 6**. All the classifiers achieved an accuracy ranging from 80 to 90%. In a subsequent step, the four algorithms (ML models) were tested on the group 2 (test = 27 participants) to verify the generalization of the results on an independent sample of participants This allowed us to demonstrate that all the models have a good generalization, as the classification accuracies remain stable at over 90%. The ML results on both the training (n = 60) and test (n = 27) sets are represented in **Table 6**.

However, in real world settings, the examiner is required to successfully distinguish malingering from real depression. Thus, the classifications were repeated, entering only liars and depressed participants as classes. In other words, we made a two-class classification and ignored truth-teller participants. The features were selected using the same method described above and then entered in the four classifiers (ML models). The selected features were the following: DS, 2DS-d, 2DS-c, VAS, IT 2EX-d, IT 2EX-c, RT 2EX-c, MD-time, MD-time DS, MD-time 2DS-d, MD-time DS&EX-c, MD-time VAS, a_y, and y-flip DS. Classification accuracies, which were revealed to be stable around 90%, are reported in **Table 7**.

As ML models are difficult to interpret; a decision tree model has been run (45). The decision tree model gives a more simple idea about the hypothetical decision rules, on which the classifications results are based. This is one of the simplest— if not *the* simplest—classifier in terms of transparency of the operations computed by the algorithm, and it permits easy highlighting of the classification logic [even if it is not the most efficient method; (46)]. The structure of the tree is reported in **Figure 3**. This model is basically built on two rules. The first rule takes into account the number of symptoms declared by the subject in the 2DS-c questions. If the participant reports fewer than 3.5 symptoms, he/she is classified as a truth-teller; else, the second rule is considered. According to the second rule, if the subject takes, on average, more than 4,048 ms to compute the response to the 2DS-d questions, he/she is either a depressed patient or a liar. This simple algorithm reaches an accuracy of 75% in the training group (correctly identifying 51 subjects out of 60), generalizing with an accuracy of 85.2% in the test group.

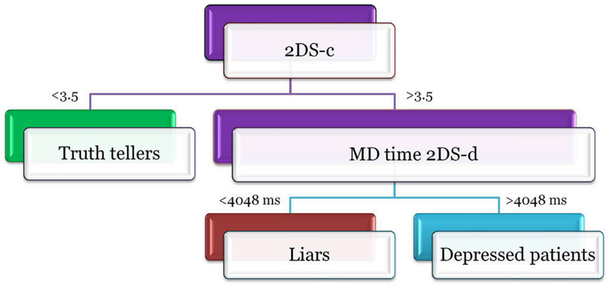


FIGURE 3 | The figure reports the structure of the decision tree model. Participants declaring <3.5 symtomps on 2DS-c questions are classified as truth-tellers. Participants declaring more than 3.5 symtomps on 2DS-c questions are depressed patients if and only if they take more than 4048 ms to compute the response to the 2DS-d questions, otherwise they are classified as liars.

Multivariate Analysis: Alternative Models

One of the most discussed topics in lie detection concerns the resistance to countermeasures (47). If the participant knows how lie detectors work, he/she may enact a series of strategies to reduce its efficacy. For example, an alteration of RTs is enough to beat aIAT or CIT (48, 49), as this is the only parameter on which they are based. In order to prevent countermeasures, the kinematic analysis of mouse movements offers a significant advantage: it is not based simply on RTs but on numerous and articulated parameters that, together, contribute to determine the truthfulness of the subject's response (26). In other words, it would be very difficult for the participants to alter all of the parameters at the same time and keep them under control. Moreover, the large number of features allows the building of alternative classification models. In this way, the examinee cannot know in advance which features are entered in the prediction model and, accordingly, which are the features to keep under control during the test. To fix this point, we developed two alternative machine-learning models, entering in the classifiers a subsets of predictors different from those above used. The six features selected above are the best to optimize the classifier's performance. However, other sub-optimal sets of features can work well in the classification. A first set of alternative predictors contained the five features most correlated to the dependent variable: DS (r = 0.55), 2DS-c (r = 0.53), VAS (r = 0.52), DS&EX-c (r = 0.45), MD-time DS (r = 0.42). A second subset of predictors included only features related to complex questions about depressive

TABLE 8 | Classification of participants using two set of alternative predictors.

Classifier	Accuracy in 10-fold-cross validation (n = 60) (%)	Accuracy in test set (n = 27) (%)
SUBSET OF PREDICTORS 1		
Naïve Bayes	85	88.9
SMO	83.3	88.9
LMT	80	92.6
Random Forest	80	88.9
SUBSET OF PREDICTORS 2		
Naïve Bayes	81.6	96.3
SMO	78.3	92.6
LMT	76.7	85.2
Random Forest	75	92.6

Alternative models were computed using four different ML classifiers (Naïve Bayes, SMO, LMT, Random Forest). Accuracies in 10-fold cross-validation and in test set are reported.

symptoms (2DS-c and 2DS-d), which are the stimuli aimed to increase liars' cognitive load: 2DS-c, 2DS-d, IT 2DS-c, IT 2DS-d, MD-time 2DS-c, MD-time 2DS-d, RT 2DS-c, RT 2DS-d, MD 2DS-c, MD 2DS-d, AUC 2DS-c, AUC 2DS-d, x-flip 2DS-c, x-flip 2DS-d, y-flip 2DS-c, and y-flip 2DS-d. The results obtained from the alternative models are reported in **Table 8**. It can be noticed that the accuracies remain stable at around 90%, supporting the reliability of this method for the identification of liars.

DISCUSSION AND CONCLUSION

The present study investigated the accuracy of a new deception detection technique in the correct identification of participants who malingered depressive symptoms. To this aim, a tool based on mouse tracking was used while participants were required to answer simple or complex questions concerning both symptoms of depression and the experimental situation.

The main result is striking: the individuals who malingered depressive symptoms were correctly identified by the algorithm with an accuracy of up to 96%. In addition, the current study also underlined that: (i) the mouse trajectory of the liars visually clearly differed from the ones of the truth-tellers (regardless of whether the latter were really depressed); (ii) the group of individuals that malingered depression reported a higher number of depressive and non-depressive symptoms; (iii) the ML classifiers recognized the complex questions within the key features for a correct classification and (iv) liars are also faster than the really depressed subjects—but slower than the healthy truth-tellers—to perform the mouse-based task, as the algorithm identified, as a critical variable for the discrimination, the time to reach the point of maximum deviation during the mouse response.

Importantly, similar results were obtained testing four different ML models (Naïve Bayes, SMO, LTM, Random Forest). This denotes that the results are not highly dependent on the selected algorithm. Furthermore, the main results are obtained using highly selected features, raising the suspicion that they cannot be generalized using different features. This is of outstanding importance, as the number of depressive symptoms (DS) and the number of very atypical symptoms (VAS) were included in the feature selection within the main analysis. Because both DS and VAS could be obtained using simpler tests, as, for instance, the M-test explained in the introduction, one may wonder about the advantage of using the current mouse tracking techniques and whether the current results remain stable even if DS and VAS were removed from the features used for the classification. Critically, these concerns were dampened by the results obtained using alternative ML models, which includes only one (DS) or none of these features within the features selected for the classification. In addition, DS has not been used alone but within the complex sentences. As these alternative models achieved very high classification accuracies as well, this rules out the hypotheses that the current results were driven by the selected features and also sustains the hypothesis that the proposed tools are not easily fooled by coaching. Indeed, the high number of parameters that could be considered to build up the best classifiers and the great variability in the features that could be selected by each classifier makes the new tool ideally suited to be almost impossible to be deceived. Thus, the results reported in the current paper are robust to the ML method and feature selection changes.

It is also worth noting that the tool is based on both mouse tracking movements and the technique of unexpected questions. Contrarily to previously used tests (for instance, the *M*-test), the current algorithm is able to detect the number of symptoms reported by each individual only relying on the use of the complex sentences (2DS-c, 2DS-d), as revealed by the alternative model, and thus excludes potential features that are more easy to be faked, as, for instance, the number of symptoms (DS) and the number of atypical symptoms (VAS).

Concerning the number of symptoms, it should be noted that malingering participants tend to report most of the symptoms which are presented during the task, both typical symptoms of depression (DS) and atypical symptoms (VAS) characterizing other mental disorders. In other words, liars reported being affected by a higher number of psychiatric symptoms than those people who were genuinely depressed. This result is line with literature reporting that the qualitative and quantitative analysis of symptom characteristics is a crucial method to identify simulators (16). It is well known that malingering is often characterized by a positive response to suggested symptoms and a tendency to endorse many symptoms indiscriminately (50). Indeed, malingerers believe that endorsing a symptom will increase the appearance of psychopathology and that more symptoms will be construed as a more severe disorder. On the other hand, genuine patients report only the symptoms that they are really experiencing, resulting in a lower number and more-common symptoms. For this reason, common strategies to detect a malingered response pattern consist in verifying the endorsement of rare and improbable symptoms [e.g., this is how the SIMS works; (18)] or the over-endorsement of symptoms.

The second important piece of evidence concerns the mouse dynamics features. As emerged from the univariate analyses and the algorithm's features selection, the most significant differences between the three experimental groups are in the time to compute the response (RT) and the time to reach the point of maximum deviation (MD-time). In more detail, depressed subjects take more time to respond than the subjects of the other two experimental conditions (liars and truth-tellers) for both simple and complex questions. This result reinforces the evidence available in literature that depression is characterized by psychomotor and ideomotor retardation [diminished ability to think or concentrate, or indecisiveness; (1)]. In other words, depressed people are differentiated from liars not on the basis of the cognitive load, which is higher in liars than in depressed people (who are responding truthfully), but on the basis of the psychomotor retardation, which is a key feature of truly depressed individuals. On the contrary, according to lie-detection literature, liars are slower than healthy truth-tellers, as the greater cognitive load due to the act of lie results in more time needed to compute the response (51).

It is important here to underline that, despite the fact that in literature, it is already known that individuals that feign depression usually report a higher number of symptoms compared to really depressed patients (50) and despite it is already known that lying takes time (28) and thus that liars are, in general, consistently slower than truth-tellers, this study enriches the literature by providing an automatic algorithm that allows combination of the two pieces of information. Critically, this enabled the identification of three different profiles: the non-depressed truth-tellers are characterized by a low number of reported symptoms and by quick answers; the depressed truth-tellers are characterized by a good number of reported symptoms and are very slow in answering and the liars are characterized by a very high number of reported symptoms, and their reaction times are slower than those of the non-depressed truth-tellers but quicker than those of depressed patients.

Finally, two drawbacks are worth highlighting. First, in this paper, the clinician and the machine-learning algorithm performance in detecting malingering has not been compared. Indeed, individuals were selected if they already had a diagnosis of depression. In addition, healthy participants assigned to the truth-tellers or liars groups never underwent a psychiatric examination but were screened using a self-report questionnaire. Thus, we cannot draw definitive conclusions on the superiority of machine learning compared to clinical assessment in detecting the malingering of depression. To date, we can only hypothesize the superiority of machine learning based on previous literature (14, 15). Secondly, eventual cognitive difficulties in patients with depression have not been taken into account. Thus, particular attention should be given to the application of this tool to patients with cognitive disabilities. Those patients could show difficulties both in processing the meaning of the complex questions and in giving the response (they are more likely to drag the mouse while attempting to make a decision, causing distorted trajectories and longer reaction times), obtaining a worse performance than liars. Therefore, the examiner should take into account that the cognitive functioning of the examinee could influence the task performance and thus alter the classifier result. Thus, further studies are needed before this algorithm could be applied to a real-world forensic setting. In particular, this study highlights an urgent need to compare the performance of clinicians and machine learning in detecting malingering, taking into consideration the cognitive difficulties the real patients might be suffering.

In conclusion, we provided evidence that the current algorithm, through an accurate feature selection procedure, can accurately identify up to 96% of the liars. This methodology, compared with the ones currently available in the literature and described in the introduction, has the following advantages: first, this tool is not possible to be cheated on, as there are too many parameters that are taken into consideration; secondly, specialized clinicians are not required to administer it and interpret the results, thus enhancing the possibility of wide use, such as by insurers; thirdly, a single test could be sufficient to understand whether or not an individual is malingering a multifaceted disorder like depression. On the contrary, previous instruments for detecting deception, such as aIAT and CIT, allowed the investigation of a single symptom instead of the disorder itself. Despite the fact that this mouse tracker-based tool has been developed and tested to identify individuals who feigned depression, the same technique could be potentially adapted to allow a wider use and generalization to other psychiatric disorders such as anxiety disorders and PTSD or physical disturbances such as whiplash. Before the translational application to real world forensic setting, further studies are needed to compare the performance of the machine-learning algorithm with the performance of clinicians in detecting malingering.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The ethics committee for psychological research of the University of Padova approved the experimental procedure (Unique Number: 276B8771D4B0F6FDC748E0ABE46D460C). All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AVAILABILITY OF DATA AND MATERIALS

The dataset used and analyzed during the current study is available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

MM, GS, and CS: Conceived the experiment; MM and AT: Designed the experimental task; MM and AT: Healthy subjects data acquisition; AT, SF, GT, MV, PD, GP, and TM: Depressed patients data acquisition; MM and GS: Data analysis; MM, GS, and CS: Data interpretation; MM, GS, and CS: Drafting of the manuscript. All the authors revised the manuscript critically and gave the final approval of the version to be published.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2018.00249/full#supplementary-material

REFERENCES

1. American Psychiatric Association. *DSM V. Diagnostic and Statistical Manual of Mental Disorders*. Arlington: American Psychiatric Publishing. (2013).
2. Cuijpers P, Smit F. Subclinical depression: a clinically relevant condition? *Tijdschr Psychiatr.* (2008) **50**:519–528.
3. Druss BG, Rosenheck RA, Sledge WH. Health and disability costs of depressive illness in a major U.S. corporation. *Am J Psychiatry* (2000) **157**:1274–8. doi: 10.1176/appi.ajp.157.8.1274
4. Ferrari RM. *Breve Manuale di Invalidità Civile*. (2016). Available Online at: http://www.raggiungere.it/attachments/article/387/Breve%20Manuale%20di%20Invalidit%C3%A0%20Civile.pdf
5. Andreani A. Tabella danno biologico di lieve entità. (2017). Available Online at: https://www.avvocatoandreani.it/servizi/calcolo_danno_biologico.php
6. Hayes J, Grieve R. Faked depression: comparing malingering via the internet, pen-and-paper, and telephone administration modes. *Telemed E Health* (2013) **19**:714–6. doi: 10.1089/tmj.2012.0278
7. Sullivan K, King J. Detecting faked psychopathology: a comparison of two tests to detect malingered psychopathology using a simulation design. *Psychiatry Res.* (2010) **176**:75–81. doi: 10.1016/j.psychres.2008.07.013
8. Rogers R. (2008). *Clinical Assessment of Malingering and Deception*. ed R. Rogers. Guilford Press.
9. Greve KW, Ord JS, Bianchini KJ, Curtis KL. Prevalence of malingering in patients with chronic pain referred for psychologic evaluation in a medico-legal context. *Arch Phys Med Rehabil.* (2009) **90**:1117–26. doi: 10.1016/j.apmr.2009.01.018
10. Mittenberg W, Patton C, Canyock EM, Condit DC. Base rates of malingering and symptom exaggeration. *J Clin Exp Neuropsychol.* (2002) **24**:1094–1102. doi: 10.1076/jcen.24.8.1094.8379
11. Young G. Psychological injury and law malingering in forensic disability-related assessments: prevalence 15 ± 15%. *Psychol Inj Law* (2015) **8**:188–199. doi: 10.1007/s12207-015-9232-4
12. Adetunji BA, Basil B, Mathews M, Williams A, Osinowo T, Oladinni O. Detection and management of malingering in a clinical setting. *Prim psychiatry* (2006) **13**:61–69.
13. Ferrara SD, Ananian V, Baccino E, Boscolo–Berto R, Domenici R, Hernàndez-Cueto C, et al. A novel methodology for the objective ascertainment of psychic and existential damage. *Int J Legal Med.* (2016) **130**:1387–99. doi: 10.1007/s00414-016-1366-8
14. Rosenhan D. On being sane in insane places. *Science* (1973) **179**:250–8. doi: 10.1126/science.179.4070.250
15. Rosen J, Mulsant BH, Bruce ML, Mittal V, Fox D. Actors' portrayals of depression to test interrater reliability in clinical trials. *Am J Psychiatry* (2004) **161**:1909–11. doi: 10.1176/ajp.161.10.1909
16. Sartori G, Orrù G, Zangrossi A. Detection of Malingering in personal injury and damage ascertainment. In: Ferrara SD, Boscolo-Berto R, Viel G, editors. *Personal Injury and Damage Ascertainment under Civil Law*. Springer (2016). pp. 547–58.
17. Beaber RJ, Marston A, Michelli J, Mills MJ. A brief test for measuring malingering in schizophrenic individuals. *Am J Psychiatry* (1985) **142**:1478–81. doi: 10.1176/ajp.142.12.1478
18. Smith GP, Burger GK. Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). *J Am Acad Psychiatry Law* (1997) **25**:183–9.
19. Storm J, Graham JR. Detection of coached general malingering on the MMPI−2. *Psychol. Assess.* (2000) **12**:158–165. doi: 10.1037/1040-3590.12.2.158
20. Agosta S, Sartori G. The autobiographical IAT: a review. *Front. Psychol.* (2013) **4**:519. doi: 10.3389/fpsyg.2013.00519
21. Allen JJB. Clinical applications of the Concealed Information Test. In: Verschuere B, Ben-Shakhar G, Meijer E. editors. *Memory Detection. Theory and Application of the Concealed Information Test.* Cambridge: Cambridge University Press (2011). pp. 231–252.
22. Sartori G, Agosta S, Gnoato F. High accuracy detection of malingered whiplash syndrome. In: *International Whiplash Trauma Congress.* (Miami, FL) (2007).
23. Duran ND, Dale R, McNamara DS. The action dynamics of overcoming the truth. *Psychon Bull Rev.* (2010) **17**:486–491. doi: 10.3758/PBR.17.4.486
24. Monaro M, Fugazza FI, Gamberini L, Sartori G. How human-mouse interaction can accurately detect faked responses about identity. In: Gamberini L, Spagnolli A, Jacucci G, Blankertz B, Freeman J editors. *Symbiotic Interaction. Symbiotic 2016. Lecture Notes in Computer Science,* Vol 9961. Cham: Springer. (2017). pp. 115–124.
25. Monaro M, Gamberini L, Sartori G. Identity verification using a kinematic memory detection technique. In: Hale K, Stanney K editors, *Advances in Neuroergonomics and Cognitive Engineering. Advances in Intelligent Systems and Computing,* Vol. 488. Cham: Springer (2017). pp. 123–132.
26. Monaro M, Gamberini L, Sartori G. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE* (2017) **12**:e0177851. doi: 10.1371/journal.pone.0177851
27. Freeman JB, Dale R, Farmer TA. Hand in motion reveals mind in motion. *Front Psychol.* (2011) **2**:59. doi: 10.3389/fpsyg.2011.00059
28. Suchotzki K, Verschuere B, Van Bockstaele B, Ben-Shakhar G, Crombez G. Lying takes time: a meta-analysis on reaction time measures of deception. *Psychol. Bull.* (2017) **143**:428–453. doi: 10.1037/bul0000087
29. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch Gen Psychiatry* (1961) **4**:561–571. doi: 10.1001/archpsyc.1961.01710120031004
30. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. The reusable holdout: preserving validity in adaptive data analysis. *Science* (2015) **349**:3–6. doi: 10.1126/science.aaa9375
31. Monaro M, Gamberini L, Zecchinato F, Sartori G. False identity detection using complex sentences. *Front Psychol.* (2018) **9**:283. doi: 10.3389/fpsyg.2018.00283
32. Bertolotti G, Michielin P, Vidotto G, Zotti AM, Sanavio E. Depression questionnaire (DQ). In: Nezu AM, Ronan GF, Meadows EA, McKlure KS, editors. *Practitioner's Guide to Empirical Based Measures of Depression*. Norwell, MA: Kluwer Academic; Plenum Publishers (2000). pp. 45–47.
33. Fagiolini A, Dell'osso L, Pini S, Armani A, Bouanani S, Rucci P, et al. Validity and reliability of a new instrument for assessing mood symptomatology: the Structured Clinical Interview for Mood Spectrum (SCI-MOODS). *Int J Methods Psychiatr Res.* (1999) **8**:71–82.
34. Walczyk JJ, Igou FP, Dixon AP, Tcholakian T. Advancing lie detection by inducing cognitive load on liars: a review of relevant theories and techniques guided by lessons from polygraph-Based approaches. *Front. Psychol.* (2013) **4**:14. doi: 10.3389/fpsyg.2013.00014
35. Vrij A, Leal S, Granhag PA, Mann S, Fisher RP, Hillman J, et al. Outsmarting the liars: the benefit of asking unanticipated questions. *Law Hum. Behav.* (2009) **33**:159–166. doi: 10.1007/s10979-008-9143-y

36. Williams EJ, Bott LA, Patrick J, Lewis MB. Telling lies: the irrepressible truth? *PLoS ONE* (2013) **8**:e60713. doi: 10.1371/journal.pone.0060713
37. Freeman JB, Ambady N. MouseTracker: software for studying real-time mouse-tracking method. *Behav Res Methods* (2010) **42**:226–241. doi: 10.3758/BRM.42.1.226
38. Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.* (2009) **11**:10–18. doi: 10.1145/1656274.1656278
39. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep.* (2015) **5**:1–12. doi: 10.1038/srep10312
40. Hall MA. (1999). *Correlation-based Feature Selection for Machine Learning*. The University of Waikato, Hamilton.
41. Breiman L. Random forest. *Mach. Learn.* (2001) **45**:5–32. doi: 10.1023/A:1010933404324
42. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proceeding of the 11th Conference on Uncertainty in Artificial Intelligence*. San Mateo, CA (1995). pp. 338–345.
43. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to platt's SMO algorithm for SVM classifier design. *Neural Comput.* (2001) **13**:637–649. doi: 10.1162/089976601300014493
44. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn.* (2005) **95**:161–205. doi: 10.1007/s10994-005-0466-3
45. Quinlan JS. (1993). *C4.5*: *Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.
46. Mitchell T. Decision tree learning. In: *Machine Learning*. Mitchell T, editor. New York, NY: McGraw Hill (1997). pp. 52–78.
47. Bowman H, Filetti M, Alsufyani A, Janssen D, Su L. Countering countermeasures: detecting identity lies by detecting conscious breakthrough. *PLoS ONE* (2014) **9**:e90595. doi: 10.1371/journal.pone.0090595
48. Agosta S, Ghirardi V, Zogmaister C, Castiello U, Sartori G. Detecting fakers of the autobiographical IAT. *Appl Cogn Psychol.* (2010) **25**:299–306. doi: 10.1002/acp.1691
49. Peth J, Suchotzki K, Matthias G. Influence of countermeasures on the validity of the concealed information test. *Psychophysiology* (2016) **53**: 1429–40. doi: 10.1111/psyp.12690
50. Conroy MA, Kwartner PP. Malingering. *Appl Psychol Crim Justice* (2006) **2**:29–51.
51. Monaro M, Galante C, Spolaor R, Li QQ, Gamberini L, Conti M, et al. Covert lie detection using keyboard dynamics. *Sci Rep.* (2018) **8**:1976. doi: 10.1038/s41598-018-20462-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AH and handling Editor declared their shared affiliation.

frontiers in Psychiatry

REVIEW
published: 09 October 2018
doi: 10.3389/fpsyt.2018.00492

Verbal Deception and the Model Statement as a Lie Detection Tool

Aldert Vrij[1]*, *Sharon Leal*[1] *and Ronald P. Fisher*[2]

[1] *Department of Psychology, University of Portsmouth, Portsmouth, United Kingdom,* [2] *Department of Psychology, Florida International University, Miami, FL, United States*

We have been reliably informed by practitioners that police officers and intelligence officers across the world have started to use the Model Statement lie detection technique. In this article we introduce this technique. We describe why it works, report the empirical evidence that it works, and outline how to use it. Research examining the Model Statement only started recently and more research is required. We give suggestions for future research with the technique. The Model Statement technique is one of many recently developed verbal lie detection methods. We start this article with a short overview of the—in our view- most promising recent developments in verbal lie detection before turning our attention to the Model Statement technique.

Keywords: deception, interview, model statement, encouraging interviewees to say more, lie detection

Edited by:
Giuseppe Sartori,
Università degli Studi di Padova, Italy

Reviewed by:
Konstantinos Tasios,
National Health System, Greece
Antonietta Curci,
Università degli Studi di Bari, Italy

****Correspondence:***
Aldert Vrij
aldert.vrij@port.ac.uk

Specialty section:
This article was submitted to Forensic Psychiatry, a section of the journal Frontiers in Psychiatry

Received: *23 July 2018*
Accepted: *20 September 2018*
Published: *09 October 2018*

Citation:
Vrij A, Leal S and Fisher RP (2018) Verbal Deception and the Model Statement as a Lie Detection Tool. Front. Psychiatry 9:492. doi: 10.3389/fpsyt.2018.00492

VERBAL LIE DETECTION

DePaulo et al.'s (1) comprehensive meta-analysis of nonverbal and verbal cues to deception showed that such cues are generally weak and unreliable (2). Research has also suggested that this applies more to nonverbal cues than to verbal cues to deception: A meta-analysis about observers' ability to detect deceit when observing nonverbal and verbal cues to deception showed that when observers could only see the target person, they performed worse (52% accuracy) than when they could only hear the target person (63%) (3). This relative weakness of nonverbal cues to deceit could at least in part be explained when taking into account the strategies truth tellers and liars use when attempting to make a convincing impression on others. Truth tellers and liars employ similar strategies regarding nonverbal behavior: Both try to suppress signs of nervousness and attempt to replace them with signs that will create the impression of being honest, such as looking conversation partners into their eyes and avoiding fidgeting (scratching head, wrists etc.) (4, 5). In contrast, truth tellers and liars use different strategies regarding verbal behavior. Truth tellers are forthcoming and employ a "tell it all" strategy, whereas liars employ a "keep it simple" strategy and avoid mentioning incriminating details (6, 7). As a consequence, truthful stories often include more details than deceptive stories (1, 8).

Because researchers found nonverbal cues to be ineffective to detect deception, they refocused their efforts to focus on verbal cues. Particularly, they have tried to elicit or enhance verbal cues through specific interview techniques that exploit the different verbal strategies that truth tellers and liars employ (9). In our view, four of these efforts have shown the best results or the best potential in terms of lie detection (10, 11): (a) The Strategic Use of Evidence, (b) Assessment Criteria Indicative of Deception, (c) the Verifiability Approach, and (d) Cognitive Credibility Assessment, to which the Model Statement technique belongs. We outline these approaches briefly and refer to Vrij (10, 11) and Vrij and Fisher (12) for further details.

Strategic Use of Evidence (SUE)

The aim of the SUE technique is to exploit the different strategies truth tellers and liars employ in interviews, particularly the difference in between forthcoming (truth tellers) and avoiding mentioning incriminating details (liars) (6, 13). In a SUE interview, the investigator asks questions related to the evidence s/he possesses without making the interviewee aware of possessing this evidence (i.e., asking about someone's whereabouts without revealing that CCTV footage showed that the suspect was in a shopping mall where a robbery took place). This typically leads to truth tellers' accounts being more consistent with the available evidence than liars' accounts (14). In addition, during an interview liars sometimes start to realize that the interviewer may have some evidence against them (i.e., CCTV footage about being in the shopping mall). Liars then often change their statement and provide an innocent explanation for the evidence (i.e., admitting for the first time to have been in the shopping mall, but not admitting to have been in the shop where the robbery took place). Such changes in liars' stories are called within-statements inconsistencies and liars show more of them than truth tellers (14).

Assessment Criteria Indicative of Deception (ACID)

The ACID interview procedure is based on the Cognitive Interview, a well-established protocol to elicit more information from cooperative witnesses through enhancing three processes: Social dynamics, memory/cognition and communication (15). In ACID, truth tellers and liars provide an initial free recall followed by instructions that stimulate communication and aid memory (16). An example of communication stimulation used in ACID is transfer of control to the respondent, and three examples of memory aids used in ACID are mental reinstatement of context, recall from another person's perspective, and reverse-order recall. ACID research has shown that, amongst other findings, truth tellers report more additional information after the initial free recall than liars (16, 17).

Verifiability Approach (VA)

The VA is based on the idea that liars face a dilemma. On the one hand, liars prefer to provide many details. This makes sense because the more details someone provides, the more likely it is that s/he will be believed (18, 19). On the other hand, liars do not wish to mention too many details. The more details they provide, the more opportunity they give to investigators to check these details and to discover their lies (19). A strategy that incorporates both seemingly conflicting goals is to provide details that cannot be verified (20). Indeed, research has shown that truth tellers typically report more details that can be checked than liars (21). Checkable details are activities that someone claims to have carried out or was witnessed by a named person, or activities that someone claimed was recorded on CCTV. In addition, activities that leave a trace (mobile phone call, text, debit/credit card purchases, and receipts) are also considered checkable. The effect that truth tellers report more checkable details than liars becomes stronger when interviewees are instructed to try to include details in their statement that the investigator can verify. Following such a request, truth tellers add more checkable details in their accounts than liars (22, 23).

Cognitive Credibility Assessment (CCA)

The CCA technique comprises three elements: (i) Imposing cognitive load; (ii) Asking unexpected questions, and (iii) Encouraging interviewees to say more (24, 25).

(i) Cognitive credibility assessment: Imposing cognitive load. fMRI research has shown that in interviews lying is typically more cognitively demanding than telling the truth (26). Investigators can exploit this difference in cognitive load by making additional requests that will further increase the cognitive load truth tellers and liars experience [such as gripping an object while telling a story, (27)]. Since liars' mental resources are already depleted by the act of lying, they find it more difficult than truth tellers to cope with such additional requests (27) and the additional requests may also impair their story telling (28).

(ii) Cognitive credibility assessment: Asking unexpected questions. Liars often prepare themselves for interviews by planning answers to possible questions (7). This planning makes sense as planned answers often contain fewer cues to deceit than spontaneous answers (1). However, there is a weakness: Liars cannot know which questions will be asked. Investigators can exploit this weakness by asking a mixture of anticipated and unanticipated questions. Liars find it easier to answer the anticipated questions than the unanticipated questions, because they can give their planned answers to the former but not to the latter (29). For truth tellers, the difficulty in answering anticipated and unanticipated questions should be less pronounced. The most straightforward application of this technique is by interviewing pairs of suspects individually and comparing their answers to the expected and unexpected questions. Pairs of truth tellers showed similar overlap in their answers to expected questions as pairs of liars, but the pairs of truth tellers showed more overlap in their answers to unexpected questions than pairs of liars (30, 31). Another comparison can also be made: Comparing the overlap between expected and unexpected questions within pairs of truth tellers and within pairs of liars. Pairs of truth tellers showed a similar overlap in their answers to the expected and unexpected questions, whereas pairs of liars showed more overlap in their answers to the expected questions than in their answers to the unexpected questions (31).

(iii) Cognitive credibility assessment: Encouraging interviewees to say more. In interview settings, truth tellers typically do not provide spontaneously all the information they hold in their memory (32, 33). There are two reasons for this, a cognitive reason and a social reason.

Regarding the cognitive reason: Interviewees are unable to retrieve spontaneously all the information from their memory. Memory recall can be enhanced by using mnemonics of which asking interviewees to sketch while talking is an example (15). Sketching while narrating elicits additional information in truth tellers (34–36). Vrij et al. (37) provide four reasons for this. First, sketching is a method to mentally reinstate the context of the interviewee's experience and context reinstatement enhances memory recall. Second, sketching is a visual output which makes it more compatible with visually experienced events than the traditional oral output. Sketching facilitates recalling visual or spatial information (15), which is often the type of information interviewees discuss. Third, making a sketch is a time consuming activity. This will result in the interviewee having more time to think about the event,[1] and this enhanced thinking may improve his/her recall of the event. Fourth, the request to sketch automatically leads to obtaining spatial information because each person/object must be positioned somewhere in the location someone sketches. Spatial information is not automatically given in a verbal response, because someone can just report who were present and which objects were present without reporting their locations (38). In the only deception experiment to date in which participants were asked to sketch while narrating (37), the difference in truth tellers reporting more additional details than liars was greater in the sketch condition than in the control condition. Truth tellers are likely to have had a richer memory of the event than liars, and truth tellers' richer memory may have led them to report more new details than liars.

The second reason why truth tellers typically do not provide spontaneously all the information they know in interview settings is a social reason: People are uncertain what and how much information they are expected to provide. The Model Statement technique addresses this social reason.

THE MODEL STATEMENT TECHNIQUE

In daily life situations social rules imply that people do not report all the information they know. For example, when someone is asked by a colleague on Monday morning what s/he did during the weekend, the answer is likely to be very short: Just a few words or few sentences highlighting the main activities. Of course, interviewees will realize in formal interview settings that they need to provide more information than a few words or sentences, but they still do not know how much detail they are expected to provide. One effective way to change truth tellers' idea about how much information to provide in an interview setting is to expose them to a Model Statement, which is an example of a detailed account unrelated to the topic of the interview (39). The Model Statement works as a social comparison (40, 41) and has shown to raise the expectations amongst both truth tellers and liars about how much information they are expected to (42). A Model Statement works better than the verbal request "to provide all the details someone can remember," perhaps because the former is a concrete example whereas the latter is an abstract instruction. It is probably easier for people to follow concrete examples than abstract instructions (43).

A Model Statement does not just elicit information, it can also be used for lie detection if certain dependent variables are analyzed. In the first two Model Statement deception studies ever published, the Model Statement facilitated the elicitation of information (39, 44). However, it did so in truth tellers and liars to a similar extent, which made the technique unsuitable for lie detection purposes when "total details" was considered as output variable. This exact pattern of results has been replicated in six out of seven ensuing studies (42, 43, 45–49), but see Porter et al. (47) as an exception. In other words, the Model Statement technique elicits more information in both truth tellers and liars, but cannot distinguish between truth tellers and liars based on the total amount of information.

For the Model Statement technique to work as a lie detection tool it is important to consider the quality rather than the quantity of information that is reported. The first Model Statement deception study (39) already hinted at this: Although truth tellers and liars provided a similar amount of information after exposure to a Model Statement, the information provided by truth tellers sounded more plausible than that of liars. That the quality of details rather than the quantity of details distinguish truth tellers from liars makes sense. Both truth tellers and liars realize after exposure to a Model Statement that they are expected to provide many details (42). The amount of details is thus unlikely to distinguish between the two groups. The type of detail becomes relevant because it takes into account the different cognitive abilities of truth tellers and liars and the different strategies they use to appear convincing.

Studies to date gave insight into two types of detail that could distinguish truth tellers from liars after exposure to a Model Statement, the number of complications (37, 49) and the number of peripheral details (43) that were reported. A complication is "an occurrence that makes a situation more difficult than necessary" (37). Examples of complications are "The sailing race was canceled, because there was not enough wind" and "When we arrived at the museum it was closed"; "Initially we did not see our friend, as he was waiting at a different entrance") (37, 49). Complications occur more often in truthful statements than in deceptive statements (8, 50). In interviews, liars prefer to keep their stories simple (7), but adding complications makes the story more complex. A Model Statement increases the number of complications interviewees report, particularly in truth tellers (37, 49). Complications are often not about key aspects of the activities that someone describes, and the story can be well understood without reporting the complications. Take for example, when someone describes traveling to a holiday destination. All sorts of complications that happen en route to a holiday destination are not necessary to understand the travel to the holiday destination (someone forgot to bring a valuable item; taxi turned up late; traffic on the road; airplane delayed; late gate change at the airport). Therefore, truth tellers may leave at least some of them out when they are not exposed to a Model Statement. Liars are reluctant to provide complications in order to keep their story simple. As a result, truth tellers are more likely than liars to report more complications after being exposed to a Model Statement.

A second measure that takes truth tellers' and liars' different strategies into account is distinguishing between core or peripheral details (43). Core details are details that, if changed, can result in changes in the basic and most important part of the story; details that have no such impact are considered peripheral (51). Thus, if someone describes attending an Adele pop concert, all details about the actual concert are core details whereas information about drinks in the pub before and after the

[1] Slowing down the output process also makes it easier for the interviewer to understand the interviewee's statement, which probably enhances the communication between interviewer and interviewee.

concert, are peripheral details. Both truth tellers and liars realize that they need to provide more details after exposure to a Model Statement (49). Truth tellers, who have actually experienced an event (e.g., attending an Adele pop concert), will be able to provide more core and peripheral information, by employing a "tell it all strategy" (7). For liars, who have not experienced an event (e.g., did not attend an Adele concert), providing core information is more difficult and risky. It is difficult because they have to make up information and it is risky because the information may provide leads to investigators that they can check. Thus, liars may avoid providing too many core details in an attempt to minimize the risk of presenting incriminating information (6, 21), but may compensate this by providing peripheral details in an attempt to provide a sufficient amount of detail. In the only Model Statement deception experiment distinguishing between core and peripheral details to date, the latter assumption was supported: In the Model Statement present condition liars reported more peripheral details than truth tellers, whereas no difference in peripheral details emerged in the control condition (43).

HOW TO USE A MODEL STATEMENT IN REAL LIFE

We believe that the Model Statement technique should be used as a within-subjects technique, as employed by Leal et al. (43). Thus, first the interviewee should be invited to initially report via an open-ended question all s/he can remember about the event under investigation. This should then be followed by a Model Statement after which the interviewee should again be invited to report via an open-ended question all s/he can remember, but this time by taking into account the amount of detail s/he heard in the Model Statement. Investigators should then listen to the number of new complications reported in the second recall and the amount of new peripheral information reported in the second recall.

Three Practical Elements Merit Attention

First, use a within-subjects structure when applying the Model Statement technique. Within-subjects comparisons are better for lie detection purposes than between-subjects comparisons (52). In a between-subjects comparison, the interviewee would be asked to report the event only once and to do this after exposure to the Model Statement. The amount of information an interviewee provides depends on many factors, including his/her personality [some people talk more than others (53–55), the situation (some events are richer in detail than others) or preparedness for the interview [pre-planned answers often contain more words than spontaneous answers, e.g. (56)]. In a within-subjects comparison, it does not matter how detailed an initial answer is or how many complications someone initially provides (which is largely influenced by personality, situation and preparedness). The only relevant measure is the number of peripheral details and complications that are *added* (more likely to be influenced by veracity).

Second, the Model Statement should be unrelated to the topic of investigation so that it does not give liars the chance to "copy" the example and use it in their own statement. In our research, we use a 734 words Model Statement in which a young man describes his experiences when attending a Formula 2 motor race, commencing where the drivers go to their grid position prior to the start of the race. This is an atypical event that does not give interviewees the opportunity to copy details.

Third, our Model Statement is an authentic experience (the person really attended a Formula 2 motor race), which we think is important. True experiences sound more realistic than made-up experiences and are therefore more powerful. It becomes even worse when someone fabricates a model statement on the spot. It typically is not detailed enough and often sounds what it actually is: a made-up story. We always present the Model Statement in the format of an audiotape. However, other ways to present the Model Statement are possible. We return to this point in the next section.

FUTURE RESEARCH

Unfortunately, many lie detection techniques are taught to practitioners without solid empirical evidence to back them up, which we consider a particularly poor and potentially harmful practice (12, 57). Many research avenues for Model Statement deception research are possible. We will conclude this article by discussing five more research ideas in somewhat more detail.

First, an obvious but important research endeavor would be replication of studies that have been carried out so far, ideally by different groups of researchers in different labs. Most Model Statement research to date comes from Vrij's lab but much stronger conclusions could be drawn if Vrij's lab findings are replicated in other labs. This refers in particular to research related to complications and core/peripheral details, as research in that area is still scarce. At the same time, those researchers could then search for other variables than complications or peripheral details on which truth tellers and liars may differ after exposure to a Model Statement.

Second, research should be carried out manipulating the content of the Model Statement. Will it have an effect on interviewees' recall? People experience activities through their perceptual senses: They see, hear, touch, smell, or taste things. An event interviewees are asked to describe may contain more information about some of these perceptual sources than about others. Will it help or hinder lie detection if interviewees are exposed to a Model Statement that corresponds with their perceptual experience? For example, if the experience the interviewee talks about contains many auditory experiences, will it then be beneficial to use a Model Statement that focuses on auditory experiences? On the one hand it may help truth tellers to recall more details they have experienced through the particular sense(s) emphasized in the Model Statement but, on the other hand, it may give liars an idea what type of information to fabricate.

Third, thus far we have always used an audiotaped Model Statement. This could be played via a loudspeaker but also from a mobile phone. Alternatives are that the investigator reads out an example or that interviewees read a written text of a Model Statement. Until tested it is unclear which—if any—modality works best for discriminating between truth tellers and liars.

Fourth, from training we give in the Model Statement technique (58), we know that interviewees quickly understand that they are requested to provide more details than they initially thought they had to provide. This may result in different mental processes in truth tellers and liars. Adding information should be easier for truth tellers than for liars, as truth tellers can go back to their memory, whereas liars have to think what made-up details to add to their stories. Consequently, liars may listen less to the content of the Model Statement than truth tellers, because liars cannot listen to the Model Statement and think of details to add at the same time. If so, truth tellers and liars might be able to report back the content of the initial part of the Model Statement to an equal extent as at the initial stage both are listening to the Model Statement. However, after this stage, liars should switch off and start thinking about the details they will add. From this point onwards, we expect liars to report back less of the content of the Model Statement than truth tellers.

Fifth, a meta-analysis summarizing Cognitive Interview research showed that "report everything" instructions result in interviewees reporting more information without a reduction in accuracy (59). We expect a Model Statement also to have this effect on truth tellers—more information without a reduction in accuracy—but believe that this issue is important enough to be examined empirically.

AUTHOR CONTRIBUTIONS

AV wrote the initial article and received comments from SL and RF.

ACKNOWLEDGMENTS

The time the first author spent working on this article was funded by the Center for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1).

REFERENCES

1. DePaulo BM, Lindsay JL, Malone BE, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. *Psychol Bull.* (2003) 129:74–118. doi: 10.1037/0033-2909.129.1.74
2. DePaulo BM, Morris WL. Discerning lies from truths: behavioural cues to deception and the indirect pathway of intuition. In: Granhag PA, Strömwall LA, editors. *Deception Detection in Forensic Contexts*. Cambridge: Cambridge University Press (2004). p. 15–40.
3. Bond CF, DePaulo BM. Accuracy of deception judgements. *Personal Soc Psychol Rev.* (2006) 10:214–34. doi: 10.1207/s15327957pspr1003_2
4. Hartwig M, Granhag PA, Strömwall L, Doering N. Impression and information management: on the strategic self-regulation of innocent and guilty suspects. *Open Criminol J.* (2010) 3:10–16. doi: 10.2174/1874917801003020010
5. Vrij A, Mann S, Leal S, Granhag PA. Getting into the minds of pairs of liars and truth tellers: an examination of their strategies. *Open Criminol J.* (2010) 3:17–22. doi: 10.2174/18749178010030200017
6. Granhag PA, Hartwig M. A new theoretical perspective on deception detection: on the psychology of instrumental mind-reading. *Psychol Crime Law* (2008) 14:189–200. doi: 10.1080/10683160701645181
7. Hartwig M, Granhag PA, Strömwall L. Guilty and innocent suspects' strategies during police interrogations. *Psychol Crime Law* (2007) 13:213–27. doi: 10.1080/10683160600750264
8. Amado BG, Arce R, Fari-a F. Undeutsch hypothesis and criteria based content analysis: a meta-analytic review. *Eur J Psychol Appl Legal Context* (2015) 7:3–12. doi: 10.1016/j.ejpal.2014.11.002
9. Vrij A, Granhag PA. Eliciting cues to deception and truth: what matters are the questions asked. *J Appl Res Mem Cogn.* (2012) 1:110–7. doi: 10.1016/j.jarmac.2012.02.004
10. Vrij A. Deception and truth detection when analysing (non)verbal cues. *Appl Cogn Psychol.* (2018). doi: 10.1002/acp.3457
11. Vrij A. Verbal lie detection tools from an applied perspective. In: Rosenfeld JP, editor. *Detecting Concealed Information and Deception: Recent Developments*. San Diego, CA: Elsevier: Academic Press (2018). p. 297–321.
12. Vrij A, Fisher RP. Which lie detection tools are ready for use in the criminal justice system? *J Appl Res Mem Cogn.* (2016) 5:302–7. doi: 10.1016/j.jarmac.2016.06.014
13. Granhag PA, Hartwig M. The Strategic Use of Evidence (SUE) technique: a conceptual overview. In: Granhag PA, Vrij A, and Verschuere B, editors. *Deception Detection: Current Challenges and New Approaches*. Chichester: Wiley (2015). p. 231–51.
14. Hartwig M, Granhag PA, Luke T. Strategic use of evidence during investigative interviews: the state of the science. In: Raskin DC, Honts CR, and Kircher JC, editors. *Credibility Assessment: Scientific Research and Applications*. Amsterdam: Academic Press (2014). p. 1–36. doi: 10.1016/B978-0-12-394433-7.00001-4
15. Fisher RP, Geiselman RE. *Memory Enhancing Techniques for Investigative Interviewing: The Cognitive Interview*. Springfield, IL: Charles C. Thomas (1992).
16. Colwell K, Hiscock-Anisman CK, Fede J. Assessment criteria indicative of deception: an example of the new paradigm of differential recall enhancement. In: Cooper BS, Griesel D, and Ternes M, editors. *Applied Issues in Investigative Interviewing, Eyewitness Memory, and Credibility Assessment*. New York, NY: Springer (2013). p. 259–92. doi 10.1007/978-1-4614-5547-9_11
17. Colwell K, Hiscock-Anisman CK, Memon A, Taylor L, Prewett J. Assessment criteria indicative of deception (ACID): an integrated system of investigative interviewing and detecting deception. *J Investr Psychol Offen Profil.* (2007) 4:167–80. doi: 10.1002/jip.73
18. Bell BE, Loftus EF. Trivial persuasion in the courtroom: the power of (a few) minor details. *J Pers Soc Psychol.* (1989) 56:669–79. doi: 10.1037/0022-3514.56.5.669
19. Nahari G, Vrij A, Fisher RP. Does the truth come out in the writing? SCAN as a lie detection tool. *Law Hum Behav.* (2012) 36:68–76. doi: 10.1037/h0093965
20. Nahari G. The applicability of the verifiability approach to the real world. In: Rosenfeld P, Editor. *Detecting Concealed Information and Deception: Verbal, Behavioral, and Biological Methods*. San Diego, CA: Academic Press (2018). p. 329–50.
21. Nahari G, Vrij A, Fisher RP. Exploiting liars' verbal strategies by examining unverifiable details. *Legal Criminol. Psychol.* (2014) 19:227–39. doi: 10.1111/j.2044-8333.2012.02069.x
22. Harvey A, Vrij A, Nahari G, Ludwig K. Applying the verifiability approach to insurance claims settings: exploring the effect of the information protocol. *Legal Criminol. Psychol*. (2016) 22:47–59. doi: 10.1111/lcrp.12092
23. Nahari G, Vrij A, Fisher RP. The verifiability approach: countermeasures facilitate its ability to discriminate between truths and lies. *Appl. Cogn. Psychol.* (2014) 28:122–8. doi: 10.1002/acp.2974

24. Vrij A, Fisher R, Blank H. A cognitive approach to lie detection: a meta-analysis. *Legal Criminol Psychol.* (2017) 22:1–21. doi: 10.1111/lcrp.12088
25. Vrij A, Fisher R, Blank H, Leal S, Mann S. A cognitive approach to elicit nonverbal and verbal cues of deceit. In: van Prooijen JW, van Lange PAM, Editors. *Cheating, Corruption, and Concealment: The Roots of Dishonest Behavior.* Cambridge: Cambridge University Press (2016). p. 284–310.
26. Christ SE, Van Essen DC, Watson JM, Brubaker LE, McDermott KB. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex* (2009) 19:1557–66. doi: 10.1093/cercor/bhn189
27. Debey E, Verschuere B, Crombez G. Lying and executive control: an experimental investigation using ego depletion and goal neglect. *Acta Psychol.* (2012) 140:133–41. doi: 10.1016/j.actpsy.2012.03.004
28. Evans JR, Michael SW, Meissner CA, Brandon SE. Validating a new assessment method for deception detection: introducing a psychologically based credibility assessment tool. *J Appl Res Mem Cogn.* (2013) 2:33–41. doi: 10.1016/j.jarmac.2013.02.002
29. Lancaster GLJ, Vrij A, Hope L, Waller B. Sorting the liars from the truth tellers: the benefits of asking unanticipated questions. *Appl Cogn Psychol.* (2012) 27:107–14. doi: 10.1002/acp.2879
30. Roos af Hjelmsäter E, Öhman L, Granhag PA, Vrij A. Mapping deception in adolescents: eliciting cues to deceit through an unanticipated spatial drawing task. *Legal Criminol. Psychol.* (2014) 19:179–88. doi: 10.1111/j.2044-8333.2012.02068.x
31. Vrij A, Leal S, Granhag PA, Mann S, Fisher RP, Hillman J, et al. Outsmarting the liars: the benefit of asking unanticipated questions. *Law Hum Behav.* (2009) 33:159–66. doi: 10.1007/s10979-008-9143-y
32. Fisher RP. Interviewing cooperative witnesses. *Legal Criminol Psychol.* (2010) 15:25–38. doi: 10.1348/135532509X441891
33. Vrij A, Hope L, Fisher RP. Eliciting reliable information in investigative interviews. *Policy Insights Behav Brain Sci.* (2014) 1:129–36. doi: 10.1177/2372732214548592
34. Dando C, Wilcock R, Milne R. The cognitive interview: the efficacy of a modified mental reinstatement of context procedure for frontline police investigators. *Appl Cogn Psychol.* (2009) 23:138–47. doi: 10.1002/acp.1451
35. Leins D, Fisher RP, Pludwinsky L, Robertson B, Mueller DH. Interview protocols to facilitate human intelligence sources' recollections of meetings. *Appl Cogn Psychol.* (2014) 28:926–35. doi: 10.1002/acp.3041
36. Mattison MCL, Dando CJ, Ormerod TC. Sketching to remember: episodic free recall task support for child witnesses and victims with autism spectrum disorder. *J Autism Dev Disord.* (2015) 45:1751–65. doi: 10.1007/s10803-014-2335-z
37. Vrij A, Leal S, Fisher RP, Mann S, Dalton G, Jo E, et al. Sketching as a technique to elicit information and cues to deceit in interpreter-based interviews. *J Appl Res Mem Cogn.* (2018) 7:303–13. doi: 10.1016/j.jrarmac.2017.11.001
38. Vrij A, Mann S, Leal S, Fisher R. Is anyone there? Drawings as a tool to detect deception in occupations interviews. *Psychol Crime Law* (2012) 18:377–88. doi: 10.1080/1068316X.2010.498422
39. Leal S, Vrij A, Warmelink L, Vernham Z, Fisher R. You cannot hide your telephone lies: providing a model statement as an aid to detect deception in insurance telephone calls. *Legal Criminol Psychol.* (2015) 20:129–46. doi: 10.1111/lcrp.12017
40. Cialdini RB. *Influence: the Psychology of Persuasion.* New York, NY: William Morrow and Company (2007).
41. Festinger L. A theory of social comparison processes. *Human Relat.* (1954) 7:117–40. doi: 10.1177/001872675400700202
42. Ewens S, Vrij A, Leal S, Mann S, Jo E, Shaboltas A, et al. Using the model statement to elicit information and cues to deceit from native speakers, non-native speakers and those talking through an interpreter. *Appl Cogn Psychol.* (2016) 30:854–62. doi: 10.1002/acp.3270
43. Leal S, Vrij A, Deeb H, Jupe L. Using the model statement to elicit verbal differences between truth tellers and liars: the benefit of examining core and peripheral details. *J. Appl Res Mem Cogn.* (2018). doi: 10.1016/j.jarmac.2018.07.001
44. Bogaard G, Meijer EH, Vrij A. Using an example statement increases information but does not increase accuracy of CBCA, RM, and SCAN. *J Invest Psychol Offen Profil.* (2014) 11:151–63. doi: 10.1002/jip.1409
45. Harvey A, Vrij A, Leal S, Lafferty M, Nahari G. Insurance based lie detection: enhancing the verifiability approach with a model statement component. *Acta Psychol.* (2017) 174:1–8. doi: 10.1016/j.actpsy.2017.01.001
46. Kleinberg BAR, van der Toolen Y, Vrij A, Arntz AR, Verschuere BJ. Automated verbal credibility assessment of intentions: the model statement technique and predictive modelling. *Appl Cogn Psychol.* (2018) 32:354–66. doi: 10.1002/acp.3407
47. Porter C, Vrij A, Leal S, Vernham Z, Salvanelli G, McIntyre N. Using specific model statements to elicit information and cues to deceit in information-gathering interviews. *J Appl Res Mem Cogn.* (2017) 7:132–42. doi: 10.1016/j.jarmac.2017.10.003
48. Vrij A, Leal S, Jupe L, Harvey A. Within-subjects verbal lie detection measures: a comparison between total detail and proportion of complications. *Legal Criminol Psychol.* (2018) 23:265–79. doi: 10.1111/lcrp.12126
49. Vrij A, Leal S, Mann S, Dalton G, Jo E, Shaboltas A, et al. Using the Model Statement to elicit information and cues to deceit in interpreter-based interviews. *Acta Psychol.* (2017) 177:44–53. doi: 10.1016/j.actpsy.2017.04.011
50. Vrij A. *Detecting Lies and Deceit: Pitfalls and Opportunities.* 2nd ed. Chichester: John Wiley and Sons (2008).
51. Heuer F, Reisberg D. Vivid memories of emotional events: the accuracy of remembered minutiae. *Mem Cognit.* (1990) 18:496–506.
52. Vrij A. Baselining as a lie detection method. *Appl Cogn Psychol.* (2016) 30:1112–9. doi: 10.1002/acp.3288
53. Merckelbach H. Telling a good story: fantasy proneness and the quality of fabricated memories. *Pers Individ Dif.* (2004) 37:1371–82. doi: 10.1016/j.paid.2004.01.007
54. Nahari G, Pazuelo M. Telling a convincing story: richness in detail as a function of gender and priming. *J Appl Res Mem Cogn.* (2015) 4:363–7. doi: 10.1016/j.jarmac.2015.08.005
55. Vrij A, Akehurst L, Soukara S, Bull R. Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law Hum Behav.* (2002) 26:261–83. doi: 10.1023/A:1015313120905
56. Sporer SL, Schwandt B. Paraverbal indicators of deception: a meta-analytic synthesis. *Appl Cogn Psychol.* (2006) 20:421–46. doi: 10.1002/acp.1190
57. Vrij A, Hartwig M, Granhag PA. Reading lies: nonverbal communication and deception. *Annu Rev Psychol.* (2018).
58. Vrij A, Leal S, Mann S, Vernham Z, Brankaert F. Translating theory into practice: Evaluating a cognitive lie detection training workshop. *J Appl Res Mem Cogn.* (2015) 4:110–20. doi: 10.1016/j.jarmac.2015.02.002
59. Köhnken G, Milne R, Memon A, Bull R. The cognitive interview: a meta-analysis. *Psychol Crime Law* (1999) 5:3–28. doi: 10.1080/106831699084 14991

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychiatry

REVIEW
published: 21 November 2018
doi: 10.3389/fpsyt.2018.00617

Testing Claims of Crime-Related Amnesia

Marko Jelicic[1,2]*

[1] *Forensic Psychology Section, Department of Clinical Psychological Science, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands,* [2] *Department of Criminal Law and Criminology, Faculty of Law, Maastricht University, Maastricht, Netherlands*

Many violent offenders report amnesia for their crime. Although this type of memory loss is possible, there are reasons to assume that many claims of crime-related amnesia are feigned. This article describes ways to evaluate the genuineness of crime-related amnesia. A recent case is described in which several of these strategies yielded evidence for feigned crime-related amnesia.

Keywords: crime-related amnesia, deception, feigning, malingering, forensic neuropsychology

OPEN ACCESS

Edited by:
Cristina Scarpazza,
Università degli Studi di Padova, Italy

Reviewed by:
Stefano Zago,
Fondazione IRCCS Ca' Granda
Ospedale Maggiore Policlinico
(IRCCS), Italy
Charles Scott,
University of California, Davis,
United States

***Correspondence:**
Marko Jelicic
m.jelicic@maastrichtuniversity.nl

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *16 August 2018*
Accepted: *02 November 2018*
Published: *21 November 2018*

Citation:
Jelicic M (2018) Testing Claims of
Crime-Related Amnesia.
Front. Psychiatry 9:617.
doi: 10.3389/fpsyt.2018.00617

INTRODUCTION

A few years ago, 29 year old Randy unexpectedly appeared at the house of his parents. Because he was covered in blood, his father asked him if something happened to his girlfriend. Randy nodded, upon which his father called the emergency number. The police speeded to Randy's apartment and found his girlfriend lying on the floor in a pool of blood. She had been stabbed to death. Randy was arrested and taken to the police station. During his interrogation, he told the police that, although he did not rule out having killed his girlfriend, he had no memory for this fatal incident whatsoever.

Randy's case is not unique: a nontrivial percentage of people who are accused or convicted of violent offenses claim crime-related amnesia. About 70 years ago, Leitch (1) found that 16 out of 51 offenders (31%) convicted of homicide reported memory loss for their crime. Several decades later, Taylor and Kopelman (2) interviewed 203 men charged with both violent and non-violent offenses. Of the 34 men accused of having committed murder or manslaughter, 9 of them (26%) claimed amnesia for their crime. More recently, Pyszora et al. (3) studied the case note-notes of 207 individuals sentenced to life imprisonment. In this sample, 60 (29%) reported memory loss for their offense. By and large, it seems that about 20 to 30% of those who have committed violent crimes claim crime-related amnesia (4). It should be noted here that this form of memory loss is not only reported by violent offenders: individuals convicted for sexual and property offenses also claim amnesia for their crimes (5).

Apparently, a considerable number of people—both laypersons and professionals—believe that offenders can forget or repress a serious crime that they committed. Magnussen et al. (6) asked 1,000 Norwegians whether or not murderers who claim amnesia for their offense are telling the truth about their memory loss. Thirty nine percent of the respondents opined that such offenders are truthful about their amnesic episode. In a follow-up study, Magnussen and Melinder (7) asked 857 Norwegian licensed psychologists, most of them working in the field of clinical psychology, for their opinion about this issue. Thirty eight percent of this sample of professionals endorsed the view that murderers who claim crime-related amnesia are honest about the gap in their memory. More recently, Melinder and Magnussen (8) asked 117 psychiatrists and psychologists who served as expert witnesses in Norwegian courts whether or not murderers who report crime-related amnesia are telling the truth about their memory loss. This time, 39 percent of the respondents indicated that such offenders are truthful about their amnesia. Because these studies were all conducted in Norway, one could argue that these findings may not be generalized to countries outside Scandinavia. However, according to Lynn et al. (9), the belief that offenders can repress crime-related memories appears to be a worldwide phenomenon.

WHAT DOES SCIENCE SAY ABOUT CRIME-RELATED AMNESIA?

There are three different explanations for memory loss in offenders. The first explanation contends that, during the time of the crime, some offenders suffer from a temporary (or permanent) brain dysfunction that prevents or undermines the storage of criminal events in memory. This type of memory loss is labeled organic amnesia (4). The second explanation holds that many offenders are in an extreme emotional condition (e.g., rage) when committing a violent crime. Therefore, crime-related details would be stored in memory in the context of strong emotions. Later, when the offender has returned to a more calm state of mind, he or she would be unable to remember the crime because of a mismatch in emotional state between the encoding of crime-related events and the retrieval of such events. This type of memory loss is termed dissociative amnesia (4). When people have (dissociative) amnesia for a crime of passion, some authors prefer to speak of a "red-out" (10). The third explanation for crime-related amnesia is that a considerable number of offenders are pretending to be unable to remember crime-related details. This type of memory loss is called feigned amnesia (4).

Temporary brain dysfunction can lead to crime-related amnesia. The thalamus, hippocampus, and prefrontal cortex are all involved in the encoding and storage of information in autobiographical memory (11). Out of these three brain areas, the hippocampus is probably the most vulnerable to temporary or permanent dysfunction. Closed head injury, consumption of large quantities of alcohol, use of certain prescription or illegal drugs, low blood sugar (hypoglycemia), as well as shortage of oxygen (hypoxia) may result in a temporary deranged hippocampus (12). A considerable portion of offenders who claim crime-related amnesia report that their inability to recollect criminal events is due to alcohol consumption (5, 13). However, drinking alcohol does not necessarily lead to amnesia. In order to develop an alcohol blackout, one should drink large quantities of alcohol. This type of memory loss is assumed to be only plausible when the blood alcohol concentration (BAC) of the offender is higher than 0.25% (14). While most medications do not affect memory, prescription drugs that do have amnesic side effects include benzodiazepines and other hypnotics, antidepressants, and anticonvulsants (15). Gamma Hydroxybutyrate (GHB) is an illegal drug that can lead to temporary memory loss (16).

There are several reasons to doubt the existence of dissociative amnesia for an offense. For one thing, laboratory studies in which participants encode and store information in a particular emotional state and retrieve that information in another state have shown that a mismatch in state between the acquisition and test phase does not lead to a substantial inability to remember stimuli presented in the learning phase (17). Also, committing a (violent) crime typically means that one performs one or more actions. Research has shown that people tend to remember their own actions better than other information (18). Most importantly, dozens of studies have demonstrated that strong emotions do not undermine memory performance, but enhance memory for stressful events (19, 20). A recent Canadian study serves as a case in point. McKinnon et al. (21) investigated the richness and accuracy of people's memory for a highly traumatic event. Their participants were former passengers of a transatlantic plane flight that nearly ditched at sea. The authors found that, a few years after the incident, all participants had excellent memory for events that took place during the near-fatal flight. Based on this investigation and many other studies showing memory enhancement by strong emotions, one could reason that dissociative amnesia for an offense is, at best, scarce. This notion has also been put forward by some forensic psychologists. Centor (22), for example, stated: "My own experience, during a period of over 11 years in a forensic unit, failed to confirm even one case of psychogenic amnesia in the absence of a psychotic episode, brain damage, or acute brain syndrome" (p.240).

Crime-related amnesia clearly has benefits for people charged with serious offenses (23). To start with, one cannot provide the police with crucial details of an offense, which might obstruct police investigations. Also, sexual offenders do not have to talk about a shameful offense. In addition, having no memory for a crime suggests that the offense was impulsive and not premeditated (in homicide cases, this could lead a manslaughter instead of a murder conviction). Moreover, this type of amnesia might lead to a mitigation of criminal responsibility. Given these advantages, it seems likely that many offenders who report memory loss for their offense are actually feigning their amnesia. A famous historical example of feigned crime-related amnesia is that of Rudolf Hess. This prominent Nazi politician claimed, at the start of the Nuremberg trials, to have no recollections of his personal and political activities in the years preceding the Second World War. Hess was examined by a number of psychiatrists who unanimously declared that his memory loss was genuine. However, when after some weeks, Hess realized that, because of his amnesia, he could not respond to the allegations against him, he informed the tribunal that he had feigned his memory loss (24).

EVALUATING THE VERACITY OF CRIME-RELATED AMNESIA

As mentioned above, a dysfunctional hippocampus can lead to impaired memory storage. Therefore, when asked to evaluate the authenticity of a claim of crime-related amnesia, the first thing a forensic psychologist or psychiatrist should do is to determine if organic factors might account for the putative memory loss reported by the offender (25). To establish whether or not the offender had a deranged hippocampus because of excess consumption of alcohol, it would be wise to calculate his or her BAC level (26). This is not a hard thing to do: many BAC calculators can be found on the Internet. Given the questionable status of dissociative amnesia, crime-related amnesia reported by an offender without hippocampal dysfunction at the time of the crime should be treated with skepticism (22).

Clinical features of the memory loss reported by the offender may shed light on the genuineness of amnesia. Power (27) argued that periods of real memory loss have a gradual and blurred onset and termination. Thus, an amnesic episode with an abrupt beginning and end would be suggestive of feigned memory loss. Moreover, people with true amnesia usually have "islands of memory" (28). That is, they do not have complete memory loss, but are still able to remember elements of events that occurred during their amnesic period. Hence, absolute amnesia would be indicative of feigned memory loss, while a "patchy" amnesia suggests bona fide memory loss. Note that in people with mild head injury or alcohol intoxication, there usually is shrinkage of their amnesia (29). At first, such individuals cannot remember events that took place in the days (or sometimes weeks) before the injury or intoxication. However, as time passes by, their memories of these events gradually return. Typically, old memories return before more recent recollections, a phenomenon called "Ribot's law"—named after the nineteenth century French psychologist Théodule Ribot (30). Thus, shrinkage of amnesia is line with a genuine inability to remember certain criminal events. Schacter (31) stated that feelings-of-knowing rating might also be used as an indicator of the veracity of crime-related amnesia. Feelings-of-knowing pertain to the idea that, when unable to remember autobiographical events, one could retrieve information from memory when given the right hints or cues. Because true amnesia often goes hand in hand with a feeling-of-knowing, an offender stating that not even hypnosis or truth serum will bring back crime-related events, would be suggestive of feigned memory loss. Although Schacter's suggestion is interesting, some authors are critical about the use of feelings-of-knowing as a tool to determine the authenticity of crime-related amnesia (32). In a number of cases, clinical features of the alleged memory loss may not provide the forensic psychologist or psychiatrist with valid information pertaining to the credibility of crime-related amnesia. Research suggests that a large percentage of offenders have a history of traumatic brain injury (33). Because such offenders have intimate knowledge of genuine temporary memory loss, forensic psychologists and psychiatrists should be cautious to use clinical features of crime-related amnesia as evidence for true memory loss.

Using standard questionnaires and tests designed to measure a tendency to feign memory problems is another strategy to determine the authenticity of crime-related amnesia. An example of such a questionnaire is the Structured Inventory of Malingered Symptomatology (34). The SIMS is a self-report instrument determining feigning of psychiatric symptoms and cognitive impairments. It comprises 75 yes/no items that measure an individual's proneness to endorse bizarre and/or atypical symptoms in five different areas including amnesia. The rationale behind the instrument is that feigners do not know how genuine symptoms manifest themselves. Examples of items from the amnesia subscale are "Recently I've noticed that my memory is getting so bad that there have been entire days I cannot recall" and "At times I've been unable to remember the names and faces of close relatives so that they seem complete strangers." Each improbable item that is endorsed is scored "1." Scores on the 75 items are added up to obtain a total SIMS score. A score of 17 or higher is considered indicative of feigning of symptoms (35)—although some authors have argued that a higher cutting score should be used (36). The SIMS has acceptable psychometric properties (37). As mentioned above, the SIMS consists of items pertaining to improbable symptoms. A potential limitation of questionnaires that only list bizarre and/or improbable symptoms is that they might be easily identifiable as tests measuring feigning. For that reason, Merten et al. (38) developed the Self-Report Symptom Inventory (SRSI) to determine feigning of different psychiatric disorders and/or cognitive impairment. In contrast with the SIMS, the SRSI consists of items that ask for pseudo-symptoms and genuine symptoms. Although the SRSI seems to have promising psychometric characteristics, more research on the diagnostic accuracy of this instrument is necessary before it can be used in forensic practice.

A well-known example of a test developed to measure feigned memory impairments is the Test of Memory Malingering (TOMM). This test may also be used to investigate the veracity of crime-related amnesia (39). The TOMM is an easy memory test requiring only passive recognition. The idea behind this test is that genuine brain-disordered patients perform quite well on it. Because feigners want to convince the forensic psychologist or psychiatrist that they suffer from memory problems, they often perform substantially poorer on the TOMM than bona fide patients with memory disorders. The TOMM contains two learning trials where the examinee is shown 50 line drawings of common objects. Both trials are followed by a forced choice recognition task. A retention trial given 15 min after the second learning trial consists of the forced choice recognition task only. For each correct answer, the item is scored "1." A score below 45 on the second learning trial or the retention trial is considered indicative of feigned memory impairments. A number of studies have shown that the TOMM has good psychometric properties (40, 41). Besides the TOMM, there are other well-validated tests that can be used to evaluate an individual's tendency to feign memory problems, such as the Amsterdam Short-Term Memory Test (42) and the Word Memory Test (43).

A drawback of the above-described questionnaires and tests is that they can only be used in cases where the offender claims that his or her inability to remember crime-related details is the result of a general memory deficit due to, for instance, sleeping problems, use of certain prescription drugs or a neurological disorder. These instruments do not work in offenders who say that normally they have no memory problems, but because of excessive drinking and/or taking illegal drugs on the day of the offense they cannot remember criminal acts. In such cases, symptom validity testing might be helpful in assessing the authenticity of claims of crime-related amnesia.

Symptom validity testing (SVT) was originally created to assess the credibility of hearing problems (44). More recently, it has been used as an instrument to assess the veracity of crime-related amnesia (45, 46). SVT consists of a forced choice technique in which an offender who claims to suffer from crime-related amnesia is asked a range of questions pertaining to details of the crime and/or crime scene (47). For each question, the examinee must choose between two equally plausible answers, one of which is correct and the other is incorrect. True memory loss for a crime should result in random performance on the SVT. Or in other words, bona fide amnesia will result in ∼50% of the answers being correctly answered. If significantly more incorrect answers are given than correct answers, an offender is performing below chance level performance. This can only be achieved when one is intentionally giving incorrect answers, which is indicative of having preserved memory for criminal events. Because SVT is based on binomial statistics, the exact probability of a deviant memory performance can be quantified (see case below). Unfortunately, SVT can only be used in a limited number of cases. One needs to be able to create a substantial number of two choice questions about the crime and/or crime scene from the investigative reports. In addition, in a proper SVT procedure, only the offender and the police should have intimate knowledge of the crime. If details of the crime have been "leaked" to the offender via the media, police officers or his or her attorney, the offender might claim amnesia and at the same time legitimize an above-chance level on the SVT by referring to the media, police officers or his or her attorney.

It should be noted here that offenders who feign amnesia for a crime are lying about their memory loss. For that reason, psychophysiological and neural measures created to detect lying (48) might also be used to evaluate the authenticity of a crime-related amnesia claim. However, these measures have not yet been used in forensic practice.

CASE

This article started with the case of Randy who claimed to have no memory of the stabbing of his girlfriend. At the time of the offense, he had not consumed any alcohol or illegal drugs. Moreover, he did not take any prescription drugs and neither was he suffering from a psychiatric or somatic disorder. Therefore, it seems unlikely that he suffered from a deranged hippocampus during the fatal incident. Randy said that he had complete amnesia for the stabbing. Thus, he did not report any islands of memory. His score on the SIMS was 32, indicating a strong indication of a tendency to feign psychiatric symptoms and cognitive impairments. When the police started their investigation, they had no clear picture regarding the manner in which the offense was committed. Therefore, they asked the Dutch Forensic Institute (NFI) to reconstruct the crime by analyzing forensic evidence. Using blood spatter patterns, the wounds on the victim's body, and other physical evidence, the NFI was able to almost completely reconstruct the offense. This information was not provided to Randy or his attorney. Based on the crime reconstruction, an SVT consisting of 25 two choice questions was created. Each question was followed by a correct and an incorrect answer. These 25 questions were given to a panel of 10 forensic psychologists, who were asked to give the most plausible answer to each question. This procedure showed that five of the questions did not contain two equally plausible answering options. Thus, the final SVT consisted of 20 questions. One of these questions was: "The victim was stabbed: (a) one time in her chest, two times in her neck, or (b) two times in her chest and one time in her neck." Randy gave wrong answers to 14 of the 20 items. According to binomial statistics, the probability that his response pattern was based on random guessing was <6 percent, indicating that there is a <6 percent chance that his amnesia was genuine. Taken together, there was converging evidence that Randy had feigned his amnesia for the stabbing. The court also found his amnesia claim not credible. He was sentenced to 12 years imprisonment.

DISCUSSION

There are multiple strategies for forensic psychologists and psychiatrists to examine the veracity of crime-related amnesia claims. When asked to evaluate such claims, it would be best to use a multi-method approach (49). Especially in cases where offenders might have suffered from a deranged hippocampus at the time of the crime, forensic psychologists, and psychiatrists are advised to exercise restraint in labeling memory loss for a crime as non-credible. Only when there is converging evidence for feigning, crime-related amnesia may be deemed not authentic (25).

In order to determine whether or not the offender suffered from a deranged hippocampus at the time of the offense, a forensic psychologist or psychiatrist should have solid knowledge of neuropsychology and psychopharmacology. Although clinical features of the amnesia may yield important information about the authenticity of the memory loss reported by the offender, they cannot always be used. Because offenders may have intimate knowledge of memory loss, those who report bona-fide symptoms of amnesia may still be feigning their amnesia. Tests may shed important light on the veracity of memory loss for a crime. However, when an offender does not have a reason to feign memory problems during the forensic evaluation (e.g., an individual who claims that he or she cannot remember crime-related events because of alcohol or drug intoxication), a normal score on the SIMS, the TOMM or a related instrument does not say much about the veracity of the amnesia claim. In such cases, it would be informative to develop and administer an SVT to determine the authenticity of the memory loss reported by the offender.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

1. Leitch A. Notes on amnesia in crime for the general practitioner. *Med Press* (1948) 26:459–63.
2. Taylor PJ, Kopelman MD. Amnesia for criminal offences. *Psychol Med.* (1984) 14:581–8. doi: 10.1017/S003329170001518X
3. Pyszora NM, Barker AF, Kopelman MD. Amnesia for criminal offences: a study of life sentence prisoners. *J Forensic Psychiatry Psychol.* (2003) 14:475–90. doi: 10.1080/14789940310001599785
4. Cima M, Merckelbach H, Nijman H, Knauer E, Hollnack S. I can't remember Your Honor: offenders who claim amnesia. *Ger J Psychiatry* (2002) 5:24–34.
5. Stout RG, Farooque RS. Claims of amnesia for criminal offenses: psychopathology, substance abuse, and malingering. *J Forensic Sci.* (2008) 53:1218–22. doi: 10.1111/j.1556-4029.2008.00819.x
6. Magnussen S, Andersson J, Cornoldi C, De Beni R, Endestad T, Goodman GS, et al. What people believe about memory. *Memory* (2006) 14:595–613. doi: 10.1080/09658210600646716
7. Magnussen S, Melinder A. What psychologists know and believe about memory: a survey of practitioners. *Appl Cogn Psychol.* (2012) 26:54–60. doi: 10.1002/acp.1795
8. Melinder A, Magnussen S. Psychologists and psychiatrists serving as expert witnesses in court: what do they know about eyewitness memory? *Psychol Crime Law* (2015) 21:53–61. doi: 10.1080/1068316X.2014.915324
9. Lynn SJ, Evans J, Laurence J-R, Lilienfeld SO. What do people believe about memory? Implications for the science and pseudoscience of clinical practice. *Can J Psychiatry* (2015) 60:541–7. doi: 10.1177/070674371506001204
10. Swihart G, Yuille J, Porter S. The role of state-dependent memory in "red outs". *Int J Law Psychiatry* (1999) 22:199–212.
11. Parkin AJ. *Memory and Amnesia: An Introduction, 2nd edn.* Oxford: Blackwell (1997).
12. Billingsley-Marshall R, Simos PG, Papanicolaou AC. Limbic amnesia. In: Papanicolaou AC, editor. *The Amnesias. A Clinical Textbook of Memory Disorders.* Oxford: Oxford University Press (1996). p. 130–55.
13. Bourget D, Bradford JM. Sex offenders who claim amnesia for their alleged offense. *Bull Am Acad Psychiatry Law* (1995) 23:299–307.
14. White AM. What happened? Alcohol, memory blackouts, and the brain. *Alcohol Res Health* (2003) 27:186–96.
15. Chavant F, Favrelière S, Lafay-Chebassier C, Plazanet C, Pérault-Pochat MC. Memory disorders associated with consumption of drugs: updating through a case/noncase study in the French PharmacoVigilance Database. *Br J Clin Pharmacol.* (2011) 72:898–904. doi: 10.1111/j.1365-2125.2011.04009.x
16. Barker JC, Harris SL, Dyer JE. Experiences of gamma hydroxybutyrate (GHB) ingestion: a focus group study. *J Psychoactive Drugs* (2007) 39:115–29. doi: 10.1080/02791072.2007.10399870
17. Eich E. Searching for mood dependent memory. *Psychol Sci.* (1995) 6:67–75. doi: 10.1111/j.1467-9280.1995.tb00309.x
18. Madan CR, Singhal A. Using actions to enhance memory: effect of enactment, gestures, and exercise on human memory. *Front Psychol.* (2012) 3:507. doi: 10.3389/fpsyg.2012.00507
19. McGaugh JL. *Memory and Emotion: The Making of Lasting Memories.* New York, NY: Columbia University Press (2003).
20. McNally RJ. *Remembering Trauma.* Cambridge, MA: Harvard University Press (2003).
21. McKinnon MC, Palombo DJ, Nazarov A, Kumar N, Khuu W, Levine B. Threat of death and autobiographical memory: a study of passengers from flight AT236. *Clin Psychol Sci.* (2015) 3:487–502. doi: 10.1177/2167702614542280
22. Centor A. Criminals and amnesia: comment on Bower. *Am Psychol.* (1982) 37:240. doi: 10.1037/0003-066X.37.2.240
23. Christianson SA, Merckelbach H. Crime-related amnesia as a form of deception. In: Granhag PA, Strömwall LA, editors. *The Detection of Deception in Forensic Contexts.* New York, NY: Cambridge University Press (2004). p. 195–217.
24. Gilbert GM. *Nuremberg Diary.* New York, NY: Farrar, Straus and Company (1947).
25. Peters MJV, Van Oorsouw K, Jelicic, M, Merckelbach H. Let's use those tests! Evaluations of crime-related amnesia claims. *Memory* (2013) 21:599–607 doi: 10.1080/09658211.2013.771672
26. Van Oorsouw K, Merckelbach H. Detecting malingered memory problems in the civil and criminal arena. *Legal Criminological Psychol.* (2010) 15:97–114. doi: 10.1348/135532509X451304
27. Power DJ. Memory, identification and crime. *Med Sci Law* (1977) 17:32–9. doi: 10.1177/002580247701700212
28. Whitty CWM, Zangwill OL. Traumatic amnesia. In: Whitty CWM, Zangwill OL, editors. *Amnesia, 2nd edition*. London: Butterworths (1977). p. 118–35.
29. Richardson JTE. *Clinical and Neuropsychological Aspects of Closed Head Injury, 2nd edition.* London: Psychology Press (2001).
30. Haber L, Haber RN. Criteria for the admissibility of eyewitness testimony of long past events. *Psychol Public Policy Law* (1998) 4:1135–59. doi: 10.1037/1076-8971.4.4.1135
31. Schacter DL. Amnesia and crime: how much do we really know? *Am Psychol.* (1986) 41:286–95. doi: 10.1037/0003-066X.41.3.286
32. Porter S, Birt AR, Yuille JC, Hervé HF. Memory for murder: a psychological perspective on dissociative amnesia in legal contexts. *Int J Law Psychiatry* (2001) 24:23–42.
33. Hughes N, Williams WH, Chitsabesan P, Walesby RC, Mounce LTA, Clasby B. The prevalence of traumatic brain injury among young offenders in custody: a systematic review. *J Head Trauma Rehab.* (2015) 30:94–105. doi: 10.1097/HTR.0000000000000124
34. Smith GP, Burger GK. Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). *J Am Acad Psychiatry Law* (1997) 25:180–9.
35. Merckelbach H. Smith GP. Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. *Arch Clin Neuropsychol.* (2003) 18:145–52. doi: 10.1093/arclin/18.2.145
36. Wisdom NM, Callahan JL, Shaw TG. Diagnostic utility of the Structured Inventory of Malingered Symptomatology to detect malingering in a forensic sample. *Arch Clin Neuropsychol.* (2010) 25:118–25. doi: 10.1093/arclin/acp110
37. van Impelen A, Merckelbach H, Jelicic M, Merten T. The Structured Inventory of Malingered Symptomatology (SIMS): a systematic review and meta-analysis. *Clin Neuropsychol.* (2014) 28:1336–65. doi: 10.1080/13854046.2014.984763
38. Merten T, Merckelbach H, Giger P, Stevens A. The Self-Report Symptom Inventory (SRSI): a new instrument for the assessment of distorted symptom endorsement. *Psychol Inj Law* (2016) 9:102–11. doi: 10.1007/s12207-016-9257-3
39. Tombaugh T. *Test of Memory Malingering (TOMM).*New York, NY: Multi-Health Systems (1996).
40. Teichner G, Wagner MT. The Test of Memory Malingering (TOMM): normative data from cognitively intact, cognitive impaired, and elderly patients with dementia. *Arch Clin Neuropsychol.* (2004) 19:455–64. doi: 10.1016/S0887-6177(03)00078-7
41. Vallabhajosula B, Van Gorp WG. Post-Daubert admissibility of scientific evidence on malingering of cognitive deficits. *J Am Acad Psychiatry Law* (2001) 29:207–15.
42. Schagen S, Schmand B, De Sterke S, Lindeboom J. Amsterdam Short-Term Memory test: a new procedure for the detection of feigned memory deficits. *J Clin Exp Neuropsychol.* (1997) 19:43–51. doi: 10.1080/01688639708403835
43. Green P, Iverson GL, Allen LM. Detecting malingering in head injury litigation with the Word Memory test. *Brain Injury* (1999) 13:813–9. doi: 10.1080/026990599121205
44. Pankratz L. Symptom validity testing and symptom retraining: procedures for the assessment and treatment of functional sensory deficits. *J Consult Clin Psychol.* (1997) 47:409–10. doi: 10.1037/0022-006X.47.2.409
45. Denney RL. Symptom validity testing of remote memory in a criminal forensic setting. *Arch Clin Neuropsychol.* (1996) 11:589–603. doi: 10.1093/arclin/11.7.589
46. Frederick RI, Carter M, Powel J. Adapting symptom validity testing to evaluate suspicious complaints of amnesia in medicolegal evaluations. *Bull Am Acad Psychiatry Law* (1995) 23:227–33.

47. Merten T, Merckelbach H. Forced-choice tests as single-case experiments in the differential diagnosis of intentional symptom distortion. *J Exp Psychopathol.* (2013) 4:20–37. doi: 10.5127/jep.023711

48. Meijer EH, Verschuere B. Detection deception using psychophysiological and neural measures. In: Otgaar H, Howe M, editors. *Finding the Truth in the Courtroom. Dealing With Deception, Lies, and Memories.* Oxford: Oxford University Press (2018). p. 209–24.

49. Giger P, Merten T, Merckelbach H. Detection of feigned crime-related amnesia: a multi-method approach. *J Forensic Psychol Pract.* (2010) 10:440–63. doi: 10.1080/15228932.2010.489875

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

REVIEW
published: 21 December 2018
doi: 10.3389/fpsyt.2018.00700

A Review of Approaches to Detecting Malingering in Forensic Contexts and Promising Cognitive Load-Inducing Lie Detection Techniques

Jeffrey J. Walczyk, Nate Sewell and Meghan B. DiBenedetto*

Psychology and Behavioral Sciences, Louisiana Tech University, Ruston, LA, United States

Malingering, the feigning of psychological or physical ailment for gain, imposes high costs on society, especially on the criminal-justice system. In this article, we review some of the costs of malingering in forensic contexts. Then the most common methods of malingering detection are reviewed, including those for feigned psychiatric and cognitive impairments. The shortcomings of each are considered. The article continues with a discussion of commonly used means for detecting deception. Although not traditionally used to uncover malingering, new, innovative methods are emphasized that attempt to induce greater cognitive load on liars than truth tellers, some informed by theoretical accounts of deception. As a type of deception, we argue that such cognitive approaches and theoretical understanding can be adapted to the detection of malingering to supplement existing methods.

Keywords: malingering detection techniques, cognitive malingering detection, theory of mind, forensic psychiatry, inducing cognitive load

Edited by:
Cristina Scarpazza,
Università degli Studi di Padova, Italy

Reviewed by:
Merylin Monaro,
Università degli Studi di Padova, Italy
Ilaria Zampieri,
IMT School for Advanced Studies Lucca, Italy

***Correspondence:**
Jeffrey J. Walczyk
walczyk@latech.edu

Specialty section:
This article was submitted to Forensic Psychiatry, a section of the journal Frontiers in Psychiatry

Received: *12 April 2018*
Accepted: *03 December 2018*
Published: *21 December 2018*

Citation:
Walczyk JJ, Sewell N and DiBenedetto MB (2018) A Review of Approaches to Detecting Malingering in Forensic Contexts and Promising Cognitive Load-Inducing Lie Detection Techniques. Front. Psychiatry 9:700. doi: 10.3389/fpsyt.2018.00700

The present article is partly a review of methods of detecting malingering. Previous reviews of malingering detection methods include Sartori et al. (1) as well as Sartori et al. (2). The present review adds uniquely to the literature by highlighting recent cognitive-based methods of lie detection and relevant theory potentially applicable to malingering detection.

The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) defines malingering as "the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives" [(3), p. 726]. Although the concept of malingering has existed for centuries, it was not until the mid-1900's that the term "malingering" was introduced to refer to soldiers who feigned illness or disability in order to avoid military service (4). The term's usage has broadened to include other incentives, such as avoiding work, gaining financial advantage, avoiding arrest, evading criminal prosecution, mitigating sentencing, receiving medication, or gaining admission to a hospital for shelter (3, 5). Despite a clear definition, the detection of malingering is elusive. For instance, Rogers and Shuman (6) found that the use of DSM criteria results in the accurate identification of only 13.6–20.1% of actual malingerers (true positives). However, 79.9–86.4% of individuals are misclassified as malingerers (false positives) using the same criteria. The accurate detection of malingering is thus a pressing societal issue.

NEGATIVE EFFECTS OF MISCLASSIFICATION/BURDEN ON THE CRIMINAL-JUSTICE SYSTEM

In addition to the problem of not identifying individuals who are malingering, there are also very serious consequences for misclassifying malingering when an individual's presentation is genuine (false positives). Labeling an individual as a malingerer can be stigmatizing, which carries negative connotations and can negatively impact individuals for the remainder of their lives (7). In addition, many clinicians avoid diagnosing malingering for fear of legal consequences. Because of the difficulty of arriving at an accurate diagnosis, these clinicians fear they will be sued and are, therefore, reluctant to apply the label (8).

In the criminal-legal realm, malingering has a negative impact on the proper execution of justice. Failure to detect malingering in cases of insanity or incompetency can delay prosecution for months or years and often results in unnecessary hospitalizations. It also provides malingerers with the opportunity to be moved from secure facilities, such as jails or prisons, to psychiatric facilities with more comfortable environments and from which escape is easier (9). Prison inmates also feign psychiatric or cognitive symptoms in order to transfer to medical centers where they can gain access to pain medication and have greater contact with female staff (10, 11). Some researchers have expressed concern regarding the inappropriate use of antipsychotic medications administered to inmates who successfully feign psychosis. In addition to being very costly, such medications can cause harmful side-effects such as dystonias, diabetes, high blood cholesterol, and metabolic syndrome (12).

In summary, despite ongoing advances in malingering detection, many individuals successfully malinger mental, cognitive, and physical disorders in order to gain financial compensation, avoid work, gain access to medications, and avoid prison. This places a large financial burden on society, negatively impacts the efficient operation of the healthcare system, and increases medical costs. The creation or discovery of new and effective malingering detection methods has the potential to significantly reduce the burden of malingering on the criminal justice system and on society generally.

CURRENT ASSESSMENTS OF MALINGERING

Measuring Malingering Detection Accuracy

The detection of malingering is typically done using standardized assessments as this approach gives clinicians access to the most current and scientifically-based methods (13). Malingering detection accuracy is assessed by evaluating each measure's *sensitivity, hit rate, positive predictive power* (PPP), *and negative predictive power* (NPP). Sensitivity refers to the ability of a measure to accurately identify individuals who have the condition the measure is designed to detect. Specificity is the ability of a measure to identify individuals for whom the condition is not present. The hit rate is the total proportion of accurately identified cases, i.e., the true positives plus the true negatives (14–16). PPP is the percentage of individuals detected as malingering who are actually malingering, while the NPP is the percentage of honest individuals (17, 18).

Psychiatric and Cognitive Malingering Detection Strategies

Rogers et al. validated 10 strategies for the detection of feigning within the domain of mental disorders (19, 20). These strategies fall into two broad categories: *unlikely presentations* and *amplified presentations* (21). **Table 1** provides a description of these 10 strategies by category (unlikely or amplified). Examination of these strategies reveals a common thread. Compared to those genuinely suffering from psychiatric disorders, malingerers present symptoms and other patterns of behavior that are deviant from what is typical, are rare, or exaggerated. In other words, compared with those actually suffering from disorders, malingerers seek to create in the minds of clinicians an impression of their affliction that often will overshoot the mark by not agreeing with actual symptom patterns of genuine cases.

Strategies for the detection of cognitive malingering differ from those used to detect the malingering of mental disorders, as they focus more on performance accuracy, although the detection of unusual response patterns applies to both domains (22–24). The two categories that classify detection strategies for cognitive malingering are *excessive impairment* and *unexpected patterns* (6, 20). **Table 2** provides a description of each strategy by category. Responses detected by these strategies include performance failures on items that are typically achievable even by those with actual cognitive impairment and the detection of failure rates that are statistically unlikely. As before, a common thread across these different kinds of malingering, psychiatric or cognitive, is that malingerers seek to create false impressions in mental health professionals and often will miss the mark.

Assessments of Psychiatric Malingering

Structures Interview of Reported Symptoms (SIRS)

SIRS is a comprehensive assessment for detecting feigned mental disorders, specifically an interview-based measure that consists of 172 items. A primary strength of the SIRS is its incorporation of multiple mental disorder detection strategies, including many of those identified in **Tables 1**, **2**. Its primary scales include: Rare Symptoms, Symptom Combinations, Improbable and Absurd Symptoms, Blatant Symptoms, Subtle Symptoms, Selectivity of Symptoms, Severity of Symptoms, and Reported vs. Observed Symptoms [RO; (21)]. Five additional scales comprise the supplemental scales, producing a total of thirteen detection strategies, resulting in a particularly robust instrument. Items on the SIRS include Detailed Inquiries regarding symptomology and their levels of severity. Repeated Inquiries assess response consistency; and General Inquiries, which are designed to probe for specific symptoms, symptom patterns and general psychological disturbances (25).

The SIRS is the most commonly used and best-validated assessment in the forensic detection of malingering (11, 21). Although some research has suggested that the SIRS has low vulnerability to coaching, it is reported to produce lower

TABLE 1 | Detection strategies for malingered mental disorders.

Category	Detection strategy	Strategy description
Unlikely presentation	Rare symptoms	Focuses on endorsed symptoms that are reported very infrequently by genuine clinical patients. Malingerers often overreport these rare psychological problems
	Quasi-rare symptoms	Focuses on symptoms and characteristics that occur infrequently in normative (non-clinical) samples
	Improbable symptoms	Focuses on endorsed symptoms that are much more extreme than Rare Symptoms. This includes symptoms of a preposterous nature
	Symptom combinations	Focuses on symptoms and characteristics that commonly occur in genuine clinical patients, but that rarely occur in the combinations endorsed by malingerers
	Spurious patterns of psychopathology	An elaboration of Symptom Combinations. Utilizes particular scale configurations that detect patterns which are characteristic of malingering, but uncommon in genuine patients
Amplified presentation	Indiscriminant symptom endorsement	This strategy is based on the finding that some malingerers tend to endorse a large number of symptoms in comparison to genuine clients
	Symptom severity	In comparison to even severely impaired genuine patients, malingerers are more likely to endorse a large number of symptoms which they describe as being “unbearable” or “extreme”
	Obvious symptoms	This strategy focuses on the finding that in contrast to genuine patients, malingerers tend to endorse symptoms that clearly indicate a serious mental disorder
	Reported vs. observed symptoms	This strategy compares an individual’s account of their symptoms to clinical observations. Malingering is often identified by clear discrepancies between endorsed symptoms and clinical observations
	Erroneous stereotypes	Focuses on common misconceptions that individuals have regarding symptoms commonly associated with mental disorders. Malingerers often overendorse these erroneous stereotypes

Adapted from Rogers (19), and Rogers and Shuman (6).

TABLE 2 | Detection strategies for malingered cognitive impairment.

Category	Detection strategy	Strategy description
Excessive impairment	Floor effect	Uses very simple items that genuine patients are likely to successfully complete. Malingerers are likely to overestimate the difficulty of the items and consequently provide incorrect answers
	Forced-choice testing	Assesses for performances on cognitive tests that are lower than would be expected. Malingerers are identified by extremely poor performances
	Symptom Validity Testing	Utilizes a forced-choice format and assesses for below-chance performance and error rates that are statistically extremely unlikely
Unexpected patterns	Magnitude of error	Genuine patients often make predictable errors. This strategy identifies malingerers by focusing on high proportions of unexpected errors
	Performance curve	With continually increasing item difficulty, a predictable pattern emerges that is typically a negative curve when plotted on a graph. Malingerers are less likely to take item difficulty into account and therefore typically produce a different pattern
	Violation of learning principles	Based on established learning principles, this strategy identifies malingerers by comparing expected results to those that violate basic learning principles
Unexpected patterns (limited validation)	Consistency of comparable items	Compares performance with that of genuine patients by focusing on predictable patterns on comparable items within the same test. Malingering is likely to result in atypical patterns of performance. Requires rigorous testing across diverse samples, which is often unavailable for many assessments
	Psychological sequelae	Patients with genuine brain injuries often manifest additional symptoms. This strategy tests whether these additional symptoms differentiate malingerers from genuine patients. Caution should be used as malingerers may be able to recognize common sequelae
	Atypical presentation	This strategy assesses for unexpected findings (e.g., substantial performance variations on similar tests). Should be used with great caution as it lacks a firm conceptual basis

Adapted from Rogers (19) and Rogers and Bender (20).

specificity estimates than those reported in the official manual and has a higher rate of classifying true patients as malingerers than indicated by previous estimates [as cited in (26)]. Finally, many settings are inadequately equipped to utilize the SIRS given that the administration is complex, and the length of the interview can take significantly longer than the administration time of 30–40 min suggested by Rogers et al. (27–29).

Structured Inventory of Malingered Symptomology (SIMS)

The SIMS (30) is a paper-and-pencil screening devise for detecting malingering. Its items were drawn and revised from validity items of existent instruments and others were derived from research on attributes typical of malingerers. A 75-item scale, its subscales include psychosis, amnesic disorders, neurological impairment, affective disorders, and low intelligence. The SIMS yields a total score and subscale scores for each of the five subscales. Based on research with college students who were instructed either to malingerer or respond honestly, compared to other measures of malingering (e.g., the F and K scales of the MMPI), the SIMS total score has the highest sensitivity for detecting malingering (95.6%). Still, its validity in detecting malingering in more authentic contexts is largely unknown.

Minnesota Multiphasic Personality Inventory-2 (MMPI-2)

The MMPI-2 is a 567-item self-report measure designed to assess personality characteristics and psychopathology, although it is also used extensively outside of mental health and medical settings (31–35). The MMPI-2 has several validity scales designed to evaluate the accuracy with which test takers respond to test items and to predict distorted presentations. These include scales to detect the under- or over-reporting of symptoms.

Validity scales, developed to uncover malingering on the MMPI-2, include the F Scale (Infrequency), Fb scale (Back Infrequency), Fp Scale (Infrequency-Psychopathology), FBS (Symptom Validity), and Gough's Dissimulation Scale [Ds; (19, 36)]. Although the F Scale achieved the highest effect sizes among the various validity scales in two meta-analyses (37, 38), some researchers consider it to be inadequate considering it was designed only to detect atypical responding, which may also occur as a result of confusion regarding test items, a low reading level, or pathological interpretation of personal experiences (21). Many of the items on the F and Fb scales do not accurately distinguish between feigning and honest responding, with the Fb scale demonstrating poorer performance than the F scale (36, 39). Rogers and Neumann (36) concluded that both F and Fb are flawed scales for the detection of malingering. The most effective scales can misclassify 5–15% of individuals who attempt to malinger (40, 41). Heinze (42) reported even higher rates of false positives, stating that between 12 and 55% of individuals with genuine mental disorders have been identified by the MMPI-2 as malingerers.

Millon Clinical Multiaxial Inventory MCMI-III

The MCMI-III is 175 item self-report scale (true/false items) that takes about 30 min to complete (43). With a focus on personality disorders, its 28 subscales comprise the following categories: Modifying Indices (including validity items), Clinical Personality Patterns, Severe Personality Pathology, Severe Syndrome, and Clinical Syndrome. Atypical patterns, extreme scores, or high invalidity can suggest malingering (43).

Miller Forensic Assessment of Symptoms (M-FAST)

The M-FAST (44) is a brief screening measure designed to detect malingered mental illness in forensic settings by assessing individual response styles (45–47). The M-FAST contains 25 items, including 15 true or false questions, 5 Likert items, 2 yes/no questions, and 3 items designed to detect discrepancies between responses and observations (45, 47).

The M-FAST utilizes similar detection strategies as the SIRS, with four of its seven scales employing the same detection strategies (Reported vs. Observed, Extreme Symptomology, Rare Combinations, Unusual Hallucinations). It also contains three additional scales: Unusual Symptom Course, which assesses the reported speed of onset of mental illness; Negative Image, which capitalizes on the tendency of malingerers to believe that they should be viewed negatively by others; and Suggestibility, which relies on the likelihood that malingerers will endorse symptoms they believe will make them appear mentally ill (9, 48–50). However, a third of the scales on the M-FAST have low internal consistency, resulting in low reliability for these scales. Vitacco et al. (49) found problems with homogeneity for the individual M-FAST scales and lower utility estimates compared to the total score. In addition, they found that the M-FAST produced an unacceptably high rate of false positives (10%) using the total scale scores.

Assessments of Cognitive Malingering

Tests of Memory Malingering (TOMM)

The TOMM is a recognition memory test that utilizes symptom validity testing (SVT), forced-choice, and floor-effect detection strategies. As a forced-choice SVT, the TOMM presents the respondent with two alternatives per test item, allowing for a 50% chance of choosing correctly. Scores falling significantly below this probability level suggest malingering (51, 52). As noted in **Table 2**, the floor-effect strategy involves the presentation of cognitive tasks which malingerers incorrectly believe impaired individuals are incapable of completing accurately (19). The TOMM contains 50 items and consists of two memory learning trials, with each trial followed by an assessment of recognition memory (53, 54). The respondent is initially shown a series of 50 line drawings, followed by a recognition assessment in which each drawing is presented alongside a foil. The subject is asked to identify the previously presented drawing and is given feedback regarding the correctness of the response (54–56). If the respondent does not achieve a correct score during the second trial on at least 45 items, a Retention Trial is administered. Malingering should be suspected if the respondent earns a score of 45 or less on the second trial or the Retention Trial (53, 56). Some researchers have reported lower hit rates with the TOMM than with other measures. Unfortunately, high face validity enables a large number of respondents to perceive it correctly as an assessment of malingering (57).

Rey Fifteen-Item Test (FIT)

The FIT utilizes the floor effect detection strategy but without a forced-choice design (6, 19, 58, 59). The FIT presents a memory task that appears difficult but is actually easy. The individual is shown 15 different items consisting of letters, numbers, and geometric shapes for a brief period and then asked to recall and reproduce as many of the items as possible (58–61). The fifteen items are presented in five rows containing three items each. The first row presents the numbers 1, 2, and 3; the second presents the roman numerals I, II, and III; the third presents a square, a triangle, and a circle; the fourth presents the letters A, B, and C (Capitalized); and the fifth presents the letters a, b, and c, all in lowercase (61). A cut-off score of nine is most commonly used (54), although some have suggested the use of lower cut-off scores of eight or less to accommodate those with true impairment (62). Schretlen et al. (63) concluded that the FIT has several limitations and that patients with genuine impairment often perform poorly on the test, while many malingerers score above the recommended cut-off score. A number of studies have shown that forced choice recognition tests are more useful in identifying cognitive malingering than the standard FIT. Clinicians should also note that the FIT does not meet the Daubert standard, which outlines criteria for the admissibility of scientific evidence in court (64–66).

Word Memory Test (WMT)

The WMT is a forced-choice test of malingering. In addition to the forced-choice detection strategy, it also utilizes the following: (a) violation of learning principles, (b) floor effect, (c) symptom validity testing, and (d) the performance curve (6, 19, 57), all noted in **Table 2**. The learning principle it utilizes is the advantage of recognition memory performance over recall, which malingerers may not account for in their efforts to deceive. The WMT is more effective than other measures of feigning in its use of this detection strategy, yielding large effect sizes (19).

Regarding administration, the respondent is presented with 20 pairs of semantically-related words during two learning trials. Immediately following these presentations, the Immediate Recognition trial begins in which each of the 40 words is paired with a foil and the individual is asked to select the correct target word. After 30 min, the delayed recognition trial is given, and target words are paired with new foils. Four separate effort tests, designed to evaluate verbal memory, are then given, including the multiple choice, paired associates, delayed free recall, and long delayed free recall subtests. Scoring is accomplished by comparing the number of words recognized consistently across the immediate and delayed trials. A score of 82.5% or below is the cut-off (54, 55, 67). Although simulated malingerers perform worse than participants instructed to perform at their best on the WMT, coaching and the use of sophisticated simulators has resulted in less accurate detection of malingering with this instrument (67, 68). Pella et al. (59) warn that the WMT may be particularly vulnerable to coaching compared to other instruments, resulting in a high rate of false negatives.

LIE DETECTION

Despite advances in malingering detection technology, current methods are far from adequate, with high rates of false positives, false negatives, and a susceptibility to coaching. Perhaps the detection of malingering can be facilitated by incorporating developments from the field of lie detection given that malingering is high-stakes deception. Current methods of lie detection are reviewed, with an emphasis on innovative cognitive-based approaches.

Human Lie Detectors

Although lying is common in everyday life (69, 70), people are amazingly poor lie detectors. Individuals accurately judge lying at or slightly above chance levels but are a bit better at identifying truth telling (71–75). Although one might assume that professional lie-catchers (e.g., police officers, customs officers, judges, mental health professionals) have better accuracies at detecting lies, the majority of studies show that they do not (73, 76–79). Rather than having to depend on unreliable human lie detectors, we now review some prominent and emerging technologies potentially applicable to ferreting out malingering, many with minimal dependence on human lie detectors.

Arousal-Based Approaches

Control Question Technique

The *polygraph* is a scientific instrument that continuously records psycho-physiological arousal as assessed by pulse rate, blood pressure, respiration rate, and/or skin conductivity, which has been applied to the detection of deception. The most common questioning procedure used with it is the Control Question Technique (CQT; 79). In a typical test, a respondent is given a pretest interview for gathering information that provides the basis for control questions. Once questions are constructed, the examiner will preview them with the respondent to ensure that they are understood and will not surprise the respondent when asked later. During the examination, irrelevant questions are asked such as "What is your age?," along with the control questions that most people tend to lie to. For example, "Have you ever stolen anything from your place of employment?" Finally, relevant questions, probing the issue central to the exam, are asked (e.g., "Did you rape … on January 7th?"). The questions usually elicit brief answers. A guilty liar, it is hypothesized, will show more arousal to relevant questions than to control questions, whereas an innocent, honest respondent will show more arousal to control questions (80). Law enforcement and federal agencies in the United States use the CQT as a screening device for hiring and retaining employees and as a tool for criminal investigations. The CQT has been used to verify victim's statements, evaluate the veracity of witnesses, and to exonerate suspects. Still, test results are largely inadmissible in US courtrooms (81).

A major criticism of polygraph-based techniques, especially the CQT, regards their generally poor validity. Specifically, the CQT produces a high rate of false positives, that is, the labeling of honest individuals as liars (81–84). Researchers have also found that respondents can easily be trained to evade detection by using mental and physical distraction techniques known as countermeasures (81, 84).

Concealed Information Test (CIT)/Guilty Knowledge Test (GKT)/Concealed Knowledge Test (CKT)

Partly in response to the validity concerns with the CQT, the CIT, also known as the CKT and GKT, was proposed. It is a questioning paradigm that can be used with the polygraph to uncover the false denials of respondents by exposing whether they possess guilty knowledge or concealed information, presumably resulting from their participation in a crime or some other experience (80). During a typical CIT, the respondent is presented with multiple-choice questions, each having one relevant alternative (correct answer) and several neutral alternatives (plausible distractors). The latter should be chosen such that an innocent person could not discriminate them from the relevant alternative (80). An example of a relevant question is “How was the victim killed?,” with the response alternatives of “shot,” “stabbed,” “struck,” “strangled,” or “poisoned.” This question could be re-asked multiple times, along with other questions probing different aspects of a crime scene. The respondent need not answer. If heightened arousal occurs consistently to relevant responses, then the respondent may be concealing information as the perpetrator. The CIT assumes that innocent respondents could not have acquired guilty knowledge indirectly and that guilty respondents encoded guilty knowledge and have retained it (85).

Some validity concerns with the CQT were resolved in the CIT, including more standardization of the procedure, more appropriate control alternatives, fewer false positives, and a stronger theoretical basis (80). Also, beyond the psychophysiological measures of the polygraph, concealed information has been uncovered with the diverse cues of response time (86–90), event-related potentials (91–93), and pupil dilation (94). Also, the CIT has been used to expose the simulation of amnesia (95). Still, the CIT is limited in the deception it can uncover to the false denials of those possessing concealed knowledge.

Cognitive Load-Inducing Approaches

Cognitive load refers to the demands made on the limited pools of attention and working memory resources for performing mental tasks (96, 97). Some recent, novel, and promising techniques for detecting deception, and possibly malingering, rather than viewing deception as a physiological/emotional event as does the CQT, view it as a cognitive act that generally imposes greater cognitive load on respondents than honesty does. In support, Vrij and Mann (98) reported that telling complex, high-stakes lies increased cognitive loads, with liars exerting significantly more effort to control their speech than did truth tellers. As further neurological support, brain imaging studies using fMRI (functional magnetic resonance imaging) scanners, which reveal brain activity during task performance, suggest that deception activates higher brain centers associated with cognitive demand, particularly in the frontal lobe (99, 100). If lying is more cognitively demanding than truth telling, deception should reveal itself in longer times needed to answer questions, more inconsistencies and hesitancy in answering logically interrelated questions, greater pupil dilation, more activity in the brain’s prefrontal cortex, more blinking, and in other signs of heightened cognitive load.

Cognitive load-inducing lie detection techniques, only some of which can be reviewed due to their sheer number, seek to enhance the mental effort of liars compared to truth tellers, in effect, making it mentally harder to deceive than to be honest (101, 102). Once refined and validated, such techniques may accurately expose malingering in forensic settings, perhaps used in conjunction with existing methods.

Asking Surprise Questions/Soliciting Surprise Drawings

Asking surprise questions of respondents can increase cognitive load on liars. For instance, Vrij et al. (103) instructed pairs of participants to lie or tell the truth about whether they had lunch together. All pairs then prepared for an interview that followed, which included anticipating likely questions. During the interview, general and unanticipated questions were asked, some of the latter probing minor details like these: “What was the color of the shirt your partner wore?” “Who sat closest to the door?” Inconsistencies across such questions from members of each pair allowed observers to classify liars and truth tellers beyond chance, as did discrepancies across surprise pictures that members were asked to draw of the layout of the restaurant. Although researchers did not measure the cognitive loads produced by the surprise questions or drawings directly, we regard them as cognitive load-inducing because deceptive participants likely had to think more than truth tellers to guess at how their partners might respond to the questions to ensure their answers and drawings would be consistent (104).

These results are promising. Still, asking surprise, detailed-oriented questions has limitations. Once knowledge of this lie detection technique disseminates, liars may include spatial and other obscure details into their deceptive narratives in anticipation of surprise questions. Also, memory for minor details can go unnoticed by truth tellers (105). Thus, if respondents claim “I can’t remember” to detail-oriented questions, they may be answering honestly. Similar concerns apply to drawing pictures. Even so, refinement of these techniques may overcome these concerns.

Having to Maintain Eye Contact

Having to maintain eye contact with another can selectively heighten cognitive load and anxiety in liars. In support, Vrij et al. (106) directed some participants to lie to interview questions while others told the truth. Some of the participants were also directed to maintain continuous eye contact with the interviewer. Interviews were videotaped and observers of the recordings were more accurate at discriminating liars from truth tellers when eye contact had to be maintained, suggesting that doing so induces higher load and anxiety in liars than in truth tellers, perhaps because eye contact is distracting to liars who need to focus their attention inwardly to construct plausible deceptions.

One likely countermeasure, as knowledge of this load-inducing technique spreads, would be to practice lying while maintaining eye contact with another, which might reduce liar-truth teller differences. Even so, combined with other techniques, it may be useful in revealing malingering in forensic contexts.

Rather than heightening cognitive load through surprise or by imposing a concurrent task (e.g., maintaining eye contact), the two techniques described next (aIAT, TARA) add to cognitive load by creating response interference in deceivers by having them respond quickly and accurately to some items intermixed with others they may want to lie to. Such techniques also allow automated lie detection, not dependent on unreliable human observers.

Autobiographical Implicit Association Test (aIAT)

Based on the Implicit Association Test of Greenwald et al. (107), the aIAT is designed to determine whether respondents possess actual autobiographical memories, for instance, of a true alibi at the time of a crime. This computerized, forced choice assessment confronts respondents with five blocks of sentences to be classified (108). In block 1, respondents classify sentences with verifiable truths as true or false. In block 2, target sentences probing specific episodic truths (guilty if true; innocence if false) about them are likewise classified. Blocks 3 and 5 are crucial. In block 3, true and guilty sentences are intermixed and classified with the same response key. In block 5, true and innocent sentences are likewise classified together. An index, D, which penalizes for incorrect responses, is largely based on subtracting block 3 response times from those of block 5. Positive D scores are expected of guilty respondents, negative D score of innocent because of the interference in guilty respondents caused by the incongruence of combining truthful and innocent responses in block 5.

The aIAT has an impressive 91% accuracy rate in identifying those possessing genuine autobiographical memories (108) and has proven effective in uncovering the malingering of whiplash (109), and unveiling phantom limb pain (110). Clearly those genuinely affected by cognitive or psychiatric impairments should have many life memories of experienced symptomology that can be probed. Still, the aIAT has some limitations. It does not allow for ascertaining the truths of answers to specific or open-ended questions (e.g., When did you first notice your memory problems?). Also, research has not adequately explored whether countermeasures, such as deliberately slowing on some blocks and speeding up on others, could reduce deception detection (108). Another limitation, the aIAT requires the possession of true identity information that can be contrasted with faked identity information. In the case of those seeking to fake their identities in the field, such information may be unavailable to examiners (111).

The Timed Antagonistic Response Alethiometer (TARA)

Like the IAT and aIAT, TARA (112) involves a multi-block classification task. This computer-administered, response time-based method of lie detection assumes that, following instructions to minimize errors, incompatible tasks take longer to execute than compatible ones. Statements are presented on a computer screen that respondents must quickly classify as true or false. At first, control statements with verifiable truths (e.g., Rocks are hard. Mozart wrote novels.) are presented. In blocks that follow, target statements probing truths specific to the individual (I am male. I am a citizen of Egypt.) are presented. When target and control statements are combined within the same block, dishonest respondents experience response interference and the longest response times, having to perform the incompatible responses of deception and truthfulness. TARA correctly classified liars and truth tellers with an accuracy of 85%.

TARA differs from the aIAT in some important ways. TARA uses two categorizations (true, false) rather than four, uses only one critical block rather than two, and identifies lying from truths based on absolute RTs in the critical block. The latter requires comparison with a matched control group, a limitation of this technique (108). Still, TARA has potential to uncover a variety of deception types, including malingering. However, like the aIAT, it does not support the verification of open-ended responses or an answer to a particular question, nor has it been applied to detect deception involving a specific issue such as participation in a crime. Also, the effects on detection accuracy of the extensive practice of deception, deliberate slowing on certain blocks, or the use of other countermeasures are unknown.

Detecting Faked Identities With Unexpected Questions and Mouse Movements

The aIAT and TARA use key press response times to uncover deception. In order to discover faked identities in a way not reliant on possessing accurate identity information about respondents, Monaro et al. (111) explored the use of computer mouse movements in responding yes/no as the cues to deception in conjunction with asking unexpected questions. Measuring mouse movements allows a much richer set of behavioral cues, such as acceleration and trajectory, not easily controlled via countermeasures. Investigators assigned participants either to rehearse their true identities or rehearse then lie based on fake identities. Expected questions (i.e., concerning rehearsed information, such as birth month) and unexpected questions (e.g., one's Zodiac sign), were asked, the latter hypothesized to be constructed impromptu under high cognitive load. Detecting an impressive 95% accurately, fakers took longer, especially in responding to unexpected questions, and had longer response trajectories, among other differences.

Asking unexpected question, [see (113)] combined with mouse movements, has much potential, for instance, in detecting faked depression (114). However, would it be effective in uncovering malingerers who have faked depression or other psychiatric disorders for years? Also, the guidelines for generating unexpected questions are unclear. For example, a truth teller, not inclined toward superstition, might lack quick access in memory to their Zodiac sign. Verifying answers to open-ended questions is not possible as well. Even so, it is interesting to consider the kinds of unexpected questions that might blindside malingerers (e.g., Does your impairment affect you when driving?) and expose them.

Time-Restricted Integrity Confirmation (TRI-Con)

Walczyk et al. (115) proposed a cognitive load-inducing technique, *Time Restricted Integrity-Confirmation* (TRI-Con), with potential to uncover different kinds of deception including malingering. It is based on a cognitive theory of high-stakes deception called Activation-Decision-Construction-Action Theory (ADCAT), summarized later. Like the aIAT and TARA, TRI-Con can be largely automated via computer-administration and scoring and selectively enhances the cognitive load on liars by adhering to seven guidelines during lie detection examinations (115, 116).

The guidelines are: (a) Respondents are prompted about the focus of the question set to follow (e.g., "The next 15 questions concern your activities and whereabouts at the time of the crime."). By priming relevant episodic and semantic truths, prompts reduce respondents' need to search memory to tell the truth, making cognitive load indices clearer cues of when respondents are constructing lies. As with reviewing questions before a polygraph exam, prompting also reduces the emotional surprise accompanying blindsiding respondents with questions that probe sensitive issues. (b) Still, the specific questions are not disclosed until asked during an exam, thus surprising respondents cognitively and reducing the chance that deceptive answers were prepared and rehearsed. (c) Questions, both yes/no and open-ended, are written when possible to be unclear regarding what truths are sought until fully asked, which should reduce respondents' chance of preparing deceptive answers as questions are being asked. (d) To obtain clearer assessment of the cognitive load needed to answer completely, questions are written to be answerable, as much as possible, with one or a few words. (e) Respondents are instructed to answer as quickly as possible to discourage and expose attempts to deceive. The high cognitive load of rapid responding to surprise questions may also increase cue leakage in the form of voice pitch elevation, pupil dilation, increased blinking, and long response times because of the limited opportunity for liars to monitor and control their own behavior (75, 117, 118) and may increase accidental blurting of the truth (119). (f) Without adequate preparation, liars' deceptive accounts should be incomplete. Questions are asked and then re-asked, along with logically interrelated questions, to increase liars' cognitive load and provoke inconsistencies (120). (g) Behavioral baselines for ground-truth answers are established for all cognitive load indices for comparison with levels of these cues of answers suspected of deception. This practice controls for individual differences in behavioral base rates and improves the accuracy of lie detection (71).

Given the inaccuracy of human lie detectors (71, 72), automatable techniques of lie detection, such as TRI-Con, TARA, and the aIAT, provide auspicious alternatives. For instance, with TRI-Con questions can be recorded and asked by a computer. Using microphone-headsets, answer response times can be precisely measured to the millisecond level of precision. Connected modern eye-tracking systems can concurrently measure pupil dilation, eye movements/fixations, and blinking rate. Voice pitch elevation can be detected using the appropriate software, etc.

Following the guidelines above, studies have shown the effectiveness of TRI-Con for uncovering deceptive answers to yes/no and open-ended questions. Walczyk et al. (115) instructed adults to lie or tell the truth to questions about various aspects of their lives such as employment history and their performance on standardized tests. Using response time as the cue, discriminant analyses allowed classification of liars and truth tellers above chance. Likewise, Walczyk et al. (116) tested TRI-Con again by asking participants to lie or tell the truth about their lives and included a rehearsal condition in which participants prepared deceptive answers, a likely load-reducing countermeasure. The consistency of answers across interrelated questions was added as a cue. Liars and truth tellers were classified up to 89% accurately. Analyses also showed that the countermeasure of rehearsing deception is detectable. Also, Walczyk et al. (121) tested TRI-Con in a forensically-relevant context. "Witnesses" observed actual crime videos, then later told the truth or lied, rehearsed or unrehearsed, when interviewed about them. The cognitive cues were response time, answer consistency, eye movements, and pupil dilation. Discriminant analyses allowed classification of the three conditions 69% accurately, 33% expected by chance. Truth tellers generally had moderate response times, the fewest inconsistencies, and the most eye movements. Regarding the latter findings, liars appeared to move their eyes less to avoid visual distraction that would have heighten cognitive loads as they focused attention inwardly to construct lies. Walczyk et al. (122) observed similar results for participants who lied or told the truth concerning their participation in a mock crime. Across these studies, low rates of false positives were observed, recalling that high rates are a perennial problem with the CQT (81).

Although TRI-Con has potential for the detection of malingering, it too is susceptible to the countermeasure of rehearsal. The good news is that load-reducing techniques can be combined. TRI-Con already involves surprise questions. Respondents can be further instructed to maintain eye contact with someone present. Surprise drawing can be added after the exam to solicit non-verbal information. Other load-inducing techniques can be added. Combining several load-inducing techniques within lie detection exams and assessing several indices of cognitive load should make the detection of malingering hard to foil.

ACTIVATION-DECISION-CONSTRUCTION-ACTION THEORY (ADCAT)

A major criticism of the polygraph-based CQT is its lack of a valid theoretical foundation (80, 81). Similarly, most existent load-inducing techniques assume that lying is more cognitively demanding than truth telling. Our discussion of the countermeasure of rehearsing deception, however, suggests that this is not always true. No coherent theory underlies most of these techniques. TRI-Con is an exception, based on ADCAT, a theory of high-stakes deception. ADCAT, with some tweaking, might account for malingering. Such a theory, once validated, could suggest cues of when malingering has taken place and new ways of detecting it. The most recent version of ADCAT, Walczyk et al. (123), is summarized below, with an emphasis on its application to malingering.

ADCAT accounts for how individuals respond deceptively to solicitations of the truth, such as a question, under high stakes. A high stakes social context is one in which being honest with targets (those soliciting truths) would likely prove very costly to respondents in the non-attainment of goals important to them. High-stakes contexts include a perpetrator interrogated by a detective concerning an alibi or a psychiatrist assessing a sane perpetrator regarding his fitness to stand trial.

ADCAT specifies four psychological components involved in most instances of deception. Each elaborates on underlying cognitive processes. ADCAT incorporates established concepts of cognitive science, including *working memory* and *executive functioning* (123). Of central importance is *Theory of Mind* (ToM), which involves the inferences individuals make regarding the mental states of targets. First-order ToM inferences in deception entail the false beliefs that liars are trying to create in others (e.g., "I want this psychiatrist to believe that I cannot distinguish right from wrong."). More abstract and cognitively demanding second-order ToM inferences concern, for instance, malingerers' guesses of what targets will expect in them if their deceit is believed ("How should I behave and what should I say to come across as legally insane?," (124, 125). As noted, malingerers are often wrong in these guesses. Both types of inferences are heavily involved in all four components.

Activation Component

The first component of ADCAT, *Activation*, involves the retrieval of the truth following targets' solicitations of accurate information. For instance, a police detective might ask a perpetrator who is feigning memory loss whether she can remember even a small fragment regarding where she was when the crime occurred (123). Based on the social context and roles targets play, ToM inferences are made regarding why, for instance, the detective is seeking the information, to what use sharing the truth will be put, etc. (e.g., "This detective suspects me and wants to build a case to charge me."). Most truths are automatically activated by a question but occasionally must be searched for in LTM if they have not been accessed in a long time or may need to be newly constructed in WM, both of which can add to the cognitive load of truth telling.

Decision Component

Typically with the truth now active in WM (126, 127), the second component, *Decision*, will execute. It describes how respondents choose whether and then how to deceive. With the help of ToM inferences, respondents will first evaluate what the likely overall gain/loss is of sharing the truth vis-à-vis the non-attainment of important goals such as staying out of prison or maintaining their disability income. Such evaluations are made intuitively when deception is impromptu but can be more deliberate when high-stakes truth solicitations are anticipated. These calculations involve intuitively combining estimates of the likelihoods of salient outcomes with their *subjective utilities*, that is, the personal value of the outcomes to respondents (128). The more negative the expected overall loss, the more likely a deception will be considered (123). In such a case, one or more context-appropriate deceptions will be evaluated in terms of their overall likely gain/cost vis-à-vis their believability and how well each helps respondents to achieve their goals. Again, first- and second-order inferences are crucial to accurately evaluate the likely impact of deceptions on targets.

The deception with the highest expect gain, if any, will be chosen, which can vary from sharing a truth with an important detail withheld (lie of omission) to a bald-faced lie (complete fabrication). The preference for respondents will be to minimize the deception needed to attain their goals (129). The decision to deceive intrinsically adds to cognitive load (115), an implication of which is that surprising respondents with questions will require them to decide impromptu whether to lie, enhancing the mental work of deception and related cognitive cues.

Construction Component

During the third stage, *Construction*, the specific deception chosen is elaborated as needed to go undetected and achieve respondents' other goals. The cognitive load imposed varies with the type of deception. A false denial or a lie of omission can impose minimal load whereas constructing a bald-faced lie, for instance, a false alibi for what happened at the time of the crime, can impose the greatest. Especially for the latter, second-order ToM inferences must be made to ensure that a lie is internally consistent, consistent with what targets' know or are likely to find out, and detailed enough to be believable (123). A chance to prepare deceptions in advance of delivery will make them more believable, internally consistent, etc., and allow respondents to anticipate likely questions from targets (130, 131). A relevant question for the detection of malingering during this component concerns what kinds of ToM inferences do malingers typically make to mislead mental health professionals in forensic contexts. Little research has addressed this question. Asking surprise and complex questions of respondents suspected of malingering under TRI-Con concerning lesser known actual symptoms of disorders might trip up malingerers, producing long response times and other signs of cognitive load compared to those actually afflicted.

Central to the construction component is the *plausibility principle*, which specifies the order of steps respondents generally will take to construct believable deceptions, especially the bald-faced variety. Respondents will (a) first attempt to modify the truth, related episodic memories, or other personally experienced memories based on second-order ToM inferences of what targets will believe (102, 123, 129, 131). Because recently accessed memories are more retrievable, they will be preferred to distant memories (132). If respondents have no such memories, for instance, because malingerers have never actually suffered from a particular mental disorder, they may (b) use schemata or scripts of what is typical within that context to provide the basis of the deception (132–134). If such schemata are unavailable, again due to limited life experience or if relevant schemata are inaccessible, respondents will (c) construct deceptions using assorted information accessible from LTM as cued by the social context, which imposes the highest cognitive load. To summarize, the plausibility principle predicts that cognitive load will increase when going from *a* to *c* as the basis of a deception and lie plausibility will tend to decrease. However, the opportunity to prepare and rehearse deceptions, for example, a false presentation of being insane, in advance of delivery is a countermeasure that can lower the cognitive load experienced by liars, even below that of truth tellers (116, 135). On a positive note, the use of such rehearsal may be detectable by cognitive load indices falling below levels of truth telling (116).

Action Component

During the final component, *Action*, respondents deliver the lies they have prepared, or will generate impromptu, to targets. In general, they will attempt to control physical movements and appear relaxed but may self-regulate too much because of inaccurate ToM beliefs they hold about the actual behavior of truth tellers. Many liars naively implicitly assume that honest individuals are relaxed and do not experience recall failures or make other mistakes in conveying truths (72, 104). As noted, the cognitive load of delivery will decrease for well-rehearsed lies, but will increase when social contexts are unfamiliar and complex. Cognitive load will also increase when, for instance, malingerers are surprised by truth solicitations, which allows little time for lie preparation (115).

Because deception is typically chosen only when honesty blocks goal attainment, truth telling is usually more practiced and automatic (129). Especially for well-rehearsed truths, conveying corresponding deceptions can impose a cognitive load during delivery, requiring active suppression of the truth (100, 116, 136, 137). In addition to this source of cognitive load, lies told in high-stakes situations are highly motivated, which can heighten the fear of being caught as well as the cognitive load of delivery (138). ADCAT hypothesizes that impromptu liars will manage the increased load of deception by minimizing eye contact (106), reducing eye movements (122), reducing body movements, occasionally scanning the environment for lie construction hints, and implementing time-buying strategies like asking for a question to be repeated or pausing before and during delivery of the lie (123).

APPLYING COGNITIVE LOAD-INDUCING TECHNIQUE AND ADCAT TO DETECT MALINGERING

Only sketched above, ADCAT advances understanding of the behavioral manifestations of deception by providing a detailed cognitive account of the processes individuals engage in as they choose deception, construct lies, and deliver them to targets (123). Professionals interested in advancing the cognitive detection of malingering are encouraged to learn more about this and other cognitive accounts of deception [see (84, 139)]. Malingering is high-stakes deception in which malingerers must actively inhibit the truth (e.g., a lack of mental illness) and decide which deceptive presentation of behavior to construct and practice. Interestingly, constructing presentations of feigned mental disorders may be more cognitively complex than constructing, for instance, alibis based on complete fabrications. Second-order ToM inferences are likely extensively made as malingerers study the kinds of symptoms typical of those afflicted with particular disorders and how the disorders are assessed. ADCAT helps clarify when cognitive load-inducing approaches for detecting malingering are likely to be effective. For instance, ADCAT recognizes the preparation and rehearsal of high-stakes deception before delivery as the most likely foil of such approaches and recommends that respondents be blindsided with memory tasks for accessing truths. These include asking unanticipated and complex questions, soliciting surprise drawings, or accessing memories for events in unusual ways like recounting an alibi in reverse-chronological order (140). In such cases, the cognitive load of deception should exceed that of honesty. Surprisingly, most researchers who have tested load-inducing approaches have not given much attention to the countermeasure of rehearsal.

The customary methods for malingering detection we reviewed rely on atypical levels or combinations of symptoms, unusual performance on cognitive tasks, or other behavioral anomalies. Sadly, their rates of false positives and false negatives tend to be high. As an alternative, we encourage those interested in detecting malingering in forensic contexts to consider combining several cognitive load-inducing approaches like those we discussed. For instance, TRI-Con automates many aspects of lie detection, involves surprise questions, and can include maintaining eye contact and other load-inducing techniques. The non-load-inducing cognitive methods of lie detection of Criteria Based Content Analysis (CBCA) and Reality Monitoring (RM) can be added as well, which assume that liars fabricate information when constructing lies (73, 141). Both attempt to differentiate memories of real experiences from fabrications by assessing for features of authentic experiences such as sensory details, the reporting of unexpected complications, thoughts or feelings experienced, contextual information, temporal details, and affective information (98, 141). Under TRI-Con, asking respondents surprise and complex questions about details like these related to their disorders, their time of onset, or how they made respondents feel might expose significantly higher cognitive loads in malingerers than in genuine patients as longer response times, elevated pitch, dilated pupils, less eye movement, or as slower and longer mouse movements (111). In conclusion, the more that malingering is understood cognitively, the more that innovative methods of lie deception detection like TARA, the aIAT, and TRI-Con can be refined to supplement existing assessments.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

1. Sartori G, Orrù G, Zangrossi A. Detection of malingering in personal injury and damage ascertainment. In: Ferrara S, Boscolo-Berto R, Viel G, editors. *Personal Injury and Damage Ascertainment under Civil Law*. Cham: Springer (2016).
2. Sartori G, Zangrossi A, Orrù G, Monaro M. Detection of malingering in psychic damage ascertainment. In: Ferrara SD, editor. *P5 Medicine and Justice*. Cham: Springer (2017). p. 330–341.
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Washington, DC: American Psychiatric Publishing (2013).
4. McCaffrey RJ, Weber M. A clinical approach to evaluating malingering in forensic neuropsychological evaluations. *Span J Neuropsychol.* (2000) 2:21–36.
5. Resnick PJ, Knoll JL. Malingered psychosis. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford (2008). p. 34–53.

6. Rogers R, Shuman DW. *Fundamentals of Forensic Practice: Mental Health and Criminal Law*. New York, NY: Springer (2005). Available online at: http://link.springer.com/chapter/10.1007/0-387-25227-4_2
7. Reynolds CR, Horton AM. Clinical acumen, common sense, and data-based decision making in the assessment of dissimulation during head injury litigation. In: Reynolds CR, Horton AM, editors. *Detection of Malingering During Head Injury Litigation.* New York, NY: Springer (2012). p. 351–69. Available online at: http://0-dx.doi.org.dewey2.library.denison.edu/10.1007/978-1-4614-0442-2
8. Adetunji BA, Basil B, Mathews M, Williams A, Osinowo T, Oladinni O. Detection and management of malingering in a clinical setting. *Primary Psychiatr.* (2006) 13:61.
9. Soliman S, Resnick PJ. Feigning in adjudicative competence evaluations. *Behav Sci Law* (2010) 28:614–29. doi: 10.1002/bsl.950
10. Denney RL. Assessment of malingering in criminal forensic neuropsychological settings. In: Boone KB, editor. *Assessment of Feigned Cognitive Impairment: A Neuropsychological Perspective*. New York, NY: Guilford Press (2007). p. 428–50. Available online at: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10201023
11. Reid WH. Malingering. *J Psychiatr Pract.* (2000) 6:226–8. doi: 10.1097/00131746-200007000-00008
12. McDermott BE, Sokolov G. Malingering in a correctional setting: the use of the structured interview of reported symptoms in a jail sample. *Behav Sci Law* (2009) 27:753–65. doi: 10.1002/bsl.892
13. Heilbronner RL, Sweet JJ, Morgan JE, Larrabee GJ, Millis SR, Conference Participants. American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *Clin Neuropsychol.* (2009) 23:1093–129. doi: 10.1080/13854040903155063
14. Kane AW. Psychology, causality, and court. In: Young G, Nicholson K, Kane AW, editors. *Psychological Knowledge in Court*. Boston, MA: Springer (2006). p. 13–51. Available online at: http://link.springer.com/chapter/10.1007/0-387-25610-5_2
15. Larrabee GJ, Berry DTR. Diagnostic classification statistics and diagnostic validity of malingering assessment. In: Larrabee GJ, editor. *Assessment of Malingered Neuropsychological Deficits*. New York, NY: Oxford University Press (2007). p. 14–26.
16. Millis SR, Putnam SH, Adams KM, Ricker JH. The california verbal learning test in the detection of incomplete effort in neuropsychological evaluation. *Psychol Assessment* (1995) 7:463. doi: 10.1037/1040-3590.7.4.463
17. Berry DTR, Schipper LJ. Detection of feigned psychiatric symptoms during forensic neuropsychological examinations. In: Larrabee GJ, editor. *Assessment of Malingered Neuropsychological Deficits*. New York, NY: Oxford University Press (2007). p. 226–63.
18. Rosenfeld B, Sands SA, Van Gorp WG. Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Arch Clin Neuropsychol.* (2000) 15:349–59. doi: 10.1093/arclin/15.4.349
19. Rogers R. Structured interviews and dissimulation. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford (2008). p. 301–22.
20. Rogers R, Bender SD. Evaluation of malingering and related response styles. In: Weiner IB, Otto RK, editors. *Handbook of Psychology, Vol. 11: Forensic Psychology*. 2nd ed. New Jersey, NJ: John Wiley and Sons (2013). p. 517–40.
21. Rogers R, Correa AA. Determinations of malingering: evolution from case-based methods to detection strategies. *Psychiatr Psychol Law* (2008) 15:213–23. doi: 10.1080/13218710802014501
22. Rogers R, Gillard ND, Berry DTR, Granacher RPJ. Effectiveness of the MMPI-2-RF validity scales for feigned mental disorders and cognitive impairment: a known-groups study. *J Psychopathol Behav Assessment* (2011) 33:355–67. doi: 10.1007/s10862-011-9222-0
23. Rogers R, Harrell EH, Liff CD. Feigning neuropsychological impairment: a critical review of methodological and clinical considerations. *Clin Psychol Rev.* (1993) 13:255–74. doi: 10.1016/0272-7358(93)90023-F
24. Rogers R. Detection strategies for malingering and defensiveness. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford (2008). p. 14–35.
25. Heilbronner RL. Structured Interview of Reported Symptoms (SIRS). In: Kreutzer JS, DeLuca J, Caplan B, editors. *Encyclopedia of Clinical Neuropsychology*. New York, NY: Springer (2011). p. 2417–8. Available online at: http://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_848
26. Singh J, Avasthi A, Grover S. Malingering of psychiatric disorders: a review. *German J Psychiatr.* (2007) 10:126–32.
27. Edens JF, Poythress NG, Watkins-Clay MM. Detection of malingering in psychiatric unit and general population prison inmates: a comparison of the PAI, MS, and SIRS. *J Personal Assessment* (2007) 88:33–42. doi: 10.1080/00223890709336832
28. Rogers R, Gillis JR, Bagby RM, Monteiro E. Detection of malingering on the Structured Interview of Reported Symptoms (SIRS). *Psychol Assessment* (1991) 3:673–7. doi: 10.1037/1040-3590.3.4.673
29. Seron X. Lying in neuropsychology. *Clin Neurophysiol.* (2014) 44:389–403. doi: 10.1016/j.neucli.2014.04.002
30. Smith GP, Burger GK. Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). *J Am Acad Psychiatr Law Online* (1997) 25:183–9.
31. Ben-Porath YS. *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press (2012).
32. Butcher JN, Williams CL. Personality assessment with the MMPI-2: historical roots, international adaptations, and current challenges. *Appl Psychol Health Well-Being* (2009) 1:105–35. doi: 10.1111/j.1758-0854.2008.01007.x
33. Greene RL. Malingering and defensiveness on the MMPI-2. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford (2008). p. 159–81.
34. Heinze MC, Purisch AD. Beneath the mask: use of psychological tests to detect and subtype malingering in criminal defendants. *J Foren Psychol Pract.* (2001) 1:23. doi: 10.1300/J158v01n04_02
35. Pope KS, Butcher JN, Seelen J. *The MMPI, MMPI-2 and MMPI-A in Court: A Practical Guide for Expert Witnesses and Attorneys*. 3rd ed. Washington, DC: American Psychological Association (2006).
36. Rogers R, Neumann CS. Conceptual issues and explanatory models of malingering. In: Halligan PW, Bass CM, Oakley DA, editors. *Malingering and Illness Deception*. New York, NY: Oxford University Press (2003). p. 71–82.
37. Berry DTR, Baer RA, Harris MJ. Detection of malingering on the MMPI: a meta-analysis. *Clin Psychol Rev.* (1991) 11:585–98. doi: 10.1016/0272-7358(91)90005-F
38. Rogers R, Sewell KW, Goldstein AM. Explanatory models of malingering. *Law Hum Behav.* (1994) 18:543–52. doi: 10.1007/BF01499173
39. Greene RL. *The MMPI-2: An Interpretive Manual*. Boston, MA: Allyn and Bacon (2000).
40. Pelfrey WV. The relationship between malingerers' intelligence and MMPI-2 knowledge and their ability to avoid detection. *Int J Offend Ther Comp Criminol.* (2004) 48:649–63. doi: 10.1177/0306624X04265085
41. Raine M. Helping advocates to understand the psychological diagnosis and assessment of malingering. *Psychiatr Psychol Law* (2009) 16:322–8. doi: 10.1080/13218710802389457
42. Heinze MC. Developing sensitivity to distortion: utility of psychological tests in differentiating malingering and psychopathology in criminal defendants. *J Foren Psychiatr Psychol.* (2003) 14:151–77. doi: 10.1080/1478994031000077961
43. Millon T, Millon C, Davis RD, Grossman S. *Millon Clinical Multiaxial Inventory-III (MCMI-III): Manual*. Upper Saddle River, NJ: Pearson/PsychCorp (2009).
44. Miller HA. *M-FAST: Miller-Forensic Assessment of Symptoms Test Professional Manual*. Odessa, FL: Psychological Assessment Resources (2001).
45. Guriel J, Yañez T, Fremouw W, Shreve-Neiger A, Ware L, Filcheck H, et al. Impact of coaching on malingered postraumatic stress symptoms on the M-FAST and the TSI. *J Foren Psychol Pract.* (2004) 4:37–56. doi: 10.1300/J158v04n02_02
46. Miller HA. The Miller-Forensic Assessment of Symptoms Test (M-FAST) test generalizability and utility across race literacy, and clinical opinion. *Criminal Justice Behav.* (2005) 32:591–611. doi: 10.1177/0093854805278805
47. Smith GP. Brief screening measures for the detection of feigned psychopathology. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford (2008). p. 323–39.

48. Guy LS, Miller HA. Screening for malingered psychopathology in a correctional setting utility of the Miller-Forensic Assessment of Symptoms Test (M-FAST). *Criminal Justice Behav.* (2004) 31:695–716. doi: 10.1177/0093854804268754
49. Vitacco MJ, Rogers R, Gabel J, Munizza J. An evaluation of malingering screens with competency to stand trial patients: a known-groups comparison. *Law Hum Behav.* (2007) 31:249–60. doi: 10.1007/s10979-006-9062-8
50. Miller HA. Examining the use of the M-FAST with criminal defendants incompetent to stand trial. *Int J Offend Ther Comp Criminol.* (2004) 48:268–80. doi: 10.1177/0306624X03259167
51. Rogers R. Current status of clinical methods. In: Rogers R, editor. *Clinical Assessment of Malingering and Deception*. 3rd ed. New York, NY: Guilford Press (2012). p. 391–410.
52. Young G. *Malingering, Feigning, and Response Bias in Psychiatric/Psychological Injury: Implications for Practice and Court*. Vol. 56. Dordrecht: Springer (2014). Available online at: http://link.springer.com/10.1007/978-94-007-7899-3
53. Ashendorf L, Constantinou M, McCaffrey RJ. The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Arch Clin Neuropsychol.* (2004) 19:125–30. doi: 10.1016/S0887-6177(02)00218-4
54. Lezak MD, Howieson DB, Bigler ED, Tranel D. *Neuropsychological Assessment*. 5th ed. OUP. New York, NY (2012).
55. Grote CL, Hook JN. Forced-choice recognition tests of malingering. In: Larrabee GJ, editor. *Assessment of Malingered Neuropsychological Deficits*. New York, NY: Oxford University Press (2007). p. 44–79.
56. Tombaugh TN. The Test of Memory Malingering (TOMM): normative data from cognitively intact and cognitively impaired individuals. *Psychol Assessment* (1997) 9:260–8. doi: 10.1037/1040-3590.9.3.260
57. Tan JE, Slick DJ, Strauss E, Hultsch DF. How'd they do it? *Malingering strategies on symptom validity tests Clin Neuropsychol.* (2002) 16:495–505. doi: 10.1076/clin.16.4.495.13909
58. Guidotti Breting LM, Sweet JJ. Freestanding cognitive symptom validity tests: use and selection in mild traumatic brain injury. In: Bush SS, Carone DA, editors. *Mild Traumatic Brain Injury: Symptom Validity Assessment and Malingering*. New York, NY: Springer (2013). Co. p. 145–57. Available online at: http://ezproxy.uvu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=trueanddb=e025xnaandAN=477408andsite=eds-live
59. Pella RD, Hill BD, Singh AN, Hayes JS, Gouvier WD. Noncredible performance in mild traumatic brain injury. In: Reynolds CR, Horton AM, editors. *Detection of Malingering During Head Injury Litigation*. Boston, MA: Springer (2012). p. 121–50. Available online at: http://link.springer.com/10.1007/978-1-4614-0442-2_3
60. Frederick RI. A review of Rey's strategies for detecting malingered neuropsychological impairment. *J Foren Neuropsychol.* (2002) 2:1–25. doi: 10.1300/j151v02n03_01
61. Williams JM, Jones K. Factitious responding and malingered memory disorder. In: Reynolds CR, Horton AM, editors. *Detection of Malingering During Head Injury Litigation*. New York, NY: Springer (2012). p. 169–99. Available online at: http://0-dx.doi.org.dewey2.library.denison.edu/10.1007/978-1-4614-0442-2_5
62. Arnett PA, Hammeke TA, Schwartz L. Quantitative and qualitative performance on Rey's 15-item test in neurological patients and dissimulators. *Clin Neuropsychol.* (1995) 9:17–26. doi: 10.1080/13854049508402052
63. Schretlen D, Brandt J, Krafft L, Van Gorp W. Some caveats in using the Rey 15-Item Memory Test to detect malingered amnesia. *Psychol Assessment* (1991) 3:667–72. doi: 10.1037/1040-3590.3.4.667
64. Strauss E, Sherman EMS, Spreen O. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Oxford: Oxford University Press (2006).
65. Vallabhajosula B, Van Gorp W. Post-Daubert admissibility of scientific evidence on malingering of cognitive deficits. *Arch Clin Neuropsychol.* (2000) 15:847. doi: 10.1093/arclin/15.8.847
66. Vickery CD, Berry DTR, Inman TH, Harris MJ, Orey SA. Detection of inadequate effort on neuropsychological testing. *Arch Clin Neuropsychol.* (2001) 16:45–73. doi: 10.1016/s0887-6177(99)00058-x
67. Green P, Lees-Haley PR, Allen LM. The Word Memory Test and the validity of neuropsychological test scores. *J Foren Neuropsychol.* (2002) 2:97–124. doi: 10.1300/J151v02n03_05
68. Dunn TM, Shear PK, Howe S, Ris MD. Detecting neuropsychological malingering: effects of coaching and information. *Arch Clin Neuropsychol.* (2003) 18:121–34. doi: 10.1016/S0887-6177(01)00188-3
69. DePaulo BM, Kashy DA. Everyday lies in close and casual relationships. *J Personal Soc Psychol.* (1998) 74:63–79. doi: 10.1037/0022-3514.74.1.63
70. DePaulo BM, Kashy DA, Kirkendol SE, Wyer MM, Epstein JA. Lying in everyday life. *J Personal Soc Psychol.* (1996) 70:979–95. doi: 10.1037/0022-3514.70.5.979
71. Bond CF, DePaulo BM. Accuracy of deception judgments. *Personal Soc Psychol Rev.* (2006) 10:214–34. doi: 10.1207/s15327957pspr1003_2
72. Bond CF Jr, DePaulo BM. Individual differences in judging deception: accuracy and bias. *Psychol Bull.* (2008) 134:477–92. doi: 10.1037/0033-2909.134.4.477
73. Vrij A. *Detecting Lies and Deceit: Pitfalls and Opportunities*. 2nd ed. Hoboken, NJ: Wiley (2008).
74. Vrij A, Mann S. Telling and detecting lies in a high-stake situation: the case of a convicted murderer. *Appl Cogn Psychol.* (2001) 15:187–203.
75. Zuckerman M, DePaulo BM, Rosenthal R. Verbal and nonverbal communication of deception. In: Berkowitz L, editor. *Advances in Experimental Social Psychology*. Vol. 14. New York, NY: Academic Press (1981). p. 1–59.
76. Bull, P. (2009). Detecting deceit: current issues. *Int Dev Investig Interview* 12:190–206. doi: 10.1350/ijps.2010.12.1.167
77. Ekman P, O'Sullivan M. Who can catch a liar? *Am Psychol.* (1991) 46:913–20.
78. Nicholson K, Martelli MF. Malingering: overview and basic concepts. In: Young G, Kane AW, Nicholson K, editors. *Causality of Psychological Injury: Presenting Evidence in Court*. New York, NY: Springer Science+Business Media, LLC (2007). p. 375–409.
79. Vrij A, Mann S. Who killed my relative? Police officers' ability to detect real-life high-stakes lies. *Psychol Crime Law* (2001) 7:119–32. doi: 10.1080/10683160108401791
80. Lykken DT. *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. New York, NY: McGraw-Hill (1998).
81. National Research Council. *The Polygraph And Lie Detection. Committee to Review the Scientific Evidence on the Polygraph.* Washington, DC: National Academies Press (2003). Available online at: http://site.ebrary.com/id/10032342
82. Boaz TL, Perry NW, Raney G, Fischler IS, Shuman D. Detection of guilty knowledge with event-related potentials. *J Appl Psychol.* (1991) 76:788–95. doi: 10.1037/0021-9010.76.6.788
83. Ekman P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY: W.W. Norton (1992).
84. Gombos VA. The cognition of deception: the role of executive processes in producing lies. *Genet Soc Gen Psychol Monogr.* (2006) 132:197–214. doi: 10.3200/MONO.132.3.197-214
85. Elaad E. Detection of guilty knowledge in real-life criminal investigations. *J Appl Psychol.* (1990) 75:521–9. doi: 10.1037/0021-9010.75.5.521
86. Seymour TL, Fraynt BR. Time and encoding effects in the concealed knowledge test. *Appl Psychophysiol Feedback* (2009) 34:177–87. doi: 10.1007/s10484-009-9092-3
87. Seymour TL, Kerlin JR. Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. *Appl Cogn Psychol.* (2008) 22:475–90. doi: 10.1002/acp.1375
88. Seymour TL, Seifert CM, Shafto MG, Mosmann AL. Using response time to measures to assess "guilty knowledge." *J Appl Psychol.* (2000) 85:30–7. doi: 10.1037/0021-9010.85.1.30
89. Varga M, Visu-Petra L, Miclea M, Buş I. The RT-based concealed information test: an overview of current research and future perspectives. *Procedia Soc Behav Sci.* (2014) 127:681–5. doi: 10.1016/j.sbspro.2014.03.335
90. Verschuere B, Crombez G, Degrootte T, Rosseel Y. Detecting concealed information with reaction times: validity and comparison with the polygraph. *Appl Cogn Psychol.* (2010) 24:991–1002. doi: 10.1002/acp.1601
91. Rosenfeld J. *Detecting Concealed Information and Deception: Recent Developments*. London: Academic Press (2018).

92. Rosenfeld JP, Biroschak JR, Furedy JJ. P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. *Int J Psychophysiol.* (2006) 60:251–9. doi: 10.1016/j.ijpsycho.2005.06.002

93. Rosenfeld JP, Cantwell B, Nasman VT, Wojdac V, Ivanova S, Mazzeri L. A modified, event-related potential-based guilty knowledge test. *Int J Neurosci.* (1988) 24:157–61. doi: 10.3109/00207458808985770

94. Dionisio DP, Granholm E, Hillix WA, Perrine WF. Differentiation of deception using pupillary response as an index of cognitive processing. *Psychophysiology* (2001) 38:205–11. doi: 10.1111/1469-8986.3820205

95. Allen JJ. Clinical applications of the concealed knowledge test. In: Verschuere B, Ben-Shakhar G, Meijer E, editors. *Memory Detection: Theory and Application of the Concealed Information Test*. Cambridge, UK: Cambridge University Press (2011). p. 231–52.

96. Sweller J. Cognitive load during problem solving: effects on learning. *Cogn Sci.* (1988) 12:257–85. doi: 10.1016/0364-0213(88)90023-7

97. van Merriënboer JJG, Sweller J. Cognitive load theory and complex learning: recent developments and future directions. *Edu Psychol Rev.* (2005) 17:147–77. doi: 10.1007/s10648-005-3951-0

98. Vrij A, Mann S. Criteria-based content analysis: an empirical test of its underlying processes. *Psychol Crime Law* (2006) 12:337–49. doi: 10.1080/10683160500129007

99. Langleben DD, Loughead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, et al. Telling truth from lie in individual subjects with fast event-related fMRI. *Hum Brain Mapping* (2005) 26:262–72. doi: 10.1002/hbm.20191

100. Mohamed FB, Faro SH, Gordon NJ, Platek SM, Ahmad H, Williams JM. Brain mapping of deception and truth telling about an ecologically valid situation: functional MR imaging and polygraph investigation—initial experience. *Radiology* (2006) 238:679–88. doi: 10.1148/radiol.2382050237

101. Vrij A, Fisher R, Mann S, Leal S. Detection of deception: cognitive load. In: Cutler B, editor. *Encyclopedia of Psychology and Law*. Thousand Oaks, CA: SAGE Publications (2008). p. 195–6. Available online at: http://eprints.port.ac.uk/11201/

102. Walczyk JJ, Roper KS, Seemann E, Humphrey AM. Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Appl Cogn Psychol.* (2003) 17:755. doi: 10.1002/acp.914

103. Vrij A, Leal S, Granhag PA, Mann S, Fisher RP, Hillman J, et al. Outsmarting the liars: the benefit of asking unanticipated questions. *Law Hum Behav.* (2009) 33:159–66. doi: 10.1007/s10979-008-9143-y

104. DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Harris C. Cues to deception. *Psychol Bull.* (2003) 129:74–112. doi: 10.1037/0033-2909.129.1.74

105. Loftus EF. Memory distortions: problems solved and unsolved. In: Garry M, Hayne H, editors. *Do Justice and Let the Sky Fall: Elizabeth Loftus and Her Contributions to Science, Law, and Academic Freedom*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers (2007). p. 1–14.

106. Vrij A, Mann S, Leal S, Fisher R. "Look into my eyes": can an instruction to maintain eye contact facilitate lie detection? *Psychol Crime Law* (2010) 16:327–48. doi: 10.1080/10683160902740633

107. Greenwald AG, McGhee DE, Schwartz JKL. Measuring individual differences in implicit cognition: the implicit association test. *J Personal Soc Psychol.* (1998) 74:1464–80. doi: 10.1037/0022-3514.74.6.1464

108. Sartori G, Agosta S, Zogmaister C, Ferrara SD, Castiello U. How to accurately detect autobiographical events. *Psychol Sci.* (2008) 9:772–80. doi: 10.1111/j.1467-9280.2008.02156.x

109. Sartori G, Agosta S, Gnoato F. High accuracy detection of malingered whiplash syndrome. *Presentation at the International Whiplash Trauma Congress*. Miami, FL (2007).

110. Ferrara SD, Ananian V, Baccino E, Boscolo–Berto R, Domenici R, Hernàndez-Cueto C, et al. A novel methodology for the objective ascertainment of psychic and existential damage. *Int J Legal Med.* (2016) 130:1387–99. doi: 10.1007/s00414-016-1366-8

111. Monaro M, Gamberini L, Sartori G. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE* (2017) 12:e0177851. doi: 10.1371/journal.pone.0177851

112. Gregg A. When vying reveals lying: the timed antagonistic response alethiometer. *Appl Cogn Psychol.* (2007) 21:621–47. doi: 10.1002/acp.1298

113. Monaro M, Gamberini L, Zecchinato F, Sartori G. False identity detection using complex sentences. *Front Psychol.* (2018) 9:283. doi: 10.3389/fpsyg.2018.00283

114. Monaro M, Toncini A, Ferracuti S, Tessari G, Vaccaro MG, De Fazio P, et al. The detection of malingering: a new tool to identify made up depression. *Front Psychiatr.* (2018) 9:249. doi: 10.3389/fpsyt.2018.00249

115. Walczyk JJ, Schwartz JP, Clifton R, Adams B, Wei M, Zha P. Lying person-to-person about life events: a cognitive framework for lie detection. *Personnel Psychol.* (2005) 58:141–70. doi: 10.1111/j.1744-6570.2005.00484.x

116. Walczyk JJ, Mahoney K, Doverspike D, Griffith-Ross D. Cognitive lie detection: response time and consistency of answers as cues to deception. *J Business Psychol.* (2009) 24:33–49. doi: 10.1007/s10869-009-9090-8

117. Buller DB, Burgoon JK. Interpersonal deception theory. *Commun Theory* (1996) 6:203–42. doi: 10.1111/j.1468-2885.1996.tb00127.x

118. Burgoon JK, Buller DB. Interpersonal deception theory. In: Baxter LA, Braithewaite DO, Dawn O, editors. *Engaging Theories in Interpersonal Communication: Multiple Perspectives*. Thousand Oaks, CA: Sage Publications (2008). p. 227-39.

119. Lane JD, Wegner DM. The cognitive consequences of secrecy. *J Personal Soc Psychol.* (1995) 69:237–53. doi: 10.1037/0022-3514.69.2.237

120. Granhag PA, Hartwig M. A new theoretical perspective on deception detection: on the psychology of instrumental mind-reading. *Psychol Crime Law* (2008)14:189–200. doi: 10.1080/10683160701645181

121. Walczyk JJ, Griffith DA, Yates R, Visconte S, Simoneaux B, Harris LL. Lie detection by inducing cognitive load: eye movements and other cues to the false answers of "witnesses" to crimes. *Criminal Justice Behav.* (2012) 39:887–909. doi: 10.1177/0093854812437014

122. Walczyk JJ, Griffith DA, Yates R, Visconte S, Simoneaux B. Eye movements and other cognitive cues to rehearsed and unrehearsed deception when interrogated about a mock crime. *Appl Psychol Criminal Justice* (2013) 9:1.

123. Walczyk JJ, Harris LL, Duck TK, Mulay D. A social-cognitive framework for understanding serious lies: activation-decision-construction-action theory. *New Ideas Psychol.* (2014) 34:22–36. doi: 10.1016/j.newideapsych.2014.03.001

124. Apperly IA, Back E, Samson D, France L. The cost of thinking about false beliefs: evidence from adults' performance on a non-inferential theory of mind task. *Cognition* (2008) 106:1093–108. doi: 10.1016/j.cognition.2007.05.005

125. Kobayakawa M, Tsuruya N, Kawamura M. Theory of mind impairment in adult-onset myotonic dystrophy type 1. *Neurosci Res.* (2012) 72:341–6. doi: 10.1016/j.neures.2012.01.005

126. Baddeley A. Exploring the central executive. *Q J Exp Psychol.* (1996) 49:5–28. doi: 10.1080/713755608

127. Baddeley A. The episodic buffer: a new component of working memory? *Trends Cogn Sci.* (2000) 4:417–23. doi: 10.1016/S1364-6613(00)01538-2

128. Stanovich KE. *Decision Making and Rationality in the Modern World*. New York, NY: Oxford University Press (2010).

129. Levine TR, Kim RK, Hamel LM. People lie for a reason: three experiments documenting the principle of veracity. *Commun Res Rep.* (2010) 27:271–85. doi: 10.1080/08824096.2010.496334

130. Strömwall LA, Hartwig M, Granhag PA. To act truthfully: nonverbal behavior and strategies during a police interrogation. *Psychol Crime Law* (2006) 12:207–19. doi: 10.1080/10683160512331331328

131. Strömwall LA, Willén RM. Inside criminal minds: offenders' strategies when lying. *J Investig Psychol Offend Profil.* (2011) 8:271–81. doi: 10.1002/jip.148

132. Leins DA, Fisher RP, Ross SJ. Exploring liars' strategies for creating deceptive reports. *Legal Criminol Psychol.* (2013) 18:141–51. doi: 10.1111/j.2044-8333.2011.02041.x

133. Sporer SL, Schwandt B. Paraverbal indicators of deception: a meta-analytic synthesis. *Appl Cogn Psychol.* (2006) 20:421–46. doi: 10.1002/acp.1190

134. Sporer SL, Schwandt B. Moderators of nonverbal indicators of deception: a meta-analytic synthesis. *Psychol Public Policy Law* (2007) 13:1–34. doi: 10.1037/1076-8971.13.1.1

135. O'Hair HD, Cody MJ, McLaughlin ML. Prepared lies, spontaneous lies, machiavellianism, and nonverbal communication. *Hum Commun Res.* (1981) 7:325–39. doi: 10.1111/j.1468-2958.1981.tb00579.x

136. Osman M, Channon S, Fitzpatrick S. Does the truth interfere with our ability to deceive? *Psychonom Bull Rev*. (2009) 16:901–6. doi: 10.3758/PBR.16.5.901

137. Pennebaker JW, Chew CH. Behavioral inhibition and electodermal activity during deception. *J Personal Soc Psychol.* (1985) 49:1427–33. doi: 10.1037/0022-3514.49.5.1427

138. Ekman P, Frank MG. Lies that fail. In: Lewis M, Saarni C, editors. *Lying and Deception in Everyday Life*. New York, NY: Guilford Press (1993). p. 184–200.

139. McCornack SA, Morrison K, Paik JE, Wisner AM, Zhu X. Information manipulation theory 2: a propositional theory of deceptive discourse production. *J Language Soc Psychol.* (2014) 33:348–77. doi: 10.1177/0261927X14534656

140. Vrij A, Mann S, Fisher R, Leal S, Milne B, Bull R. Increasing cognitive load to facilitate lie detection: the benefit of recalling an event in reverse order. *Law Hum Behav*. (2008) 32:253–65. doi: 10.1007/s10979-007-9103-y

141. Johnson MK, Raye CL. Reality monitoring. *Psychol Rev*. (1981) 88:67–85. doi: 10.1037/0033-295X.88.1.67

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

OPINION
published: 17 January 2019
doi: 10.3389/fpsyt.2018.00774

Separating the Wheat From the Chaff: Guidance From New Technologies for Detecting Deception in the Courtroom

*Judee K. Burgoon**

Center for the Management of Information, University of Arizona, Tucson, AZ, United States

Keywords: lying, deception detection, linguistics, multimodal deception, nonverbal deception

OPEN ACCESS

Edited by:
Cristina Scarpazza,
University of Padova, Italy

Reviewed by:
Xunbing Shen,
Jiangxi University of Traditional Chinese Medicine, China
Pasquale Capuozzo,
University of Padova, Italy

***Correspondence:**
Judee K. Burgoon
judee@email.arizona.edu

Specialty section:
This article was submitted to Forensic Psychiatry, a section of the journal Frontiers in Psychiatry

Received: *13 September 2018*
Accepted: *24 December 2018*
Published: *17 January 2019*

Citation:
Burgoon JK (2019) Separating the Wheat From the Chaff: Guidance From New Technologies for Detecting Deception in the Courtroom. Front. Psychiatry 9:774. doi: 10.3389/fpsyt.2018.00774

The courtroom is among the most challenging contexts for detecting deception. Testimony has been carefully scripted and rehearsed in advance. Witnesses may proffer answers that create rather than reduce ambiguity. Questioning may skew attention away from a defendant's transgressions. Character testimony may malign opposition witnesses while painting a sanitized picture of the defendant. Amid so many inconsistent depictions, facts and opinions, jurists face a significant challenge in separating valid wheat from invalid chaff.

Jurists currently must render decisions unaided by the latest lie detection technologies such as fMRI (1), which places respondents in a loud, magnetized tube; EEG, ERP or fNIRS (2, 3), which connect wires to respondents' head or hands to detect brain waves; computer vision techniques that extract facial expressions from videotapes (4); instruments that discern voice pitch, tempo and fluency from audio recordings (5); or software that identifies linguistic patterns (6). Or, questioning techniques like the Concealed Information Test (7) and the autobiographical Implicit Association Test (8), in which respondents are questioned about aspects of a crime while their response latency is gauged. Jurists must rely on their own observational acumen, what they see and hear.

Nevertheless, we can learn what these technologies and techniques have unearthed that is applicable to courtroom deceit, focusing especially on indicators that prevaricators are less likely to attend to or control. Here deceit references the whole gamut of what is said, not said, and how it is said, both non-verbally and verbally. There is no silver bullet, no single indicator, that will invariably gauge a speaker's veracity (9), but by taking a holistic approach that bundles indicators together (10) and combines them across modalities (11) and by looking for temporal changes from baseline to later responding (12), it is possible to improve detection accuracy over that of the unaided human jurist (13).

NON-VERBAL SIGNALS

The various non-verbal indicators of deceit can be grouped according to what they signify: (1) arousal and emotional activation, (2) cognitive difficulty, (3) memory access, (4) attempted control of unbidden behavior, and (5) self-presentation (14, 15).

Emotions and Arousal

The first avenue of spotting telltale signs has been to look for outward signs of anxiety, fear, shame and other negative emotions (16). These are thought to be involuntary and uncontrollable or uncontrolled autonomic responses. Microexpressions of emotions such as contempt have been touted as reliable (17). But, among the many shortcomings of microexpressions is that they are not readily observable at normal courtroom interaction distances and are extremely infrequent [see

(18)]. Better to watch for *macro*-expressions (19), which can leak feigned sadness or inappropriately felt happiness, especially during high-intensity fear (20, 21). However, because people work to manage their facial expressions, these are often not the best place to look.

More helpful are some indicators of arousal [(22); but cf. (4)]. Subtle behaviors like restive foot movements and postural shifts reveal unease but may not be visible when suspects and witnesses take the stand. Close up one might be able to observe blinks and pupil dilation, which are tied to arousal (23). But, absent videotaped recordings available for slow-motion review, these would be difficult to spot in the courtroom. More visible are what I have labeled face-adaptors and lip-adaptors—behaviors that function to alleviate discomfort. The former are things like rubbing one's check or neck and twisting hair. The latter are lip movements such as biting, licking, scrunching and tongue-showing that are associated with states of nervousness or serious concentration. These are less likely to be controlled by liars.

Also telling are vocal indicators: higher voice pitch, increased vocal tension, and more hesitations and speech errors (24). The fallacy of relying too heavily on arousal and emotion indicators is that such arousal behaviors are not associated exclusively with deception. Witnesses and innocent suspects may exhibit these indicators just because testifying in a courtroom is anxiety-inducing, resulting in them being judged deceptive—a false-positive—whereas guilty suspects may be sufficiently coached and rehearsed as to be judged truthful—a false-negative. A jurist's level of discernment must be highly tuned to navigate between the revealing and the concealing signals.

Cognitive Difficulty

Many scholars have argued that a more fruitful direction in identifying valid and reliable indicators is to focus not on misleading signs from affect but on cognitive difficulty (25). These indicators derive from the assumption that lying is harder than truth-telling and will produce outward signs of such difficulty [(26); but cf. (27)]. Kinesic behaviors include blinks, avoidance of eye contact, reduction in illustrator gestures, and cessation of gesturing. Vocalic indicators include delayed responding to questions, shorter responses, and more speech errors (5, 24). All of these indicators are detectable in the courtroom and are among the most reliable ones available. Questioning that would be easy for truth-tellers to answer but difficult for deceivers (e.g., Who else might have had reason to commit X?) are most likely to elicit them.

Mental Processes

An extension of the cognitive approach is to consider what memory processes are implicated with lying. Liars engaged in serious lies—the type present in courtrooms—not only must access the truth and decide *if* to lie, but also conduct a cost-benefit analysis of different forms of deceit, choose the type of lie to be expressed, decide how to enact the lie non-verbally and verbally, and anticipate receivers' responses (28). A meta-analysis by Christ et al. (29) established that lying entails 8 of 13 brain regions and 173 deception-related foci that are more active for deceptive than truthful responses. These included accessing working memory, inhibitory control, and task switching (i.e., interspersing truthful with deceptive details). These mental gymnastics need not entail extreme mental effort to produce indicators of these executive processes. Longer between-turn and within-turn pauses (30) along with non-fluencies, gaze aversion and temporary cessation of gestures are likely to be most relevant. Again, these indicators may be indistinguishable from other cognitive difficulty indicators, so it becomes essential to evaluate the nature and difficulty of the questions they answer.

Behavioral Control

The aforementioned research frequently points to liars reducing postural, gestural, head and facial activity to the point of crossing a thin line between appearing composed and appearing wooden, rigid and unnatural (31). This generalized inhibition and rigidity across trunk, limbs, head, and face may reflect overcontrol of felt arousal and negative emotions (32, 33). Even when told about this trend, liars fail to increase their movement (34). Thus, attempted control does not succeed fully and may be one of the best classes of deception indicators because truth-tellers, in an effort to maximize their credibility, are likely to become more, not less, animated.

Self-Presentation

Scholars and practitioners alike have opined that deceivers attempt to project a demeanor of honesty and believability. This is more likely to occur when deceivers have opportunities to plan, rehearse, or adapt how they appear and sound (35). Especially they may adopt a veneer of facial and vocal pleasantness and calm. In the courtroom, judgments must factor into account the likelihood that witnesses and suspects are well-practiced in responding to anticipated questions. Smooth, fluent presentations therefore may or may not be indicative of truthfulness. The longer respondents are on the stand, the more they will be able to detect jurors' belief in their testimony and adapt responses accordingly. Veracity judgments formed early should be more informative than ones formed later.

VERBAL SIGNALS

Turning next to automatically analyzed linguistic indicators, seven clusters taken from Burgoon et al. (36) are likely to matter in the forensic context.

Quantity and Specificity

Deceivers tend toward shorter statements (fewer words and phrases) and less specific sensory, spatial and temporal details (37, 38). But when respondents are highly motivated and when accounts have been rehearsed over and over, this difference may evaporate (36). Here is where questioning strategies that elicit specific details can challenge liars while aiding truth tellers with their recall. "Was it daylight or twilight?" "What did the immediate vicinity look like?" And so on. Asking respondents to take a second look from a different perspective—perhaps the viewpoint of a bystander—has two advantages (39). First, deceivers who are fabricating an account will not have new details to present and may fear that inventing new ones risks contradicting previous statements, a risk compounded by any mental strain they are experiencing. Second, truth-tellers are often eager to be helpful, even adopting a Sherlock Holmes or Agatha Christie mantle of offering yet more recalled details.

Diversity

A key tip-off of veracity is how varied a speaker's vocabulary is. This feature is somewhat beyond liars' control. They can't spontaneously broaden their lexicon. And deceivers are especially likely to repeat the same phraseology over and over. In the face of deeper questioning, liars' primary strategy is to stick to their same prepared cover-story, whereas truth tellers principally try to be honest, leading to more varied responding (40). Repetitiveness is thus less common among truth-tellers.

Ambiguity/Hedging/Uncertainty

Vagueness, equivocation and hedging language such as weak modal verbs ("might have"), tentative words ("maybe"), and passive voice ("Mistakes were made") are more common in fraudulent statements (37, 38, 41), especially during unprepared remarks (36). The caveat is that liars may also pepper their remarks with linguistic markers of certainty to project confidence.

Personalism

This is a tricky one because researchers and practitioners have pointed to the "I" and "me" personal pronouns for indications of whether or not speakers take ownership of what they are saying. Liars recounting an accused rape may use personal pronouns ("I did this," "I did that") until the time frame of the actual event then shift to impersonal language. But this indicator varies wildly across written statements, interviews and in-person communication, making it unreliable; second person ("you") pronouns and impersonal pronouns have an equally checkered record (42).

Immediacy

Responding in the "here and now" (using linguistic immediacy) is often associated with truthfulness and non-immediacy, with deceit (43). However, that doesn't always hold in the courtroom. Shifts in verb tense from past to present ("I *go* golfing" instead of "I *went* golfing") produce less precision and more uncertainty in answer to the question, "Where were you last weekend?" In other cases, more immediate language is associated with truthfulness. Parents of missing children who fraudulently appeal for the return of their already-dead children may make statements like, "She *was* such a sweet girl." The validity of language immediacy as a veracity indicator depends on whether verb tense matches what is expected. For example, the question, "What did you do next?" calls for past tense; the question, "What are you thinking right now?" calls for present tense. If the tense is a mismatch with the question, it warrants a deeper dig. Amount of advance time for planning one's account can also increase the immediacy of language (44).

Emotion

Here I am talking about whether the language in use carries emotional overtones. This is also an indicator with an irregular history. It has been proposed in some quarters [e.g., (45)] that deceivers' fear, guilt, and shame creep into their choice of language. There is good evidence that compared to truth-tellers, deceivers' speech is either devoid of emotion language or includes more negatively valenced terms (42). But in other quarters, liars have adopted a more persuasive, pleasant stance (46). Fraudulent responses during a quarterly earnings call included more extremely pleasant adverbs and adjectives (36). Again, context is a critical guide as to whether liars might be motivated to paint a rosy picture.

IMPLICATIONS FOR DECEPTION DETECTION IN THE COURTROOM CONTEXT

The complexities of deception indicators might lead one just to rely on gut judgments of veracity. That has its merits (47). But there are still ways to separate the truthful wheat from the deceptive chaff. Signs of a frozen demeanor, occasionally peppered with face-and lip-adaptors, invite a closer look, particularly earlier during testimony. Close attention to voice and language choices that are not easily feigned can be particularly revealing. Comparing what are likely prepared or rehearsed remarks to extemporaneous ones will expose the most revelatory verbal and non-verbal indicators. And, questioning strategies that require multiple retellings of a narrative can further draw out behavior to be analyzed.

In sum, discoveries from emerging detection technologies and interviewing methods represent a new torch illuminating the search for the truth, the whole truth and nothing but the truth.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Portions of this research were supported by funding from the U. S. Air Force Office of Scientific Research under the U. S. Department of Defense University Research Initiative (Grant #F49620-01-1-0394), from the Defense Academy for Credibility Assessment (Grant #IIP-0701519), and from the National Science Foundation (Grants #0725895 and #IIP-0701519) (Principal Investigator: JB). The views, opinions, and/or findings in this report are those of the author and should not be construed as an official Department of Defense position, policy, or decision.

ACKNOWLEDGMENTS

I thank the cadre of graduate students in communication and management information systems at the University of Arizona and in communication at Michigan State University for their invaluable contributions to the programs of research funded by the agencies above on credibility assessment, interpersonal communication, and deception.

REFERENCES

1. Langleben DD, Moriarty JC. Using brain imaging for lie detection: where science, law, and policy collide. *Psychol Pub Policy Law* (2013) 19:222–34. doi: 10.1037/a0028841
2. Meijer EH, Verschuere B, Gamer M, Merckelbach H, Ben-Shakhar G. Deception detection with behavioral, autonomic, and neural measures: conceptual and methodological considerations that warrant modesty. *Psychophysiology* (2016) 53:593–604. doi: 10.1111/psyp.12609
3. Sai L, Zhou X, Ding XP, Fu G, Sang B. Detecting concealed information using functional near-infrared spectroscopy. *Brain Topogr.* (2014) 27:652–62. doi: 10.1007/s10548-014-0352-z
4. Burgoon JK, Metaxas D, Bourlai T, Elkins A. Social signals of deception and dishonesty. In: Vinciarelli A, Pantic M, Magnenat-Thalmann N, Burgoon JK, editors. *Social Signal Processing.* Cambridge: Cambridge University Press (2017). p. 404–28.
5. Scherer K, Schüller B, Elkins A. 6 Computational analysis of vocal expression of affect: trends and challenges. In: Vinciarelli A, Pantic M, Magnenat-Thalmann N, Burgoon JK, editos. *Social Signal Processing.* Cambridge: Cambridge University Press (2017). p. 56–68.
6. Lee CC, Welker RB, Odom MD. Features of messages that support automatable linguistics-based indicators for deception detection. *J Inform Sys.* (2009) 23:5–24. doi: 10.2308/jis.2009.23.1.24
7. Verschuere B, Ben-Shakhar G, Meijer E. (Eds.). *Memory Detection: Theory and Application of the Concealed Information Test.* Cambridge, UK: Cambridge University Press (2011).
8. Sartori G, Agosta S, Zogmaister C, Ferrara SD, Castiello U. How to accurately detect autobiographical events. *Psychol Sci.* (2008) 19:772–80. doi: 10.1111/j.1467-9280.2008.02156.x
9. Hartwig M, Bond CF Jr. Lie detection from multiple cues: a meta-analysis. *Appl Cogn Psychol.* (2014) 28:661–6. doi: 10.1002/acp.3052
10. Twyman NW, Proudfoot JG, Blair RM, Elkins AC, Derrick DC. Robustness of multiple indicators in automated screening systems for deception detection. *J Manage Inform Sys.* (2015) 32:215–45. doi: 10.1080/07421222.2015.1138569
11. Porter S, ten Brinke L. The truth about lies: what works in detecting high-stakes deception? *Legal Criminol Psychol.* (2010) 15, 57–75. doi: 10.1348/135532509X433151
12. Hoque ME, McDuff DJ, Picard RW. Exploring temporal patterns in classifying frustrated and delighted smiles. *IEEE Tran Affect Comput.* (2012) 3:323–34. doi: 10.1109/T-AFFC.2012.11
13. Matsumoto D, Hwang HC. Clusters of nonverbal behaviors differ according to type of question and veracity in investigative interviews in a mock crime context. *J Police Criminal Psychol.* (2017) 33:302–15. doi: 10.1007/s11896-017-9250-0
14. Burgoon JK. Nonverbal measurement of deceit. In: Manusov V, editor. *The Sourcebook of Nonverbal Measures: Going Beyond Words.* Hillsdale, NJ: Erlbaum (2005). p. 237–50.
15. Sporer S. Bodily communication and deception. In: Müller C, Cienki A, Fricke E, Ladewig SH, McNeill D, Tessendorf S, editors. *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction.* Berlin; Boston, MA: DE Gruyter Mouton (2013). p. 1913–21.
16. Ekman P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition).* New York, NY: WW Norton (2009).
17. Ekman P, Friesen WV. *Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues.* Englewood Cliffs, NJ: Prentice Hall (1975).
18. Burgoon JK. Opinion: microexpressions are not the best way to catch a liar. *Front Psychol.* (2018) 9:1672. doi: 10.3389/fpsyg.2018.01672
19. Yan WJ, Wu Q, Liang J, Chen YH, Fu X. How fast are the leaked facial expressions: the duration of micro-expressions. *J Nonverb Behav.* (2013) 37:217–30. doi: 10.1007/s10919-013-0159-8
20. Porter S, ten Brinke L, Wallace B. Secrets and lies: involuntary leakage in deceptive facial expressions as a function of emotional intensity. *J Nonverb Behav.* (2012) 36:23–37. doi: 10.1007/s10919-011-0120-7
21. ten Brinke L, Porter S, Baker A. Darwin the detective: observable facial muscle contractions reveal emotional high-stakes lies. *Evolut Hum Behav.* (2012) 33:411–16. doi: 10.1016/j.evolhumbehav.2011.12.003
22. Sporer SL, Schwandt B. Moderators of nonverbal indicators of deception: a meta-analytic synthesis. *Psychol Pub Policy Law* (2007) 13:1–34. doi: 10.1037/1076-8971.13.1.1
23. Leal S, Vrij A. Blinking during and after lying. *J Nonverb Behav.* (2008) 32:187–94. doi: 10.1007/s10919-008-0051-0
24. Sporer SL, Schwandt B. Paraverbal indicators of deception: a meta-analytic synthesis. *Appl Cogn Psychol.* (2006) 20:421–46. doi: 10.1002/acp.1190
25. Grazioli S, Johnson PE, Jamal K. A cognitive approach to fraud detection. *J Forensic Account.* (2006) 7:65–8. doi: 10.2139/ssrn.920222
26. Vrij A. *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice.* Chichester: John Wiley and Sons (2009).
27. Burgoon JK. When is deceptive message production more effortful than truth-telling? A baker's dozen of moderators. *Front Psychol.* (2015) 6:1965. doi: 10.3389/fpsyg.2015.01965
28. Walczyk JJ, Harris LI, Duck TK, Mulay D. A social-cognitive framework for understanding serious lies: activation-decision-construction-action theory. *New Ideas Psychol.* (2014) 34:22–36. doi: 10.1016/j.newideapsych.2014.03.001
29. Christ EC, van Essen DC, Watson JM, Brubaker LE, McDermott KB. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cereb Cortex* (2009) 19:1557–66. doi: 10.1093/cercor/bhn189
30. Sporer S. Deception and cognitive load: expanding our horizon with a working memory model. *Front Psychol.* (2016) 7:420. doi: 10.3389/fpsyg.2016.00420
31. Twyman NW, Elkins A, Burgoon JK. A rigidity detection system for the guilty knowledge test In: *Proceedings of the 44th Annual Hawaii International Conference on System Sciences.* Maui:(CD-ROM), Computer Society Press (2011).
32. Pentland SJ, Twyman NW, Burgoon JK, Nunamaker JF, Diller CBR. A video-based screening system for automated risk assessment using nuanced facial features. *J Manage Inform Sys.* (2017) 34:970–93. doi: 10.1080/07421222.2017.1393304
33. Twyman NW, Elkins A, Burgoon JK, Nunamaker JF Jr. A rigidity detection system for automated credibility assessment. *J Manag Infor Sys.* (2014) 31:173–201. doi: 10.2753/MIS0742-1222310108
34. Vrij A, Semin GR, Bull R. Insight into behavior displayed during deception. *Hum Commun Res.* (1996) 22:544–62. doi: 10.1111/j.1468-2958.1996.tb00378.x
35. Burgoon JK. Interpersonal deception theory. In: Levine TR, editor. *Encyclopedia of Lying and Deception.* Thousand Oaks, CA: Sage (2014).
36. Burgoon JK, Mayew WJ, Giboney JS, Elkins AC, Moffitt K, Dorn B, et al. Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls *J Lang Soc Psychol.* (2016) 35:123–57. doi: 10.1177/0261927X15586792
37. Hwang HC, Matsumoto D, Sandoval V. Linguistic cues of deception across multiple language groups in a mock crime context. *J Invest Psychol Offend Profil.* (2016) 13:56–69. doi: 10.1002/jip.1442
38. ten Brinke L, Porter S. Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law Hum Behav.* (2012) 36, 469–477. doi: 10.1037/h0093929
39. Köhnken G, Milne R, Memon A, Bull R. The cognitive interview: a meta- analysis. *Psychol Crime Law* (1999) 5:3–27. doi: 10.1080/10683169908414991
40. Clemens F, Granhag PA, Strömwall LA. Counter-interrogation strategies when anticipating questions on intentions. *J Invest Psychol Offend Profil.* (2013) 10:125–38. doi: 10.1002/jip.1387
41. Humpherys S, Moffitt K, Burns M, Burgoon JK, Felix W. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis Support Syst.* (2011) 50:585–94. doi: 10.1016/j.dss.2010.08.009
42. Hauch V, Blandón-Gitlin I, Masip J, Sporer SL. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception *Pers Soc Psychol Rev.* (2015) 19:307–42. doi: 10.1177/1088868314556539
43. Zhou L, Burgoon JK, Twitchell D, Nunamaker JF Jr. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision Negotiation* (2004) 13:81–106. doi: 10.1023/B:GRUP.0000011944.62889.6f

44. Chan S, Bull R. The effect of co-offender planning on verbal deception. *Psychiatry Psychol Law* (2014) 21:457–64. doi: 10.1080/13218719.2013.835703

45. Frank M, Ekman P. The ability to detect deceit generalizes across different types of high-stake lies. *J Pers Soc Psychol.* (1997) 72:1429–39. doi: 10.1037/0022-3514.72.6.1429

46. Masip J, Bethencourt M, Lucas G, Segundo MSS, Herrero C. Deception detection from written accounts. *Scand J Psychol.* (2012) 53:103–11. doi: 10.1111/j.1467-9450.2011.00931.x

47. Gigerenzer G. *Gut Feelings: The Intelligence of the Unconscious*. New York, NY: Penguin Group (2007).

Conflict of Interest Statement: The author declares she is affiliated with Discern Science International, a for-profit entity that develops systems for credibility assessment.

ORIGINAL RESEARCH
published: 22 January 2019
doi: 10.3389/fpsyt.2018.00748

Accuracy, Confidence, and Experiential Criteria for Lie Detection Through a Videotaped Interview

Antonietta Curci, Tiziana Lanciano, Fabiana Battista, Sabrina Guaragno and Raffaella Maria Ribatti*

Department of Education, Psychology, Communication, University of Bari "Aldo Moro", Bari, Italy

An individual's ability to discriminate lies from truth is far from accurate, and is poorly related to an individual's confidence in his/her detection. Both law enforcement and non-professional interviewers base their evaluations of truthfulness on experiential criteria, including emotional and expressive features, cognitive complexity, and paraverbal aspects of interviewees' reports. The current experimental study adopted two perspectives of investigation: the first is aimed at assessing the ability of naïve judges to detect lies/truth by watching a videotaped interview; the second takes into account the interviewee's detectability as a liar or as telling the truth by a sample of judges. Additionally, this study is intended to evaluate the criteria adopted to support lie/truth detection and relate them with accuracy and confidence of detection. Results showed that judges' detection ability was moderately accurate and associated with a moderate individual sense of confidence, with a slightly better accuracy for truth detection than for lie detection. Detection accuracy appeared to be negatively associated with detection confidence when the interviewee was a liar, and positively associated when the interviewee was a truth-teller. Furthermore, judges were found to support lie detection through criteria concerning emotional features, and to sustain truth detection by taking into account the cognitive complexity and the paucity of expressive manifestations related with the interviewee's report. The present findings have implications for the judicial decision on witnesses' credibility.

Keywords: lie detection, detection accuracy, confidence, experiential criteria, interview

OPEN ACCESS

Edited by:
Cristina Scarpazza,
Università degli Studi di Padova, Italy

Reviewed by:
Chris N. H. Street,
University of Huddersfield,
United Kingdom
Georgios Tzeferakos,
National and Kapodistrian University
of Athens, Greece

***Correspondence:**
Antonietta Curci
antonietta.curci@uniba.it

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

***Received:** 14 September 2018*
***Accepted:** 18 December 2018*
***Published:** 22 January 2019*

Citation:
Curci A, Lanciano T, Battista F,
Guaragno S and Ribatti RM (2019)
Accuracy, Confidence, and
Experiential Criteria for Lie Detection
Through a Videotaped Interview.
Front. Psychiatry 9:748.
doi: 10.3389/fpsyt.2018.00748

INTRODUCTION

One of the main challenges in police investigation and legal proceedings is to assess whether an interviewed suspect, defendant or witness is offering a deceitful account of relevant facts. To corroborate an interviewee's claims, police, and jurors might rely on extrinsic sources of evidence, such as documents, phone tapping, CCTV, GPS-tracked movements, etc. When such external sources are not available—such as in cases of family abuses and maltreatments—interviewers can only focus on the intrinsic qualities of interviews and derive from these qualities some experiential criteria to detect lies.

Classical studies on lie detection have demonstrated that the ability of laypeople to discriminate lies from truth "*is only slightly better than flipping a coin*" [(1), p. 284]: DePaulo et al. (2) combined the results of more than 1,300 estimates of the relationship between behaviors and deceit to identify behavioral cues of deceit. The authors concluded that simply relying on non-verbal behavior to discriminate truth from lies is insufficient, and further evidence is needed to definitively establish if someone is lying or not. In their comprehensive meta-analysis on deception detection accuracy, Bond and DePaulo (3) synthesized the results from 206 documents and 24,483 judges and found that people achieve an average of 54% correct lie-truth discrimination, correctly classifying 47% of lies as deceptive, and 61% of truths as non-deceptive. This proportion only increases marginally for professional lie-catchers: Vrij (4) reviewed studies on deception detection accounting for an accuracy rate of 55.91% for law enforcement personnel, although professionals' evaluation might be biased by overconfidence in their judgements (5). Finally, a meta-analysis by DePaulo et al. (6) yielded a correlation of 0.04 between accuracy ratings and confidence in lie detection, indicating that, even when people feel overconfident in their evaluation, there is no guarantee of detection accuracy.

LEGAL CRITERIA FOR THE EVALUATION OF WITNESS REPORTS: EXAMPLES ACROSS JURISDICTIONS

In spite of scientific evidence, the legal system is forced to identify some reliable criteria for lie-truth discrimination. Legal criteria have been variously set up across different jurisdictions, aiming to provide triers of fact with standards to evaluate witness truthfulness and decide on witness credibility. Most importantly, witness "demeanor" is the crucial aspect that judges and jurors are instructed to consider. This aspect does not refer to the content of evidence, but, as defined in the classical Goffman's studies, it concerns "*deportment, dress, bearing*" (7), and includes every visible and/or audible expression manifested by the witness in front of the Court or any interviewer, either fixed or variable, voluntary or involuntary, simple or complex (8).

Across different national contexts, guidelines and Court rulings have supplied specific instructions on how to evaluate a witness' demeanor. For instance, in the US, the 2017 Manual of Model Criminal Jury Instructions of the Ninth Circuit Jury Instructions Committee recommends jurors to consider some intrinsic features of witness testimonies, such as the witness' manner of testifying and the intrinsic reasonableness of witness reports. In Canada, a recent ruling (*Breed v. Breed, 2016, NSSC 42*) referred to specific aspects of testimonies, such as the consistency of external and internal reports ("*what are the inconsistencies and weaknesses in the witness' evidence, which include internal inconsistencies, prior inconsistent statements, inconsistencies between the witness' testimony, and the documentary evidence, and the testimony of other witnesses*"), accuracy and quantity of details ("*sufficient power of recollection to provide the court with an accurate account*"), and exposure modality ("*Was the evidence provided in a candid and straight forward manner, or was the witness evasive, strategic, hesitant, or biased*"). The consideration of internal and external consistency-of-witness accounts is among the 14 rules of thumb listed by Douglas in a paper presented at the 2004 Australian Institute of Judicial Administration Tribunal's Conference (9). In Europe, there are many examples of the criteria adopted by Courts in evaluating witness evidence. The judgement on *Berezovsky v Abramovich*, set up in UK in 2007, includes, among these criteria, confidence ("*witnesses can easily persuade themselves that their recollection of what happened is the correct one,*" p. 14), specificity of reported details ("*careful and thoughtful answers, which were focused on the specific issues about which he was being questioned,*" p. 18), and memory consistency ("*I found Mr. Berezovsky's evidence (and that of his witnesses) in relation to this issue to be vague, internally inconsistent,*" p. 23). In 1988, the Spanish Supreme Court held a sentence (Sentencia del TS, Sala de lo Penal, de 28 septiembre 1988, RJ 7070) focusing on external corroboration ("*verisimilitude*") and over time consistency of witness accounts ("*persistence in incrimination*") (10). The Italian Supreme Court (Cassazione) has underlined the importance of judges making a critical evaluation of a witness' evidence, expressly in cases of victims of sexual abuse, by paying special attention to consistency both across different interviews, and with other witnesses of the same crime. However, decision *no. 37988 of September 13*th*, 2016* leaves a "*large margin of appreciation regarding the methods for controlling witness credibility in specific cases.*" In sum, this brief juridical review shows that the legal evaluation of witness testimonies is generally based on a subjective evaluation by judges and jurors of the truthfulness on witness reports (11).

EXPERIENTIAL GROUNDING OF LEGAL CRITERIA

As surprising it may be, judges and jurors evaluate witness evidence based upon categories which correspond to what laypeople usually consider as indicators of truthful/deceptive behavior. In other words, the purity of legal judgement seems to be grounded in the subjective experience and commonsense of triers of fact. In the following section we will review a large body of evidence concerning the psychological processes underpinning the legal criteria for the assessment of truthfulness of witness accounts. Paralleling the jurisprudential review, we will enucleate a set of psychological categories related with lie/truth deception, which might be equated to the legal criteria above presented.

Emotional Features and Expressive Indices

The so-called emotional approach to lie detection emphasizes that lying is much more arousing than merely telling the truth. When lying, individuals experience a range of internal states (physiological and psychological) associated with specific behavioral indices (12–14). As a forensic instance, during a police interview, a suspect trying to propose a false alibi might experience fear of being caught; shame or guilt might be associated with violation of moral standards implied by lying; finally, a liar might also experience excitement, satisfaction or happiness for getting away with lies (15). Individuals are thus forced to mask the associated physiological and behavioral reactions, so that deception detection can be based on an accurate analysis of these patterns escaping control. It is thus not surprising that, among the criteria recommended across jurisdictions to discriminate lies from truthful accounts, the legal systems consistently encourage an analysis of the witness "demeanor," including all emotional manifestations implied in a witness testimony.

The pioneering work by Paul Ekman has emphasized the role of emotion identification in deception detection. Derived from the Darwinian evolutionary principles, Ekman's neurocultural theory of emotions considers the expression of emotions as a universal signaling system for organisms to communicate conspecifics in the presence of a predator or other critical cues for the survival of the individual and the species (15, 16). This signaling system includes physiological reactions and behavioral indices, many of which are conveyed by facial expressions. Culture teaches humans how to manage emotions in social contexts, by intensifying, de-intensifying or also dissimulating a given expressive pattern [i.e., display rules; (17)]. Emotion identification responds to the evolutionary need to ensure survival from danger, hence individuals learn to decode emotional signals from interactions with their conspecifics. Through daily life experiences, laypeople refine their capacity to identify others' behavioral manifestations and any form of emotional expression. Different theoretical accounts and empirical findings have emphasized the cultural variability of the production and perception of emotions (18, 19). Despite these different positions, on the whole, the emotion-based approach to deception can explain the reluctance of law enforcement and other professional lie-catchers to undergo extensive lie detection training (20): individuals generally rate themselves as sufficiently expert to correctly identify universal emotional signals; they also consider perceived indicators of deception based on non-verbal behavior as sufficiently accurate as actual indicators of deception (21). It follows that a professional lie-catcher adds his/her experience in lie identification to the competence attained from daily life experiences. However, as the above-mentioned studies by DePaulo et al. (2), and Bond and DePaulo (3) pointed out, the laypeople's ability to discriminate lies from truth based upon non-verbal signals is only slightly above chance.

Cognitive Complexity

The cognitive approach to lie detection is based on the empirical observation that, during a face-to-face interview, lying is much more cognitively demanding than telling the truth (4, 22). Simulating an episodic event or a story requires access to executive control processes involved in suppressing the truth, searching for information in long-term memory, and packing a lie in working memory (23). More specifically, the liar is asked to perform several cognitive tasks consuming high resources: (1) to produce a lie that is plausible and coherent with what the listener knows or may find out, (2) to keep in mind his/her inventions to report consistent statements in the future, (3) to monitor his/her reactions not to look deceptive, as well as the listener's reactions to make sure the listener does not distrust him/her, and (4) to suppress the truth (24, 25).

The cognitive approach to lie detection supports a consideration of the intrinsic characteristics of verbal reports to discriminate lies from truthful accounts. Recent studies have underlined that this approach downplays the role of other cognitive processes intervening in deception and does not include an adequate consideration of individual differences (22, 23, 26). However, the brief jurisprudential review referred to above shows that judges and jurors are generally instructed to pay attention to internal and external consistency in witness narratives, associated sense of confidence, quantity, and specificity of reported details, and intrinsic reasonableness and plausibility of witness' accounts. Such aspects are usually considered genuine proxies of accuracy by the empirical literature on autobiographical memory in forensic settings.

To illustrate, Peace and Porter (27) compared the properties of genuine vs. fabricated memories of a traumatic experience, and showed that, over a 6-month period, genuine accounts were more consistent, detailed, rich of contextual, and emotional information, and rated as more plausible than fabricated narratives. However, liars might also be highly motivated to keep consistent reports to protect either themselves or somebody else from the unwanted consequences of legal proceedings. At odds with the beliefs of laypeople and law professionals, consistency might also be indicative of lying rather than truth telling, especially in cases of repeated assessment of a suspect (28, 29).

Moreover, the sense of confidence exhibited by the interviewee has been shown to have a powerful persuasive effect on jurors (30). Experimental studies have met judicial case studies concerning innocent people being accused, tried, convicted, imprisoned, and sometimes executed for crimes they did not commit, following the testimony of individuals high in self-confidence or interacting with highly confident co-witnesses (31, 32). The persuasive effect of confidence has, however, been found to be moderated by a number of factors, such as the extension of witness reports (33), format of questioning (34), role of the interviewer [i.e., prosecutor vs. defense attorney, (35)], information provided to jurors to enhance skepticism (36), and reliance on an expert witness (37).

Finally, the phenomenological richness of details of reports sustains the interviewer's feeling that the interviewee's mental representations exactly correspond to events which really occurred in the past, very different from events only imagined, beliefs or semantic knowledge (38–40). The former representations have been shown to display greater clarity, more visual details, and more details for smell, sound, taste, location, time, and setting than imagined events (41, 42). The extensive meta-analysis by Oberlader et al. (43) summarizes the results of 56 English- and German-language studies, including studies adopting Criteria-Based Content Analysis [CBCA; (44)]. The authors concluded that a content analysis of reports concerning really experienced events—such as sexual abuse and violent offenses—qualitatively differ from deceitful accounts. However, the use of content analysis tools on witness reports is problematic in that systems of categories do not have the same validity when applied to either children or adults [see (45, 46)], require specific training in clinical psychology and psychological assessment, and are not easy to handle by jurors and judges in legal proceedings (47).

Paraverbal Aspects

Paraverbal cues are related with the emotional features of deceptive behaviors. Their investigation has been carried out not only in police and legal settings, but also in the workplace as a strategy for getting employment, advancements, or to avoid punishment (48, 49). Paraverbal behavior concerns the way the interviewee communicates his/her accounts during a face-to-face interaction, and, according to Sporer (23), reveals the interviewee's nervousness associated with fabricating a deceptive report.

In the above-mentioned meta-analysis by DePaulo et al. (2), only two of the whole set of paraverbal indicators considered by studies were found to be significantly and positively associated with deception, i.e., pitch and vocal tension, while taking time was found to be shorter in deceptive statements than in truthful ones. Given relevant differences in samples, methods, and construct operationalization across studies, results from that meta-analysis were rather contradictory. A new meta-analysis by Sporer and Schwandt (50) was run on a small subset of paraverbal behaviors, i.e., message duration, number of words, speech rate, response latency, unfilled pauses, filled pauses, speech errors, repetitions, and pitch. In this study, the authors also included a broad set of moderators, i.e., the interviewee's preparation, motivation, content of the deceptive message, sanctioning, degree of interaction between experimenter and participant and type of experimental design, and operationalizations used. In this study too, the pattern of effect sizes was rather heterogeneous: only pitch, response latency and speech errors positively related with deception, while message duration was negatively associated with deception. However, the results were significantly influenced by all moderators, indicating that the interviewee's individual characteristics largely influence the interviewer's ability to discriminate lies from truth based on paraverbal indices. Moreover, lie detection seems to be based on paraverbal behaviors especially in low familiarity situations, while individuals preferably rely on verbal indices when facing highly familiar situations (51). Finally, a very recent meta-analysis by Hauch et al. (1) on the effects of training interviewers on detection abilities reported a medium effect size on lie accuracy for verbal cues, while training on paraverbal behaviors, alone or in association with other non-verbal cues, only resulted in marginal effects.

AIM AND HYPOTHESES

The above-reviewed studies indicated that the individual's capacity to discriminate lies from truth is far from accurate and poorly related with the individual's confidence in his/her detection (6). As highlighted above, both law enforcement and non-professional interviewers base their evaluations of truthfulness on some experiential criteria which can be matched with categories largely investigated in psychological studies on lie and deception, i.e., emotional features, cognitive complexity, and paraverbal aspects of interviewees' reports. However, as noted above, research work on these issues has demonstrated that such criteria are, at the very least, questionable (1, 3, 28, 29, 32, 47, 51). Nevertheless, legal systems across different jurisdictions consistently recommend relying on them in assessing the truthfulness of a witness and his/her credibility. To our knowledge, no studies so far have attempted a systematic investigation of the psychological underpinnings of these criteria in a controlled context such as a lab setting. Furthermore, a quantifiable control on the judges' individual characteristics and expectations in supporting the accuracy and confidence of lie/truth detection is lacking in previous research work.

Following the above-reviewed studies, we designed an experimental study aimed at providing a better understanding of the criteria adopted in lie/truth detection, and to relate these criteria with accuracy and confidence in lie/truth identification. To this end, we adopted two perspectives of investigation in a mixed model design: we were indeed interested not only in the capacity of naïve people (including in this category Court judges, jurors, and professional interviewers) at lie/truth detection from a videotaped interview, but also in assessing the interviewee's detectability as a liar or truth-teller by our judges; additionally, we investigated the associations of these criteria with lie/truth detection not only from the judge's perspective, but also from the perspective of the interviewee. The combination of these two perspectives of investigation enabled us to control for the judge's dispositional preference toward one or more criteria.

In employing an experimental manipulation, we sought to improve control issues for our study, without decreasing the generalizability of our results [see the meta-analysis by Hartwig and Bond, (52), on the stability of lie detection across different contexts]. We thus administered to a sample of naïve "judges" a random sequence of videotaped interviews of liar vs. truth-teller "interviewees," and we tested the hypothesis that the judges' ability at lie/truth detection will be moderately accurate (1–3), and poorly associated with the judges' confidence in their evaluation (6); furthermore, we predicted that accuracy for truth identification will exceed that of lie detection (3). We finally expected that judges would support lie detection through experiential criteria concerning emotional and expressive features, cognitive complexity, and paraverbal aspects conveyed by the interviewees (2, 15, 23, 50). We also explored the associations of these criteria with both the judge's ability at lie/truth detection and with the evaluation of deceitfulness/truthfulness assigned to each interview.

METHOD

Design

The study adopted a mixed model design with the videotaped Interview Condition (Liar vs. Truth-Teller) as a between-subjects (fixed effect) factor, and a random effect for a sample of judges evaluating interviewees' behavior. The dependent variables were: (a) detection accuracy, (b) confidence in detection accuracy, (c) and the experiential criteria adopted to support detection.

Samples

The sample of judges consisted of 50 Italian volunteers (25 women), aged between 20 and 36 ($M = 24.54$; $SD = 3.41$), with an average level of education of 13.96 years ($SD = 1.62$). Each judge watched and listened to 50% of the videotaped interviews randomly selected from a pool of 20 interviews (see below), distributed across the two conditions (Liar vs. Truth-Teller, from 3 to 7 videos for each condition, 10 in total).

The sample of interviewees consisted of 20 Italian volunteers (10 women), aged between 21 and 28 ($M = 23.50$; $SD = 1.91$), with an average level of education of 13.75 years ($SD = 1.33$). Participants were matched for the two experimental interview conditions (10 Liar vs. 10 Truth-Teller). Each participant was administered an interview that was recorded to be subsequently shown to judges. Each videotaped interview was watched and listened to by 50% of the sample of judges.

All participants of both samples were recruited among students and experimenters' acquaintances. There was neither kinship, nor friendship, nor familiarity between the two samples: Each judge was unknown to each interviewee, and vice versa. Data were collected anonymously, and participants were preliminarily presented with an informed consent form. The Ethical Committee of the Department of Education, Psychology, Communication, University of Bari approved the study.

Measures and Procedure

Videotaped Interviewees Sample

The 20 interviewees were invited to participate in an experiment on the cognitive processing involved in an interview. Two separate sessions (Writing session and Videotaped Interview session) were arranged, and participants were randomly assigned to one of two videotaped Interview Conditions (Liar vs. Truth-Teller).

Writing Session

All participants were previously contacted by email or telephone and invited to provide a brief written story. Specifically, in the Liar condition, participants were asked to fabricate a fake holiday that supposedly happened in the last 12–18 months (e.g., *"Lie to me about your last holiday. So, for example, if your last holiday was to Paris where you visited galleries, we ask you to make up a holiday to a place you have never been to - for example, Barcelona - and lie to us about what you did - for example, went out with friends and swam with dolphins -")*. In the Truth-Teller condition, participants were asked to describe a holiday they really had in the last 12–18 months. Participants were invited to send us the texts by email, in order to prepare a base for the subsequent interview session.

Videotaped Interview Session

The day after the writing session, each participant was asked to sit relaxed in the lab and to talk facing a camera placed in front of him/her. Participants were informed that they were to talk about the holiday they had written of the day before, by talking to the video camera. It was also specified that the experimenter who performed the video interview did not know if the holiday they were recounting was real or simulated. As a consequence, participants were asked to be as credible as possible. This session was conducted by an experimenter who was not involved in the previous phase, and was unknown to participants. It consisted of two phases:

1. *Free telling phase.* Participants were asked to just relax and sit in front of a camera and to talk freely about the holiday they wrote of the day before (Liar vs. Truth-Teller condition), for about one minute and a half. They were asked to describe their trip, without worrying about timing; and were stopped when they had talked for long enough.
2. *Questions phase.* Each participant was interviewed in order to specify some details which had already been provided in the story or to give additional contextual details (weather, delay, scheduled event). For example, if an interviewee had said that he/she had a 1-week trip on Paris with his/her parents, he/she would be asked: *"Could you please tell me about some of the neighborhoods you visited?"* or *"How was the weather the first day you arrived in Paris?"* This phase lasted for about one and a half minutes.

Once both sessions were completed (about 3 min in total), participants were debriefed and thanked.

Judges' Sample

The sample of 50 judges was recruited by being asked whether they were willing to participate in an experimental study on evaluating a videotaped interview. Each judge was tested in a unique session, sitting in a quiet room, and requested to watch on a computer screen a random sequence of 10 videotaped interviews taken from the whole pool and distributed across the two conditions of the design (Liar vs. Truth-Teller, from 3 to 7 videos for each condition, 10 in total). We employed an unequal number of truthful and lying videotaped interviews to avoid the judges' expectation that half of the interview would be lies.

The judges were then asked: (a) to detect to which interview condition the interviewee was assigned (Liar vs. Truth-Teller), (b) to evaluate the level of confidence in their detection on an 11-point scale (Confidence score, 0 = "not at all"; 10 = "very much"), and (c) indicate the criteria they adopted to support detection through answering an open-ended question.

Coding System for Criteria of Lie/Truth Detection

Each judge's answer concerning the criteria adopted to support detection was transcribed verbatim and a coding system was applied, based on the psychological categories presented in the Intro. The authors identified four main categories of criteria, comparable with the psychological constructs presented above. The first category includes general *emotional features* of the interview, i.e., the judge emphasized the interviewee's ability to emotionally involve the viewer, the interviewee's calmness vs. nervousness, the coherence between story content and emotions expressed, the coherence between behavioral indices and emotions expressed. The second category includes the judges' mentions of specific *expressive indices*, such as the interviewee's direction of the gaze; mimic, and facial expressions (smiles, stillness, lip movements); body gestures (touching your nose, scratching elbow, etc.); and physical characteristics of the interviewee (appearance, bodily attitudes). The third category refers to the *cognitive complexity* of the story, i.e., the judge stated whether the interviewee's account appeared consistent, truthful, detailed, and vivid. The last category refers to the *paraverbal*

TABLE 1 | Descriptive statistics for judges' sample level ($N = 50$).

Indices	Total interviews	Liar interview condition	Truth-Teller interview condition	Paired samples t-test (df = 49) [mean difference 95% CI]
	M (SD)	M (SD)	M (SD)	
Detection accuracy	0.53 (0.15)	0.46 (0.21)	0.60 (0.24)	−3.01** [−0.23, −0.05]
Detection confidence	6.95 (1.09)	6.93 (1.23)	6.90 (1.17)	0.21 [−0.23, 0.29]
Emotional features	0.17 (0.10)	0.15 (0.11)	0.19 (0.13)	−1.83 [−0.08, 0.003]
Expressive indices	0.19 (0.10)	0.21 (0.11)	0.17 (0.14)	2.02* [.00, 0.09]
Cognitive complexity	0.36 (0.15)	0.36 (0.22)	0.37 (0.18)	−0.43[−0.08, 0.05]
Paraverbal aspects	0.28 (0.13)	0.29 (0.17)	0.28 (0.15)	0.33 [−0.04, 0.06]

*$p < 0.05$; **$p < 0.01$.

aspects of the report, i.e., the judges explicitly mentioned the exposure clarity, fluency of the speech vs. hesitation, reactivity and/or readiness of response, latency times, confidence, and/or spontaneity in the exhibition, voice tone, participation vs. acting, and linear vs. fragmented exposition. One point was assigned for each criterion mentioned. Two trained coders—who were blind to each other's results—independently scored half of the total 500 judges accounts (50 judges × 10 videotaped interviews). The interrater reliability was high for such a scoring ($r_{\text{Emotional features}} = 0.94$; $r_{\text{Expressive indices}} = 0.96$; $r_{\text{Cognitive complexity}} = 0.90$; $r_{\text{Paraverbal aspects}} = 0.90$). (See **Appendix** for an example of the coding system).

RESULTS

Judges' Level

Descriptive and Correlational Analyses

For each judge, we analyzed his/her lie/truth detection capacity and the criteria adopted to support detection. To this end, we computed the following indices: (a) Detection Accuracy was obtained by averaging the accuracy scores for each observed interview in the two conditions of the design (0 = "error"; 1 = "correct"; range 0–1 for liar and truth-teller conditions, respectively); (b) Detection Confidence was obtained by averaging the Confidence scores for each watched interview separately for the two conditions of the design (range 0–10 for both liar and truth-teller conditions); (c) frequencies of occurrence of each category of experiential criteria were transformed into proportions; for each condition of the design (Liar vs. Truth-teller) we computed the total occurrence of each category across all videotaped interviews shown to the judge, and divided that sum by the maximum occurrence of categories for that judge. This computation takes into account the individual's distribution of category occurrences, normalizing for the individual's propensity to prolixity. **Table 1** showed descriptive analyses for the judges' level. Overall, results showed that judges report a medium level of Accuracy and a medium-high level of Confidence in detecting liar vs. truth-teller interviewees, and a low occurrence for the categories of experiential criteria, with Cognitive complexity and Paraverbal aspects as the highest experiential criteria mentioned to support detection. The parametric paired t-test revealed a significant effect of condition (Liar vs. Truth-teller) on the measure of Detection Accuracy, in that it seemed to be slightly easier for our judges to accurately detect truthful rather than deceitful interviews. Additionally, the t-test showed a significantly higher occurrence of Expressive indices in the evaluation of liars than truth-teller interviewees. The significant effect on the index of Detection Accuracy was further explored to evaluate if it was due to a different base rate of truthful videos presented to our judges as compared with lying ones. To this end, the entire sample of judges was divided into three subsamples viewing, respectively 3–4 vs. 5 vs. 6–7 truthful videotaped interviews, and separate t-test analyses were run on the measure of Detection Accuracy for each of the three subsamples. The effect of condition vanished when the base rate of truthful interviews was ≤50% (ts < |1.80|, *n.s.*), but it remains significant for the subgroup of judges viewing 6–7 truthful videotaped interviews [$t_{(14)} = -3.01$, $p < 0.01$].

Table 2 shows Pearson's zero-order correlations of all the indices described above for the judges' sample. Interestingly, for the liar condition, Detection Accuracy was negatively associated with Detection Confidence, whilst for the truth-teller condition the two indices were positively associated (see also **Figure 1**). Additionally, Confidence scores for the two conditions were strongly positively associated. Detection Accuracy in the liar condition was positively related to the occurrence of Emotional features of the reported story. By contrast, Detection Accuracy in the truth-teller condition was positively associated with the Cognitive complexity category and negatively associated with Expressive indices.

Interviewees' Level

Descriptive and Correlational Analyses

This set of analyses reverses the logics of those described in the previous section, since our aim in designing the study was also to assess the interviewee's detectability as either lying or truthful by our judges. For each videotaped interview, we computed the following indices: (a) Detection Accuracy score was obtained by averaging the accuracy scores across judges (0 = "error"; 1 = "correct"; range 0-1); (b) Detection Confidence was obtained by averaging the Confidence scores across judges (range 0–10); (c) frequencies of occurrence of each category of experiential criteria were transformed into proportions; we computed the total occurrence of each category across judges,

TABLE 2 | Pearson's correlations for judges' sample level ($N = 50$).

Indices	Detection accuracy liar	Detection accuracy truth-teller	Detection confidence liar	Detection confidence truth-teller
Detection accuracy—truth-teller	0.00			
Detection confidence—liar	−0.32*			
Detection confidence–truth-teller		0.30*	0.71**	
Emotional features—liar	0.30*		−0.33*	
Emotional features—truth-teller		−0.04		0.12
Expressive indices—liar	0.09		0.05	
Expressive indices—truth-teller		−0.35*		−0.11
Cognitve complexity—liar	−0.22		0.08	
Cognitive complexity—truth-teller		0.43**		0.14
Paraverbal aspects—liar	0.02		0.07	
Paraverbal aspects—truth		−0.16		−0.18

*$p < 0.05$; **$p < 0.01$.

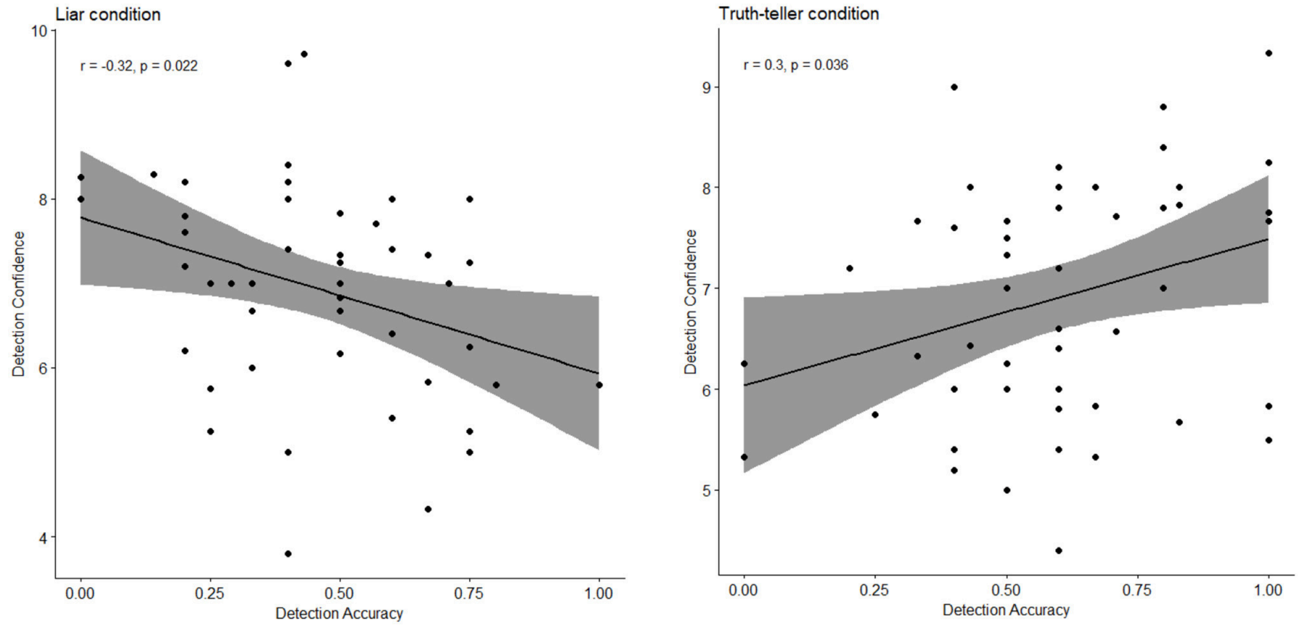


FIGURE 1 | Scatterplot of correlations Detection Accuracy-Detection Confidence measures for judges' sample level ($N = 50$), for the liar **(Left),** and truth-teller conditions **(Right)**.

and divided that sum by the maximum occurrence of categories mentioned by all judges for that interview. **Table 3** showed descriptive analyses for the interviewees' level. Overall, these results substantially confirmed those obtained for the judges' level. Medium levels of Accuracy and medium-high levels of Confidence were observed for the detection of both liar and truth-teller interviewees. Additionally, a low occurrence of the categories of experiential criteria was found to be associated with detection, with Cognitive complexity and Paraverbal aspects as the highest criteria mentioned. An independent samples t-test was run on all indices considered in the study (Detection Accuracy, Confidence, and experiential criteria), with the design condition (Liar vs. Truth-teller) as a between-subjects factor. Given the limited sample size ($n = 10$ videotaped interviews for condition), the non-parametric bootstrapping method was used as a robust estimation of t-test. Bootstrapping provided a confidence interval (CI) around the mean difference, which is significant if the interval between the upper limit (UL) and lower limit (LL) of a bootstrapped 95% CI does not contain zero, which means that the difference between the two groups is different from zero. Albeit not significant (CI includes 0), Detection Accuracy appeared to be slightly higher in the truth-teller condition than in the liar condition, confirming the findings for the judges' level. Furthermore, a significantly higher occurrence of Emotional features was observed in the evaluation of truth-teller interviewees. Bootstrapped Pearson's zero-order correlations were run among indices on the total sample of interviewees, and separately for the two interview conditions

TABLE 3 | Descriptive statistics for total sample of interviewees and for the two interview conditions ($N = 20$; 1,000 Bootstrapped Samples).

Indices	**Total sample ($N = 20$)**	**Liar condition ($n = 10$)**	**Truth-Teller condition ($n = 10$)**	**Independent samples *t*-test (df = 18) [mean difference 95% CI]**
	M (SD)	**M (SD)**	**M (SD)**	
Detection accuracy	0.53 (0.19)	0.46 (0.21)	0.61 (0.15)	−1.83 [−0.30, 0.01]
Detection confidence	6.95 (0.41)	6.99 (0.53)	6.92 (0.25)	0.41 [−0.26, 0.46]
Emotional features	0.17 (0.05)	0.15 (0.05)	0.20 (0.05)	−2.14* [−0.09, −0.00]
Expressive indices	0.20 (0.06)	0.22 (0.07)	0.18 (0.05)	1.46 [−0.01, 0.09]
Cognitive complexity	0.34 (0.06)	0.33 (0.06)	0.35 (0.06)	−0.80 [−0.07, 0.03]
Paraverbal aspects	0.28 (0.05)	0.29 (0.06)	0.27 (0.04)	1.28 [−0.02, 0.07]

*$p < 0.05$.

(Liar vs. Truth-teller; **Table 4**). For the interviewee's level, the association between Detection Accuracy and Confidence, albeit non-significant, is consistent with the judges' level, i.e., negative for the liar condition and positive for the truth-teller condition. However, none of the associations among the categories was found to reach the significance level.

Receiver-Operating-Characteristic (ROC) Analysis

Generally speaking, a receiver-operating-characteristic (ROC) analysis (53) is used to determine the diagnostic performance of a test to discriminate diseased cases from normal cases (i.e., diagnostic accuracy). The accuracy of the test depends on how well the test separates the two categories or conditions (diseased vs. normal). Analogously, in our data, we adopted the ROC analysis to determine the accuracy of judges' detection (the above-referred as test) in discriminating truthful from deceitful videotaped interviews (the above-referred as diseased vs. normal cases). Our measure on which diagnostic accuracy was tested corresponds to the judges' raw detection whether the interviewee belongs to either a liar condition or a truth-teller condition, regardless if the detection was correct or not. This measure was obtained by summing up all detection scores provided by the 25 judges for each videotaped interview (score 0 = liar's detection, score 1 = truth-teller's detection, regardless of the correctness of such a detection). This aggregate Raw Detection index ranged from 0 to 25, with higher values indicating a prevalence of truth-tellers' detection, and lower values indicating a prevalence of liars' detection.

The Raw Detection index was employed as a measure indicating the diagnostic power that an interviewee falls into one condition (1 = truth-teller) or the other (0 = liar). In our data, the truth-teller condition was employed as a state variable indicating the "true category" to which the interviewee belongs. The value of the state variable indicates which category should be considered positive (in our case 1 = truth-teller). Higher values indicate a greater probability of positive category (1 = truth-teller).

Diagnostic accuracy is measured by the Area Under the Curve (AUC), which takes values from 1 (perfectly accurate discrimination) to 0 (perfectly inaccurate discrimination). In general, an AUC of 0.5 suggests no discrimination, from 0.7 to 0.8 is considered acceptable, from 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding (54). The ROC analysis ran on the present data yielded an AUC of 0.61, indicating a little above typical power to discriminate truthful from deceitful videotaped interviews.

Multilevel Analyses

The following analyses were run to control for the judges' individual variability in accuracy, confidence, and experiential criteria for lie/truth detection. To this end, we tested a random intercept model with the package *lme4* (55) for multilevel analysis through R (56). This analysis was separately applied to the indices of Detection Accuracy and Detection Confidence, and to the proportion of occurrence of the four categories of criteria for each interview shown to each judge (Emotional features, Expressive indices, Cognitive complexity, and Paraverbal aspects). Only for the Detection Accuracy index, given the dichotomous nature of the dependent variable (0 = "error"; 1 = "correct"), the estimated model was a logistic regression.

As a general procedure, we first estimated a model with fixed factors only and we included the Interview condition (Liar vs. Truth-Teller) as a fixed effect variable. Along with this factor of the design, we also intended to control for the congruence between demographic characteristics of judges and interviewees: given that judges and interviewees do not differ as to their age [$t_{(68)} = 1.28$, *n.s.*], we only considered gender congruence as an additional fixed factor in our models (dichotomous indicator, 0 = non-congruent; 1 = congruent). We finally included the individual variability of judges evaluating interviewees' behavior as a random factor (judges' ID). An a priori power analysis was applied through the *lme4*, *simglm* (57), and *paramtest* (58) R packages, on a model with two fixed dichotomic factors, and a random factor with $\sigma^2 = 0.50$. With a medium effect size = 0.50 for the two fixed factors, $p < 0.05$, a total of 500 observations (50 judges × 10 interviewees), and simulated samples = 100, the analysis yields a power >0.75.

The fit of our models was estimated by applying the *car* R package to obtain the Wald test statistics (59). The AIC and BIC indices were computed to enable a comparison between the model with only fixed effects (Interview condition and gender congruence) with the model with both fixed and random effects (judges' ID). As **Table 5** shows, the only significant fixed effect was found for the Interview condition on the measure of Detection Accuracy, in that truth-teller interviewees were more

TABLE 4 | Pearson's correlations for total sample of interviewees and for the two interview conditions ($N = 20$; 1,000 Bootstrapped Samples).

Indices	Total sample ($N = 20$)		Liar condition ($n = 10$)		Truth-Teller condition ($n = 10$)	
	Detection accuracy	Detection confidence	Detection accuracy	Detection confidence	Detection accuracy	Detection confidence
Detection confidence	−0.21 [−0.65, 0.50]		−0.34 [−0.79, 0.53]		0.25 [−0.26, 0.77]	
Emotional features	−0.06 [−0.36, 0.35]	0.16 [−0.25, 0.60]	−0.39 [−0.81, 0.28]	0.20 [−0.48, 0.72]	−0.13 [−0.72, 0.69]	0.35 [−0.20, 0.84]
Expressive indices	−0.16 [−0.50, 0.22]	0.01 [−0.63, 0.52]	0.08 [−0.40, 0.59]	0.14 [−0.58, 0.76]	−0.29 [−0.74, 0.27]	−0.56 [−0.87, 0.05]
Cognitive complexity	−0.02 [−0.40, 0.41]	0.18 [−0.23, 0.54]	−0.32 [−0.83, 0.38]	0.29 [−0.44, 0.82]	0.18 [−0.31, 0.77]	0.05 [−0.54, 0.60]
Paraverbal aspects	0.27 [−0.19, 0.62]	−0.37 [−0.72, 0.13]	0.52 [−0.20, 0.85]	−0.58 [−0.89, 0.05]	0.23 [−0.34, 0.69]	0.18 [−0.78, 0.79]

accurately identified than liars ($\beta = 0.62$, z = 3.38, $p < 0.001$; Wald test = 13.02, $p < 0.001$; AIC = 684.31; BIC = 696.96). For none of the dependent variables entered in the model the effect of gender congruence judge-interviewee was found to be significant. Furthermore, for none of our dependent variables, controlling for the judges' individual variability resulted in a significant improvement of the model [AIC and BIC were lower for the model with fixed effects only; (60)]. Finally, in order to rule out any confounding due to the interviewees' variability, the multilevel models were also run by including a random intercept for interviewees' ID. The last columns of **Table 5** display the fit indices for the models with both fixed effects (Interview condition and gender congruence) and interviewees' variability as random factor. As shown in the table, the inclusion of the random factor does not improve the fit of the models (see AIC and BIC indices in the last column of **Table 5**), thus confirming the stability of our results also after controlling for interviewees' variability.

In sum, truth detection appeared to be slightly easier than lie detection, regardless of the peculiar individual characteristics of the judge evaluating the interviewee's behavior. Furthermore, the associated sense of confidence in detection and—surprisingly—the adoption of the experiential criteria to support detection accuracy resulted as being completely unaffected by the individual's variability. Overall the present results are consistent with results reported in the preceding sections of the present paper, in spite of the different measurement models adopted (i.e., for Detection Accuracy, average measure across judges vs. dichotomous items in multilevel modeling).

DISCUSSION

In the current study, we aimed to assess the ability of naïve judges to discriminate deceitful vs. truthful reports by watching a videotaped interview (the judges' level) in a lab context, along with the interviewees' detectability as liars or truth-tellers from the judges (the interviewees' level). We also aimed to identify the criteria adopted by lay people to justify lie/truth detection and relate them with detection accuracy and confidence. To accomplish our goals we adopted a multilevel approach, requiring a sophisticated data-analytic methodology for an experimental design. Two sets of analyses were conducted to account for the structure of our data (judges and interviewees levels). Overall, results are consistent for the two levels of investigation: Lie/truth detection was found to be moderately accurate across judges and across interviewees, but a slightly higher accuracy was observed for detection of truthful accounts than deceitful ones; furthermore, judges appeared to be moderately confident that their detection was accurate.

The accuracy-confidence link showed an interesting pattern of results across the liar vs. truth-teller conditions: when naïve people are faced with a deceitful report, detection—although accurate—appears to be negatively associated with confidence; contrariwise, naïve people seem more confident when accurately identifying a truthful report. In other words, "I am not too sure when I detect a lie, even if it is really a lie." It thus seems that detecting a lie has a greater "cost" in terms of confidence, for a kind of "conservative attitude" when people have to identify an unknown other as a liar. In that our results are consistent with DePaulo et al. conclusion, that judges are more confident when they are evaluating actual truths (accurate truth detection) as compared to when they are evaluating actual lies (accurate lie detection) (6).

As regards the criteria, results showed that the interviewee's physical characteristics, his/her mimic, and facial expressions, his/her gaze direction and body gestures were the indices most mentioned to detect a liar than a truth-teller interviewee. The interviewee's nervousness, and the incoherence between story content, behavioral indices, and emotions expressed were the criteria most frequently adopted to accurately detect a liar; consistency of the reports, richness of details, and vividness and poor expressive manifestations were the most recurrent criteria to accurately detect a truth-teller interviewee. Finally, as shown by the ROC analyses, the strength to which deceitful vs. truthful reports were discriminated from each other was modest. Jointly considered, our findings sustain the idea that people are not accurate in detecting the deception of unknown people (61), and that they occasionally support detection through experiential criteria concerning internal features of the witness accounts (2, 3).

TABLE 5 | Multilevel analysis on the measures of the study, testing the effect of interview condition, and gender congruence (fixed effects) and judges' and interviewees' individual variability (random intercept models).

	Fixed effects only					Fixed and random effects (Judges' ID)			Fixed and random effects (Interviewees' ID)		
	Interview condition (Liar vs. Truth-teller)	Gender congruence judge-interviewee	Wald test (df =1)	AIC	BIC	Wald test (df = 1)	AIC	BIC	Wald test (df = 1)	AIC	BIC
Detection accuracy	0.62 ($z = 3.38$***)	−0.22 ($z = -1.21$)	13.02***	684.31	696.96	12.87***	686.31	703.17	4.92*	667.13	683.99
Detection confidence	−0.08 ($t = -0.42$)	0.01 (t =0.05)	0.18	2125.86	2142.71	0.07	2080.31	2101.38	0.17	2133.96	2155.03
Emotional features	0.05 ($t = 1.83$)	0.00 (t =0.13)	3.38	131.65	148.51	3.38	150.26	171.33	2.18	149.66	170.73
Expressive indices	−0.03 ($t = -1.47$)	0.03 ($t = 1.16$)	3.52	12.27	29.13	3.76	27.87	48.94	1.71	30.80	51.87
Cognitive complexity	0.01 (t =0.24)	−0.02 ($t = -0.55$)	0.36	453.34	470.20	0.35	451.64	472.71	0.33	471.37	492.44
Paraverbal aspects	−0.02 ($t = -0.59$)	−0.01 ($t = -0.21$)	0.39	319.48	336.34	0.33	330.05	351.12	0.27	330.05	351.12

*$p < 0.05$, ***$p < 0.001$.

AIC, Akaike's Information Criterion; BIC, Bayesian Information Criterion.

The lme4 package returns z-tests (logistic model) and t-tests for fixed effects, and estimated variance for random effects; the Wald test for the models has a chi-square distribution.

Our study shows that individuals selectively choose either emotional or cognitive indices to identify lies vs. truthful interviews. A possible explanation for this asymmetry might be that individuals are naturally trained to detect emotional signals as cues of deception (15, 62), so that they justify their feeling that the interviewee is truthful by relying upon an evaluation of emotional signals. This explanation is in line with neuropsychological and neurobiological studies which have underlined the role of specific neural circuits in deception detection (63, 64). Among those circuits, the amygdala and the anterior cingulate cortex have been shown to be activated in social judgement tasks, when decoding of emotional signals is particularly relevant for interpersonal cooperation, communication, social business, and management, and for the ultimate goal of individuals' and species survival (65–67). When individuals are alerted by emotional signals that a speaker is lying, reliance upon those signals disrupts the usual cognitive processing of verbal messages (20). It follows that, while an accurate evaluation of truthful interviews is supported by a controlled analysis of cognitive features of verbal accounts, lie detection is preferentially anchored to decoding emotional indices. Our study reveals that naïve judges keep a sort of implicit knowledge of this differential processing of lies and truthful reports, and this knowledge is reflected in the legal criteria suggested across different jurisdictions to evaluate witness' truthfulness and decide on witness credibility.

An important strength of our study is that the experimental approach enabled a sizeable control on the judges' individual dispositions and expectations when deciding on witness truthfulness. Previous meta-analyses have shown that the judges' individual variability does not play a crucial role on detection accuracy (61). However, as underlined by Aamodt and Custer (68), there is a paucity of studies available to assess the relationship between the individual's characteristics and accuracy in detecting deception. On this issue, a surprising outcome of the multilevel analysis is that judges' individual variability did not in any way affect the adoption of each one of the categories of experiential criteria to support lie/truth detection. In another words, the final decision concerning whether to believe a witness or not does not display any regularity with regards to the judge's individual disposition/bias nor with regards to the similarity between judges and interviewees (operationalized as gender congruence in our study). Among our findings, the only relevant effect concerns detection being more accurate for truth-teller than liar interviewees, but the amount of variance attributable to the judge's individual tendencies is worthless as compared with the variance due to the interview condition. These findings might lead us to conclude that detecting lies is generally more difficult than identifying truth. However, a more in-depth exploration of this difference in detection of truthful video interviews as compared with lying ones showed that the effect remains significant only when the base rate of truthful interviews exceeds lying ones. These findings might be accounted for by the so-called "veracity effect," which predicts that detection accuracy is a linear function of message veracity, so that the probability for a judge of giving an accurate truth identification increases as long as the proportion of honest messages increases

(69). This effect depends on the fact that people have a kind of "truth bias" (3), so that they are more prone to believe to others since they consider them essentially truthful (70). This bias is even underestimated in experimental settings as compared with real life interactions, where individuals are naïvely prone to accept deceptive messages as truth.

The present results have noteworthy implications in the forensic domain. To illustrate, gender congruence between jurors and witnesses can be an influential factor with respect to the composition of juries, especially in crimes such as rape or sexual aggression, where the victim and the defendant are the only people present on the crime scenario (71–75). Following these findings, we introduced gender congruence in our multilevel model to explore its role in predicting accuracy, confidence, and experiential criteria adopted by judges for lie/truth detection. However, our results prove that this factor is ineffective in lie/truth discrimination, hence could be neglected if the main task required of jurors were lie/truth detection. It should however be considered that our conclusions are based on an experimental paradigm in which naïve judges are required to decide whether an interviewee is truthful/liar when narrating a holiday narrative. This artificial setting cannot fully emulate the emotional and cognitive requirements of a sex crime trial. Another important point regards the role of judges with respect to a witness whose truthfulness has to be assessed. In the Italian legal system, as in other countries in which the witness undergoes classical cross-examination, the role of the judge and jurors is—except in specific instances—that of passive observers while attorneys and prosecutors run the witness' interview directly interacting with him/her. However, it is up to judges and/or jurors to draw conclusions on the witness' truthfulness, and credibility. The experimental setting adopted in our study, through the administration of a videotaped interview to a sample of naïve judges, attempts to emulate as far as possible the real context of a criminal proceeding, in which the interaction between judges and interviewees is generally precluded. Judges are thus forced to only focus on the intrinsic qualities of interviews and base on them lie/truth detection.

Findings from the present study highlight the experiential grounding of the legal criteria identified across jurisdictions to support the legal decision on witness credibility. The content analysis run on the answers provided by judges to the open-ended question yields a category system including references to emotional and expressive features of the interviewee's accounts, indices of cognitive complexity of reports, and paraverbal aspects concerning the story-telling regulation by the interviewee (2, 15, 23, 50). Each of these categories captures some features of the general concept of witness "demeanor" (7, 8), which triers of fact are requested to consider. The mention of these categories in the judges' responses was consistently assessed in our study, accounting for an experiential base for the jurisprudential criteria recommended across different national contexts. However, as the review in the introductory sections of the present paper shows, these criteria are largely disputed across scientific studies. People rate themselves as sufficiently expert at identifying lies from the interlocutor's physiological pattern and expressive behavior, but the laypeople's ability at lie/truth discrimination based upon non-verbal signals has been demonstrated as being only slightly above chance (2, 3). Narrative proxies of accuracy are controversial across studies, in that consistency, confidence, and phenomenological richness might also characterize deceitful and/or only imagined accounts (29, 32, 76, 77). In sum, the present findings confirm once more that, despite triers of fact struggling to apply jurisprudential principles and professional guidelines, the basis for the legal evaluation of witness evidence across jurisdictions is experiential and, as such, mainly unwarranted.

The results of the current study should be considered in the light of limitations and future perspectives. First, the composition of our two samples reduces the chance of massively generalizing our findings to a real lie/truth detection context: Our sample of judges did not include individuals belonging to categories especially concerned with witness' assessment (e.g., professional judges, jurors, police detectives, federal law enforcement, investigators, etc…), and interviewees' calmness and quietness when sitting in a "sterile" lab environment do not fully reproduce the real emotional state of a witness testifying in a criminal proceeding. Furthermore, the age and educational range both judges and interviewees is quite limited and this might compromise the generalizability of our findings. To illustrate, it has been shown that the ability to detect lie through visual information conveyed by facial expressions is attenuated in elderly as compared with young people (78, 79). It follows that our approach needs to be replied on samples of elder adults, which can be very often involved in criminal trials as victims of maltreatments or financial exploitation, perpetrators of crimes as internet frauds and sexual abuses, or professional judges. Second, and related with the first point, while in our study we controlled for gender congruence between judges and interviewees, the age limitation of our sample prevented us from assessing a possible effect of age congruence: the issue of age matching needs to be carefully considered in future replications, since studies do not converge on it, either showing no age-matching effect (80) or a significant effect only for young people (79). Third, our participants were instructed to give their accounts for about one minute and a half, and this temporal limitation might have influenced their ability at deception detection. Fourth, our study did not enable a direct interaction between judges and the interviewees, and the manipulation through a videotaped session hardly resembles a realistic situation of a court hearing. However, the issue of generalization does not represent a serious flaw in the study: As the recent meta-analysis by Hartwig and Bond (52) concludes, lie detectability is substantially stable for multiple cues, including in those cues both lab sessions and forensic settings. Finally, we used a straightforward questioning during the interviews, and avoided using suggestive techniques or imposing additional cognitive load on our interviewees. As studies adopting the cognitive-based approach to lie detection have variously shown, liars might find it particularly difficult to deceive when asked to maintain eye contact with the interviewer, expose their version of facts in reverse order, or answer unexpected questions (25). Future studies should implement the procedure by adding further constraints to the interview session and subsequent assessment.

Despite the aforementioned limits, the current work has several strengths. First, the number of observations was high ($n = 500$, 50 judges × 10 interviews) and based on full audiovisual modality (face, body, and speech), enabling the judge to evaluate many behavioral manifestations as usually done in naturalistic settings. Second, through a two-level design, the study provides a strong experimental control on individual variability in lie/truth detection from both the perspective of the judge and the interviewees (81). Third, our study investigated levels of accuracy and confidence and the criteria adopted for the evaluation of both deceitfulness and truthfulness, and the results we have obtained, as outlined before, clearly demonstrate that the two processes do not match completely. Fourth, the current work took into account not only indices traditionally associated with lie detection such as accuracy and confidence, but also the explanations on which naïve judges generally base their lie/truth detection. The current findings, although not exhaustive, thus represent a meaningful step forward in understanding the experiential base for the legal criteria adopted by courts to decide on witnesses' credibility.

In sum, the present study provides a contribution to the field of investigation of lie/truth detection, by showing through a lab manipulation that judges' assessment of witness' truthfulness and credibility rests upon experiential criteria, respectively focusing on the emotional features of the liar's account and on the cognitive complexity and scarcity of expressive manifestations of the truth-teller's account. The legal decision concerning witness' credibility is ultimately grounded on the experiential evaluation of judges and jurors, and, as such, it might gain benefit from the informative value of scientific evidence.

AUTHOR CONTRIBUTIONS

AC and TL designed the experiment, analyzed the data, and edited the manuscript. FB, SG, and RR collected the data and scored the protocols. All authors listed have made substantial, direct and intellectual contributions to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2018.00748/full#supplementary-material

REFERENCES

1. Hauch V, Sporer SL, Michael SW, Meissner CA. Does training improve the detection of deception? A meta-analysis. *Commun Res.* (2016) 43:283–343. doi: 10.1177/0093650214534974
2. DePaulo BM, Lindsay JJ, Malone BE, Muhlenbruck L, Charlton K, Cooper H. Cues to deception. *Psychol Bull.* (2003) 129:74–118. doi: 10.1037/0033-2909.129.1.74
3. Bond CF Jr, DePaulo BM. Accuracy of deception judgments. *Person Soc Psychol Rev.* (2006) 10:214–34. doi: 10.1207/s15327957pspr1003_2
4. Vrij A. *Detecting Lies and Deceit: Pitfalls and Opportunities*. Chichester: John Wiley and Sons (2008).
5. Porter S, McCabe S, Woodworth M, Peace KA. Genius is 1% inspiration and 99% perspiration… or is it? An investigation of the impact of motivation and feedback on deception detection. *Legal Criminol Psychol.* (2007) 12:297–309. doi: 10.1348/135532506X143958
6. DePaulo BM, Charlton K, Cooper H, Lindsay JJ, Muhlenbruck L. The accuracy-confidence correlation in the detection of deception. *Person Soc Psychol Rev.* (1997) 1:346–57. doi: 10.1207/s15327957pspr0104_5
7. Goffman E. The nature of deference and demeanor. *Am Anthropol.* (1956) 58:473–502. doi: 10.1525/aa.1956.58.3.02a00070
8. Stone M. Instant lie detection. Demeanor and credibility in criminal trials. *Criminal Law Rev.* (1991) 821–30.
9. Douglas J. How should tribunals evaluate the evidence? In: *Paper presented at 7th Annual Tribunal's Conference of the Australian Institute of Judicial Administration.* Brisbane, QLD (2004).
10. Arce R, Seijó A, Novo M. Testimony validity: a comparative study of legal and empirical criteria. *Psychol Spain* (2010). 14:1–7.
11. Rosenthal JR. Suggestibility, reliability, and the legal process. *Dev Rev.* (2002) 22:334–69. doi: 10.1016/S0273-2297(02)00002-3
12. Ekman P. Lie catching and microexpressions. In: Martin C, editor. *The Philosophy of Deception*. New York, NY: Oxford University Press (2009). p. 118–33. doi: 10.1093/acprof:oso/9780195327939.003.0008
13. Ekman P, Friesen WV. Detecting deception from the body or face. *J Pers Soc Psychol.* (1974) 29:288–98. doi: 10.1037/h0036006
14. Riggio RE, Friedman HS. Individual differences and cues to deception. *J Pers Soc Psychol.* (1983) 45:899–915. doi: 10.1037/0022-3514.45.4.899
15. Ekman P. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage.* Revised ed. New York, NY: Norton (2009).
16. Ekman P. Universals and cultural differences in facial expressions of emotion. In: Cole JK, editor. *Nebraska Symposium on Motivation.* Lincoln, NE: University of Nebraska Press (1972). p. 207–83.
17. Ekman P, Friesen WV. The repertoire of nonverbal behavior: categories, origins, usage, and coding. *Semiotica* (1969) 1:49–98. doi: 10.1515/semi.1969.1.1.49
18. Barrett LF, Lindquist KA, Gendron M. Language as context for the perception of emotion. *Trends Cogn Sci.* (2007) 11:327–32. doi: 10.1016/j.tics.2007.06.003
19. Gendron M, Roberson D, van der Vyver JM, Barrett LF. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion* (2014) 14:251–62. doi: 10.1037/a0036052
20. Frank MG, Feeley TH. To catch a liar: challenges for research in lie detection training. *J Appl Commun Res.* (2003) 31:58–75. doi: 10.1080/00909880305377
21. Vrij A, Semin GR. Lie experts' beliefs about nonverbal indicators of deception. *J Nonverbal Behav.* (1996) 20:65–80.
22. Blandón-Gitlin I, Fenn E, Masip J, Yoo AH. Cognitive-load approaches to detect deception: searching for cognitive mechanisms. *Trends Cogn Sci.* (2014) 18:441–4. doi: 10.1016/j.tics.2014.05.004
23. Sporer SL. Deception and cognitive load: expanding our horizon with a working memory model. *Front Psychol.* (2016) 7:420. doi: 10.3389/fpsyg.2016.00420
24. Vrij A, Granhag PA, Porter SB. Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychol Sci Public Interest* (2010) 11:89–121. doi: 10.1177/1529100610390861
25. Vrij A, Fisher R, Blank H. A cognitive approach to lie detection: a meta-analysis. *Legal Criminol Psychol.* (2017) 22:1–21. doi: 10.1111/lcrp.12088
26. Lane SM, Vieira KM. Steering a new course for deception detection research. *J Appl Res Mem Cogn.* (2012) 1:136–8. doi: 10.1016/j.jarmac.2012.04.001
27. Peace KA, Porter S. Remembrance of lies past: a comparison of the features and consistency of truthful and fabricated trauma narratives. *Appl Cogn Psychol.* (2011) 25:414–23. doi: 10.1002/acp.1708
28. Stromwall LA, Granhag PA. Children's repeated lies and truths: effects on adults' judgments and reality monitoring scores. *Psychiatry Psychol Law* (2005) 12:345–56. doi: 10.1375/pplt.12.2.345

29. Vredeveldt A, van Koppen PJ, Granhag PA. The inconsistent suspect: a systematic review of different types of consistency in truth tellers and liars. In: Bull R, editor. *Investigative Interviewing*. New York, NY: Springer (2014). p. 183–207. doi: 10.1007/978-1-4614-9642-7_10
30. Wells GL, Lindsay RC, Ferguson TJ. Accuracy, confidence, and juror perceptions in eyewitness identification. *J Appl Psychol.* (1979) 64:440–8. doi: 10.1037/0021-9010.64.4.440
31. Thorley C. Blame conformity: innocent bystanders can be blamed for a crime as a result of misinformation from a young, but not elderly, adult co-witness. *PLoS ONE* (2015) 10:e0134739. doi: 10.1371/journal.pone.0134739
32. Thorley C, Kumar D. Eyewitness susceptibility to co-witness misinformation is influenced by co-witness confidence and own self-confidence. *Psychol Crime Law* (2017) 23:342–60. doi: 10.1080/1068316X.2016.1258471
33. Brewer N, Burke A. Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law Hum Behav.* (2002) 26:353–64. doi: 10.1023/A:1015380522722
34. Kebbell MR, Johnson SD. Lawyers' questioning: the effect of confusing questions on witness confidence and accuracy. *Law Hum Behav.* (2000) 24:629–41. doi: 10.1023/A:1005548102819
35. Brigham JC, Wolfskeil MP. Opinions of attorneys and law enforcement personnel on the accuracy of eyewitness identifications. *Law Hum Behav.* (1983) 7:337–49. doi: 10.1007/BF01044736
36. Penrod S, Cutler B. Witness confidence and witness accuracy: assessing their forensic relation. *Psychol Publ Pol Law* (1995) 1:817–45. doi: 10.1037/1076-8971.1.4.817
37. Sporer SL, Penrod SD, Read JD, Cutler BL. Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychol Bull.* (1995) 118:315–27. doi: 10.1037/0033-2909.118.3.315
38. Johnson MK, Bush JG, Mitchell KJ. Interpersonal reality monitoring: judging the sources of other people's memories. *Soc Cogn.* (1998) 16:199–224.
39. Johnson MK, Hashtroudi S, Lindsay DS. Source monitoring. *Psychol Bull.* (1993) 114:3–28. doi: 10.1037/0033-2909.114.1.3
40. Johnson MK, Raye CL. Reality monitoring. *Psychol Rev.* (1981) 88:67–85. doi: 10.1037/0033-295X.88.1.67
41. Johnson MK, Foley MA, Suengas AG, Raye CL. Phenomenal characteristics of memories for perceived and imagined autobiographical events. *J Exp Psychol Gen.* (1988) 117:371–6. doi: 10.1037/0096-3445.117.4.371
42. Suengas AG, Johnson MK. Qualitative effects of rehearsal on memories for perceived and imagined complex events. *J Exp Psychol Gen.* (1988) 117:377–89. doi: 10.1037/0096-3445.117.4.377
43. Oberlader VA, Naefgen C, Koppehele-Gossel J, Quinten L, Banse R, Schmidt AF. Validity of content-based techniques to distinguish true and fabricated statements: a meta-analysis. *Law Hum Behav.* (2016) 40:440–57. doi: 10.1037/lhb0000193
44. Steller M, Köhnken G. Criteria-based statement analysis. In: Raskin DC, editor. *Psychological Methods in Criminal Investigation and Evidence*. New York, NY: Springer (1989). p. 217–45.
45. Amado BG, Arce R, Fariña F. Undeutsch hypothesis and criteria-based content analysis: a meta-analytic review. *Eur J Psychol Appl Legal Context* (2015) 7:3–12. doi: 10.1016/j.ejpal.2014.11.002
46. Amado BG, Arce R, Fariña F, Vilariño M. Criteria-Based Content Analysis (CBCA) reality criteria in adults: a meta-analytic review. *Int J Clin Health Psychol.* (2016) 16:201–10. doi: 10.1016/j.ijchp.2016.01.002
47. Masip J. Deception detection: state of the art and future prospects. *Psicothema* (2017) 29:149–59. doi: 10.7334/psicothema2017.34
48. Hart CL, Fillmore DG, Griffith JD. Deceptive communication in the workplace: an examination of beliefs about verbal and paraverbal cues. *Individ Differ Res.* (2010) 8:176–83.
49. Reinhard MA, Scharmach M, Müller P. It's not what you are, it's what you know: experience, beliefs, and the detection of deception in employment interviews. *J Appl Soc Psychol.* (2013) 43:467–79. doi: 10.1111/j.1559-1816.2013.01011.x
50. Sporer SL, Schwandt B. Paraverbal indicators of deception: a meta-analytic synthesis. *Appl Cogn Psychol.* (2006) 20:421–46. doi: 10.1002/acp.1190
51. Reinhard MA, Sporer SL, Scharmach M, Marksteiner T. Listening, not watching: situational familiarity and the ability to detect deception. *J Pers Soc Psychol.* (2011) 101:467–87. doi: 10.1037/a0023726
52. Hartwig M, Bond CF Jr. Lie detection from multiple cues: a meta-analysis. *Appl Cogn Psychol.* (2014) 28:661–76. doi: 10.1002/acp.3052
53. Swets JA. Measuring the accuracy of diagnostic systems. *Science* (1988) 240:1285–93. doi: 10.1126/science.3287615
54. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* (2010) 5:1315–6. doi: 10.1097/JTO.0b013e3181ec173d
55. Bates DM, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* (2015) 67:1–48. doi: 10.18637/jss.v067.i01
56. R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: URL http://www.R-project.org/
57. LeBeau B. *simglm: Simulate Models Based on the Generalized Linear Model* (2018). Available online at: https://cran.r-project.org/web/packages/simglm/index.html
58. Hughes J. Simulating power with the paramtest package (2017). Available online at: https://cran.r-project.org/web/packages/paramtest/index.html
59. Fox J, Weisberg S. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks, CA: Sage Publications (2018).
60. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res.* (2004) 33:261–304. doi: 10.1177/0049124104268644
61. Bond CF Jr, DePaulo BM. Individual differences in judging deception: accuracy and bias. *Psychol Bull.* (2008) 134:477–92. doi: 10.1037/0033-2909.134.4.477
62. Sebanz N, Shiffrar M. Detecting deception in a bluffing body: the role of expertise. *Psychon Bull Rev.* (2009) 16:170–5. doi: 10.3758/PBR.16.1.170
63. Sip KE, Lynge M, Wallentin M, McGregor WB, Frith CD, Roepstorff A. The production and detection of deception in an interactive game. *Neuropsychologia* (2010) 48:3619–26. doi: 10.1016/j.neuropsychologia.2010.08.013
64. Wu D, Loke IC, Xu F, Lee K. Neural correlates of evaluations of lying and truth-telling in different social contexts. *Brain Res.* (2011) 1389:115–24. doi: 10.1016/j.brainres.2011.02.084
65. Damasio AR. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philos Transac R Soc Lond B Biol Sci.* (1996) 351:1413–20. doi: 10.1098/rstb.1996.0125
66. Grèzes J, Frith C, Passingham RE. Brain mechanisms for inferring deceit in the actions of others. *J Neurosci.* (2004) 24:5500–5. doi: 10.1523/JNEUROSCI.0219-04.2004
67. Quarto T, Blasi G, Maddalena C, Viscanti G, Lanciano T, Soleti E, et al. Association between ability emotional intelligence and left insula during social judgment of facial emotions. *PLoS ONE* (2016) 11:e0148621. doi: 10.1371/journal.pone.0148621
68. Aamodt MG, Custer H. Who can best catch a liar?: a meta-analysis of individual differences in detecting deception. *Forensic Examiner* (2006). 15:6–11.
69. Levine TR, Kim RK, Sun Park H, Hughes M. Deception detection accuracy is a predictable linear function of message veracity base-rate: a formal test of Park and Levine's probability model. *Commun Monogr.* (2006) 73:243–60. doi: 10.1080/03637750600873736
70. Levine TR. Truth-default theory (TDT) a theory of human deception and deception detection. *J Lang Soc Psychol.* (2014) 33:378–92. doi: 10.1177/0261927X14535916
71. Bottoms BL, Peter-Hagene LC, Stevenson MC, Wiley TR, Mitchell TS, Goodman GS. Explaining gender differences in jurors' reactions to child sexual assault cases. *Behav Sci Law* (2014) 32:789–812. doi: 10.1002/bsl.2147
72. Dunlap EE, Lynch KR, Jewell JA, Wasarhaley NE, Golding JM. Participant gender, stalking myth acceptance, and gender role stereotyping in perceptions of intimate partner stalking: a structural equation modeling approach. *Psychol Crime Law* (2015) 21:234–53. doi: 10.1080/1068316X.2014.951648
73. Grubb A, Harrower J. Attribution of blame in cases of rape: an analysis of participant gender, type of rape and perceived similarity to the victim. *Aggress Violent Behav.* (2008) 13:396–405. doi: 10.1016/j.avb.2008.06.006
74. Quas JA, Bottoms BL, Haegerich TM, Nysse-Carris KL. Effects of victim, defendant, and juror gender on decisions in child sexual assault cases 1. *J Appl Soc Psychol.* (2002) 32:1993–2021. doi: 10.1111/j.1559-1816.2002.tb02061.x

75. Steffensmeier D, Hebert C. Women and men policymakers: does the judge's gender affect the sentencing of criminal defendants? *Soc Forces* (1999) 77:1163–96. doi: 10.1093/sf/77.3.1163

76. Goodwin KA, Kukucka JP, Hawks IM. Co-witness confidence, conformity, and eyewitness memory: an examination of normative and informational social influences. *Appl Cogn Psychol.* (2013) 27:91–100. doi: 10.1002/acp.2877

77. Schooler JW, Eich E. *Memory for Emotional Events*. New York, NY: Oxford University Press (2000).

78. Stanley JT, Blanchard-Fields F. Challenges older adults face in detecting deceit: the role of emotion recognition. *Psychol Aging* (2008) 23:24–32. doi: 10.1037/0882-7974.23.1.24

79. Sweeney CD, Ceci SJ. Deception detection, transmission, and modality in age and sex. *Front Psychol.* (2014) 5:590. doi: 10.3389/fpsyg.2014.00590

80. Ruffman T, Murray J, Halberstadt J, Vater T. Age- related differences in deception. *Psychol Aging* (2012) 27:543–9. doi: 10.1037/a00 23380

81. Law MK, Jackson SA, Aidman E, Geiger M, Olderbak S, Kleitman S. It's the deceiver, not the receiver: no individual differences when detecting deception in a foreign and a native language. *PLoS ONE* (2018) 13:e0196384. doi: 10.1371/journal.pone.0196384

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 23 January 2019
doi: 10.3389/fpsyt.2018.00745

Validity of the Reaction Time Concealed Information Test in a Prison Sample

Kristina Suchotzki, Aileen Kakavand and Matthias Gamer*

Department of Psychology, University of Würzburg, Würzburg, Germany

Detecting whether a suspect possesses incriminating (e.g., crime-related) information can provide valuable decision aids in court. To this means, the Concealed Information Test (CIT) has been developed and is currently applied on a regular basis in Japan. But whereas research has revealed a high validity of the CIT in student and normal populations, research investigating its validity in forensic samples in scarce. This applies even more to the reaction time-based CIT (RT-CIT), where no such research is available so far. The current study tested the application of the RT-CIT for an imaginary mock crime scenario both in a sample of prisoners ($n = 27$) and a matched control group ($n = 25$). Results revealed a high validity of the RT-CIT for discriminating between crime-related and crime-unrelated information, visible in medium to very high effect sizes for error rates and reaction times. Interestingly, in accordance with theories that criminal offenders may have worse response inhibition capacities and that response inhibition plays a crucial role in the RT-CIT, CIT-effects in the error rates were even elevated in the prisoners compared to the control group. No support for this hypothesis could, however, be found in reaction time CIT-effects. Also, performance in a standard Stroop task, that was conducted to measure executive functioning, did not differ between both groups and no correlation was found between Stroop task performance and performance in the RT-CIT. Despite frequently raised concerns that the RT-CIT may not be applicable in non-student and forensic populations, our results thereby do suggest that such a use may be possible and that effects seem to be quite large. Future research should build up on these findings by increasing the realism of the crime and interrogation situation and by further investigating the replicability and the theoretical substantiation of increased effects in non-student and forensic samples.

Keywords: concealed information test, deception, lying, reaction times, inmates, forensic sample

Edited by:
Giuseppe Sartori,
Università degli Studi di Padova, Italy

Reviewed by:
Laura Visu-Petra,
Babeş-Bolyai University, Romania
Galit Nahari,
Bar-Ilan University, Israel

****Correspondence:***
Kristina Suchotzki
kristina.suchotzki@uni-wuerzburg.de

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *30 July 2018*
Accepted: *17 December 2018*
Published: *23 January 2019*

Citation:
Suchotzki K, Kakavand A and Gamer M (2019) Validity of the Reaction Time Concealed Information Test in a Prison Sample. Front. Psychiatry 9:745. doi: 10.3389/fpsyt.2018.00745

INTRODUCTION

Valid lie detection tests would provide valuable means in police interrogations and court, yet unfortunately most lie detection test that have been developed so far are not endorsed by the scientific community. For instance, the Comparison Question Test [also called the Control Question Test, CQT; (1)] has been strongly criticized for its lack of an adequate control condition and its high rate of false positives [i.e., truthful suspects being determined as deceptive; (2, 3)]. Nevertheless, the CQT is the most popular and most commonly applied deception detection test being used by police and secret service in many countries worldwide (e.g., USA and Israel), and in some even admissible as evidence in court (e.g., Belgium). For several years now, scientists have raised their concern about this method and proposed it being replaced with evidence-based tools (4, 5). One of those proposed methods to replace the CQT is the so-called Concealed Information Test [CIT; (6)]. Developed by one of the earliest critics of the CQT, the Concealed Information Test (CIT) does not aim to detect deception, but rather whether a suspect possesses certain incriminating knowledge (therefore, the test had been originally termed the Guilty Knowledge Test). In the CIT, the suspect is presented with a question that only someone with critical crime-knowledge can answer, for instance: "What was the color of the bag that was stolen?" The suspect then receives several possible neutral answers, among which the correct one is hidden, for instance: "Yellow," "Green," "Blue," "Red," and "Black." Depending on the CIT version and the dependent measure that is used, the suspect may be instructed to simply listen to those answers or to respond "No" to each of them. The CIT relies on the idea that only a knowledgeable suspect will recognize the correct answer. Note here that therefore the test will never come to the conclusion that a certain suspect is guilty, only that (s)he may be knowledgeable of certain crime aspects. Where this knowledge comes from (e.g., committing the crime, observing the crime, hearsay) needs to be determined in further interrogations. Crucially, it has been found that such recognition leads to measurable changes in different autonomic indices, as for instance an increase in skin conductance, and a decrease in heart rate and respiration for the critical crime knowledge compared to the other neutral answer alternatives (7). No such changes should be observable in an unknowledgeable suspect, for which all alternatives should be equally likely. As the most recent meta-analysis has shown, CIT validity is very promising as evident in a very high effect size (Cohen's d) for the differentiation between knowledgeable and unknowledgeable test subjects [$d = 1.55$, $d = 0.89$, and $d = 1.11$ for skin conductance, heart rate and respiration, respectively; (8)].

More recently, it has been shown that behavioral measures such as reaction times (RTs) also show some promise for CIT applications (9, 10). Note that in order to ensure attention to the stimuli, the CIT was for this purpose adapted by asking participants to respond "No" to each of the critical and neutral answer alternatives and to respond "Yes" to a number of designated (crime unrelated) target items (usually via button presses). Using this adapted RT-CIT version also results in a very high effect size, this time calculated as the RT difference between critical and neutral items [$d = 1.30$; for a meta-analysis see (11)]. The main advantage of RT measures in deception detection is their ease of application. For example, they do not require sophisticated equipment (one laptop suffices) or scoring procedures. They do, however, also have a number of potential disadvantages, one of them being that they may not be as easy applicable in populations that differ from the typically studied student and normal populations. Populations such as forensic ones may be less familiar with computerized testing and probably being generally slower may obscure or even eliminate RT CIT-effects. There are also theoretical considerations that may suggest that RT CIT-effects could differ between normal and forensic populations. Whereas the autonomic CIT has been shown to mostly rely on orienting toward familiar or significant stimuli (12–14), there are indications that in the RT-CIT, the requirement to suppress the automatic "Yes" response toward crime related items may also crucially contribute to the effect [i.e., response inhibition; (15–17)]. Importantly, research suggests that response inhibition capacities may be impaired in forensic populations, as well as impulsivity (a trait that has been discussed as being related to response inhibition) increased (18, 19). Thus, instead of being obscured or diminished in forensic populations, the response inhibition account would rather predict the RT CIT-effect to be increased in forensic populations due to an increased difficulty to suppress the unwanted truthful "Yes" response toward critical items. Being the first to employ the RT-CIT in a forensic sample, the current experiment aimed to explore those two contradicting predictions.

METHODS

Participants

In total, 30 male inmates of a youth detention center in the federal state of Baden-Württemberg in Germany volunteered to take part in the study. The study conformed to the principles expressed in the Declaration of Helsinki. All provided written informed consent. Inclusion criteria for the male control group were, based on the sample of inmates, an age between 16 and 25 years and no education higher than "mittlere Reife" (10 years of formal education, approximately equivalent to the General Certificate of Secondary Education, GCSE). Participants for the control group were recruited through paper and online advertisement ($n = 6$) and via a contact to a vocational school ($n = 26$). All participants from the control group provided written informed consent, and in case they were younger than 18, written informed consent was obtained from the parents. Data of one control participant were exculded because of his higher education. Data of three inmates and six control participants were excluded because they had <50% trials for one item type in the CIT after exclusion of trials exceeding the response deadline, error trials and RT outliers (see below). The mean age of the remaining 27 inmates was 20.15 years ($SD = 2.14$ years). The mean age of the remaining 25 control participants was 18.88 years ($SD = 3.17$ years). There was no significant age difference between both groups, $t_{(41.74)} = 1.68$, $p = 0.101$, $d = 0.47$.

Procedure

Testing took place in a quiet room in the youth detention center, in the vocational school building, or at the University. Participants first answered a questionnaire asking for the following demographical data: age, mother tongue, origin, if origin was not German, how long they had already been in Germany, education, type of current employment, and handedness. They then received the instruction that they would see a picture story on the screen of a laptop and they should try to imagine experiencing the depicted scenario. Participants were told to imagine they had to go to the doctor and were in the *waiting room*. They would be alone there and would see a forgotten *handbag*. They would seize the opportunity and look inside the bag. There they would find an *identity card with the name Maria*. They would continue their search and find a *ring* that they would decide to steal. They would still continue and find a *smartphone* that they would also take. Then they would quickly leave the waiting room. Words marked in italics refer to the pictures (i.e., photographs) that were depicted on the screen. Pictures were taken from the internet and can be obtained from the authors upon request (sharing them with the data is not possible due to copyright issues). Participants then saw a short summary of their imaginary activity on the screen: "You were in a WAITING ROOM and stole a RING and a SMARTPHONE from the HANDBAG of MARIA." Note that the words printed in capital letters were the ones that were later used as critical items in the CIT. The experimenter then asked the participants to repeat those crime details to her, to ensure correct memory of those. Although such an explicit encoding procedure might differ from typical field situations where crime related information is rather encoded incidentally, we chose to use such a procedure to ensure that potential group differences in CIT detection efficacy were not related to group differences in memory for critical items. Now participants were informed that they were suspects of this theft and that they should therefore undergo a lie detection test. For this lie detection test, they further had to memorize five additional words (i.e., the target items). Those words were presented on paper and participants were asked afterwards to write them down to also ensure memory for those. If those were not written down correctly, the words were presented again and this was repeated until all words were remembered correctly. Participants were then instructed to do their best to hide their knowledge of the crime during the following lie detection test. Participants received the instructions for the CIT on the laptop screen. Those instructions specified that they would see words on the screen, one after the other. For each word they should judge as fast as possible, whether they recognized it or not. Importantly, they should only respond "Yes" to the words from the paper list and "No" to all other words. They should further try to always respond as fast and correctly as possible. Responses had to be given via the keyboard (see details below). Participants then performed the CIT. After the CIT, participants were asked to repeat the details from the picture story to the experimenter. They were then asked how motivated they were during the lie detection test (from 1 to 10), how difficult they experienced the test (from 1 to 10) and whether they used any specific strategies to pass the test. They were also asked whether they took any medication or suffered from a physical or mental illness. The experimenter additionally noted a subjective estimation of their German language proficiency (from 1 to 6, 1 being the best according to the German grading system). After this, participants received the instructions for the Stroop task, again on the laptop screen. Those instructions specified that participants would be presented with words in different colors. Their task was to indicate the color of each word while ignoring its meaning. As an example, it was explained that if the word RED would be presented in GREEN color, participants should say "GREEN." Participants were also told to respond as fast and correctly as possible, as their reaction time would be measured. They were also told that incorrect or too slow responses would result in a black "X" being presented on the screen. Participants then performed the Stroop task. After the Stroop task, participants received another Questionnaire in which they were asked how motivated they were during the Stroop test (from 1 to 10), how difficult they experienced the test (from 1 to 10) and in case they belonged to the control group, whether they were ever found guilty of a crime and if so, what this crime was. Finally, as a measure of trait impulsivity, participants were asked to fill in the Barratt Impulsiveness Scale [BIS-11; (20)]. The BIS-11 comprises 30 items and results in overall values between 30 and 120 with higher values indicating higher trait impulsivity.

Concealed Information Test

The Concealed Information Test (CIT) was programmed and presented with Inquisit 4. In the CIT, the Question "DO YOU RECOGNIZE THIS WORD" was always presented central in the upper part of the screen. Reminder labels for the two possible responses, "YES" and "NO" were always presented on the left and right lower part of the screen. The position of those labels and thereby the assignment to the "a" and "l" keys on a standard QUERTZ keyboard was counterbalanced between participants. In total, 30 different CIT items were presented centrally on the screen (5 target items, 5 critical items, and 20 neutral items). Note that words instead of pictures were used. A list of all used items can be found on https://osf.io/c5us4/. Each item was presented six times, resulting in 180 trials in total (plus 2 neutral buffer items at the beginning of each test block that were not analyzed). Items were presented in completely randomized order, yet in two blocks each containing each item three times. Between both blocks, participants could take a self-paced break. Each item was presented until a response was given and the inter-trial varied between 500 and 1,000 ms. If participants did not respond after 4,000 ms, the item also disappeared and the words "Too slow!" were presented in red centrally on the screen. No error feedback was given.

Stroop Task

The Stroop task was presented with Inquisit 4 and the script was taken from the Millisecond test library (http://www.millisecond.com/download/library/). The English instructions and stimuli were translated from English to German und adapted in the experiment script. Responses were given verbally and recorded with the speech recognition function of Inquisit 4. In the Stroop task, the words "red," "green," "blue," and "yellow" were always presented centrally on the screen in one of the four colors. Each color was presented 20 times, 10 times congruent with the corresponding word and 10 times incongruent with one of the other three words (which were chosen randomly). Colors were presented in completely randomized order. They were presented until a response was given and the inter-trial was 200 ms. If participants did not respond after 2,500 ms, the word also disappeared and the next trial started. In case of incorrect responses, error feedback was given in the form of a black "X" presented for 400 ms centrally on the screen.

RESULTS

Data were analyzed with R and raw data as well as analysis scripts can be accessed on https://osf.io/c5us4/. To compare demographics between both groups, Fisher's Exact Test for Count Data was used, testing the null hypothesis that the odds ratio is equal to one. Analysis steps for the CIT were as follows. First, trials exceeding the response deadline were excluded (2.78%). Mean error rates were computed separately for probes and irrelevant items and analyzed with a two (Group: inmates vs. control) × 2 (Item: critical vs. neutral) mixed ANOVA. Before conducting the same 2 × 2 ANOVA on RTs, error trials (9.40%) and RT outliers (2.40%; RTs >2.5 *SD*s from the mean per subject and item type) were removed. For the analysis of the Stroop task the preprogrammed standard script as implemented in the experimental task taken from http://www.millisecond.com/download/library/ was used. Here, error trials (3.39%) were removed, before mean RTs were computed separately for congruent and incongruent trials and analyzed with a two (Group: inmates vs. control) × 2 (Congruency: congruent vs. incongruent) repeated measures ANOVA.

For ANOVA effects, η_p^2 was calculated as a measure of effect size. For follow-up *t*-tests, the standardized mean difference *d* was calculated, with 0.20, 0.50, and 0.80 as thresholds for "small," "moderate," and "large" effects (21). When *d* was computed for dependent samples, it was corrected for inter-correlations (22, 23).

Demographics and Questionnaire

An overview of the demographic data is given in **Table 1**.

Ratings of the estimated German language proficiency (from 1 to 6, 1 being the best according to the German grading system, rated by the experimenter) as well as the number of remembered crime-related items and the motivation and perceived difficulty of the CIT and the Stroop task can be found in **Table 2**.

TABLE 1 | Demographic data of inmates and control group.

	Inmates	Control	*p*
ORIGIN			1
Germany	23	21	
Other	4	4	
MIGRATION BACKGROUND			1
Yes	16	15	
No	11	10	
MOTHER TONGUE			1
German	17	15	
Other	10	10	
HIGHEST EDUCATION			<0.001***
No diploma	2	9	
"Hauptschule" (9 years of formal education)	20	6	
"Realschule" (10 years of formal education)	5	10	
APPRENTICESHIP			0.046*
Yes	0	4	
No	27	21	
CURRENT EMPLOYMENT			0.344
School	7	9	
Apprentice	19	13	
Employed	1	3	
HANDEDNESS			1
Left	3	2	
Right	24	23	

*p-values reported two-tailed. *p < 0.05, ***p < 0.001.*

TABLE 2 | Questionnaire data of inmates and control group.

	Inmates	Control	*t*	*df*	*p*	*d*
Language proficiency	1.30 (0.47)	1.04 (0.20)	2.61	35.86	0.013*	0.72
Number of remembered Items	4.70 (0.72)	4.68 (0.63)	0.13	49.79	0.900	0.04
Motivation CIT	8.44 (1.70)	8.24 (2.05)	0.39	46.75	0.698	0.11
Difficulty CIT	4.67 (2.39)	4.60 (2.10)	0.11	49.88	0.916	0.03
Motivation Stroop	8.44 (1.72)	8.64 (1.71)	0.41	49.75	0.682	0.11
Difficulty Stroop	4.63 (2.68)	4.72 (2.30)	0.13	49.74	0.896	0.04

*Standard deviations are given in brackets. p-values reported two-tailed. *p < 0.05.*

Results CIT

The mean error rate for all four conditions can be found in **Table 3**. The 2 × 2 ANOVA on the error rate revealed a significant main effect of Group, $F_{(1,50)} = 6.06$, $p = 0.017$, $n_p{}^2 = 0.11$, with a higher error rate for the inmates compared to the control group. It also revealed a significant main effect of Item, $F_{(1,50)} = 24.43$, $p < 0.001$, $n_p{}^2 = 0.33$, with a higher error rate for critical compared to neutral items. These effects were qualified by a significant interaction of Group x Item, $F_{(1,50)} = 5.90$, $p = 0.019$, $n_p{}^2 = 0.11$, with a larger CIT-effect (i.e., differences between critical and neutral items) in the inmates $t_{(26)} = 4.30$, $p < 0.001$, $d = 0.83$, compared to the control group, $t_{(24)} = 2.66$, $p = 0.014$, $d = 0.53$.

The mean RTs for all four conditions can be found in **Table 3**. The 2 × 2 ANOVA on the RTs revealed only a significant main effect of Item, $F_{(1,50)} = 144.24$, $p < 0.001$, $n_p{}^2 = 0.74$, with longer RTs for critical compared to neutral items. Neither the main effect of Group, $F_{(1,50)} = 0.07$, $p = 0.792$, $n_p{}^2 < 0.01$, nor the interaction of Group x Item, $F_{(1,50)} = 1.78$, $p = 0.188$, $n_p{}^2 = 0.03$, were statistically significant.

Results BIS-11 and Stroop Task

The mean BIS-11 value in the inmates group was $M = 65.89$ ($SD = 8.42$) and $M = 65.92$ ($SD = 8.24$) in the control group, with no significant difference between both groups, $t_{(49.84)} = 0.01$, $p = 0.989$, $d = 0.00$. Cronbach's α for the BIS-11 was not very high with 0.66. There were also no significant differences between both groups in any of the BIS-11 subscales, all *p*'s > 0.05. The mean RTs (in ms) for all four conditions in the Stroop Task were $M = 735.27$ ($SD = 97.57$) in the inmates group in the congruent trials, $M = 919.11$ ($SD = 128.52$) in the inmates group in the incongruent trials, $M = 807.95$ ($SD = 112.44$) in the control group in the congruent trials, and $M = 968.20$ ($SD = 138.84$) in the control group in the incongruent trials. The 2 × 2 ANOVA revealed only a significant main effect of Congruency, $F_{(1, 50)} = 252.35, p < 0.001, n_{\mathrm{p}}^2 = 0.83$, with longer RTs for incongruent compared to congruent trials. Neither the main effect of Group, $F_{(1, 50)} = 3.73, p = 0.059, n_p{}^2 = 0.07$, nor the interaction of Group × Congruency, $F_{(1, 50)} = 1.19, p = 0.281, n_p{}^2 = 0.02$, were statistically significant.

Correlations

Correlations between both CIT-effects, Stroop effects and participants' scores in the BIS-11 are shown in **Table 4**. As can be seen, there was only a significant correlation between the CIT-effects in the error rate and the RTs, but no significant correlations between those and the Stroop effects or the BIS-11 values. Note that based on the suggestion of a reviewer, we also checked the intercorrelations between CIT-effects and Stroop effects and the BIS-11 subscales (while controlling for multiple testing due to the exploratory nature of those analyses), which also revealed no significant correlations.

DISCUSSION

The aim of the current study was to explore the applicability of the RT-CIT in a sample different from the samples usually investigated in experimental research. This is particularly important as the latter differ fundamentally from the ones in which a CIT would ultimately be applied on and even currently is in field investigations in Japan. Nevertheless, studies examining the CIT in forensic samples are very scarce and particularly for the RT-CIT even non-existing. In the current study, we therefore recruited inmates of a youth detention center to complete an imaginary mock crime and afterwards an RT-CIT. As a control group, we recruited a sample that we tried to match as closely as possible regarding age and education background. Note that thereby also the control group differs from the student samples usually investigated in psychological research.

TABLE 3 | Mean error rates and RTs in all four experimental CIT conditions.

	Error rate (in %)		Reaction times (in ms)	
	Inmates	Control	Inmates	Control
Critical items	11.64 (13.30)	4.41 (6.35)	1033.02 (204.55)	1016.76 (239.43)
Neutral items	1.00 (1.62)	0.78 (1.35)	744.49 (118.41)	785.91 (172.40)
$d_{\mathrm{CIT-effect}}$	0.83	0.53	1.86	1.48

Standard deviations are given in brackets. $d_{CIT-effect}$ *values refer to the difference between critical and neutral items within each group.*

TABLE 4 | Correlations (*r*) between CIT-effects, Stroop-effects and BIS-11.

Measure	ER CIT- effect	RT CIT-effect	Stroop effect	BIS-11
ER CIT-effect	–	–	–	–
RT CIT-effect	0.51***	–	–	–
Stroop effect	0.04	−0.14	–	–
BIS-11	−0.17	−0.16	0.15	–

*p-values reported two-tailed. *** = p <0.001.*

The first notable result is that in both samples, the RT-CIT produced medium to large effects in error rate and RTs. Effects were larger in the RTs than in the error rate, which is in accordance with results usually obtained with the RT-CIT [e.g., (10, 24–26)]. This result is of course very promising for applied contexts and speaks against the argument that the RT-CIT may not be applicable in samples that are less familiar with computerized tests. Note here that one adaptation that we made is that instead of the typically used response deadlines of 800 or 1,000 ms (9, 10, 25, 26), we used a longer response deadline of 2,500 ms. This was primarily done to ensure that the RT-CIT would also be applicable in participants with generally slower responding. The use of short response deadlines does therefore not seem mandatory to obtain stable RT-CIT effects and the mean RTs in our samples indicate that a shorter response deadline may still have been applicable. Such a shorter response deadline would also be desirable as it makes it harder for suspects to strategically slow down responses and employ so-called countermeasures (see also below).

The second notable result is that at least in the error rates, CIT-effects were even stronger in the inmate group compared to the control group. Although numerically also the case for the RTs, this difference did not become significant. This allows a number of possible explanations. First, the absence of significant group differences in the RTs may simply represent a power issue and may not necessarily indicate a genuine dissociation between both measures. However, even though we cannot ensure an absence of group differences in RTs, our data at least indicate that such group differences seem to be larger for error rates as compared to RTs. Second, the current pattern of results might indicate differences between both groups in their speed accuracy trade-off. Thus, control participants might have concentrated more on avoiding errors even at the expense of longer response latencies than inmates. Whereas, the generally higher error rate for the inmates compared to the control group substantiates this notion, the absence of reversed general effects for RTs speaks against such shift of the response criterion. Of course, we also cannot exclude from our data that the difference between both groups in the error rate may constitute a chance finding, and a replication of our finding, preferably by a different research group, would be highly desirable. Note also that as mentioned above, our control group was deliberately designed to be closely matched to our inmate group, as we wanted to isolate differences related to the forensic background of the inmates and minimize differences related to age or education. One would, however, expect differences to be even larger between forensic samples and the ones typically tested in experimental research, a hypothesis that would be worth pursuing in future research. Such research should also incorporate a formal assessment of IQ, instead of only assessing education levels.

Importantly, our data provides no support for the hypothesis that differences in response inhibition capacities or impulsivity may explain larger CIT-effects in our forensic sample. While based on previous findings it is not so surprising that we did not find any correlation between our behavioral measure of executive functioning (i.e., the Stroop task) and our impulsivity measure [i.e., the BIS-11; (27–29)], it was unexpected that we even failed to observe differences in those measures between both groups. One explanation here may be that despite our matching not having succeeded perfectly (with differences in education and language proficiency), groups were still very similar. Also here, increasing group differences between the forensic and the control group may increase differences in executive functioning and impulsivity traits between both groups. The absence of a correlation between the BIS-11 and the Stroop effect with both CIT-effects, respectively, does, however, question the hypothesis that differences in those constructs may explain any differences in the size of CIT-effects. Note that this is against theoretical accounts and previous results indicating a substantial contribution of failures of response inhibition to deception and the RT CIT-effect (16, 17). It is, however, noteworthy that despite the popularity of this account, results so far are still mixed [see e.g., (30, 31)] and one fundamental challenge that has still received insufficient attention would be to better isolate which of the different facets of executive functioning [working memory vs. response inhibition vs. task switching (15, 32)] or even response inhibition [e.g., interference inhibition vs. action cancelation; measured with e.g., Stroop or Stop-Signal tasks; (33)] is the one that actually contributes to the CIT-effect.

As mentioned above, our findings seem promising for applied contexts, although it should be kept in mind here that so far, the CIT is only rarely applied and accepted in court. An exception is Japan where ∼5,000 CIT examinations are carried out by the police each year (34). However, CIT examinations are based on recordings of autonomic nervous system activity in Japan and not on behavioral measures as in the current study. Yet even with the autonomic CIT, experimental research in forensic samples (35, 36) or field investigations in such populations (37–39) are still very rare. Filling this gap seems important for two reasons. First, it would provide information on the validity of the CIT in the population in which it is actually applied, providing the basis for a more informed debate on whether this test should be applied and, as supported by many CIT researchers (4) replace currently used invalid lie detection methods (e.g., the CQT). Second, it would be very interesting from a theoretical perspective, as it has been argued that the autonomic and the RT-CIT differ with regard to their underlying psychological mechanisms [orienting vs. response inhibition (16)]. Following this line of arguments, one would expect the autonomic CIT to be less affected by the specific population than the RT-CIT. Another interesting question to pursue would be to what degree different populations may differ with regard to their potential countermeasure use. Countermeasures are deliberate strategies taken by suspects in order to systematically influence their test outcome and increase their chance of being classified innocent (40). The likelihood and the ability to successfully employ countermeasures may be dependent on many variables (e.g., experience with the CIT and/or computer-based testing, education) and may therefore differ between populations. On a related note, it has also often been hypothesized that people with psychopathic personality traits, whose prevalence is higher in forensic samples, may have better deception skills (41–45), which may result in smaller CIT effects or an increased likelihood to successfully implement countermeasures. Future research should therefore also aim to employ assessments of psychopathy.

One of the limitations of the current study is certainly the use of an imaginary instead of an actual mock crime scenario. The reasons that we employed an imaginary one were to be independent of the specific locations the experiment was run at (e.g., the detection center and the school) and ethical considerations, as we did not want to give the impression of furthering "illegal" behavior in a forensic population, even if it was only a role play (as is usually the case in mock crimes). Future research should, however, aim at increasing the realism of the crime and interrogation situation, in order to obtain information to what degree for instance a larger emotional involvement may impact crime-related memory in forensic populations (46). Such a more ecologically valid crime could for instance involve an actual mock crime, which should of course be very carefully instructed as role play in a prison sample. The same is true for increasing the realism of the interrogation situation, in which the experimenter could be introduced as actual police interrogator, which for instance conducts the test for training purposes.

To sum up, the current study provides a first crucial step toward an investigation of the RT-CIT in a forensic population. It indicates the usability of the RT-CIT in such a population, with even some support that effects may even be stronger. Further research should continue this challenge by investigating the replicability of those effects as well was their theoretical substantiation.

ETHICS STATEMENT

The ethics committee of the Department of Psychology of Wuerzburg University usually does not require ethical approval for single studies using well-established (also slightly adapted) experimental protocols and procedures that have obtained ethical approval before (as is the case in our study). The study was discussed and approved by the responsible at the JVA Adelsheim, in which we recruited part of our sample.

AUTHOR CONTRIBUTIONS

KS was involved in the study conception and the design, the analysis and interpretation of the data, and the writing of the manuscript. AK was involved in the recruitment of the participants, the collection and interpretation of the data, and the critical revision of the manuscript. MG was involved in the study conception and the design, the interpretation of the data, and the critical revision of the manuscript.

ACKNOWLEDGMENTS

We thank Jürgen Naber and Arne Ringlstetter, and the JVA Adelsheim for their support with our data acquisition. This publication was funded by the German Research Foundation (DFG) and the University of Wuerzburg in the funding programme Open Access Publishing.

REFERENCES

1. Reid JE. A revised questioning technique in lie-detection tests. *J Crim Law Criminol.* (1947) 37:542–7. doi: 10.2307/1138979
2. Ben-Shakhar G. A critical review of the Control Questions Test (CQT). In: Kleiner M, editor. *Handbook of polygraph testing*. Waltham, MA: Academic Press (2002). p. 103–26.
3. Lykken DT. *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. New York, NY: Plenum Press (1998).
4. Meijer EH, Verschuere B, Gamer M, Merckelbach H, Ben-Shakhar G. Deception detection with behavioral, autonomic, and neural measures: conceptual and methodological considerations that warrant modesty. *Psychophysiology* (2016) 53:593–604. doi: 10.1111/psyp.12609
5. National Research Council. *The Polygraph and Lie Detection. Committee to Review the Scientific Evidence on the Polygraph. Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academic Press (2003).
6. Lykken DT. The Gsr in the detection of guilt. *J Appl Psychol.* (1959) 43:385–8. doi: 10.1037/h0046060
7. Ambach W, Gamer M. Physiological measures in the detection of deception and concealed information. In: Rosenfeld JP, editor. *Detecting Concealed Information and Deception*. London: Academic Press (2018). p. 3–34. doi: 10.1016/B978-0-12-812729-2.00001-X
8. Meijer E, Klein Selle N, Elber L, Ben-Shakhar G. Memory detection with the Concealed Information Test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology* 51:879–904. doi: 10.1111/psyp.12239
9. Seymour TL, Seifert CM, Shafto MG, Mosmann AL. Using response time measures to assess "guilty knowledge". *J Appl Psychol.* (2000) 85:30–7. doi: 10.1037/0021-9010.85.1.30
10. Verschuere B, Crombez G, Degrootte T, Rosseel Y. Detecting concealed information with reaction times: validity and comparison with the polygraph. *Appl Cogn Psychol.* (2010) 24:991–1002. doi: 10.1002/acp.1601
11. Suchotzki K, Verschuere B, Van Bockstaele B, Ben-Shakhar G, Crombez G. Lying takes time: a meta-analysis on reaction time measures of deception. *Psychol Bull.* (2017) 143:428–53. doi: 10.1037/bul0000087
12. klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Orienting versus inhibition in the Concealed Information Test: different cognitive processes drive different physiological measures. *Psychophysiology* (2016) 53:579–90. doi: 10.1111/psyp.12583
13. Klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Unraveling the roles of orienting and inhibition in the Concealed Information Test. *Psychophysiology* (2017) 54:628–39. doi: 10.1111/psyp.12825
14. Verschuere B, Ben-Shakhar G. Theory of the Concealed Information Test. In: Verschuere B, Ben-Shakhar G, and Meijer E, editors. *Memory Detection: Theory and Application of the Concealed Information Test* Cambridge: Cambridge University Press (2011). p. 128–48. doi: 10.1017/CBO9780511975196.008
15. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn Psychol.* (2000) 41:49–100. doi: 10.1006/cogp.1999.0734
16. Suchotzki K, Verschuere B, Peth J, Crombez G, Gamer M. Manipulating item proportion and deception reveals crucial dissociation between behavioral, autonomic and neural indices of concealed information. *Hum Brain Mapp.* (2015) 36:427–39. doi: 10.1002/hbm.22637
17. Verschuere B, De Houwer J. Detecting concealed information in less than a second: response latency-based measures. In: Verschuere B, Ben-Shakhar G, and Meijer E, editors. *Memory Detection: Theory and Application of the Concealed Information Test* Cambridge: Cambridge University Press (2011). p. 46–63. doi: 10.1017/CBO9780511975196.004
18. Morgan AB, Lilienfeld SO. A meta-analytic review of the relation between antisocial behavior and neuropsychological measures of executive function. *Clin Psychol Rev.* (2000) 20:113–36. doi: 10.1016/S0272-7358(98)00096-8
19. Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH. Fifty years of the Barratt Impulsiveness Scale: an update and review. *Pers Individ Differ.* (2009) 47:385–95. doi: 10.1016/j.paid.2009.04.008
20. Patton JH, Stanford MS. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol.* (1995) 51:768–74. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1
21. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, MI: Lawrence Erlbaum (1988).
22. Dunlap, WP, Cortina, JM, Vaslow, JB, Burke, MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods* (1996) 1:170–7. doi: 10.1037//1082-989X.1.2.170
23. Morris SB, Deshon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods* (2002) 7:105–25. doi: 10.1037/1082-989X.7.1.105
24. Hu X, Evans A, Wu H, Lee K, Fu G. An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychol.* (2013) 142:278–85. doi: 10.1016/j.actpsy.2012.12.006
25. Visu-Petra G, Bus I, Miclea M. Detecting concealed information from a mock crime scenario by using psychophysiological and RT-based measures. *Cogn Brain Behav.* (2011) 15:19–37.
26. Visu-Petra G, Miclea M, Visu-Petra L. Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Appl Cogn Psychol.* (2012) 26:342–51. doi: 10.1002/acp.1827
27. Caswell AJ, Morgan MJ, Duka T. Inhibitory control contributes to "motor"-but not "cognitive"-impulsivity. *Exp Psychol.* (2013) 60:324–34. doi: 10.1027/1618-3169/a000202
28. Dougherty DM, Marsh-Richard DM, Hatzis ES, Nouvion SO, Mathias CW. A test of alcohol dose effects on multiple behavioral measures of impulsivity. *Drug Alcohol Depend.* (2008) 96:111–20. doi: 10.1016/j.drugalcdep.2008.02.002
29. Reynolds B, Ortengren A, Richards JB, de Wit H. Dimensions of impulsive behavior: personality and behavioral measures. *Pers Individ Differ.* (2006) 40:305–15. doi: 10.1016/j.paid.2005.03.024
30. Suchotzki K, Crombez G, Debey E, Van Oorsouw K, Verschuere B. *In vino* veritas? Alcohol, response inhibition and lying. *Alcohol Alcohol.* (2014) 50:74–81. doi: 10.1093/alcalc/agu079
31. Verschuere B, Schuhmann T, Sack AT. Does the inferior frontal sulcus play a functional role in deception? A neuronavigated theta-burst transcranial magnetic stimulation study. *Front Hum Neurosci.* (2012) 6:284. doi: 10.3389/fnhum.2012.00284
32. Christ SE, Essen DC, Watson JM, Brubaker LE, McDermott KB. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta analyses. *Cereb Cortex* (2009) 19:1557–66. doi: 10.1093/cercor/bhn189
33. Sebastian A, Pohl MF, Klöppel S, Feige B, Lange T, Stahl C, et al. Disentangling common and specific neural subprocesses of response inhibition. *Neuroimage* (2013) 64:601–15. doi: 10.1016/j.neuroimage.2012.09.020
34. Osugi A. Daily application of the concealed information test: Japan. In: Verschuere B, Ben-Shakhar G, and Meijer EH, editors. *Memory Detection: Theory and Application of the Concealed Information Test.* Cambridge: Cambridge University Press (2011). p. 253–75. doi: 10.1017/CBO9780511975196.015
35. Verschuere B, Crombez G, De Clercq A, Koster EH. Psychopathic traits and autonomic responding to concealed information in a prison sample. *Psychophysiology* (2005) 42:239–45. doi: 10.1111/j.1469-8986.2005.00279.x
36. Verschuere B, Crombez G, Koster EHW, De Clercq A. Antisociality, underarousal and the validity of the Concealed Information Polygraph Test. *Biol Psychol.* (2007) 74:309–18. doi: 10.1016/j.biopsycho.2006.08.002
37. Elaad E. Detection of guilty knowledge in real-life criminal investigations. *J Appl Psychol.* (1990) 75:521–9. doi: 10.1037/0021-9010.75.5.521

38. Elaad E, Ginton A, Jungman N. Detection measures in real-life criminal Guilty Knowledge Tests. *J Appl Psychol.* (1992) 77:757–67. doi: 10.1037/0021-9010.77.5.757

39. Suzuki R, Nakayama M, Furedy JJ. Specific and reactive sensitivities of skin resistance response and respiratory apnea in a Japanese concealed information test (CIT) of criminal guilt. *Can J Behav Sci.* (2004) 36:202–19. doi: 10.1037/h0087230

40. Ben Shakhar G. Countermeasures. In: Verschuere B, Ben Shakhar G, and Meijer E, editors *Memory Detection: Theory and Application of the Concealed Information Test.* Cambridge: Cambridge University Press (2011). p. 200–14. doi: 10.1017/CBO9780511975196.012

41. Assadi SM, Noroozian M, Pakravannejad M, Yahyazadeh O, Aghayan S, Shariat SV, et al. Psychiatric morbidity among sentences prisoners: prevalence study in Iran. *Br J Psychiatry* (2006) 188:159–64. doi: 10.1192/bjp.188.2.159

42. Hare RD. *Manual for the Revised Psychopathy Checklist.* 2nd ed. Toronto, ON, Canada: Multi-Health Systems (2003).

43. Hare RD, Forth AE, Hart SD. The psychopath as prototype for pathological lying and deception. In: Yuille JC, editor. *Credibility Assessment*. Dordrecht: Springer (1989). p. 25–49. doi: 10.1007/978-94-015-7856-1_2

44. Ullrich S, Paelecke M, Kahle I, Marneros A. Categorical and dimensional assessment of psychopathy in German offenders. Prevalence, gender differences and aging. *Der Nervenarzt.* (2003) 74:1002–8. doi: 10.1007/s00115-003-1495-4

45. Verschuere B, Crombez G, Koster E, Uzieblo K. Psychopathy and physiological detection of concealed information: a review. *Psychol Belgica* (2006) 46:99–116. doi: 10.5334/pb-46-1-2-99

46. Peth J, Vossel G, Gamer M. Emotional arousal modulates the encoding of crime-related details and corresponding physiological responses in the Concealed Information Test. *Psychophysiology* (2012) 49:381–90. doi: 10.1111/j.1469-8986.2011.01313.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

REVIEW
published: 05 February 2019
doi: 10.3389/fpsyt.2019.00024

Broadening the Use of the Concealed Information Test in the Field

Izumi Matsuda, Tokihiro Ogawa and Michiko Tsuneoka*

National Research Institute of Police Science, Tokyo, Japan

Japan is the only country where the polygraph with the concealed information test (CIT) is widely applied to criminal investigations. The CIT can reveal whether an examinee has knowledge of specific details of a crime. Furthermore, the CIT can extract crime-relevant information that investigative organizations have not yet uncovered. This article introduces how Japanese polygraphers take advantage of the CIT in criminal investigations. We also describe how polygraphs with the CIT are currently used in court. Then we propose statistical discrimination methods that can be easily applied to CIT interpretation in the field. Appropriate application of the statistical values is discussed. We hope that this article will facilitate more active use of the CIT outside Japan.

Keywords: concealed information test (CIT), statistical discrimination, field application, memory detection, searching CIT

OPEN ACCESS

Edited by:
Cristina Scarpazza,
University of Padova, Italy

Reviewed by:
Kristina Suchotzki,
Universität Würzburg, Germany
Antonietta Curci,
Università degli Studi di Bari, Italy

****Correspondence:***
Izumi Matsuda
izumimmatsuda@gmail.com

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *14 August 2018*
Accepted: *15 January 2019*
Published: *05 February 2019*

Citation:
Matsuda I, Ogawa T and Tsuneoka M (2019) Broadening the Use of the Concealed Information Test in the Field. Front. Psychiatry 10:24. doi: 10.3389/fpsyt.2019.00024

Many people regard the polygraph as a deception detection technique. However, the polygraph using the concealed information test (CIT) does not aim to detect deception: rather, it aims to detect crime-relevant memory. The CIT can assess whether an examinee knows details of a crime, despite saying "I don't know." The CIT also can provide clues about crime details that the investigative organization has not yet grasped. However, despite its effectiveness, the CIT is widely used only in Japan. In this article, we aim to address this situation and facilitate more active use of the CIT. We first introduce how Japanese polygraphers take advantage of the CIT. We then propose simple scoring methods and their possible thresholds, which can be easily applied in the field.

POLYGRAPH AS A MEMORY DETECTION TEST

The term *polygraph* generally refers to a test conducted with a polygraph device. In forensic situations, a polygraph measures autonomic responses to questions related to a crime. Autonomic responses, such as skin conductance and respiration, have high signal-to-noise ratios and can easily be measured outside controlled laboratory settings, unlike central measures such as electroencephalograms (1). Thus, in the field of criminal investigations, autonomic responses are still preferred to central responses (2).

There are several question techniques for the polygraph. Worldwide, the most commonly used technique is the control question test or comparison question test (CQT) (3). In the CQT, an examiner asks crime-relevant questions (e.g., "Did you rob the Mart last night?"), comparison questions (e.g., "Did you ever take something that did not belong to you?"), and neutral questions ("Did you live in the United States?"). The CQT aims to reveal whether an examinee has lied about the crime-relevant question by comparing the physiological responses for the crime-relevant and comparison questions.

The CIT, or the guilty knowledge test, is another question technique for the polygraph, although it does not directly aim to detect deception. The CIT assesses the examinee's memory of a particular crime detail (4, 5). For a question about the crime detail (e.g., the accessory that was stolen from the Mart), the examiner typically shows five items as possible answers (e.g., "a necklace?" "an earring?" "a watch?" "a brooch?" "a ring?"), including one correct (i.e., actually crime-relevant) item. These items are selected so that persons who do not know the crime detail cannot distinguish the crime-relevant item from the irrelevant items. The perpetrator can distinguish the crime-relevant item, but may attempt to avoid revealing this to the examiner, to conceal his or her involvement in the crime. Therefore, the CIT is conducted when the examinee claims that he or she does not know which is the crime-relevant item among the items. The examiner infers that the examinee in fact recognizes the crime-relevant item, despite his or her statement to the contrary, when the responses to the crime-relevant item differ from those to the crime-irrelevant items. Typically, greater skin conductance, suppressed respiration, slower heart rate, and smaller pulse volume are observed for the relevant item than for the irrelevant items [for reviews, see (1, 6)].

The validity of the CIT has been confirmed by laboratory studies. Elaad (7) conducted a meta-analysis of laboratory CIT studies and found that the weighted average of the false positive rates was 4.1%, and that of the false negative rates was 19.4%. A recent meta-analysis showed discrimination performance of each measure: the areas under the receiver operating characteristic (ROC) curve of skin conductance response, respiration, and heart rate were 0.848, 0.770, and 0.735, respectively (8). The CIT has been found to achieve high discrimination performance, with particularly low false positive rates (9).

THE CIT IN JAPAN

Despite the validity of the CIT described above, it is rarely used in real criminal investigations worldwide. One potential reason is that many practitioners have not known how to apply the CIT in the field. In this section, we introduce the field use of the CIT in Japan, where the CIT has been widely used for criminal investigations.

In Japan, the CIT is the only polygraph application used in criminal investigations. The CQT is not currently used at all. About 100 polygraph examiners deal with about 5,000 cases per year (10). These examiners administer the polygraph after completing a 3 month training course at the Forensic Science training center, affiliated with the National Research Institute of Police Science.

Figure 1 outlines how the polygraph is conducted in Japan. A consenting examinee receives the polygraph. At the beginning of the test, the examiner interviews the examinee to check what the examinee says about his or her knowledge of the crime. If the examinee says that he or she knows some crime details, the examiner will not perform CITs on these details.

Then the examiner attaches sensors to the examinee. In Japan, the examiner usually records several physiological measures: an electrocardiography (ECG), respiratory movement, skin conductance, and pulse wave. The ECG is used for computing heart rate. The pulse waves recorded with different filter settings are used for computing the normalized pulse volume (11).

The examiner conducts a so-called card test as a demonstration of a CIT. Typically, the examinee is asked to select one playing card from several playing cards with different numbers (e.g., 3, 4, 5, 6, and 7) and to memorize the number on it. Then the examiner asks the examinee which number he or she selected by presenting the numbers one by one, with an inter-stimulus interval of about 20–30 s. This process shows the examinee how the following CITs will be conducted. Additionally, through this card test the examiner can observe how the examinee physiologically responds to the item that he or she recognized.

Next, the examiner conducts the CITs. One CIT question usually consists of four to six items, one of which is supposed to be related to the crime. Before conducting the CIT, the examiner shows the examinee the CIT question and all included items and confirms the following three points. First, whether the examinee understands the meaning of the question and the items. If the examinee seems to have trouble with understanding the question or the items, the examiner adds explanations or replaces words with easier ones. Second, whether the examinee claims to know which item is crime-relevant. If the examinee says, prior to the test, that he or she can identify the crime-relevant item, the examiner does not conduct the test for that question. Finally, whether the examinee says that he or she is concerned about any items. For example, in the above CIT on the stolen accessory, if an examinee bought a watch a few days before, he may show a large response to the item "watch," even though the examinee has no crime-related knowledge. If the examinee says that he or she is concerned about a certain item, the examiner often replaces it to another item or discards the question.

In the CIT, the examiner vocally, and sometimes visually, presents each item, with the inter-stimulus interval of about 20–30 s. After all items have been presented, a short break is inserted if needed. This process is usually repeated 3–5 times, changing the order of the items to remove possible confounding effects due to the presentation order. After the CIT, the examiner often asks to the examinee whether he or she has any concerns about the test.

Based on the responses to the items, the examiner examines whether responses to a specific item are different from those to other items. If the examiner observes differences in responses between items, the examiner will infer that the examinee recognizes a specific item as crime-relevant.

Typically, the examiner conducts 4–7 CIT questions (12), each of which deals with different crime-relevant information. For example, in a theft case, in addition to the CIT on the stolen item, the examiner may conduct CITs on the time the crime happened, the crime scene, and the placement of the stolen item at the scene.

WHAT THE CIT CAN REVEAL IN CRIMINAL INVESTIGATIONS

As described above, the CIT examines whether the examinee recognizes a crime-relevant item that only a person associated

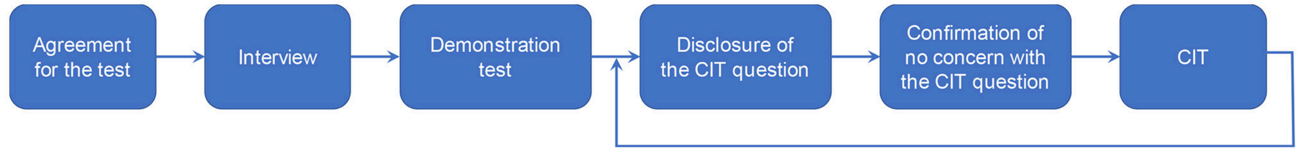

Agreement for the test
Interview
Demonstration test
Disclosure of the CIT question
Confirmation of no concern with the CIT question
CIT

FIGURE 1 | Flowchart of the polygraph in Japan.

with the crime could possibly know. More concretely, the CIT is conducted in Japan (1) to reveal whether the examinee knows a specific criminal detail, (2) to obtain new crime-relevant information, and (3) to reveal whether the examinee's statement is true.

Whether the Examinee Knows a Criminal Detail

This is the most typical usage of the CIT, an example of which is described in section Polygraph as a Memory Detection Test. Consider that there a crime-relevant fact has been obtained through an investigation (e.g., a ring was stolen). If it is assumed that only a person related to the crime could know this crime-relevant fact, the CIT can be used to examine whether the examinee does indeed know the fact. If the CIT result indicates that the examinee knows the fact, the investigators will extend the investigation to reveal the reason (e.g., because the examinee committed the theft or was an accomplice).

New Crime-Relevant Information

The CIT also can reveal crime-relevant information that even investigative organizations have not yet discovered. This type of the CIT is called a searching CIT. The searching CIT is conducted in the same way as the usual CIT. However, in the searching CIT, the examiner does not know which item is crime-relevant. For example, consider a case that a woman is missing. In this case, the examiner might conduct a CIT on the woman's location. The examiner may ask "Is she in City A? City B? City C? City D? City E? Another city?" to the examinee and compare responses among items. If the responses differ between City C and other items, the examiner infers that the examinee knows that she is in City C. In this case, the investigators can focus their search on City C to find her. In this way, the result of the searching CIT can be used to find new evidence and streamline investigations. Osugi (10, 13) reported other practical examples in which the searching CIT has been applied.

Credibility of the Examinee's Statement

The CIT also can be used to infer whether the examinee's statement is true or not. Osugi (13) reported this example: an examinee who sold a stolen ring insists that he found the ring on the road. To determine whether this statement is true, the examiner can conduct a CIT consisting of other possibilities (e.g., "You received the stolen ring from someone without paying anything," "You paid money to get the stolen ring," "You stole the ring yourself and did it alone," "You stole the ring together with an accomplice," "You got the ring in some other way"). If differential responding is not observed for any items, the examinee's statement that he found the ring on the road would be evaluated as true. In contrast, if differential responding is observed for a specific item in the CIT, his statement would be considered false. The CIT can assess not only whether the statement is true, but also what the truth is, as the examinee remembers it. This type of CIT also can be used to examine eyewitness or victim statements. However, few research has been conducted on this topic; future research is expected to support this usage of CIT.

The Difference Between Laboratory and Field CIT

As shown above, the CIT is used in the field in Japan to reveal examinees' recognition of the details of a crime. This approach differs from that used in typical laboratory CIT studies, which usually integrate responses among all CIT questions and conclude whether the examinee is guilty or innocent (14–16). Ben-Shakhar and Elaad (17) reported that discrimination performance was much higher for integrating responses from 12 different CIT questions repeated once, than for integrating responses from one CIT question repeated 12 times.

However, in the field, it is sometimes difficult to find enough crime details that have not been publicly announced. Thus, Japanese examiners actively use the searching CIT (10). Since the crime-relevant item is not identified in the searching CIT, integrating multiple CIT questions is impossible.

Moreover, it is difficult to assume that a person relevant to a crime remembers all the details. He or she may forget or genuinely not know some details. For example, the CIT in a theft case may reveal that the examinee knows the time the crime happened and the crime scene, but does not recognize the placement of the stolen item at the scene. This suggests the possibility that the examinee only drove a perpetrator to the crime scene.

Analyzing CIT questions individually can reveal what the examinee knows and what he or she does not know about the crime. Such an approach is sometimes much more informative in criminal investigations than integrating the CIT questions to conclude whether the examinee is guilty or innocent. However, it should be noted that this approach requires a sufficient number of repetitions of each CIT question to maintain high discrimination performance (18, 19).

CIT IN COURTS

In Japan, the results of the polygraph are usually used by investigative organizations as tools to assess whether and how the examinee is related to the crime. The results are rarely dealt with in court: a few of the about 5,000 cases are discussed each year. However, the Supreme Court admitted polygraph results as an evidence in 1968. Recent legal literature has noted that the probative value of the CIT result can be relatively high if the CIT is correctly conducted to examine the defendant's knowledge of facts that only the perpetrator could know (20). That is, the CIT result that the defendant knows the crime-relevant fact can be one reference information for the judge to decide whether he/she is guilty.

We checked court precedents relevant to the polygraph for the last 10 years. In many cases, legal professionals have focused on whether differential responding to the crime-relevant fact could be explained other than via a memory obtained through perpetration. For example:

- The defendant might have had an opportunity to encounter the fact through interrogation and rumors.
- The defendant might have had prior concerns about the fact because of personal reasons irrelevant to the crime.
- The defendant might have speculated about the fact.

These possibilities can detract from the probative value of the CIT for demonstrating the defendant's knowledge about the crime-relevant fact. As we mentioned above, the examiner conducts the CIT after confirming that the examinee says that he or she has no concerns about any of the items. The examiner should properly denote this confirmation process in the report.

In criminal investigations, the CIT can also be used to extract new information that the investigators had not previously known about (section New Crime-Relevant Information), and to examine the credibility of the examinee's statement (section Credibility of the Examinee's Statement). When differential responding is observed for a specific item in these CITs, later criminal investigations try to obtain new facts or statements underpinning the results. However, if such new facts or statements are not obtained, these CIT results would be rarely discussed in court.

A REMAINING TASK FOR THE CIT IN JAPAN

In Japan, the CIT has been widely used in criminal investigations and sometimes discussed in court. However, there are issues that remain to be solved. One issue is related to the process for assessing physiological differences. Below, we introduce the current judgment method in Japan and discuss statistical judgment in the following sections.

Current Judgment Method in Japan

Japanese polygraphers primarily judge differences in autonomic responses by visual inspection. Osugi (10) explained this judgment process as follows: the examiner ascertains whether the examinee showed differential responses based on the charts, the difference between the mean responses to crime-relevant and irrelevant items, and the consistency of the response differences across repetitions. It has been repeatedly confirmed that the discrimination performance of this judgment is sufficiently high (21–23). The latest study was conducted by Ogawa et al. (23), where 36 Japanese polygraphers blindly judged experimental CIT data from 152 examinees by visual inspection. Eighty examinees performed a mock crime before the CIT, while 72 examinees did nothing. Of the cases, 20.4% were judged as inconclusive. Excluding the inconclusive cases, the hit rate was 86.4%, and the correct rejection rate was 94.5%.

This high performance of visual inspection judgments could be attributable to its flexibility for inter- and intra-individual response differences. Which autonomic measures clearly respond to the relevant item differs across individuals (24). Furthermore, an examinee's reactivity can change between the first half and the second half of the polygraph, because of habituation and fatigue. Visual inspection enables the examiner to flexibly adjust the measures to consider the examinee's response tendency at that time.

However, visual inspection is sometimes regarded as subjective and dependent on the skill and experience of the examiner (3, 25). Introducing statistical judgment methods will make the CIT more objective and scientifically valid, even if the performance does not increase (2, 13). Increased objectivity will enhance the probative power of CIT results in court.

Requirements of Statistical Methods for Field Use

Recently, researchers have proposed many statistical classification methods [(24, 26, 27) for a review, see (2)]. However, the chosen statistical method for interpreting CITs in the field should meet the following requirements.

(1) Simplicity. The examiner may have to explain the judgment process in court. A simple method is required so that law and citizen judges can understand easily.
(2) Low false positive rate. In criminal investigations, at least in Japan, attempts are made to avoid false charges as much as possible. Although the low likelihood of false positives constitutes a major advantage for the CIT (9), measures should be taken to minimize the occurrence of false positive cases, while maintaining the relatively small number of inconclusive and false negative cases.
(3) Manageability for missing measures. In the field, the examiner sometimes cannot use some measures for analysis. For example, the rate of electrodermal non-responsivity is about 25% (28). A statistical method that can flexibly deal with such a situation is preferable.
(4) Avoidance of database use. Autonomic responses are influenced by age, sex, season, time of day, and so on (28). A database that would be appropriate for all examinees is thus difficult to envision at present.

DISCRIMINATION BASED ON EFFECT SIZE AND RANDOMIZATION

Considering the above four conditions, Matsuda et al. (29) proposed the use of the *d* value for effect size (30–32) and the *p*-value of the randomization test (33). Both *d* and *p*-values can be simply computed and require no database. In this section, we first explain how to compute d values (section Known-Solution CIT) and p-values (section Searching CIT). We then introduce the performances of d and p-values as compared with that of a traditional method (i.e., Lykken scoring) according to Matsuda et al. (29) (section Summary of the Threshold). We also compared these performance data with those of recent machine learning methods.

Effect Size

Consider a CIT consisting of five items, each of which is presented five times, which measures heart rate, skin conductance, respiration, and normalized pulse volume. That is, the number of responses to the crime-relevant item is five (i.e., $n_1 = 5$) and the number of responses to crime-irrelevant items is 20 (i.e., $n_2 = 20$) for each measure. The difference between the mean of the responses to the crime-relevant item and the mean of the responses to the irrelevant items is divided by a standard deviation, which is the effect size d. The standard deviation for computing the effect size has several calculation methods (34). Here, we calculate the effect size d by the following pooled standard deviation (s_p) using the unbiased variance of the responses to the relevant item (s_1^2) and that to the irrelevant item (s_2^2):

$$s_p = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

In general, when the examinee recognizes the relevant item, the relevant item elicits greater skin conductance, but slower heart rate, depressed respiration, and smaller normalized pulse volume than the irrelevant items. Thus, ds of heart rate, respiration, and normalized pulse volume are multiplied by −1.

The effect size d is computed for each measure. To integrate the results of all measures, we simply average their d values so far. If some measures are missing, we can average the d values across the remaining measures.

Randomization Test

The randomization test calculates the probability that the response difference between relevant and irrelevant items is obtained randomly. If the response difference can be obtained randomly, it means that we might obtain a similar response difference by randomizing the correspondence between the responses and the items. The procedure of the randomization test is shown in **Figure 2**. We assume a CIT consisting of five items × five repetitions and measuring heart rate, skin conductance, respiration, and normalized pulse volume. As shown in **Figure 2A**, five out of the 25 values for each measure are randomly selected and relabeled as the responses to the relevant item; the remaining 20 values are relabeled as the responses to the irrelevant items. Then the difference is computed between the mean of the values relabeled as relevant and the mean of the values relabeled as irrelevant. This process is repeated up to thousands of times (here, 1,000 times). Thus, we obtain 1,000 generated response differences. Regarding skin conductance, as shown in **Figure 2B**, if the real difference is the xth largest among the generated response differences, the p-value is calculated as $x/1{,}000$ (e.g., if $x = 50$, $p = 0.05$). Regarding heart rate, respiration, and normalized pulse volume, if the real difference is the xth smallest among the generated response differences, the p-value is calculated as $x/1{,}000$. Unlike the t test, the randomization test does not assume population parameters (35), which would be preferable for the CIT, whose sample size is rather small.

The method of integrating the results of each measure is shown in **Figure 2C**. At first, the p-value of each measure is multiplied across all measures. This is the original multiplied p-value. In contrast, we can calculate the p-value for each of the 1,000 repetitions by ranking the generated response difference at a certain repetition among 1,000 generated response differences. We then multiply these p-values across all measures. Thus, 1,000 multiplied p-values are generated. If the original multiplied p-value is the xth smallest among the generated multiplied p-values, the integrated p-value is $x/1{,}000$. If some measures are missing, we can multiply the p-values across the remaining measures.

Performance of *d* and *p*

Matsuda et al. (29) assessed the performance of d and p-values using the dataset of Ogawa et al. (23). The dataset consists of experimental CIT data from 152 examinees. Eighty of the examinees stole a ring in a mock crime, and 72 did not. The CIT consisted of five accessory names, including "ring," each of which was presented five times to examinees. During the CIT, respiration line length, skin conductance, heart rate, and normalized pulse volume were measured. For more details about the dataset, see Matsuda et al. (32), which is written in English.

Matsuda et al. (29) computed the integrated d and p-values for each CIT, in addition to the integrated Lykken score. Lykken score is a traditional scoring method (4) that assigns 2 to the largest response and 1 to the second-largest response in a block of repetitions, and then summarizes the scores across all blocks. The Lykken scores were integrated across all measures by averaging. The area under the ROC curve was 0.92, 0.92, and 0.90, for the integrated d value, the integrated p-value, and the integrated Lykken score, respectively. However, the ROC curve showed that maintaining a low false positive rate is more difficult for the integrated Lykken score than for the integrated d and p-values. This is probably because the Lykken score necessarily assigns scores even if the examiner observes no salient response in the block.

Recently, many machine learning methods have been proposed. We applied typical machine learning methods to the same dataset used in Matsuda et al. (29) using the Classification Learner App in MATLAB R2018a. This app automatically calculates the performance of various classifiers by protecting against overfitting using cross-validation. We computed the area under the ROC curve of decision trees, discriminant analysis, logistic regression, support vector machine, nearest neighbors, and ensemble classification. The area under the ROC curve was 0.85 (decision tree), 0.92 (discriminant analysis), 0.92 (logistic regression), 0.91 (support vector machine), 0.91 (nearest neighbors), and 0.92 (ensemble classification). The performances of the machine learning methods are almost the same as those of d and p-values. The calculation of d and p-values is simpler

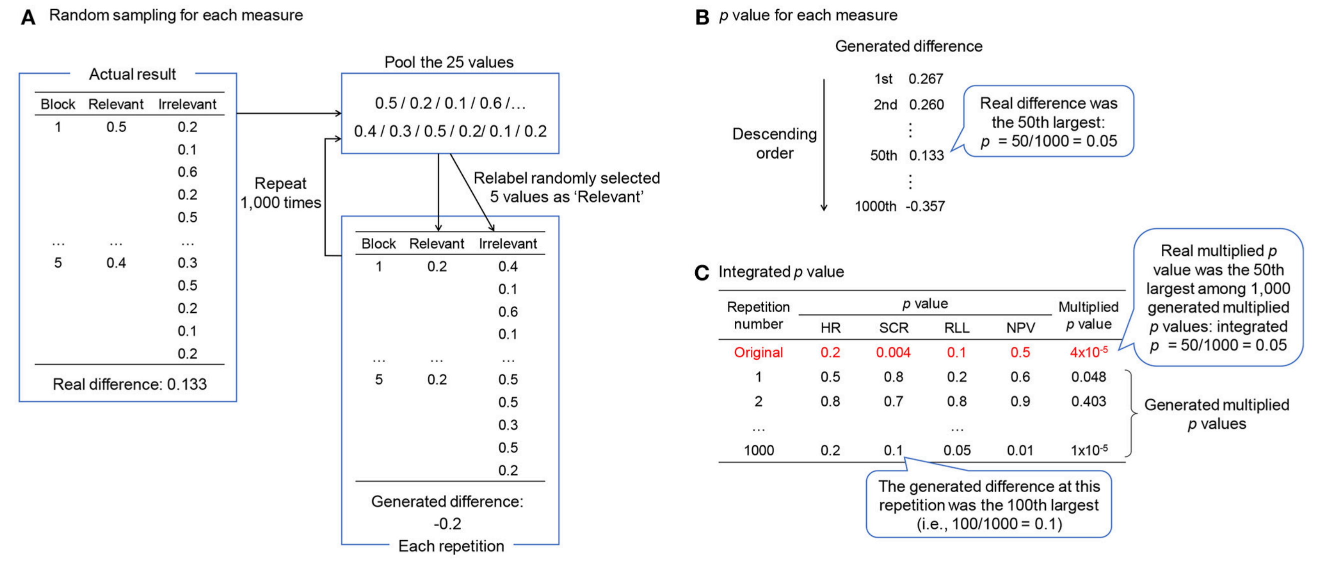

A Random sampling for each measure
Actual result
Block Relevant Irrelevant
1 0.5 0.2
0.1
0.6
0.2
0.5
… … …
5 0.4 0.3
0.5
0.2
0.1
0.2
Real difference: 0.133
Repeat 1,000 times
Pool the 25 values
0.5 / 0.2 / 0.1 / 0.6 /…
0.4 / 0.3 / 0.5 / 0.2/ 0.1 / 0.2
Relabel randomly selected 5 values as 'Relevant'
Block Relevant Irrelevant
1 0.2 0.4
0.1
0.6
0.1
… … …
5 0.2 0.5
0.5
0.3
0.5
0.2
Generated difference: -0.2
Each repetition
B p value for each measure
Generated difference
Descending order
1st 0.267
2nd 0.260
50th 0.133
1000th -0.357
Real difference was the 50th largest: p = 50/1000 = 0.05
C Integrated p value
Repetition number
p value
HR SCR RLL NPV
Multiplied p value
Original 0.2 0.004 0.1 0.5 4x10-5
1 0.5 0.8 0.2 0.6 0.048
2 0.8 0.7 0.8 0.9 0.403
… …
1000 0.2 0.1 0.05 0.01 1x10-5
Real multiplied p value was the 50th largest among 1,000 generated multiplied p values: integrated p = 50/1000 = 0.05
Generated multiplied p values
The generated difference at this repetition was the 100th largest (i.e., 100/1000 = 0.1)

FIGURE 2 | The procedure of the randomization test. **(A)** How to compute the generated response difference in each repetition. **(B)** How to compute the *p*-value for each measure. **(C)** How to integrate *p*-values across all measures. HR, heart rate; SCR, skin conductance response; RLL, respiration line length; NPV, normalized pulse volume.

than these machine learning methods. Moreover, the machine learning methods require a database to estimate parameters, whereas the *d* and *p*-values do not. Thus, *d* and *p*-values are currently more useful for field CIT.

DISCRIMINATION THRESHOLD

As shown above, the performances of the effect size *d* and the randomization test *p* were sufficiently high. However, in the field, we should decide on thresholds for these statistical values to enable practitioners to judge whether the responses differ or not for each CIT. In this section, we show reference information for deciding thresholds in the case where the crime-relevant item is designated in advance and the case where it is unknown. We use the same dataset used by Ogawa et al. (23) described in section Performance of *d* and *p*: 80 recognizing and 72 unrecognizing examinees received the CIT with five items, which was presented five times.

Known-Solution CIT

The known-solution CIT assesses whether an examinee recognizes the crime-relevant information that the investigative organization has already grasped. In this section, response differences between relevant and irrelevant items are scored as *d* or *p*-values.

d Value

Figure 3A shows the percentage of the recognizing and unrecognizing examinees whose *d* values of the CIT are in the range of < -0.2, -0.2–0, 0–0.2, 0.2–0.4, 0.4–0.6, or > 0.6, respectively[1]. The dashed yellow line shows the ratio of the examinees whose *d* scores are in each range to all examinees. The solid red line shows the ratio of the recognizing examinees to all examinees whose *d* scores are in each range. The blue chain line shows the ratio of the unrecognizing examinees to all examinees whose *d* scores are in each range.

Figure 3A shows that, for each measure, over 80% of examinees whose *d* values were > 0.6 did indeed have recognition of the relevant item. For heart rate, over 80% of the examinees whose *d* values were < 0 did not have recognition of the relevant item. As shown in the extreme right of **Figure 3A**, 100% of the examinees whose integrated *d* values were > 0.4 did indeed have recognition of the relevant item. More than 80% of the examinees whose integrated *d* values were < 0 did not have recognition of the relevant item.

If we judge the case of an integrated $d > 0.4$ as *recognized*, $0 < d < 0.4$ as *inconclusive*, and $d < 0$ as *unrecognized*, the inconclusive rate is 44.7%. Without the inconclusive cases, the hit rate is 89.1% and correct rejection rate is 100%. We can reduce the inconclusive cases by judging the case of integrated $d > 0.3$ as *recognized*, $0.1 < d < 0.3$ as *inconclusive*, and $d < 0.1$ as *unrecognized*. In this case, the inconclusive rate is 20.4%, the hit rate is 86.4%, and the correct rejection rate is 94.6%.

p Value

Figure 3B shows the percentage of the recognizing or unrecognizing examinees whose *p*-values are in the range

[1] We chose these horizontal axis ranges of **Figure 3** considering the following two points: (1) The range should not be too wide to observe the change of the ratio of the recognizing/unrecognizing examinees according to the increase of the *d*/*p* values; (2) The range also should not be too narrow to include a sufficient number of examinees.

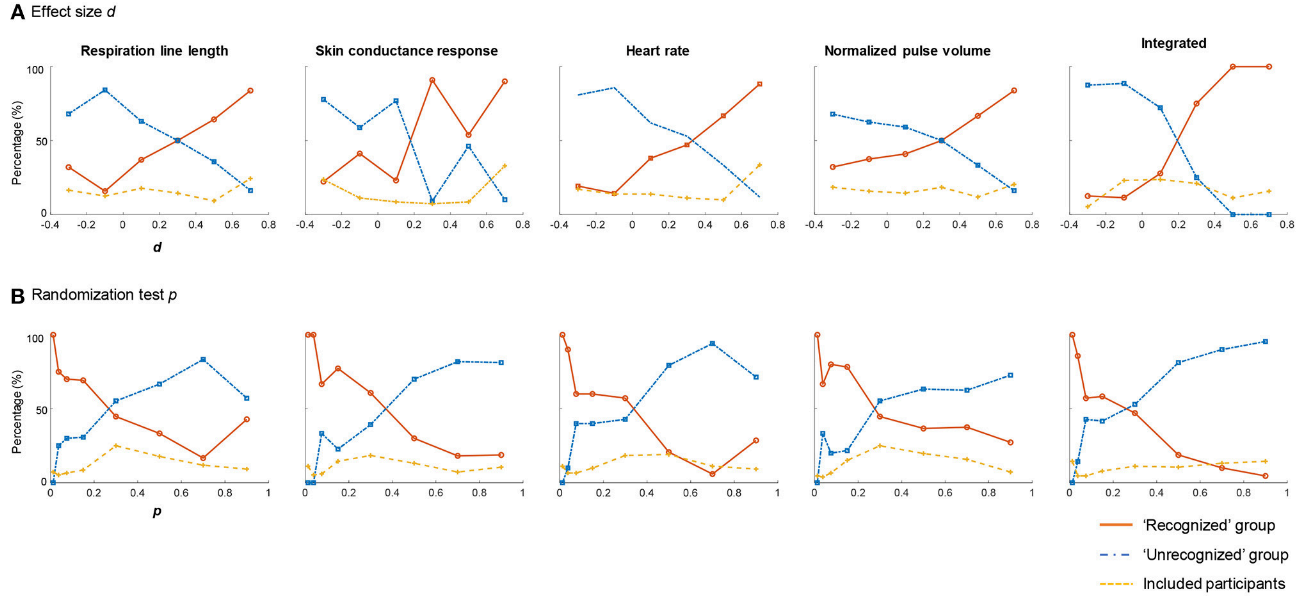


FIGURE 3 | The statistical values for the known-solution CIT. **(A)** The percentage of the recognized or unrecognized examinees whose *d* values are in the range of < −0.2, −0.2–0, 0–0.2, 0.2–0.4, 0.4–0.6, or > 0.6, respectively. **(B)** The percentage of the recognized or unrecognized examinees whose *p*-values are in the range of 0–0.025, 0.025–0.05, 0.05–0.1, 0.1–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, or 0.8–1, respectively. The solid red solid line shows (the number of recognizing examinees whose *d* or *p* scores are in each range)/(the number of examinees whose *d* or *p* scores are in each range). The blue chain line shows (the number of unrecognizing examinees whose *d* or *p* scores are in each range)/(the number of examinees whose *d* or *p* scores are in each range). The dashed yellow dash line shows (the number of examinees whose *d* or *p* scores are in each range)/(the number of all examinees).

of 0–0.025, 0.025–0.05, 0.05–0.1, 0.1–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, or 0.8–1, respectively. The dashed yellow line shows the ratio of the examinees whose *p*-values are in each range to all examinees. The solid red line shows the ratio of the recognized examinees to all examinees whose *p*-values are in each range. The blue chain line shows the ratio of the unrecognized examinees to all examinees whose *p*-values are in each range.

As shown in **Figure 3B**, for each measure, 100% of the examinees whose *p*-values were < 0.025 did indeed have recognition of the relevant item. As shown in the extreme right of **Figure 3B**, more than 90% of the examinees whose integrated *p*-values were > 0.6 did not have recognition of the relevant items.

If we judge the case of the integrated $p < 0.025$ as *recognized*, $0.025 < p < 0.6$ as *inconclusive*, and $p > 0.6$ as *unrecognized*, the inconclusive rate is 40.8%. Without the inconclusive cases, the hit rate is 94.1% and the correct rejection rate is 100%. If we want to reduce the inconclusive cases by judging the case of the integrated $p < 0.05$ as *recognized*, $0.05 < p < 0.4$ as *inconclusive*, and $p > 0.4$ as *unrecognized*, the inconclusive rate is 24.3%, the hit rate is 88.5%, and the correct rejection rate is 98.2%.

Figure 3 also indicates that the integration of multiple measures dramatically improves the discrimination performance. The integrated *d* and *p*-values clarify the difference between the recognized and unrecognized groups and reduce the range judged inconclusive, where the percentages of the recognized and unrecognized examinees are competing.

Searching CIT

In the searching CIT, an examiner assesses whether an examinee recognizes any of the items in a CIT question as crime-relevant. Thus, the examiner has to compare responses among all items. If the maximum response is sufficiently great and reliable, the examiner judges that the examinee recognizes the item as crime-relevant.

In this section, we examine the thresholds of the *d* and *p*-values for the searching CIT. We used the same dataset described in the above section but assume that the relevant item is unknown: we calculate five *d* or *p*-values for a CIT question assuming that each of the five items is the relevant item. We compare the five values to judge whether the examinee recognizes any item, and, if so, which item is recognized.

d Value

Figure 4A shows histograms of the integrated *d* values for the searching CIT. The first panel shows the integrated *d* for the actual relevant item in the recognized group. The second panel shows the maximum integrated *d* among the five items in the recognized group. The third panel shows the maximum integrated *d* among the five items but where the item is actually irrelevant in the recognized group. The fourth panel shows the maximum integrated *d* among the five items in the unrecognized group. In the searching CIT, we must avoid two types of false positive cases: the case that recognizing examinees are judged as recognizing an irrelevant item, and the case that unrecognizing examinees are judged as recognizing a certain item. The threshold to avoid the first type of false positive is suggested by comparing

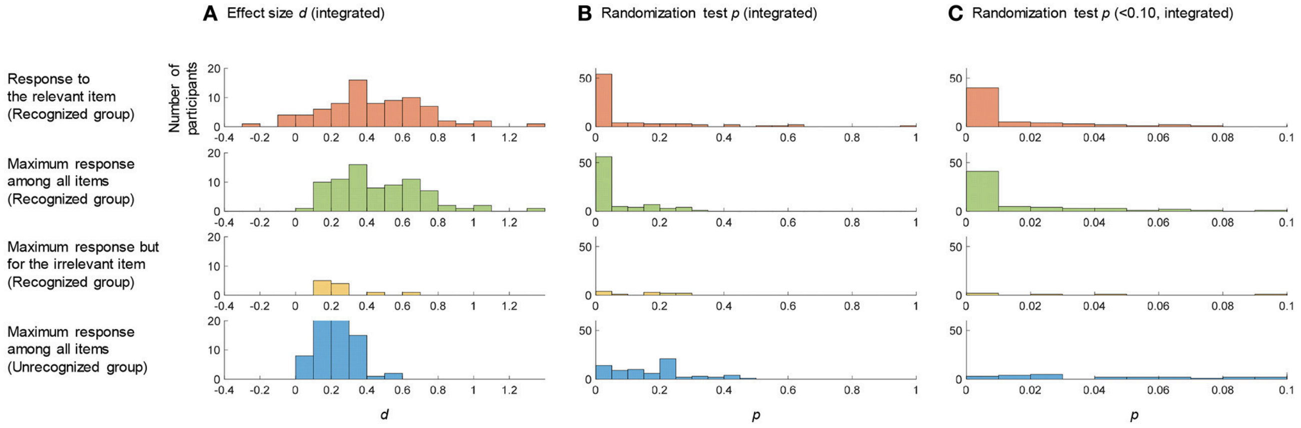


FIGURE 4 | The statistical values for the searching CIT. The first panel shows the histogram of the integrated *d* **(A)** or *p* **(B,C)** for the actual relevant item in the recognized group. The second panel shows the histogram of the maximum/minimum integrated *d/p* among all items in the recognized group. The third panel shows the histogram of the integrated *d/p* that is maximum/minimum among all items but the item is actually irrelevant in the recognized group. The fourth panel shows the histogram of the maximum/minimum integrated *d/p* among all items in the unrecognized group.

the first panel with the third panel, and the threshold to avoid the second type of false positive is suggested by comparing the first panel with the fourth panel. **Figure 4A** shows that both types of false positive cases can be avoided by the threshold of 0.6. If we judge the case of the maximum integrated $d > 0.6$ as *recognized*, $0.2 < d < 0.6$ as *inconclusive*, and $d < 0.2$ as *unrecognized*, the inconclusive rate is 54.6%, the hit rate is 63.9%, and the correct rejection rate is 100%.

p Value

Figures 4B,C show histograms of *p*-values for the searching CIT. The first panel shows the integrated *p* for the relevant item in the recognized group. The second panel shows the minimum integrated *p* among the five items in the recognized group. The third panel shows the minimum integrated *p* among the five items but where the item is actually irrelevant in the recognized group. The fourth panel shows the minimum integrated *p* among the five items in the unrecognized group. The comparison between the first panel and the third/fourth panel of **Figure 4C** reveals that we can avoid false positive cases with a threshold of 0.01. If we judge the case of the minimum integrated $p < 0.01$ as *recognized*, $0.01 < p < 0.2$ as *inconclusive*, and $p > 0.2$ as *unrecognized*, the inconclusive rate is 44.1%, the hit rate is 79.6%, and the correct rejection rate is 91.7%.

Summary of the Threshold

These results will provide reference information to judge the examinee's recognition based on effect size *d* and randomization test *p*. In the known-solution CIT, an examinee would recognize the relevant item if its integrated *d* is more than 0.4 or *p* is < 0.025. In the searching CIT, an examinee would recognize a certain item if its integrated *d* is more than 0.6 or *p* is < 0.01. The *d* value evaluates the response difference quantitatively, whereas the *p*-value evaluates the difference stochastically. Therefore, we would do well to consider both the *d* and *p*-values when judging the examinee's recognition.

Of course, before applying these thresholds to CIT in the field, we must verify them with other datasets. We believe that the proposed statistical judgment methods can be applied to the field datasets, because autonomic responses are essentially the same between laboratory and field CITs (13, 36). However, the magnitude of response differences is sometimes larger in the field than in the laboratory (13). We must therefore confirm whether the thresholds proposed above have sufficient discrimination performance when we apply them to the field datasets.

REDUCING INCONCLUSIVE CASES

Although the above section shows high discrimination performance using *d* and *p*-values, it also demonstrates that the inconclusive rates were relatively high, particularly in the searching CIT. To reduce the number of inconclusive cases, we would have to add new measures to the current autonomic measures (2). Recent studies have indicated that facial information, such as eye movement, pupil size, blinks, and facial skin temperature, are promising as new CIT measures (37–41). Some facial information can be recorded using current polygraph devices in Japan (42), but can also be remotely sensed by camera. Remote sensing can dramatically reduce the discomfort of attaching sensors to the examinee. In contrast, voice information obtained by the examinee's responses to each item has rarely been analyzed (43), and could be recorded without attaching sensors. Adding these remote sensing techniques is a new direction in the use of CIT in the field. However, it is important to pay attention to how the examiner informs the examinee about physiological recordings that he or she cannot perceive.

CONCLUSION

Although many people think of the polygraph as a deception detection technique, the polygraph based on the CIT should be regarded as memory detection technique. The CIT can reveal what an examinee knows and what he or she does not know about a crime. The CIT can also reveal, through the examinee's memory, new crime-relevant information that the examiner and investigators did not previously know about. Furthermore, the CIT can be used for assessing the credibility of examinees' statements. Correct understanding of the CIT will change the role of the polygraph in criminal investigations. The development of statistical judgment methods will make the CIT more objective and promote its use outside Japan.

The CIT is a scientifically valid method and can reveal how the examinee is related to the crime through his or her memory. Although the CIT has much potential, Japan is the only country in which it has been widely used. We hope that this paper will encourage more practitioners to try CIT in their fields.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This study was supported by grant (KAKENHI17K04513) to IM.

ACKNOWLEDGMENTS

This manuscript is partially based on a keynote lecture at National Forensic Service, Republic of Korea. We thank Ellen Daldoss from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript.

REFERENCES

1. Ambach W, Gamer M. Physiological measures in the detection of deception and concealed information. In: *Rosenfeld JP, editor. Detecting Concealed Information and Deception*. London: Academic Press (2018). p. 3–33.
2. Matsuda I, Nittono H, Allen JJB. The current and future status of the concealed information test for field use. *Front Psychol.* (2012) 3:532. doi: 10.3389/fpsyg.2012.00532
3. Raskin DC, Kircher JC. Validity of polygraph techniques and decision methods. In: *Raskin DC, Honts CR, Kircher JC, editors. Credibility Assessment: Scientific Research and Applications. San Diego, CA: Academic Press* (2014). p. 63–129. doi: 10.1016/B978-0-12-394433-7.00003-8
4. Lykken DT. The GSR in the detection of guilt. *J Appl Psychol.* (1959) 43:385–8. doi: 10.1037/h0046060
5. Verschuere B, Ben-Shakhar G, Meijer E. *Memory Detection: Theory and Application of the Concealed Information Test Cambridge: Cambridge University Press.* (2011). doi: 10.1017/CBO9780511975196
6. Matsuda I, Nittono H. Physiological responses in the concealed information test: a selective review in the light of recognition and concealment. In: *Rosenfeld JP, editor. Detecting Concealed Information and Deception*. London: Academic Press (2018). p. 77–96.
7. Elaad E. The challenge of the concealed knowledge polygraph test. *Exp Evidence* (1998) 6:161–87. doi: 10.1023/A:1008855511254
8. Meijer EH, Klein Selle N, Elber L, Ben-Shakhar G. Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology* (2014) 51:879–904. doi: 10.1111/psyp.12239
9. Iacono WG. Encouraging the use of the guilty knowledge test (GKT): what the GKT has to offer law enforcement. In: *Verschuere B, Ben-Shakhar G, Meijer E, editors. Memory Detection: Theory and Application of the Concealed Information Test. Cambridge: Cambridge University Press* (2011). p. 12–23. doi: 10.1017/CBO9780511975196.002
10. Osugi A. Daily application of the concealed information test: Japan. In: *Verschuere B, Ben-Shakhar G, Meijer E, editors. Memory Detection: Theory and Application of the Concealed Information Test. Cambridge: Cambridge University Press* (2011). p. 253–75. doi: 10.1017/CBO9780511975196.015
11. Sawada Y, Tanaka G, Yamakoshi K. Normalized pulse volume (NPV) derived photo-plethysmographically as a more valid measure of the finger vascular tone. *Int J Psychophysiol.* (2001) 41:1–10. doi: 10.1016/S0167-8760(00)00162-8
12. Kobayashi T, Yoshimoto K, Fujihara S. The contemporary situation of field polygraph tests. *Jpn J Physiol Psychol Psychophysiol.* (2009) 27:5–15. doi: 10.5674/jjppp.27.5
13. Osugi A. Field findings from the concealed information test in Japan. In: *Rosenfeld JP, editor. Detecting Concealed Information and Deception*. London: Academic Press (2018). p. 97–121.
14. Meijer EH, Smulders FT, Johnston JE, Merckelbach HL. Combining skin conductance and forced choice in the detection of concealed information. *Psychophysiology* (2007) 44:814–22. doi: 10.1111/j.1469-8986.2007.00543.x
15. Gamer M, Gödert HW, Keth A, Rill H-G, Vossel G. Electrodermal and phasic heart rate responses in the guilty actions test: comparing guilty examinees to informed and uninformed innocents. *Int J Psychophysiol.* (2008) 69:61–8. doi: 10.1016/j.ijpsycho.2008.03.001
16. Klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Orienting versus inhibition in the concealed information test: different cognitive processes drive different physiological measures. *Psychophysiology* (2016) 53:579–90. doi: 10.1111/psyp.12583
17. Ben-Shakhar G, Elaad E. Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: a reexamination. *J Appl Psychol.* (2002) 87:972–7. doi: 10.1037/0021-9010.87.5.972
18. Adachi K. Statistical classification procedures for polygraph tests of guilty knowledge. *Behaviormetrika* (1995) 22:49–66. doi: 10.2333/bhmk.22.49
19. Elaad E. Validity of the concealed information test in realistic contexts. In: Verschuere B, Ben-Shakhar G, Meijer E, editors. *Memory Detection: Theory and Application of the Concealed Information Test. Cambridge: Cambridge University Press* (2011). p. 171–86. doi: 10.1017/CBO9780511975196.010
20. The Legal Training and Research Institute of Japan *Scientific Evidence and How to Used it in Court [in Japanese]*. Tokyo: Housoukai (2013).
21. Yokoi Y, Okazaki Y, Kiriu K, Kuramochi K, Ohama T. The validity of the guilty knowledge test used in field cases. *Jpn J Crimin Psychol.* (2001) 39:15–27. doi: 10.20754/jjcp.39.1_15
22. Hira S, Furumitsu I. Polygraphic examinations in Japan: applications of the guilty knowledge test in forensic investigations. *Int J Police Sci Manag.* (2002) 4:16–27. doi: 10.1177/146135570200400103
23. Ogawa T, Matsuda I, Tsuneoka M. Accuracy of concealed information test as a memory detection technique: a laboratory study [in Japanese]. *Jpn J Forensic Sci Technol.* (2013) 18:35–44. doi: 10.3408/jafst.18.35
24. Matsuda I, Hirota A, Ogawa T, Takasawa N, Shigemasu K. A new discrimination method for the concealed information test using pretest data and within-individual comparisons. *Biol Psychol.* (2006) 73:157–64. doi: 10.1016/j.biopsycho.2006.01.013
25. Ben-Shakhar G, Furedy JJ. *Theories and Applications in the Detection of Deception. New York, NY: Springer-Verlag* (1990). doi: 10.1007/978-1-4612-3282-7
26. Gamer M, Verschuere B, Crombez G, Vossel G. Combining physiological measures in the detection of concealed information. *Physiol Behav.* (2008) 95:333–40. doi: 10.1016/j.physbeh.2008.06.011
27. Matsuda I, Hirota A, Ogawa T, Takasawa N, Shigemasu K. Within-individual discrimination on the concealed information test using dynamic mixture modeling. *Psychophysiology* (2009) 46:439–49. doi: 10.1111/j.1469-8986.2008.00781.x

28. Venables PH, Mitchell DA. The effects of age, sex and time of testing on skin conductance activity. *Biol Psychol.* (1996) 43:87–101. doi: 10.1016/0301-0511(96)05183-6
29. Matsuda I, Ogawa T, Tsuneoka M. Discrimination performance of statistical values used in the concealed information test studies [in Japanese]. *Jpn J Forensic Sci Technol.* (2015) 20:59–67. doi: 10.3408/jafst.681
30. Cohen J. *Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Erlbaum* (1988).
31. Noordraven E, Verschuere B. Predicting the sensitivity of the reaction time-based concealed information test. *Appl Cogn Psychol.* (2013) 27:328–35. doi: 10.1002/acp.2910
32. Matsuda I, Ogawa T, Tsuneoka M, Verschuere B. Using pretest data to screen low-reactivity individuals in the autonomic-based concealed information test. *Psychophysiology* (2015) 52:436–9. doi: 10.1111/psyp.12328
33. Bowman H, Filetti M, Janssen D, Su L, Alsufyani A, Wyble B. Subliminal salience search illustrated: EEG identity and deception detection on the fringe of awareness. *PLoS ONE* (2013) 8:e54258. doi: 10.1371/journal.pone.0054258
34. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol. (2013)* 4:863. doi: 10.3389/fpsyg.2013.00863
35. Gadbury GL, Page GP, Heo M, Mountz JD, Allison DB. Randomization tests for small samples: an application for genetic expression data. *J R Stat Soc Ser C Appl Stat.* (2003) 52:365–76. doi: 10.1111/1467-9876.00410
36. Zaitsu W. External validity of concealed information test experiment: comparison of respiration, skin conductance, and heart rate between experimental and field card tests. *Psychophysiology* (2016) 53:1100–7. doi: 10.1111/psyp.12650
37. Pollina D, Dollins A, Senter S, Brown T, Pavlidis I, Levine J, Ryan A. Facial skin surface temperature changes during a "concealed information" test. *Ann Biomed Eng.* (2006) 34**:**1182–9. doi: 10.1007/s10439-006-9143-3
38. Peth J, Kim JS, Gamer M. Fixations and eye-blinks allow for detecting concealed crime related memories. *Int J Psychophysiol.* (2013) 88:96–103. doi: 10.1016/j.ijpsycho.2013.03.003
39. Peth J, Suchotzki K, Gamer M. Influence of countermeasures on the validity of the concealed information test. *Psychophysiology* (2016) 53:1429–40. doi: 10.1111/psyp.12690
40. Lancry-Dayan OC, Nahari T, Ben-Shakhar G, Pertzov Y. Do you know him? *Gaze dynamics toward familiar faces on a concealed information test. J Appl Res Mem Cogn.* (2018) 7:291–302. doi: 10.1016/j.jarmac.2018.01.011
41. Matsuda I, Hirota A, Ogawa T. Nasal blood flow in the concealed information test [in Japanese]. *Jpn J Physiol Psychol Psychophysiol* (in press). doi: 10.5674/jjppp.1705si
42. Hirota A, Matsuda I, Kobayashi K, Takasawa N. Development of a portable digital polygraph system (in Japanese with English abstract). *Jpn J Forensic Sci Technol.* (2005) 10:37–44. doi: 10.3408/jafst.10.37
43. Gamer M, Rill HG, Vossel G, Gödert HW. Psychophysiological and vocal measures in the detection of guilty knowledge. *Int J Psychophysiol.* (2006) 60:76–87. doi: 10.1016/j.ijpsycho.2005.05.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychiatry

OPINION
published: 12 February 2019
doi: 10.3389/fpsyt.2019.00056

Digging Further Into the Speech of Liars: Future Research Prospects in Verbal Lie Detection

Galit Nahari[1]* *and Zvi Nisin*[2]

[1] Department of Criminology, Bar-Ilan University, Ramat-Gan, Israel, [2] Investigative Psychology Section, The Division of Identification and Forensic Science, Investigation and Intelligence Branch, Israel Police, Haifa, Israel

Keywords: deception and deceptive strategies, verbal lie detection, verifiability approach, detection deception, reality monitoring, context embedded perception

OPEN ACCESS

Edited by:
Giuseppe Sartori,
University of Padova, Italy

Reviewed by:
Bennett Kleinberg,
University College London,
United Kingdom
Emiliano Ricciardi,
IMT School for Advanced Studies
Lucca, Italy

****Correspondence:***
Galit Nahari
galit.nahari@biu.ac.il

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *14 August 2018*
Accepted: *24 January 2019*
Published: *12 February 2019*

Citation:
Nahari G and Nisin Z (2019) Digging Further Into the Speech of Liars: Future Research Prospects in Verbal Lie Detection. Front. Psychiatry 10:56. doi: 10.3389/fpsyt.2019.00056

The field of verbal lie detection has grown rapidly in the past decade. Derived by the assumption that lies have different content patterns than do truths, research in this area promotes searching for content criteria to detect them. One prime content-based indicator for deception detection, which stems from the Reality Monitoring (RM) theory (1), is *richness in detail*. According to RM, truthful memories of actual events originate in perceptual experience and are embedded in the context of time and space. As such, they are expected to include more spatial and temporal contextual attributes (i.e., locations, spatial arrangement of people and objects, times, duration and sequence of events) and perceptual attributes (i.e., what the individual felt, tasted, smelled, heard, or saw when the event took place) than do false memories, which originate in self-generated thought or imagination. Derived from this prediction, the traditional use of richness in detail as an indicator of deception is based on the *number* of perceptual and contextual details in the interviewee's accounts. However, as a memory source-monitoring theory, RM does not take into consideration the intention of liars to deceive and consequently cannot explain the full scope of *richness in detail* in the field of deception (2). In contrast to false memories, where the individual has no intention to deceive but *wrongly* believes that his/her memory of an event that never happened is truthful, fabricated memories are an outcome of manipulation [and have thus been labeled "self-manipulated memories"; (2)]. Liars frequently attempt to manipulate their fabricated accounts to make them seem truthful (3–5), for example by *intentionally* adding false perceptual and contextual details (6, 7). Affecting the quantity of the details in their fabricated accounts, such strategic manipulations reduce the diagnostic efficacy of the *richness in detail* indicator. Yet, in the current paper, we aim to show that the same strategies leave traces on the quality of details. Therefore, we propose that to maximize the potential utility of the *richness in detail* indicator, it is necessary to dig deeper into the speech of liars, particularly by looking for traces of deception strategies found in the *quality* of the details. In fact, the Verifiability Approach [VA; (4, 8)] applies this notion.

THE VERIFIABILITY APPROACH (VA)

The VA (8) for lie detection was initiated based on the understanding that lies, by nature, are based on strategies. The first VA study (4) clearly demonstrated that lie detection benefits more from consideration of the *quality* of perceptual and contextual details than it does from consideration of their quantity alone.

According to the VA, the strategy employed by liars is guided by the *liars' dilemma hypothesis*. Specifically, liars perceive richness in detail as an indicator of truthfulness (9, 10) and are thus motivated to provide many details to make an impression of honesty (7, 11). On the other hand, the provision of details also puts liars at risk, as the truthfulness of the details provided can be checked. Aware of this danger [see (6, 7)], liars are inclined to avoid mentioning false details, to minimize the chances of being caught. These two contradicting motivations—for and against the provision of details—put liars in a dilemma. A strategy that resolves the conflict involves the provision of details that cannot be checked and verified.

When used by liars, this strategy of providing non-verifiable information affects the quantity and quality of the contextual and perceptual details that appear in their accounts. They "inflate" the quantity of detail by incorporating false, non-verifiable, details, and as a result provide accounts that appear closer to the RM prototype of truthful accounts (i.e., accounts rich in perceptual and contextual details). However, their strategy leaves traces in the quality of their accounts, in terms of verifiability. By assessing the quality (i.e., the verifiability of the contextual and perceptual details) rather than the quantity of details provided, it is possible to reveal the liars' strategy, and thereby indicate their lies.

In the last years, the validity of the VA, which was originally developed and tested in police interview setting [e.g., (4, 5, 12, 13)] has been examined in other settings including insurance [e.g., (14–17)], airport security [see (18, 19)], occupation [e.g., (20)], and malingering [e.g., (21, 22)]. Some of these applications were more successful than others, but mostly the VA perspectives were confirmed [for a recent review see (23)], thereby providing an empirical evidence to the profitability of looking for quality of details. Encouraged by the success of the VA, we propose that research in this field should dig further into the speech of liars, in an attempt to identify additional indications of strategies in the quality of details provided. As such, we present two new approaches, both are derived from the theoretical and empirical framework of the VA.

CONTEXT EMBEDDED PERCEPTION (CEP)

The first approach was recently proposed by the authors of the current paper Nisin and Nahari (in preparation), who suggest that the qualitative differences between perceptual and contextual details can serve as a potential generator of deception strategy. According to this approach, while perceptual information is actually experienced, and acquired directly by the senses, contextual information is virtual in its nature, and based on semantic knowledge and relative conceptualizations. For instance, we experience the perceptual aspects of an interaction with a friend through our senses: we see the friend and the clothes he is wearing, hear his speech, and feel his touch. Meanwhile, the contextual aspects of this interaction, such as its length and specific location, are based on conceptualization and knowledge. In fact, the contextual attributes are imposed on the perceptual details and frame them in time and space. Accordingly, the perceptual details (e.g., visions, smells, sounds, sensations, and tastes) can be regarded as primary data, and the contextual details (e.g., indications of where, when, and for how long those perceptual details were experienced)—as meta-data. Obviously, the truthfulness of perceptual details can be checked only when they are given by the interviewee within the framework of contextual information regarding time and space. Thus, the contextual details are those that confer the status of verifiability upon perceptual details. Considering the differences between the two types of details in light of the VA (8), liars would be expected to avoid the provision of contextual details as often as possible. Motivated to provide non-verifiable details, they would be likely to provide perceptual details without framing them in time and space, making it difficult to check their truthfulness. Truth-tellers, on the other hand, would be expected to freely provide both types of details, as they have no reason to avoid verification. Thus, the prediction yielded from this approach is that liars, when adding false details to their accounts, will strategically prefer to provide perceptual details over (or without) contextual details, while truth-tellers will provide both types of details. As such, the number of contextual details in an account can serve as a verbal lie indicator.

RESOLUTION OF VERIFIABILITY (ROV)

The second approach involves the resolution of the verifiable details provided, as determined by the immediacy in which the information they incorporate can be verify. A good example of such resolution involves the use of names, which already found significance for lie detection (24). According to the VA (4), events that occurred in the presence of another person will be considered verifiable only when that person can be traced. Once an identifiable person has been mentioned, that person can be approached to confirm the truthfulness of the reported occurrences. However, the mention of a name is not a necessary condition for rendering the person identifiable and traceable. It is reasonable to assume that an interviewee who mentions a "friend" or "neighbor," even without volunteering a name, has considered that the police will ask about the person's specific identity, especially because identified persons can serve as witnesses (i.e., prime and significant evidence) who can confirm details in the interviewee's account. Consequently, it is likely that by mentioning persons who can be traced, with or without mentioning their names, interviewees mean, or at least are aware of the fact, that they are providing verifiable details [see (8)]. The difference between the two conditions (named vs. unnamed but traceable) is in the *resolution* of the verifiable details provided: names increase the resolution of the details. The mention of an identifiable person without a name leaves the interviewees with degrees of freedom, at least temporarily, such that the verifiability of the details is neither immediate nor easy (relatively speaking) to check. Importantly, these degrees of freedom also range, as mentioning an unnamed uncle, for example, leaves less degrees of freedom than mentioning an unnamed acquaintance. These assumptions lead to the expectation that liars, when they do provide verifiable details, will strategically prefer low-resolution over high-resolution details, while truth-tellers will prefer high-resolution details. Again, as with the contextual details, the number of low-resolution verifiable details in an account can serve as a verbal lie indicator.

CONCLUSIONS

The current paper is a call for more strategy-based research in the field of verbal lie detection. By demonstrating how strategies blur the differences in *detail quantity* while sharpening the differences in *detail quality* between truthful and fabricated accounts, we stress that research should go beyond the surface of content, to look for strategies that activate verbal behaviors among liars, define the qualitative manifestations of their strategies, and then—to exploit these manifestations in indicating their lies. The VA, a new and promising paradigm for deception detection is an outcome of this research approach. Having presented two new, un-studied, strategy-based approaches to further demonstrate and extend this line of research, we propose that following this path will benefit the field both theoretically and practically.

AUTHOR CONTRIBUTIONS

GN: conceptualization and writing the first draft, revising and editing. ZN: conceptualization and review and editing.

FUNDING

This research was supported by the Israel Science Foundation grant (grant No. 372/14).

REFERENCES

1. Johnson MK, Raye CL. Reality monitoring. *Psychol Rev.* (1981) 88:67–85.
2. Nahari G. The applicability of the verifiability approach to the Real world. In: Rosenfeld JP, editors. *Detecting Concealed Information and Deception: Recent Developments.* London: Elsevier (2018). p. 329–49.
3. Granhag PA, Hartwig M. A new theoretical perspective on deception detection: on the psychology of instrumental mind-reading. *Psychol Crime Law* (2008) 14:189–200. doi: 10.1080/10683160701645181
4. Nahari G, Vrij A, Fisher R. Exploiting liars' verbal strategies by examining unverifiable details. *Legal Criminol Psychol.* (2014) 19:227–39. doi: 10.1111/j.2044-8333.2012.02069.x
5. Nahari G, Vrij A, Fisher R. The verifiability approach: countermeasures facilitate its ability to discriminate between truths and lies. *Appl Cogn Psychol.* (2014) 28:122–8. doi: 10.1002/acp.2974
6. Masip J, Herrero C. 'What would you say if you were guilty?' Suspects' strategies during a hypothetical behavior analysis interview concerning a serious crime. *Appl Cogn Psychol.* (2013) 27:60–70. doi: 10.1002/acp.2872
7. Nahari G, Vrij A, Fisher RP. Does the truth come out in the writing? Scan as a lie detection tool. *Law Human Behav.* (2012) 36:68–76. doi: 10.1037/h0093965
8. Nahari G. Reality monitoring in the forensic context: digging deeper into the speech of liars. *J Appl Res Mem Cogn.* (2018) 7:432–40. doi: 10.1016/j.jarmac.2018.04.003
9. Bell BE, Loftus EF. Trivial persuasion in the courtroom: the power of (a few) minor details. *J Pers Soc Psychol.* (1989) 56:669–79.
10. Johnson MK. Memory and reality. *Am Psychol.* (2006) 61:760–71. doi: 10.1037/0003-066X.61.8.760
11. Hartwig M, Granhag PA, Strömwall LA. Guilty and innocent suspects' strategies during police interrogations. *Psychol Crime Law* (2007) 13:213–27. doi: 10.1080/10683160600750264
12. Nahari G, Vrij A. Are you as good as me at telling a story? Individual differences in interpersonal reality-monitoring. *Psychol Crime Law* (2014) 20:573–83. doi: 10.1080/1068316X.2013.793771
13. Nahari G, Vrij A. Can someone fabricate verifiable details when planning in advance? It all depends on the crime scenario. *Psychol Crime Law* (2015) 21:987–99. doi: 10.1080/1068316X.2015.1077248
14. Harvey CH, Vrij A, Leal S, Lafferty M, Nahari G. Insurance-based lie detection: enhancing the verifiability approach with a model statement component. *Acta Psychol.* (2017) 174:1–8. doi: 10.1016/j.actpsy.2017.01.001
15. Harvey AC, Vrij A, Nahari G, Ludwig K. Applying the verifiability approach to insurance claims settings: exploring the effect of the information protocol *Legal Criminol Psychol.* (2016) 22:47–59 doi: 10.1111/lcrp.12092
16. Vrij A, Nahari G, Isitt R, Leal S. Using the verifiability lie detection approach in an insurance claim setting. *J Investig Psychol Offend Profiling* (2016) 13:183–97. doi: 10.1002/jip.1458
17. Nahari G, Leal S, Vrij A, Warmelink L, Vernham Z. Did somebody see it? Applying the verifiability approach to insurance claims interviews. *J Investig Psychol Offend Profiling* (2014) 11:237–43. doi: 10.1002/jip.1417
18. Jupe LM, Leal S, Vrij A, Nahari G. Applying the verifiability approach in an international airport setting. *Psychol Crime Law* (2017) 23:812–25. doi: 10.1080/1068316X.2017.1327584
19. Kleinberg B, Nahari G, Arntz A, Verschuere B. An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychol.* (2017) 3:21. doi: 10.1525/collabra.80
20. Jupe LM, Vrij A, Leal S, Mann S, Nahari G. The lies we live: using the verifiability approach to detect lying about occupation. *J Articles Support Null Hypothesis* (2016) 13:1–13. Available online at: https://researchportal.port.ac.uk/portal/en/publications/the-lies-we-live(56beb3f2-0c7c-4821-905f-49bbba9c0b7e).html
21. Boskovic I, Bogaard G, Merckelbach H, Vrij A, Hope L. The verifiability approach to detection of malingered physical symptoms. *Psychol Crime Law* (2017) 23:717–29. doi: 10.1080/1068316X.2017.1302585
22. Boskovic I, Gallardo CT, Vrij A, Hope L, Merckelbach H. Verifiability on the run: an experimental study on the verifiability approach to malingered symptoms. *Psychiatry Psychol Law* (2018). doi: 10.1080/13218719.2018.1483272. [Epub ahead of print].
23. Vrij A, Nahari G. The verifiability approach. In: Dickinson J, Schreiber Compo N, Carol R, McCauley M, editors. *Evidence-Based Investigative Interviewing.* London: Routledge; Taylor and Francis Group (in press).
24. Kleinberg B, Mozes M, Arntz A, Verschuere B. Using named entities for computer-automated verbal deception detection. *J Forensic Sci.* (2018) 63:714–23. doi: 10.1111/1556-4029.13645

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer BK declared a past collaboration with one of the authors GN to the handling editor.

ORIGINAL RESEARCH
published: 12 March 2019
doi: 10.3389/fpsyt.2019.00114

Attentional Avoidance for Guilty Knowledge Among Deceptive Individuals

*Kiho Kim, Go-eun Kim and Jang-Han Lee**

Clinical Neuro-Psychology Laboratory, Department of Psychology, Chung-Ang University, Seoul, South Korea

The purpose of the present study is to differentiate between innocent suspects who have knowledge of crime information and guilty suspects. The study investigated eye-movement differences among three groups: a guilty group who took part in a mock crime, an innocent-aware group who did not commit a mock crime but were exposed to the crime stimuli, and an innocent-unaware group who neither committed a mock crime nor had crime-relevant information. Each group's eye movements were tracked while all participants viewed stimuli (crime-relevant, crime-irrelevant, and neutral). The results revealed that the guilty group not only viewed all stimuli later than the other groups, they also viewed crime-relevant and crime-irrelevant stimuli for a shorter time period than the innocent-aware group; the innocent-aware group focused their attention on crime-relevant and crime-irrelevant stimuli longer than neutral stimuli, and the innocent-unaware group showed no differences in their attention focus among all types of stimuli. This present study suggests that guilty individuals show attentional avoidance from all stimuli in a lie detection situation, whereas innocent-aware and innocent-unaware individuals did not show avoidance responses.

Keywords: attentional bias, attentional avoidance, deception detection, guilty knowledge test, concealed information test, eye-movement

OPEN ACCESS

Edited by:
Cristina Scarpazza,
University of Padova, Italy

Reviewed by:
Kirsten Jordan,
University of Göttingen, Germany
Coralie Boillat,
University Psychiatric Clinic Basel,
Switzerland

****Correspondence:***
Jang-Han Lee
clipsy@cau.ac.kr

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *10 August 2018*
Accepted: *15 February 2019*
Published: *12 March 2019*

Citation:
Kim K, Kim G and Lee J-H (2019)
Attentional Avoidance for Guilty
Knowledge Among Deceptive
Individuals. Front. Psychiatry 10:114.
doi: 10.3389/fpsyt.2019.00114

INTRODUCTION

The Guilty Knowledge Test (GKT) or the Concealed Information Test (CIT) is a deception detection method and is intended to establish the existence of a specific memory trace (1–3). The GKT is based on the assumption that suspects who possess knowledge about specific crime related details will be physiologically more reactive to crime-relevant questions than crime-irrelevant questions, by utilizing a series of multiple-choice questions, each having one crime-relevant question and several control questions (4). GKT relies on a solid scientific principle, called an orienting response (OR), which is an elicited response caused by a novel stimulus or a familiar stimulus with relevance or "signal value" (5, 6), and it has been shown that guilty knowledge has an added signal value (7). That is, people who have guilty knowledge show a stronger OR to crime-relevant questions than to other questions, whereas all questions elicit equivalent responses from truth tellers. Laboratory research has reported that the GKT has high validity coefficients for the differentiation between guilty and innocent persons on the basis of autonomic measures such as skin conductance responses, respiration, heart rate, the P300 event-related potential (8–11), and can be generalized to the criminal field (12, 13).

However, the GKT also has the possibility of false positive errors because it may not work correctly when innocent suspects are exposed to crime-related information (14–17). It is not easy to keep salient features of a crime from being leaked to the innocent group, and the leakage of the critical features of the crime might put the innocent group in substantial danger because knowledge of the critical crime stimuli might be sufficient for producing differential responses to the stimuli. Therefore, it is widely known that false positive errors, where the innocent group is judged as guilty, can be controlled as long as information about the crime is not leaked to innocent group in the GKT.

Because the GKT may not work with an innocent group that has guilty knowledge, Ben-Shakhar et al. (18) attempted to identify the effects of awareness of crime-relevant information on the deception detection with the GKT. They investigated whether introducing target stimuli, to reduce false positive outcomes, occurred from the leakage of crime-relevant information to the innocent group. In their study, they introduced target items to which participants must respond while answering the GKT questions with the purpose of drawing the attention of informed innocent suspects. As a result, the informed innocent group showed relatively larger electrodermal responses to the critical stimuli than uninformed ones—but not as large as the responses of the guilty group. However, it is a hasty conclusion to suggest that the informed innocent group attended to target stimuli at a level near that of the guilty group, because the study did not directly measure the effect of the target items in drawing attention. Therefore, it remains unclear whether discrimination between informed innocent and guilty suspects is possible (19, 20).

It is known that not only physiological activity, but also attentional processes are involved in responses to guilty knowledge, as a component of the OR (21). Indeed, several authors have argued that the main function of the OR is to enhance information processing, which is achieved by not only directing the senses to the stimulus but also allocating attention toward it. Both novel and significant stimuli are associated with an allocation of attention, as measured by task interference on a concurrent reaction time task (22, 23). In addition, Verschuere et al. (24) found that guilty knowledge elicits a signal-OR and therefore demands attentional resources with a probe classification task. Therefore, it is reasonable to think that guilty knowledge demands attention. However, there has been no indication of spatial shifting of attention on guilty knowledge thus far, and it remains unclear whether participants would shift attention either toward or away from guilty knowledge. An eye-tracking technique can be an effective way to investigate the direction of attention because eye tracking is a continuous method of measuring eye movement, which allows for the direct observation of attentional engagement, shift, and a disengagement pattern (25). The eye-tracking device not only provides a highly direct measure of visual attention but also allows continuous measurement of gaze patterns.

Gaze patterns reveal complex information processing that can be explained as an attentional bias that involves both autonomic and controlled processes (26, 27). Eye tracking literature defines initial gaze fixation or first fixation latency "where one looks" as OR /initial orienting of overt attention to a stimulus (often a more automatic process); and dwell time or fixation time as the later process of "how long one looks," a rather strategically controlled process (28, 29). That is, it is likely that people who have guilty knowledge initially fixate their eyes toward crime-relevant information automatically because of the OR but subsequently show cognitive eye movements as a manifestation of strategic behavior in controlled processes. Recently, Kim et al. (30) attempted to identify whether or not liars, as compared to truth tellers, would have an attentional bias for guilty knowledge using the eye tracker. As a result, both the guilty and the innocent groups initially fixated on crime-relevant stimuli rather than on both crime-irrelevant and neutral stimuli. In addition, the guilty group showed a longer dwell time for neutral stimuli than the innocent group did, although there was no difference between the two groups for crime-relevant and irrelevant stimuli. These findings possibly indicate that the guilty group reflexively moved their eyes toward crime-relevant stimuli as an OR, but they strategically diverted their attention from these stimuli so as not to be found guilty of theft. It has been found that liars use an "avoid and escape" strategy when confronted with deceptive evidence during communication (31). It might be assumed that guilty people who have guilty knowledge show differential responding to crime-relevant information than innocent people who have guilty knowledge and innocent people who have no guilty knowledge. Therefore, there is a need to investigate in order to differentiate innocent suspects who have knowledge of crime information from guilty suspects, using attentional bias regarding crime information, by measuring eye movement.

The purpose of this study was to investigate attentional bias regarding crime information, by measuring eye movement and to reveal differences in attentional patterns between guilty and innocent-aware groups. We investigated the eye-movement differences among three groups: a guilty group who committed a mock crime, an innocent-aware group who did not commit a mock crime but were naturally exposed to guilty knowledge, and an innocent-unaware group who did not have any knowledge of the crime and did not commit a mock crime. We predicted that the guilty group would show a shorter first fixation time and a shorter dwell time toward crime-relevant stimuli than crime-irrelevant and neutral stimuli. In addition, we expected that the innocent-aware group would show a shorter first fixation time and a longer dwell time toward both crime-relevant and crime-irrelevant stimuli than neutral stimuli. Finally, we expected that there would be no differences in a first fixation time and a dwell time to all types of stimuli in the innocent-unaware group.

MATERIAL AND METHODS

Participants

A total of 60 undergraduate students from Seoul, Korea were recruited for this experiment. All participants were physically and psychologically healthy, and their state of health was checked by an interview. Participants were randomly assigned to one of the three groups: a guilty group who committed a mock crime and possessed crime-relevant knowledge, an innocent-aware group who possessed crime-relevant information even though they did not take part in the mock crime, and an innocent-unaware group who did not have any knowledge of the mock crime. Of all participants, three from the guilty group, five from the innocent-aware group, and one from the innocent-unaware group were removed as outliers—three because their dwell time results were more than 2 SD from the mean (three had unusually variable dwell time), five because their first fixation time results were more than 2 SD from the mean (five had unusually variable first fixation time) and one had almost half of the data missing due to measurement errors. Finally, the guilty group consisted of 17 participants (six males, mean age = 22.56; *SD* = 2.10), the innocent-aware group consisted of 15 participants (nine males, mean age = 23.67, *SD* = 2.82), and the innocent-unaware group consisted of 19 participants (nine males, mean age = 21.45, *SD* = 2.16).

Apparatus and Materials

Eye movements for all participants were recorded with an eye-tracking device (iView XTM Red—IV Eye Tracking System, Sensomotoric Instruments GmbH, Berlin, Germany) at a sampling rate of 60 Hz. In order to analyze the eye-movement data, we used the Begaze (SMI, Berlin, Germany) software package, which provided a variety of gaze information, such as how long they fixated their attention, where they focused, how many times they saw the specific location or stimulus, and so on. Each participant was seated at a desk, at a distance of 70 cm from a 23-inch wide monitor (1,920 × 1,080), and the eye tracker allowed the participants to naturally move their heads and eyes without any attached sensors. The eye movements that were stable for at least 80 ms within the visual angle of 1.4° were defined as a fixation (28).

Three types of stimuli were used: crime-relevant, crime-irrelevant, and neutral stimuli. Crime-relevant stimuli comprised of four items that were used in the mock crime: black USB, white envelope, purple legal seal, and black pen. Crime-irrelevant stimuli comprised of four items that were similar to the crime-relevant stimuli in shape but were not used for the mock crime: silver USB, purple postcard, unofficial seal, and pencil. Finally, neutral stimuli comprised of four items that were not exposed to participants during the experiment: Thermos, stapler, toothbrush, and felt-tipped pen. Twenty-four people other than the experimental participants rated the valence and arousal of each stimulus with a 7-point Likert scale, with 1 labeled as "very unpleasant" and "calm," and 7 labeled as "very pleasant" and "arousing." There were no differences in the mean valence and arousal rating among the three stimuli types. Each picture was 95 mm high by 130 mm wide when displayed on the screen, and the distance between their inner edges was 30 mm. The distance between the two probe positions was 105 mm (visual angle of 5.4°). The task consisted of 36 pairs, Crime-relevant & Crime-irrelevant, Crime-relevant & Neutral, and Crime-irrelevant & Neutral which were presented on one screen at the same time. The pairs were presented in a counterbalanced order between the left and right sides of the screen. A total of 72 trials were performed in two blocks.

Measures

Recognition Test

A recognition test was conducted to determine how well-participants remembered the crime-relevant and crime-irrelevant stimuli. The test consisted of 12 single-selection questions (four questions of crime-relevant stimuli, four questions of crime-irrelevant stimuli, four questions of neutral stimuli), and participants were asked to mark an X in the appropriate answer (i.e., 1: the stimuli you stole during the experiment, 2: the stimuli you did not steal during the experiment, and 3: you do not remember the stimuli or do not know the answer). Therefore, the correct answer was different for each group. One point was given if the answer was correct, if not then 0 points were given adding up to the total score of 12 points.

Procedure

Upon their arrival, participants were given a brief description of the experiments and their rights as a research participant and signed an informed consent form. Both the experiment and the informed consent was approved by the institutional review board of Chung-Ang University. Afterward, they were informed that they would take part in an experiment on detecting deception and instructed to try not to be judged as guilty. Then, they were randomly assigned to one of the three groups: guilty, innocent-aware, and innocent-unaware. The mission for the guilty group was to enter the teaching assistant's office, steal money (∼50 dollars) in a white envelope, then falsify an account book file in the black USB to cover up for stealing the money. After, they were to write out a fake receipt using a black pen, and then stamp a purple seal on the fake receipt without getting caught. The mission for the innocent-aware group was to go to the teaching assistant's office, ask someone for permission to bring eight items, including crime-relevant and crime-irrelevant stimuli, as an errand for the assistant. There was no specific mission for the innocent-unaware group. The innocent-unaware group just stayed in the laboratory for about 15 min doing nothing. After each mission was completed, all participants came back to the psychology laboratory and moved to the next room for the eye-tracking experiment. Then, we presented crime-relevant and crime-irrelevant stimuli as criminal evidence to the guilty and innocent-aware groups on the computer screen on the desk before eye tracking, whereas the innocent-unaware group did not receive such information. A total of eight stimuli were presented one by one for a 1,000 ms, and both the guilty and innocent-aware groups were informed that these stimuli were criminal evidence of a theft case in the laboratory. Therefore, the guilty group was exposed to crime-relevant knowledge, but individuals in this group knew the difference between crime-relevant and crime-irrelevant stimuli. While the innocent-aware group was exposed to crime-relevant knowledge, but individuals in this group could not differentiate between crime-relevant and crime-irrelevant stimuli. The innocent-unaware group was not exposed to crime-relevant knowledge at all. All participants were required to answer "No" when asked if they had committed a theft crime, and their eye movements were recorded while they looked at the pairs of stimuli that included crime-relevant information (free-viewing task).

Each trial started with a central cross-fixation for 1,000 ms, followed by a pair of stimulus shown side-by-side for 8,000 ms; then, a blank screen was presented for 1,000 ms for a given inter-trial interval before the start of the next trial (**Figure 1**). In order to control leftward or rightward bias, the location of the stimulus was counter-balanced (32). A total of 72 trials were conducted, and one stimulus was located on the left side on the screen and the other on the right side. All participants were required to maintain fixation until target stimuli appearance, and fixation behavior of the subjects was controlled prior to each trial, in that way the next trial only started if the subject fixated the cross-fixation cross for more than 300 ms. Participants' eye movements were recorded with an eye-tracker while they viewed stimuli displayed on the computer monitor. After the experiment, all participants were asked to perform a recognition test, were debriefed about the experiment and payment procedure, and were given 5,000 Won (~5 US dollars) as a reward. In addition, they were each asked not to share any information with anyone who might participate in the experiment in the future.

Data Analysis

SPSS 15.0 for windows was used for the analyses. The changes in participants' eye movements while they were exposed to stimuli displayed on the computer monitor were measured. An area of interest (AOI) was designated to cover each picture, and the eye movements were examined in terms of fixations recorded within an AOI. In order to investigate the total amount of time spent at each stimulus (dwell time) and the amount time until the first fixation (first fixation time) in each group, a 3 (group: guilty, innocent-aware, innocent-unaware) as a between-subject factor × 3 (stimuli: crime-relevant, crime-irrelevant, neutral) as a within-subject factor repeated measures analysis of variance (ANOVA) was conducted, and degrees of freedom were adjusted with the Greenhouse-Geisser epsilon to correct for violations of the sphericity assumption.

RESULTS

Sample Characteristics

There were no significant gender differences among the three groups, $\chi^2(2) = 3.45$, $p = 0.18$, *n.s.* In addition, there were no significant age differences among the three groups, $F_{(2,48)} = 1.42$, $p = 0.25$, *n.s.*

Recognition Test

The number of correctly remembered items in the recognition test was 11.17 out of 12 details ($SD = 0.59$) for the guilty group, 11.80 ($SD = 0.41$) for the innocent-aware group, and 11.74 ($SD = 0.56$) for the innocent-unaware group. The one-way ANOVA on the number of correctly recognized items revealed no significant effect of the factor group, $F_{(2,48)} = 0.13$, $p = 0.88$, *n.s.*, indicating that participants in all groups remembered the crime-relevant and/or crime-irrelevant stimuli well-according to each group's mission and did not differ in their recognition rates.

Dwell Time

Degrees of freedom were adjusted with the Greenhouse-Geisser epsilon to correct for violations of the sphericity assumption. The results revealed significant group × stimuli interaction, $F_{(2.64,63.32)} = 9.11$, $p < 0.01$, $\eta^2 = 0.28$, indicating that each group showed different eye-movement responses depending on the stimulus type. Further analysis revealed that the innocent-aware group showed a significant difference in dwell time among all stimulus types, $F_{(1.12,15.69)} = 8.56$, $p < 0.05$, $\eta^2 = 0.38$. Specifically, the innocent-aware group spent more time gazing at crime-relevant and crime irrelevant stimuli than neutral stimuli, $t_{(14)} = 2.80$, $p < 0.05$, $t_{(14)} = 3.23$, $p < 0.05$ (**Figure 2**). On the other hand, the guilty and innocent-unaware groups showed similar eye-movement regardless of stimulus type, $F_{(2,32)} = 1.41$,

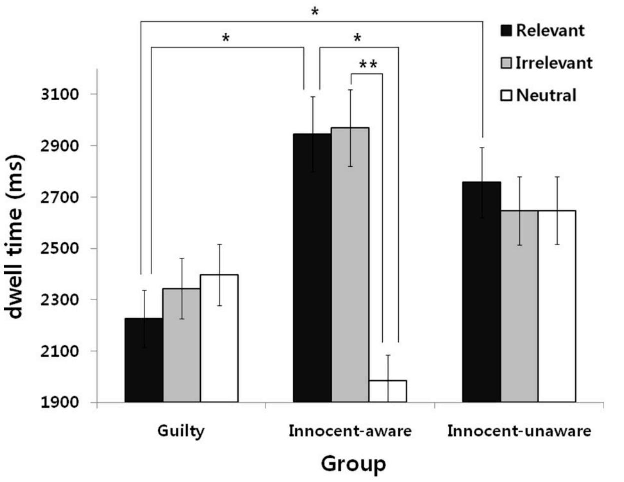


FIGURE 2 | Dwell time for the three stimulus types with respect to the subject group. Means and standard error (*$p < 0.05$, **$p < 0.01$) are shown.

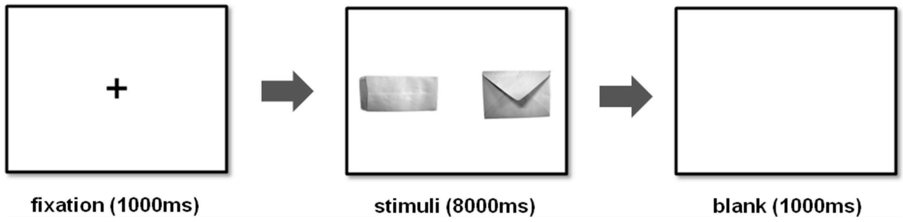


FIGURE 1 | An example of the computer screen as it appeared to the subjects during the task.

n.s., $F_{(2,36)} = 2.48$, *n.s.* In addition, we conducted analyses to compare dwell time on each stimulus type among the three groups for exploratory analysis. As a result, we found that dwell time was a significantly more for crime-relevant stimuli,t for the innocent-aware and innocent-unaware groups than for the guilty group, $t_{(30)} = 2.41$, $p < 0.05$, $t_{(34)} = 2.20$, $p < 0.05$. In addition, the innocent-aware group spent more time gazing at crime-irrelevant stimuli than the guilty group, $t_{(30)} = 2.46$, $p < 0.05$, and spent less time gazing at neutral stimuli than the innocent-unaware group, $t_{(24)} = 2.50$, $p < 0.05$.

There was a significant main effect for the stimuli, $F_{(1.32,63.32)} = 9.06$, $p < 0.01$, $\eta^2 = 0.14$, indicating that there was a statistically significant difference among stimuli in dwell time. Further analysis revealed that there was no difference between crime-relevant and crime-irrelevant stimuli (*n.s.*), while the dwell time for both crime-relevant and the crime-irrelevant stimuli were significantly longer than that for the neutral stimuli ($P < 0.05$ for both stimuli). There was no significant main effect for the group, $F_{(2,48)} = 1.28$, *n.s.*

First Fixation Time

The results revealed no significant group × stimuli interaction, $F_{(4,96)} = 1.23$, *n.s.*, and no main effect for stimuli, $F_{(2,96)} = 0.02$, *n.s.* However, there was a significant main effect for the group, $F_{(2,48)} = 3.57$, $p < 0.05$, $\eta^2 = 0.13$. A LSD *post-hoc* test revealed that the first fixation time for the guilty group was significantly longer than that for the innocent-unaware group ($p < 0.05$) and marginally longer than that for the innocent-aware group ($p = 0.06$), indicating that the guilty group showed an avoidance tendency from all stimulus types, unlike the other groups (**Figure 3**).

DISCUSSION

The purpose of the present study was to investigate the attentional bias for guilty knowledge in the guilty group in the GKT using an eye tracker. In addition, the study aimed to examine whether eye-movement measurement can compensate for the defect of the GKT, in which innocent subjects who are exposed to guilty knowledge, may be judged as guilty.

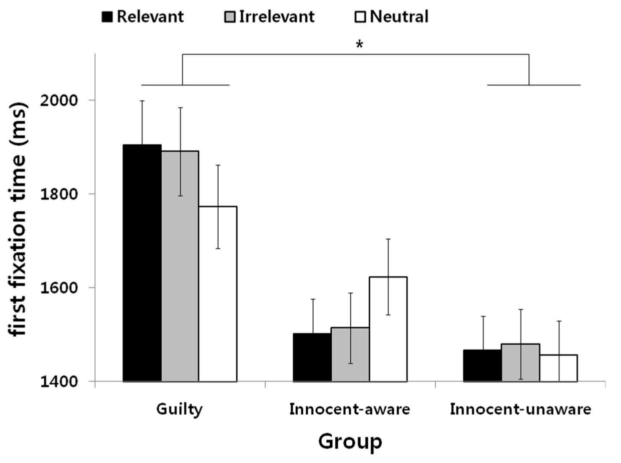


FIGURE 3 | First fixation time for the three stimulus types with respect to the subject group. Means and standard error (*$p < 0.05$) are shown.

The main finding of the present study is that the guilty group showed avoidance responses from all stimuli in a lie detection situation. Thus, our first hypothesis that the guilty group would show a shorter first fixation time and a shorter dwell time toward crime-relevant stimuli than crime-irrelevant and neutral stimuli, was rejected. In the present study, the guilty group showed no differences in dwell time for all stimulus types. They spent less time gazing at crime-relevant stimuli than the innocent-aware and innocent-unaware groups and spent less time gazing at crime-irrelevant stimuli than the innocent-aware group. Although there were no differences in dwell time, the guilty group spent less time gazing at all types of stimuli than the other two groups, indicating that they showed attentional avoidance from all stimuli. This finding is partially consistent with that of a previous study showing that guilty knowledge demands attention (24). In their experiments,Verschuere et al. (24) found that probe responses were slower in guilty knowledge trials than in neutral trials in a probe classification task, indicating general interruption of attentional performance in guilty knowledge trials, but no spatial shifting of attention. They concluded that it remains possible that participants may shift their attention away from guilty knowledge to try to avoid detection, and this result may support this prediction. We might assume that the guilty group did not involve all presented stimuli because of fear of regarding the deception detection situation *per se*. This lack of involvement is in agreement with a hesitation response during deception, which is one of the cognitively demanding tasks, such as gaze aversion (33), fewer body movements (34), and long pauses in statements between the lie detector's questions and responses.

The second important finding is that the innocent-aware group showed attentional bias toward crime-relevant and crime-irrelevant stimuli. Thus, our second hypothesis that the innocent-aware group would show a shorter first fixation time and a longer dwell time toward both crime-relevant and crime-irrelevant stimuli than neutral stimuli was partially supported. In the present study, they focused their attention on crime-relevant stimuli and crime-irrelevant stimuli longer than neutral ones. In addition, they observed crime-relevant stimuli for a longer time than the guilty group and observed neutral stimuli for less time than the innocent-unaware group. In the case of crime-relevant stimuli, such stimuli were significant to both the guilty and innocent-aware groups, but the guilty group avoided crime-relevant stimuli, unlike the innocent-aware group. This difference between the guilty and innocent-aware groups could possibly be interpreted to show the existence of the feeling of threat. The high level of threat for the crime-relevant stimuli in the guilty group under a lie detection situation may have contributed to the avoidance response, which is consistent with a previous study showing that the guilty group avoided guilty knowledge (30). In contrast, we may assume that those who had knowledge of crime information but did not commit the crime did not show an avoidance response toward crime-relevant stimuli because of the low level of threat under the lie detection test. Therefore, we assume that this indicates that mere knowledge plays an important role in allocating more attentional resources toward crime-relevant and crime-irrelevant stimuli in innocent-aware examinees, whereas actual action may have contributed to the attentional process for guilty participants.

Finally, the innocent-unaware group showed no differences in dwell time among all types of stimuli. Thus, our third hypothesis that there would be no differences in a first fixation time and a dwell time to all types of stimuli in the innocent-unaware group was partially supported. This is consistent with our prediction that there were no specific responses toward crime-relevant stimuli in the innocent-unaware group. Unlike the guilty group, the innocent-unaware group had not taken part in actual criminal action; thus, the presented stimuli might not have threatened them at all. In addition, the innocent-unaware group had no knowledge of criminal information, so there were no stimuli with significance or meaningfulness to them which might cause a threat or OR.

The present study has some implications. The guilty group, when faced with a lie detection situation showed a different pattern of attention from the innocent aware and the innocent-unaware group. This indicates that it is important to consider deception detection particularly with respect to nonverbal behaviors. Therefore, we should be careful when using detection of deception with visual stimuli since this avoidance response might cause problems, such as cheating the lie detection. In addition, attentional-avoidance patterns using eye-trackers can be used as an additional marker to distinguish deception from truth in criminal investigative settings.

The present study also has some limitations. First, it is difficult to generalize these findings to other populations and applied settings. This is because the study was conducted with undergraduate students in a mock-crime paradigm and addressed only one kind of mock-crime paradigm. Therefore, future research should be conducted with criminal suspects in a real deception detection setting and include more kinds of crimes. Second, we did not accurately and concretely assess physiological responses according to the stimulus type. Although we have controlled the valence and arousal of each stimulus type from a preliminary study, participants of the current study did not rate valence and arousal rate during this specific eye tracking experiment. Therefore, we cannot be sure that there were no differences between the stimulus types or between the groups. Future studies should rate the stimuli and measure physiological responses such as skin conductance responses and pupil sizes. Finally, although it constitutes a normal distribution, since the small sample size may elicit low statistical power, greater sample sizes may be useful in future research.

Despite these limitations, our findings may make up for the shortcomings of the GKT and provide important information on the effectiveness of the GKT as a lie detection technique. Namely, this study suggests that even if the innocent group is exposed to guilty knowledge, eye-tracking technology seems to be an effective method for distinguishing between deceptive groups and non-deceptive groups.

ETHICS STATEMENT

The experiment was approved by the institutional review board in Chung-Ang University. All of the participants signed an informed consent that had been approved by the institutional review board in Chung-Ang University.

AUTHOR CONTRIBUTIONS

KK, GK, and J-HL conceived the experiment. KK and GK designed the experimental task and data acquisition of subjects. KK data analysis. KK, GK, and J-HL data interpretation. KK, GK, and J-HL drafting of the manuscript. All the authors revised the manuscript critically and provided final approval of the version to be published.

FUNDING

National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2012-M3A2A1051124).

REFERENCES

1. Lykken DT. Psychology and the lie detection industry. *Am Psychol.* (1974) 29:725–39. doi: 10.1037/h0037441
2. Elaad E. Effects of context and state of guilt on the detection of concealed crime information. *Int J Psychophysiol.* (2009) 71:225–34. doi: 10.1016/j.ijpsycho.2008.10.001
3. Verschuere B, Ben-Shakhar G, Meijer E. Memory detection: theory and application of the concealed information test. In: Verschuere B, Ben-Shakhar G, editors. *Theory of the Concealed Information Test.* Cambridge, UK: Cambridge University Press (2011), p. 128–48.
4. Lykken DT. *A Tremor in the Blood: Uses and Abuses of the Lie Detector*. New York, NY: Plenum Trade (1998).
5. Sokolov EN. *Perception and the Conditioned Reflex*. New York, NY: MacMillan (1963).
6. Furedy JJ. The concealed information test as an instrument of applied differential psychophysiology: methodological considerations. *Appl Psychophysiol Biofeedback*. (2009) 34:149–60. doi: 10.1007/s10484-009-9097-y
7. Ben-Shakhar G. The roles of stimulus novelty and significance in determining the electrodermal orienting response: interactive versus additive approaches. *Psychophysiology*. (1994) 31:402–11. doi: 10.1111/j.1469-8986.1994.tb02448.x
8. Ben-Shakhar G, Elaad E. The validity of psychophysiological detection of information with the guilty knowledge test: a meta-analytic review. *J Appl Psychol.* (2003) 88:131–51. doi: 10.1037/0021-9010.88.1.131
9. Rosenfeld JP, Hu X, Labkovsky E, Meixner J, Winograd MR. Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. *Int J Psychophysiol*. (2013) 90:118–34. doi:10.1016/j.ijpsycho.2013.08.012
10. Meijer EH, klein Selle N, Elber L, Ben-Shakhar G. Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*. (2014) 51:879–904. doi: 10.1111/psyp.12239
11. Peth J, Suchotzki K, Gamer M. Influence of countermeasures on the validity of the Concealed Information Test. *Psychophysiology*. (2016) 53:1429–40. doi: 10.1111/psyp.12690

12. Ogawa T, Matsuda I, Tsuneoka M. The comparison question test versus the concealed information test? that was the question in japan: a comment on palmatier and rovner (2015). *Int J Psychophysiol*. (2015) 95:29–30. doi: 10.1016/j.ijpsycho.2014.09.006
13. Zaitsu W. External validity of concealed information test experiment: comparison of respiration, skin conductance, and heart rate between experimental and field card tests. *Psychophysiology*. (2016) 53:1100–7. doi: 10.1111/psyp.12650
14. Gamer M, Gödert HW, Keth A, Rill H-G, Vossel G. Electrodermal and phasic heart rate responses in the Guilty Action Test: comparing guilty examinees to informed and uninformed innocents. *Int J Psychophysiol.* (2008) 69:61–8. doi: 10.1016/j.ijpsycho.2008.03.001
15. Zvi L, Nachson I, Elaad E. Effects of coping and cooperative instructions on guilty and informed innocents' physiological responses to concealed information. *Int J Psychophysiol.* (2012) 84:140–8. doi: 10.1016/j.ijpsycho.2012.01.022
16. Zvi L, Nachson I, Elaad E. Effects of perceived efficacy and prospect of success on detection in the Guilty Actions Test. *Int J Psychophysiol.* (2015) 95:35–45. doi: 10.1016/j.ijpsycho.2014.12.010
17. Selle NK, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Orienting versus inhibition in the concealed information test: different cognitive processes drive different physiological measures. *Psychophysiology*. (2015) 53:579–90. doi: 10.1111/psyp.12583
18. Ben-Shakhar G, Gronau N, Elaad E. Leakage of relevant information to innocent examines in the GKT: an attempt to reduce false-positive outcomes by introducing target stimuli. *J Appl Psychol.* (1999) 84:651–60. doi: 10.1037/0021-9010.84.5.651
19. Ambach W, Stark R, Vaitl D. An interfering n-back task facilitates the detection of concealed information with EDA but impedes it with cardiopulmonary physiology. *Int J Psychophysiol.* (2011) 80:217–26. doi: 10.1016/j.ijpsycho.2011.03.010
20. Gamer M. Does the Guilty Actions Test allow for differentiating guilty participants from informed innocents? A re-examination. *Int J Psychophysiol*. (2010) 76:19–24. doi: 10.1016/j.ijpsycho.2010.01.009
21. Ryan JD. Hannula DE, Cohen NJ. The obligatory effects of memory on eye movements. *Memory*. (2007) 15:508–25. doi: 10.1080/09658210701391022
22. Siddle DA. Orienting, habituation, and resource allocation: an associative analysis. *Psychophysiology.* (1991) 28:245–59. doi: 10.1111/j.1469-8986.1991.tb02190.x
23. Verschuere B, Crombez G, Degrootte T, Rosseel Y. Detecting concealed information with reaction times: validity and comparison with the polygraph. *Appl Cogn Psychol.* (2010) 24:991–1002. doi: 10.1002/acp.1601
24. Verschuere B, Crombez G, Koster EHW. Orienting to guilty knowledge. *Cogn Emotion*. (2004) 18:265–79. doi: 10.1080/02699930341000095
25. Hermans D, Vansteenwegen D, Eelen P. Eye movement registration as a continuous index of attention deployment: data from a group of spider anxious students. *Cogn Emotion*. (1999) 13:419–34. doi: 10.1080/026999399379249
26. Schwedes C, Wentura D. The revealing glance: eye gaze behavior to concealed information. *Memory Cogn.* (2012) 40:642–51. doi: 10.3758/s13421-011-0173-1
27. In-Albon T, Kossowsky J, Schneider S. Vigilance and avoidance of threat in the eye movements of children with separation anxiety disorder. *J Abnormal Child Psychol.* (2010) 38:225–35. doi: 10.1007/s10802-009-9359-4
28. Armstrong T, Olatunji BO. Eye tracking of attention in the affective disorders: a meta-analytic review and synthesis. *Clin Psychol. Rev*. (2012) 32:704–23.doi: 10.1016/j.cpr.2012.09.004
29. Mogg K, Bradley BP. Time course of attentional bias for fear-related pictures in spider-fearful individuals. *Behav Res Ther*. (2006) 44:1241–50. doi: 10.1016/j.brat.2006.05.003
30. Kim K, Kim J, Lee JH. Guilt, lying, and attentional avoidance of concealed information. *Soc Behav Personal.* (2016) 44:1467–76. doi: 10.2224/sbp.2016.44.9.1467
31. Hartwig M, Granhag PA, Strömwall LA, Vrij A. Detecting deception via strategic disclosure of evidence. *Law Hum Behav*. (2005) 29:469–84. doi: 10.1007/s10979-005-5521-x
32. Nicholls MER, Orr CA, Okubo M, Loftus A. Satisfaction guaranteed: the effects of spatial biases on responses to Likert scales. *Psychol Sci.* (2006) 17:1027–8. doi: 10.1111/j.1467-9280.2006.01822.x
33. Ekman P. *Telling Lies*: *Clues to Deceit in The Marketplace, Politics and Marriage*. New York, NY: Norton and Company (2001).
34. Ekman P, Friesen WV. Hand movements. *J Commun.* (1972) 22:353–74.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 15 April 2019
doi: 10.3389/fpsyt.2019.00189

Financial Incentive Does Not Affect P300 in the Complex Trial Protocol (CTP) Version of the Concealed Information Test (CIT) in Malingering Detection. II. Uninstructed Subjects

J. Peter Rosenfeld, Elena Davydova, Elena Labkovsky and Anne Ward*

Department of Psychology, Northwestern University, Evanston, IL, United States

Well-known research showed that the skin conductance response (SCR) of the Autonomic Nervous System (ANS) in the Concealed Information Test (CIT) is usually augmented in participants who are financially and motivationally incentivized to beat the CIT. This is not what happens with Reaction Time (RT)-based CITs, P300 CITs based on the 3-stimulus protocol, nor on the P300-based complex trial protocol for detection of malingering (however these tests differ from forensic CITs). The present report follows up the Rosenfeld et al. (1, 2) study of motivated malingerers *instructed* how to beat the test, with *uninstructed* motivated (paid and unpaid) and unmotivated ("simple malingering") subjects, using episodic and semantic memory probes. The Test of Memory Malingering (TOMM) validated behavioral differences among groups. The "CIT effect" (probe-minus-irrelevant P300 differences) did *not* differ among incentive groups, although as previously, semantic memory-evoked P300s exceeded episodic memory evoked P300s. An effect of specific test-beating instructions was found to enhance the CIT effect for semantic information.

Keywords: P300 CIT, deception, motivation, incentive, complex trial protocol

OPEN ACCESS

Edited by:
Giuseppe Sartori,
University of Padova, Italy

Reviewed by:
Merylin Monaro,
University of Padova, Italy
Stefano Zago,
IRCCS Ca' Granda Foundation
Maggiore Policlinico Hospital, Italy

****Correspondence:***
J. Peter Rosenfeld
jp-rosenfeld@northwestern.edu

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *20 December 2018*
Accepted: *14 March 2019*
Published: *15 April 2019*

Citation:
Rosenfeld JP, Davydova E, Labkovsky E and Ward A (2019) Financial Incentive Does Not Affect P300 in the Complex Trial Protocol (CTP) Version of the Concealed Information Test (CIT) in Malingering Detection. II. Uninstructed Subjects. Front. Psychiatry 10:189. doi: 10.3389/fpsyt.2019.00189

INTRODUCTION

The Concealed Information Test [CIT, (3), previously known as the Guilty Knowledge Test or GKT] has been studied for half a century; [for reviews, see (4–6)]. In this test, there are at least two kinds of stimuli randomly presented regarding order to participants: The (1) *probes* are the items expected to be remembered; they are often from a crime scene in a forensic scenario—such as, a stolen diamond necklace. The (2) *Irrelevant* stimuli are other comparably valuable items (a watch, a bracelet, a broach, etc.) which are from the same category as the probe (jewelry), but are not identical to it, so are unrecognized by the thief as the stolen item. The probe *is* recognized, and therefore elicits a larger physiological response in only the knowledgeable participant. To innocent suspects, the probe is just another irrelevant so elicits a smaller or no physiological response.

The traditional responses examined in the CIT are autonomic nervous system (ANS) responses such as, Skin Conductance Response (SCR), respiration pattern, and cardiac responses. More recently, the P300 component of the event-related potential (ERP) and fMRI have been utilized [see (6–8). When P300 is used, the probes are presented infrequently, e.g., probability $= p = 0.15$, and the irrelevants are presented frequently, e.g., $p = 0.7$, and a third stimulus type—the target ($p = 0.15$)—that has a unique response requirement—is also used, mainly to assure attention].

In a recent meta-analysis, Meijer et al. (5) noted that many workers have reported that motivation and incentive typically increase the CIT effect in the SCR measure of the ANS. However, this does not happen with reaction time (RT) measures of CIT effects (9–11).

With respect to P300-based CITs, Meijer et al. (5) stated that "The bulk of CIT studies based on P300 did not use motivational instructions." We agree with this, since most of those studies were from this lab where we never reported effects of motivation on P300 in several reports. (That is, P300 amplitudes in CIT studies with incentivized subjects appear to be in the same range as they are in those studies without financial incentive). This was formally confirmed in Ellwanger et al. (12): Participants in a truth-telling group, instructed to do their best on P300 tests (involving semantic, as well as incidentally acquired, episodic memory), were compared to a motivated/incentivized "dishonest" group offered a $10 reward to "beat the test." There were no significant P300 differences found: The sensitivity of the truth tellers was 0.74, vs. 0.73 for the incentivized dishonest group. This is clear evidence that the motivational manipulation of offering a $10 reward for beating the test did not affect the CIT effect or sensitivity of the P300-based CIT. This study utilized the older "3-stimulus protocol" [3SP, (7)]. We want to emphasize, however, that the malingering protocol that detects feigned cognitive deficit about autobiographical knowledge has critical differences from the forensic CIT protocol that detects feigned ignorance of crime details, and this fact makes it difficult to generalize from malingering data to forensic CIT data. We will re-visit this issue in the discussion.

It is noted that the present and previous tests of malingering use both verbal/behavioral tests as well as P300 data, typically with a comparative aim. The verbal/behavioral tests are designed to entrap malingerers by giving them an explicit test of autobiographical memory recognition, which is easy, but appears to be more difficult, and on which they typically, but not reliably, score poorly. Because of dissatisfaction with these tests among neuropsychologists, physiological measures, especially P300, were introduced to detect malingered cognitive deficit in closed head injury (CHI) patients; (13–16). P300s are reliably evoked in response to recognized information, which has prompted their use in forensic situations, (7). It followed that P300 tests might be profitably used in detecting malingering: Malingerers may state that they forgot a learned word but if the word elicits a P300, this strongly suggests that the denied word is recognized despite the behavioral denial.

Recently, Rosenfeld et al. (1) formally observed a similar result—no effects of financial incentive manipulations on P300—using the newer, and countermeasure-resistant Complex Trial Protocol (CTP detailed below) for detection of concealed information (17). In this 2017 study (1), there were two groups. Both were motivated to beat the test and instructed specifically how to beat the test, but one group was paid for success and the other was not. Our main finding was that although there were clear, behavioral differences in the malingering *behavior* (on the Test of Memory Malingering, described below, p. 7) of the two groups, these significant effects were not reflected in the ERP data: The "Concealed Information Test (CIT) Effect"—the difference between rare critical probe and frequent irrelevant P300 amplitude– did not differ between groups. Detailed description, comparison and review of the 3SP vs. the CTP is in Rosenfeld (7). Thus, when two groups are motivated to defeat the test and instructed how best to beat it, there is no incremental effect of financial incentive on the P300 CIT effect. Indeed, it may have been the case that since both groups were motivated to beat the test and shown how to beat it, they may have been at a ceiling level of motivation.

Therefore, in the present study, we focus solely on uninstructed participants (Ps), and compare an unpaid, unmotivated "simple malingering" (SM) group to two other groups, both motivated to beat the test, with one paid to do so, and the other, unpaid. We will also compare the paid vs. unpaid, but both motivated, groups. None of the aforementioned studies examined the incremental effect of instructions specifically directed to defeating the tests by simulating malingering. This will be done here by comparing instructed groups of Rosenfeld et al. (1) with two uninstructed groups run here 1 year later on a different participant set by different experimenters.

In both Rosenfeld et al. (1) and Ellwanger et al. (12), the experimental scenario involved the simulated malingering of cognitive (memory) deficits which accompany closed head injury (CHI). As Ellwanger et al. (12) have noted, the simulating normals are not instructed to suppress *all* responses to critical/probe items, which, in contrast, *is* the case with a classical CIT scenario, making scientific comparison (of malingering and forensic scenarios) problematic. Rather, the CHI malingerer is told to imitate the performance of a real CHI patient by not making errors on *all* critical/probe items, but to only about half of them.

In the present paper as well in Rosenfeld et al. (1), we use the Test of Memory Malingering [TOMM, (18)] which is universally regarded today as the gold standard for such tests [(19, 20). See methods for more detail]. This is a familiar study-test protocol where old stimuli are first learned, after which a recognition test for learned (old) vs. new stimuli is given. For a given test item (old or new) a subject can respond either correctly/honestly, or—in a malingering fashion—dishonestly or truly incorrectly. Based on our earlier studies cited above, we expect that paid malingerers will pay closer attention to test items than unpaid subjects will, and so (a), will give more correct than incorrect responses on the TOMM, yet based on Ellwanger et al. (12) and Rosenfeld et al. (1), (b) they will *not* show an effect of financial and other incentivization on the P300 CTP test.

The background and essence of the CTP is described here: The CTP was designed to address the weaknesses of the original "3-stimulus protocol" (3SP, 17). Rosenfeld et al. suggested that the 3SP generated smaller than usual P300 responses to probes because Ps also make an explicit target decision (i.e., target vs. non-target) on every trial. Although probes do produce a P300 in guilty individuals in the 3SP, the extra job of determining if each presented item is a target weakens attention to probes, and since decision-making absorbs processing resources, it reduces the P300 response to the probe (21, 22). The CTP addresses this issue by separating probe vs. irrelevant and target vs. non-target decisions by ∼1 s. In this two-part trial, a simple "I

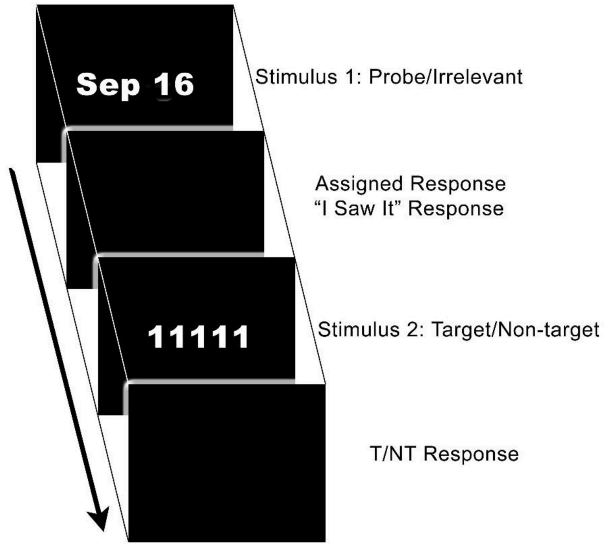


FIGURE 1 | The Complex Trial Protocol event sequence, with a date stimulus as stimulus 1 (probe or irrelevant), then the perception acknowledgment response ("I saw it"), then the target or non-target as stimulus 2, then the target or non-target response. All stimuli are presented for 300 ms each. There is a randomly varying interval of 1300–1800 ms between S1 and S2. There is a 2 s interval between the T/NT response and the next S1 for the next trial.

saw it" response is required for the first stimulus (probe or irrelevant), which is followed by a target vs. non-target decision; (see **Figure 1**, showing a date stimulus 1 [S1] and a subsequent target ["1111"]). The initial stimulus (i.e., probe or irrelevant) requires a unitary "I saw it" button response with the left hand, but the subsequent target-non-target response depends on the second stimulus type (S2), so that differing right-hand mouse buttons correspond with the target ("yes" button) and non-targets ("no" button). Also, targets and non-targets are typically from a different category than probe/irrelevants. Separating the implicit (probe vs. irrelevant) and explicit (target vs. non-target) decisions—*combined* in the 3SP—frees processing resources, resulting in larger P300 responses, and greater differences between probe and irrelevant P300s, thereby improving CM resistance (17). Comparisons of the CTP and the 3SP are detailed in Rosenfeld (7).

METHODS

Participants

The subjects were recruited from the Northwestern University Introductory Psychology Pool. Participants were mostly college freshmen and sophomores, plus a few juniors and seniors, aged 17–22. The study was consistent with ethical guidelines as it was approved by the Northwestern IRB. There were initially three groups, 2 of 21 each, and one of 22 participants. The groups were formed by random assignment to groups which is expected to assure gender and age balance across groups. The three groups had 14, 16, and 15 females. The group numbers were based on a power analysis directed at having an 80% likelihood of discovering a medium size effect with alpha = 0.05. For all 64 subjects, the mean age = 18.8, SD = 1.4. There was (1) A group told to simulate malingering (SM group) but not to try to beat the test, nor rewarded for same. (2) Two groups told to simulate malingering and encouraged to try to beat the test. One of these groups was unpaid (BtNo) and the other was paid (Bt $) to beat the test. None were instructed how to beat the test as the subjects were in Rosenfeld et al. (1). These subjects were told to duplicate performance of head injury patients by not getting every item wrong, but by answering incorrectly only about half the time. The **Supplementary Materials** give detailed instructions.

Procedures

The probe stimuli used in the CITs were the P's birthday (semantic memory) in one block, and the experimenter's name (episodic memory) in a second CIT block; block order was counterbalanced across Ps in both paid and unpaid groups. We do not mean to imply that birthday and experimenter name are the perfect exemplars of the 2 respective memory categories. Other exemplars may have different results. Here, so as to replicate Ellwanger et al. (12) and Rosenfeld et al. (1), exposure to the experimenter's name was as follows: The P was first contacted via an e-mail (sent to arrange the experimental session time) in which the experimenter's name appeared twice. When a P entered the lab, (s)he was greeted at the door with the sentence, "Hi come on in. My name is Elena. I e-mailed you about our appointment." (The entire verbatim interaction and instructions used with all Ps are seen in the **Supplementary Materials**). Instructions were given, the subject was asked to look at a list of intended irrelevant date stimuli to be used, and to circle any that were unintentionally and by chance, relevant personally—such as, the birth date of a close acquaintance. This was replaced in the list of irrelevants to be used in the date CIT. After full instructions (in **Supplementary Materials**) were given, and just before the name block in the P300-CIT, the experimenter asked the P if (s)he remembered the experimenter's name. If P did, the CIT was given. If not, the experimenter repeated her first name while holding up a card with this name. The subject then repeated the name. The P300-CIT followed, and after it, the P was then tested again on the name. All Ps responded correctly. All Ps were also asked after the birthday block if they saw the birthday; all reported that they did.

The next procedure was administration of a modified version of the Test of Memory Malingering [TOMM, (18)], a validated (23, 24) instrument strongly supported by Teichner and Wagner (25) to detect malingering. (They stated: "Results suggest that the TOMM is a useful index for detecting the malingering of memory deficits.") The TOMM is universally regarded today as the gold standard for tests of memory malingering (19, 20).

The version we used was an abbreviated version of the TOMM as suggested by Hilsabeck et al. (26). The abbreviated TOMM was used in order to assess the malingering manipulation and compare its effects among groups.

The TOMM we used involves a study-test manipulation, with 50 initial exposures of line drawings of common objects in a study block, one by one, followed after about 2 min with a test on 100 more pictures containing the randomly ordered 50 initially studied ("old") pictures randomly shuffled with 50 novel ("new") pictures. Ps were instructed to press one button if they recognized the picture, and another if they did not. Thus, there were two types of outcomes (correct and incorrect/faked) on all test trials with *Old* stimuli, and likewise for test trials with *New* stimuli. Ps were still under the malingering instruction set, and were so reminded in the TOMM. (See **Supplementary Material**).

We note that in the usual clinical version of the TOMM (18), the *test* stimuli are presented as 50 pairs, each containing an old drawing plus a new drawing. This is similar to our test, which is no more difficult than the clinical TOMM, so the norms (26) for the clinical version, probably apply here. They are that a score of 82% or more is probably from a non-malingerer, whereas, a score of 62% or less is from a malingerer.

At this point, all motivated (told to beat the test) Ps were shown their averaged probe and irrelevant P300s, so as to determine with our bootstrap software (described below) whether they were detected in their malingering or not, based on the P300 values. We illustrate the superimposed probe and irrelevant ERPs of guilty vs. Innocent participants in previous studies, and describe how large differences indicate guilt. Moreover, we tell them that the software will output the expected numbers of times in 100 samples that the probe > irrelevant P300. We also tell them that a 90 is required for a guilty diagnosis. Successful members of the paid group were paid. Then all Ps were discharged.

It is emphasized that malingering instructions were in effect during both the P300 tests, as well as during the TOMM sessions. This is detailed in the **Supplementary Material**.

Data Acquisition

P300, measured P300 peak to the subsequent negative peak ["peak to peak" or p-p as in (27)] from Fz, Cz, and Pz, was recorded, filtered, artifacted, and averaged as previously [e.g., (28)]:

EEG recording used tin electrodes on the scalp at sites Fz, Cz, and Pz. They were referenced to linked mastoids. EOG was recorded with an electrode (tin) above the right eye and also referenced to the linked mastoids. Eyeblinks were removed with the method of Semlitsch et al. (29). Any remaining eye artifacts were manually detected, marked, and all trial data containing 80uV (or more) signals in any channel were dropped. The forehead was connected to the chassis of the isolated side of the amplifier ("ground"). Signals were passed through a Mitsar 19 channel (model 201) amplifier with a.16 Hz high pass filter setting, and low pass filters at 30 Hz. Output was conveyed to a 16-bit Mitsar Analog to Digital converter sampling at 500 Hz. For analyses and displays, single sweeps and averages were digitally filtered; the filter passed frequencies from 0 to 6 Hz using a *Kaiser* filtering algorithm. A minimum of 20 sweeps per average were required for each stimulus. The average number collected across subjects was 27.6 per subject.

P300 amplitude was measured at Pz using both the "base-to-peak" (b-p) and the (p-p) methods. [The p-p method has often been confirmed as the most accurate in P300-based deception studies: See (27, 30). Both b-p and p-p methods search from 300 to 650 ms for the largest positive 100 ms segment; this is the b-p P300. The midpoint of this segment is defined as the P300 latency. The average amplitude difference of the segment from the pre-stimulus baseline is defined as the base-peak value. For p-p, the algorithm likewise searches for the largest *negative* 100 ms segment between P300 latency and 1,300 ms and then subtracts the average amplitude of that segment from that of the maximally positive segment. Our present choice of a search window was made based on the grand average of all subjects in all conditions, the procedure recommended by Keil et al. (31).

Within Individual Analysis: Bootstrapped Amplitude Difference Method

To determine if the P300 elicited by one stimulus is greater than that elicited by another *within an individual*, the bootstrap method (32) was used on the recording from Pz. The bootstrap method answers the question of whether or not the probability is more than 90 in 100 that the real difference between the average probe P300 and the average irrelevant P300 is > 0. However, for each subject, one has only one average probe P300 and one average irrelevant P300 available. Answering the question requires distinct distributions of *average* probe and *average* irrelevant P300s, and these distributions are unavailable. We thus bootstrap these distributions with the following procedure: An algorithm goes through the combined (probe-followed-by target in the CTP and probe-followed-by non-target in CTP) set (all single sweeps) and randomly draws, *with replacement*, a set of n1 probe waveforms. It averages these and computes P300 amplitude from this average using the segment selection method described for the p-p index. Next a set of n2 waveforms is drawn on a random basis *with replacement* from the set of irrelevant waves, from which an average P300 amplitude is calculated. The numbers n1 and n2 are the actual numbers of accepted probe and irrelevant sweeps for a given participant, but n2 is multiplied by a fraction (about.142 in the present report) which randomly reduces the number of irrelevant trials to within one trial of the n1. The computed irrelevant mean P300 is then subtracted from the comparable probe value, resulting in a difference value for a distribution that will contain 100 values after 100 iterations of the process just described. (*BSITERS* is the number of iterations in which probe P300 > Irrelevant P300; it must be 90 or more in this report for a knowledgeable decision). Multiple iterations yield differing probe-minus-irrelevant differences because of the sampling-with-replacement process. (We also use the mean of this 100-iteration difference distribution here as a dependent variable, *BSMEAN*).

Dependent Variables

In evaluating group effects of the critical independent variables, two different and related dependent variables were utilized here. First is the Pz p-p P300 amplitude difference from our sample in microvolts between probe and irrelevant P300 averages, that is usually large in knowledgeable, but not unknowledgeable subjects. We also use BSITERS and BSMEAN, defined above.

Group Statistical Analyses

ANOVAs and *t*-tests were used for group analyses. Effect sizes for $p < 0.2$ are reported using partial eta squares (*petasq*). These values can be benchmarked against Cohen's (33), pp. 278–280) criteria of small (0.01), medium (0.06), and large (0.14) effects, as reviewed by Richardson (34). Cohen's d is used for all *t*-tests. Guidelines for d are as follows: small (0.2), medium (0.5), and large (0.8). For 2-level independent variables in all cases of marginally significant effects ($p < 0.15$), Bayes factors [JZS BFs, with scaled r = 0.707, as in Rouder et al. (35); as obtained from http://pcl.missouri.edu/bayesfactor] are also reported here as "BFs." The BFs are mostly used to confirm the likelihood of the null hypothesis (relative to the likelihood of the alternative hypothesis) when $p > 0.05$. This cannot result from non-Bayesian analysis. We do not use them when $p > 0.25$ or <0.01. BFs are mainly used to provide relative support for the null hypothesis, however when $p > 0.25$, the probability that an observed difference is due to chance is difficult to rule out so it becomes pointless to give the BF. Likewise, if $p < 0.01$, it is increasingly gratuitous to use BFs to help confirm the alternative. These BF likelihood ratios are stated as favoring the null or the alternative hypothesis, and the associated numbers will be odds ratios favoring either hypothesis. When these ratios are close to 1.0, they cannot be interpreted as favoring either hypothesis, as one is about as likely as the other.

Despite the often cited 57-year old interpretation (36) of the BF, this factor is a *continuous measure* and "does not force an all-or-none decision, but instead reallocates belief [in null vs. alternative hypothesis] on a continuous scale." [from Schönbrodt et al. (37), p. 2]. One can never prove the Null Hypothesis, but the *continuous measure* perspective of the BF discourages arbitrary thresholds of confidence, although these are still often used. A recent treatment was provided by Kass and Raferty (38) who suggested that BF = 1–3.2 is worth a bare mention, BF = 3.2–10 is "substantial," BF = 10–100 is "strong" and BF > 100 is "decisive." In givinging a BF, we always divide null likelihood ratio by alternative if BF favors null, and we always divide alternative by null if BF favors alternative. Therefore, all our BFs are positive and equal to or > +1. When we state "The BF in this test was 2.5 in favor of the null" we mean that the null hypothesis is 2.5 times as likely as the alternative hypothesis. Likewise, "The BF in this test was 5.5 in favor of the alternative" means that the alternative is 5.5 times as likely as the null hypothesis. For higher (>2) level ANOVAs, in which the usual ANOVA yields an effect of interest with $p < 0.2$, we do Bayesian ANOVAs in JASP (https://jasp-stats.org/) in order to estimate evidence for the null relative to the alternative hypothesis.

Behavioral Results: TOMM Data

All behavioral and ERP data collected are in a SYSTAT 8.0 data file and may be obtained by contacting the senior author, jp-rosenfeld@northwestern.edu

We used the TOMM to establish (1) that malingering groups (simple, SM; motivated-paid, Bt $; and motivated-unpaid, BtNo) were malingering, as instructed, and (2) to establish that there were behavioral differences among groups attributable to the differing instructional sets heard by each group.

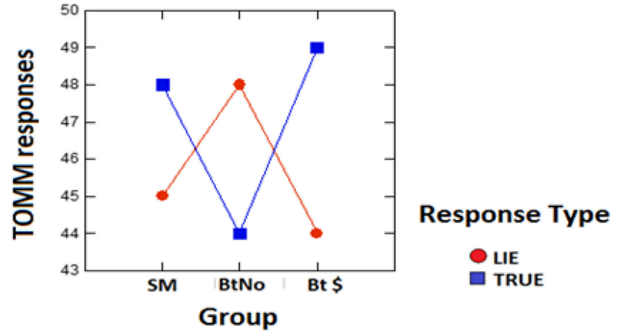


FIGURE 2 | TOMM data. Numbers of correct/honest ("TRUE" in blue) and incorrect/malingered ("LIE" in red) responses in 100 trials as a function of group.

There is no question that all three groups were malingering. Using the Hilsabeck et al. (26) norms (>82/100 correct is normal/not malingering; 62/100 or less suggests malingering), all groups were malingering since no P scored more than 59 of 100 opportunities for correct responses.

The number of correct/honest ("TRUE" in **Figure 2**) and incorrect/malingered ("LIE" in **Figure 2**) responses out of 100 total trials is shown as a function of incentive group in **Figure 2**. This figure and the subsequent statistical analyses are based on the full initial P number = 64, less four outlier subjects (one each from Groups SM and BtNo, and two from Bt $) whose correct response numbers were more than 2SD from the respective group mean.

Figure 2 shows what appears to be a complex interaction of group and response type. A 2-way, mixed 3 (groups) × 2 (response types) ANOVA was performed. The main effect of groups was ns at $F_{(2,56)} = 0.091$, $p = 0.914$. Likewise the main effect of response type was ns at $F_{(1,56)} = 1.363$, $p = 0.248$. These null effects were expected in view of the apparent interaction of group by response type. This effect was significant at $F_{(2,56)} = 3.17$, $p < 0.05$, petasq = 0.098 (medium to large). To follow up on this result, we decided to do one 2 × 2 ANOVA on a *post-hoc* basis, in which we compared only the paid vs. unpaid, both encouraged to beat the test without instructions as to how to do so. This would allow comparison with the same test done on instructed participants (also paid vs. unpaid) in Rosenfeld et al. (1). The 2 (groups, BtNo vs. Bt $) by 2 (response types, TRUE vs. LIE) ANOVA revealed no main group effect with $F_{(1,37)} = 0.118$, $p = 0.733$. Neither was there a main effect difference between numbers of TRUE vs. LIE responses; $F_{(1,37)} = 0.13$, $p = 0.721$. This also related to the significant interaction, $F_{(1,37)} = 5.805$), $p = 0.021$, BF = 2.84 in favor of the alternative (meaning that the alternative hypothesis of the interaction is 2.84 times as likely as a null effect), with petasq near large at 0.136. This significant interaction for the motivated *uninstructed* groups run here was exactly the same as the one reported in Rosenfeld et al. (1) for motivated *instructed* groups which were otherwise exactly like the *uninstructed* BtNo and Bt $ groups run here. We suggested then that the interaction is consistent with the view that the paid malingered group pays more attention to malingering instructions, by being more careful about not malingering too much. However, the present results suggest that this interaction does not depend on detailed malingering instructions, given that the specific malingering instructions were not used here, yet the same interaction was obtained.

The 2 × 2 interaction was decomposed by doing t-tests comparing TRUE and LIE responses within each group: Within the BtNo group, $t_{(18)} = 1.231$, $p = 0.234$, BF was null at 2.183. However, within the Bt $ group, $t_{(19)} = 2.427$, $p = 0.025$, BF supported alternative at 2.39. Thus, in the Bt $ group, the financial incentive was sufficient to produce the significantly greater number of truthful responses.

The results emphasizing that paid malingerers perform more accurately/honestly –*as instructed*–than unpaid malingerers is as we predicted, and as was seen in earlier studies of malingering reviewed in the introduction. The interactions and related results also confirm our manipulation regarding malingering.

By combining TOMM data from the *instructed*, motivated groups in Rosenfeld et al. (1) with the present TOMM data from *uninstructed* motivated groups, we found no effects in an ANOVA (2 groups × 2 response types, TRUE, and LIE) on combined instructed and uninstructed groups: For groups, $F_{(1,80)} = 1.132$, $p = 0.29$, BF favors null at 2.07 (i.e., the null is more than twice as likely as the alternative). For response types, $F_{(1,80)} = 0.918$, $p = 0.341$, BF favors null at 2.25 (i.e., the null is more than twice as likely as the alternative). Neither was the interaction significant; $F_{(1,80)} = 0.174$, $p = 0.678$, BF favors null at 3.02, which approximates the Kass and Raferty (38) criterion of "substantial" evidence for the null hypothesis. Thus, we saw no evidence supporting the effect of instruction on honest vs. dishonest behavioral responding.

Behavioral Results: Reaction Time Data

RTs to Probe and Irrelevant items in birthday and experimenter name conditions of the P300 CIT are shown in **Table 1** for all three groups. We had no specific predictions about the effect of motivational manipulation on RTs, other than what might be predicted from Seymour et al. (39), i.e., that probe RTs would be longer than irrelevant RTs. Moreover, we found no group differences in the instructed groups of Rosenfeld et al. (1). Thus, we performed a 2 (stimulus types; probe vs. irrelevant) × 2 (memory types; name vs. birthdate) × 3 (group; SM vs. BtNo vs. Bt $) ANOVA. The effect of group was ns; $F_{(2,46)} = 0.598$, $p = 0.554$. The effect of memory type was likewise ns; $F_{(1,46)} = 0.619$, $p = 0.435$. The interaction of group and memory type was likewise ns, $F_{(2,46)} = 0.332$, $p = 0.719$. The main effect of stimulus type was also ns; $F_{(1,46)} = 1.164$, $p = 0.286$, nor did stimulus type interact with group; $F_{(2,46)} = 0.158$, $p = 0.855$. However, the interaction of memory type and stimulus type was significant; $F_{(1,46)}$ was 10.294, $p = 0.002$, with petasq = large value of 183. The triple interaction was ns; $F_{(2,46)} = 0.733$, $p = 0.486$.

We thus, re-examined effects within memory type by first performing a 2 (stimulus types) × 3 (groups) ANOVA on birthdate data only. The results were no group effect; $F_{(2,52)} = 0.743$, $p = 0.481$. However, we did find the predictable effect of stimulus type, with $F_{(1,52)} = 8.57$, $p = 0.005$, petasq = 0.141 (large), with BF substantially favoring alternative at 6.69. The interaction was ns at $F_{(2,52)} = 0.161$, $p = 0.851$. The same analysis on the name data yielded no significant effects: For groups, $F_{(2,47)} = 0.402$, $p = 0.671$. For stimulus type, $F_{(1,47)} = 0.354$, $p = 0.555$, and the interaction was $F_{(2,47)} = 0.273$, $p = 0.763$.

TABLE 1 | Behavioral reaction times (msec) to probe and irrelevant birthdates (BD) and Experimenter Names (NM) during CTP.

GROUP	PROBE BD	IRREL BD	PROBE NM	IRREL NM
Unpaid	345.8	328.0	332.7	332.6
Paid	372.1	357.7	364.2	370.7
Simple malinger	388.7	377.7	362.8	363.5

There are thus, in agreement with others [e.g., (11)], no effects of motivational group on RT; the familiar effect of stimulus type on RT (39) holds up, but only in the birthday data.

Qualitative ERP Results

The grand average ERPs are seen in **Figure 3**, sorted by incentive groups (columns; Simple Malingering, "SM", beat test without pay, "BtNo" and beat test for pay, "Bt $") and memory types (rows; Top: experimenter's name/episodic vs. Bottom: participant's birthday/semantic). The visually obvious effects are probe P300 > Irrelevant P300, and birthdate probe-minus-irrelevant P300 > name probe-minus-irrelevant P300. **Figure 4** shows a plot of computed P300 amplitude (p-p) as a function of group: (SM, BtNo, and Bt $), stimulus type (PR: probe vs. IALL: irrelevant), and memory type (name, NM vs. birthdate, BD). (IALL is the average P300 of all irrelevant P300s).

Quantitative ERP Results

Of the 64 initially run participants in three groups, the analyses and figures below are based on 46–53 participants (depending upon whether name or birthdate stimuli were involved. Members of the SM, unpaid (BtNo) group and of the paid (Bt $) group had either birthdate and/or experimenter name data removed due to excessive artifacts, or in one case, failing to follow instructions. Thus, between-group analyses were based on at least 18 SM, 13 unpaid, and 15 paid subjects, cell sizes we have used in multiple previous studies [determined via a priori power analysis, and as reviewed in (40)].

Following up on **Figure 4**, we first did a 3-way, 2 (stimulus types, probe vs. irrelevant) by 2 (memory type, episodic vs. semantic) by 3 (groups, SM, BtNo, Bt $) ANOVA; the "Bt" notation means both groups were motivated to *beat* the *test*. As we found in Rosenfeld et al. (1) with a different group of instructed participants, there was the usual main effect of stimulus type with $F_{(1,44)} = 145.1$, $p < 0.001$, petasq = 0.767, and a main effect of memory type, $F_{(1,44)} = 22.1$, $p < 0.001$, petasq = 0.34. The interaction of stimulus type and group was ns with $F_{(2,44)} = 0.47$, $p = 0.628$, petasq = 0.02. The interaction of memory type and group was ns with $F_{(2,44)} = 1.19$, $p = 0.313$, petasq = 0.05. As in Rosenfeld et al. (1), we also saw a significant interaction of stimulus type and memory type, $F_{(1,44)} = 22.6$, $p < 0.001$, petasq = 0.34, indicating a greater effect of stimulus type for semantic than for episodic memory

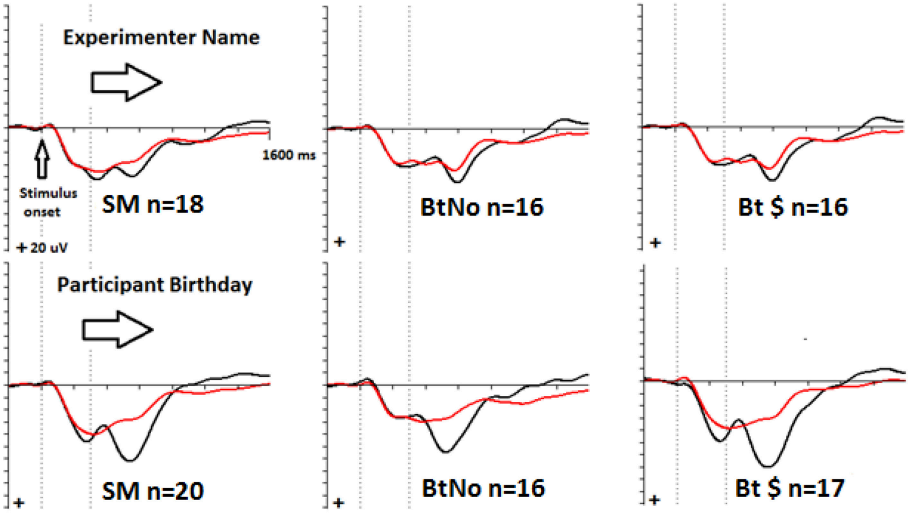


FIGURE 3 | Averaged P300 response waveforms to probes (black font) superimposed on irrelevant (red font) P300s in 3 groups in 3 columns from left to right: SM, BtNo, and Bt $. The top row shows ERPs elicited by episodic experimenter name stimuli; the bottom row shows ERPs elicited by semantic participant name stimuli. The dashed vertical lines show stimulus onset and offset in temporal order.

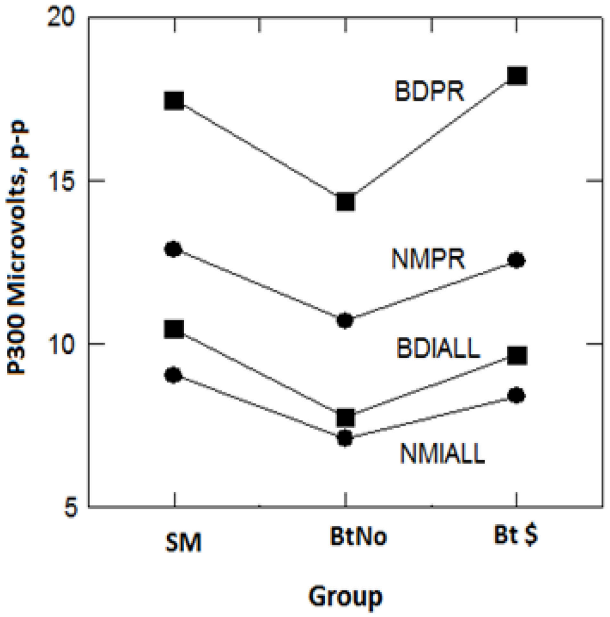


FIGURE 4 | Computed P300 (p-p) values in microvolts as a function of groups on the x-axis for 4 stimulus/memory types: NM is experimenter name, BD is participant birthdate, PR is average probe, IALL is average of all irrelevants.

type. This 2-way interaction, evident from **Figure 4**, shows that the probe-irrelevant differences were greater for the birthday (semantic) than experimenter name (episodic) stimuli across all three groups, SM, BtNo, and Bt $. This was confirmed in a follow-up ANOVA in which the dependent variable was probe-irrelevant P300 difference as a function of memory type and group. The effect of group was again ns, $F_{(2,53)} = 0.799, p = 0.455$. The critical effect of semantic vs. episodic memory type was $F_{(1,53)} = 48.94$, $p < 0.001$ with petasq = 0.46, a very large effect. The interaction of memory type and group was just short of significance, $F_{(2,53)} = 2.795, p = 0.07$.

The triple interaction was clearly not significant, $F_{(2,44)} = 0.289$, $p = 0.75$, petasq = 0.01. The main effect of group was marginally short of significance with $F_{(2,44)} = 2.85$, $p = 0.069$, BF = 1.34 (indeterminate), petasq = 0.115, probably reflecting the fact that the BtNo group showed slightly reduced P300s across all stimuli in **Figure 4** for unknown reasons. However, this non-significant effect is of minor interest in this study; our main interest concerns effects of motivational group on the *CIT effect*, i.e., the probe-irrelevant P300 (p-p) amplitude difference, and that is reflected by the non-significant interaction of stimulus type and group, described above as $p = 0.628$. This was not the case for the behavioral/TOMM data in which **Figure 2** and its analysis showed a clear difference between paid and unpaid groups: The interaction term in that analysis meant that the difference between probe-irrelevant differences was significant at $p = 0.02$, with a BF of 2.84 in favor of the alternative. To compare P300 data, we did a *post-hoc* comparison (t-test) restricted to paid vs. unpaid groups' probe-irrelevant P300 differences (name and birthday combined) from **Figure 4**. The result was $t_{(36)} = 0.438, p = 0.664$, BF favoring null at 2.67.

In view of the significant effect of memory type, we decided to do follow-up, separate analyses within memory type, and the dependent variable we used was in all cases the CIT effect, i.e., the probe- irrelevant p-p P300 difference: In these follow-up tests, we planned *a priori*, orthogonal comparisons, namely, (1) the comparison of the SM group with both combined motivated groups (paid and unpaid), and (2) the comparison of paid and unpaid groups. For the episodic experimenter name stimuli, the comparison of SM with both motivated groups combined was ns,

$t_{(47)} = 0.038$, $p = 0.968$, d = 0.012, BF = 3.4, substantial evidence (38) in favor of null. For comparison of the two motivated groups, likewise, $t_{(29)} = 0.386$, $p =$ 0.703, $d = 0.139$, BF = 2.77, which is close to substantial evidence in favor of null, and is close to the null hypothesis being three times as likely as the alternative hypothesis. For the semantic birthdate stimuli, the comparison of SM with both motivated groups combined was ns, $t_{(52)} = 0.462$, $p =$ 0.646, $d = 0.126$, BF = 3.25, which is substantially in favor of null. For comparison of both motivated groups, $t_{(32)} = 1.623$, $p = 0.114$, $d = 0.557$, BF = 1.12 in favor of null, although this low value provides clear support for neither null nor alternative hypothesis. Over all these comparisons, there is scant support for the effects of financial motivation and incentive to defeat the test on the P300-based CIT effect.

In Rosenfeld et al. (1) there were also two motivated malingering groups, one paid and one unpaid, but both were additionally instructed how to beat the test (on the same stimuli as used here). It is thus possible to combine that data set with the present one, and thereby obtain the isolated effect of instructions. **Figure 5** shows a bar graph of the five groups run in both the present and previous studies, the latter groups italicized in the following list: (1) the simple malingering (SM) group, (2) the uninstructed, unpaid group motivated to defeat the test (BtNo), (3) the uninstructed, paid group motivated to defeat the test (Bt $), *(4) the instructed, unpaid group motivated to defeat the test (BtINo), and (5) the instructed, paid group motivated to defeat the test (BtI $)*. To examine the effect of instructions, we compared the combined second and third groups (both uninstructed) with the combined fourth and fifth groups (both instructed). For the name stimuli, $t_{(68)} = 0.042$, $p = 0.967$, $d = 0.01$, BF = 4.04, substantial evidence in favor of null. However, for semantic birthdate stimuli, $t_{(72)} = 2.07$, p =0.04, $d =$ 0.505, BF = 1.48 anecdotally in favor of alternative. As **Figure 5** suggests, for semantic birthday stimuli the probe-irrelevant difference for the two instructed groups at right is greater than for the comparable uninstructed groups, second and third from the left. So while we saw no effect of financial motivation on P300, we did see an effect of test-beating instruction.

Finally, although we showed separately within instructed (1) and uninstructed (above) groups, that financial motivation does not impact the CIT effect, we can now combine data from the previous and present studies (as in **Figure 5**) to do a more powerful test on the same issue. Thus, we compared the two combined motivated paid groups from **Figure 5** with the combined motivated unpaid groups from the same figure. For the episodic name stimulus, $t_{(68)}$ =0.84, p =0.404, d =0.20, with BF substantially favoring null at 3.01. For the semantic birthday stimulus, $t_{(72)} = 1.13$, p =0.263, d =0.26, with BF favoring null at 2.4; i.e., null is 2.4 times as likely as alternative. This supports the *lack of effect of financial motivation on the CIT effect for episodic as well as semantic stimuli.*

Bootstrap-based individual diagnostic data are shown in **Figure 6**. The averaged, within-subject percentage of total iterations in 100 in which the probe P300 > Irrelevant P300 is shown on the y-axis, with incentive group, as in **Figure 5**, on the x-axis. Semantic birthdate-evoked values are at left, and episodic experimenter name-evoked values are at right.

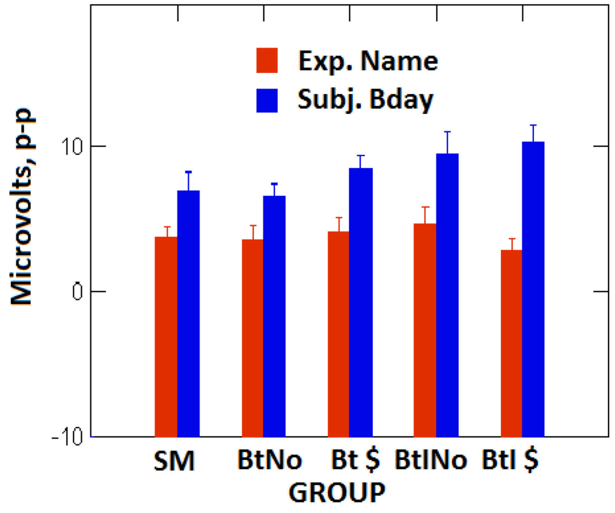


FIGURE 5 | Computed probe-minus-irrelevant P300 (p-p) difference values in microvolts as a function of 3 groups on the x-axis, as in **Figure 4**, supplemented by 2 instructed groups (BtINo and BtI $) from Rosenfeld et al. (1). Experimenter name values are in blue, participant birthdate values are in red. Error bars are S.E.M. values.

Consistent with the amplitude data described above, the hit rates are greater for semantic birthdate stimuli (at about 93% overall) than for episodic name stimuli (about 77% overall), nor does there seem to be much of a systematic main effect of group, with birthdate values slightly increasing across groups, while name values decrease. Although the y-axis ranges of both birthdate and name boxes are about the same (35–37, respectively), the error bars representing S.E.M. appear greater for name values than for birthdate values. Formal analysis of this effect is in **Table 2**, which shows variability indices in the five groups for the bootstrap iteration scores varying between 0 and 100%. Range refers to maximum score minus minimum across 100 iterations within Name (Nm) and Birthdate (Bd) conditions. The F-values are the variance ratios (distributed as F) of Bd divided by Nm. It is seen that Nm percentage variances are significantly smaller than Bd values in all four motivated groups, but not in the SM group. Likewise, there is no overlap between mean Nm and Bd range values in the motivated groups.

The first analysis of the data in **Figure 6** involved a 2 (groups) by 2 (memory types) ANOVA. The two groups compared were the SM group vs. the four combined motivated groups. As predicted there was no main effect of group, $F_{(1,83)} = 0.009$, $p = 0.927$, petasq = 0.0001. Consistent with amplitude data and visual impressions, there was a main effect of memory type, $F_{(1,83)} = 13.17$, $p < 0.001$, petasq = 0.137. There was also a significant interaction, $F_{(1,83)} = 4.87$, $p = 0.03$, petasq =0.05, confirming that the birthday percentages followed a different trend than the name values. We therefore next did separate t-tests within memory type, in which we compared SM with motivated groups as in the first ANOVA. For Birthdate (**Figure 6**, left), $t_{(23)} = 1.937$, $p = 0.065$, with the BF indeterminately favoring the alternative at 1.24 with Cohen's d = 0.55. For Name (**Figure 6**, right), $t_{(86)} = 1.202$, $p = 0.233$, with BF favoring Null at 2.051.

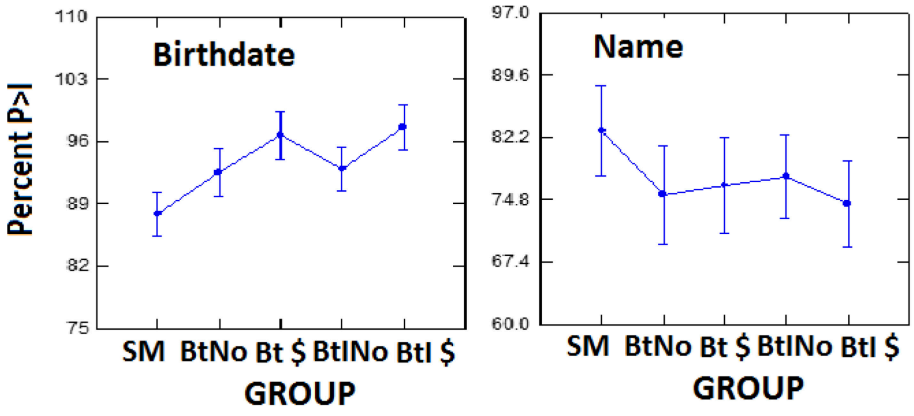


FIGURE 6 | Percent of iterations in which probe P300 > irrelevant P300 as a function of group (as in **Figure 5**. and memory type; left box for birthdate stimuli, right box for name stimuli. The error bars for each panel show the mean SEM averaged across groups, so are all the same. There were actually differences among group SEMs.

TABLE 2 | Variability of bootstrap number (number of bootstrapped iterations in which P>I in 100 trials) across motivational groups from Rosenfeld et al. (1), BtINo and BtI $ (both instructed); and present uninstructed groups: SM, BtNo, and Bt $.

Group	(*n*)	Nm range	BD range	*F*	*p*
SM	(20)	53	61	1.14	ns
BtNo	(17)	23	79	9.45	<0.01
Bt $	(17)	20	77	22.3	<0.01
BtINo	(20)	49	73	2.24	<0.05
BtI $	(20)	20	89	22.8	<0.01

Range refers to maximum minus minimum in the group from 0 to 100 for name (Nm) and birthdate (BD) stimuli. F is the variance ratio from dividing BD variance by NM variance, with associated probabilities (p).

Cohen's d = 0.358. There was thus no clear and consistent support for the notion that motivated groups perform differently than the simple malingering group. Finally, we analyzed for possible differences among the four motivated groups, separately within memory type. A 1 by 4 group ANOVA on the name data, with the dependent variable being number of P>I iterations in 100 yielded $F_{(3,66)} = 0.064$, $p = 0.979$, petasq = 0.003. For the birthdate data, $F_{(3,70)} = 1.392$, $p = 0.253$, petasq = 0.0563. On the bootstrapped iterations variable, within each memory type, there is no clear evidence of an effect of financial incentive.

DISCUSSION

A possible limitation on the conclusiveness and generality of the presently observed lack of support for motivational effects on the P300 CIT effect in the malingering scenario concerns the possible lack of statistical power available given the numbers of subjects utilized, i.e., 13, 15, or 18 per group. Although many of our previous ERP studies [see (7) for review] have utilized 12–15 subjects per cell, based on power planning analyses, and reported many robust effects, researchers used to group sizes of 20 or more may have reasonable concerns regarding some of the null ERP findings reported here. These concerns may be tempered, however, by the fact that the motivational manipulations which had negligible effects on P300 here, nevertheless had clear behavioral effects here in the same subjects. Moreover, we did make use of Bayes Factors, which allow one to quantify the relative likelihood of null and alternative statistical hypotheses. These values clearly favored the alternative hypotheses regarding the TOMM test effects, but favored the null hypotheses at near to and at substantial levels regarding P300 effects.

Another limitation on the generality of these results concerns the fact that the age range of participants was narrow (17–22). A future study can remedy this limitation by using the same methods with a sample of participants from a wider range of ages.

The present finding that financial incentive at levels that do produce behavioral effects, but do not appear to affect the P300 CIT effect in the CTP version of the P300-based CIT (for detection of malingering) is consistent with what we found previously (1, 12) using both the older 3-stimulus protocol, as well as the Complex Trial Protocol (CTP): (1) Both the older and present reports found this lack of financially motivated influence with both episodic and semantic memory stimuli, and (2) Semantic memory—evoked P300s are larger than Episodic memory-evoked P300s. In order to support the null findings on the incentive effect, which conflict with most findings on the SCR-based CIT (5), but are consistent with findings in the RT-based CIT (9, 11), it is essential (as emphasized by these authors regarding their manipulation check) to establish that the financial manipulation here produced credible behavioral effects.

We used the objective *test of memory malingering* [TOMM; (18)] to establish that: (1) both paid and unpaid groups malingered, and that (2) there were differences in malingering among groups. All three malingering groups here did indeed malinger, in that their correct response percentages were well-below the 82% cutoff for non-malingering behavior [and < 62% indicates malingering; (26)]. Furthermore, the 3 × 2 ANOVA on **Figure 2** showed that behavioral responses differed across groups as revealed in the interaction of response type and group. Moreover, the further *post-hoc* analysis of **Figure 2** yielded a significant interaction that was exactly the same as that found in Rosenfeld et al. (1) with *instructed* subjects. This established that the financial incentive did create a behavioral effect in the present paid group that differed from the effect in the present unpaid group. Furthermore, since the same interaction obtained with or without detailed instructions [used in (1)] on how to beat the test, those instructions are apparently unnecessary for the interaction to obtain.

The instructions used in Rosenfeld et al. (1) emphasized that in order to malinger effectively, (i.e., to imitate the performance of a truly head-injured person), a participant would have to score about 50% correct and 50% incorrect responses. Thus, one would need to take care not to make too many errors. We suggested in Rosenfeld et al. (1) that a paid instructed subject would be more motivated to attend to the instructions than an unpaid subject, and thus not make too many errors, which would explain why they had more correct than incorrect responses in contrast to their unpaid counterparts. However, in the present study, the specific instructions (to approach 50% accuracy) were omitted, yet the present *uninstructed* participants closely approximated the performance of the previous *instructed* participants. The present participants were simply told "Although you are, of course, normal and have NOT suffered memory loss, your goal during all today's tests is to play the role of a head injured individual who has suffered traumatic brain injury. In other words you are to try to look and act as though you have suffered memory loss due to brain damage from an accident." Apparently, more explicit instructions to approximate 50% accuracy rates were unnecessary to achieve rates near 50% accuracy, in that both the present paid and unpaid participants performed at near 50% accuracy (see **Figure 2**) i.e., between 44 and 49%, with the paid subjects showing a significant difference between correct/TRUE and incorrect/LIE responses (correct > incorrect), unlike their unpaid counterparts.

Indeed the lack of behavioral effect of specific 50% accuracy malingering instructions was further supported by the direct comparison of combined paid and unpaid instructed groups from Rosenfeld et al. (1) with the present paid and unpaid uninstructed groups: In that analysis, there were no effects in the TOMM scores, and all p's were >0.29 with all Bayes Factors supporting Null at values from 2.07 to 3.02. The lack of effect of specific 50% accuracy instructions on malingering performance was all the more interesting in view of the fact that with BD (although not NM) stimuli, the *instructions* do increase the P300 CIT effect with a medium effect size ($d = 0.505$). One explanation is that the instructions could have increased attention levels during the P300 CIT, which would lead to larger probe P300s (21). Thus, the TOMM seems to be a test of malingering, not attention, whereas, P300 is sensitive to attentional variables.

Why is it that SCR measures, but not RT-based nor P300-based CIT measures, are affected by financial incentive manipulations? As noted, the lacking effect of financial incentive could be attributed to not enough statistical power. Kleinberg and Verschuere (9) noted this possibility regarding their lacking effects of financial incentive on RT indices of the CIT effect. However, given that theirs was an internet study with many subjects, inadequate power seemed unlikely. ERP studies cannot be run at present on the internet, so we elected the *n*-values in the present study and in Rosenfeld et al. (1), based on power analysis. We supported our lack of effects with Bayes Factors (BFs) that allow statements about the likelihood ratios of null to alternative hypotheses. Given that these null effects of financial incentive on the P300 CIT effect are consistent with the results of Ellwanger et al. (12) using the 3-stimulus protocol, and of Rosenfeld et al. (1) using the complex trial protocol, we feel it reasonable to conclude that the financial incentives at levels utilized here do not appreciably influence P300-based indices of the malingering of cognitive deficit. However, effects of incentives of a magnitude used in field situations, cannot yet be ruled out. Again, these results, do not necessarily apply to the classical forensic CIT scenario.

Kleinberg and Verschuere (9) suggested that whereas, the ANS (SCR) CIT effect is more likely related to the Orienting Reflex (41), the RT CIT is instead more likely related to inhibitory processes and to response conflict. Likewise, the P300 CIT effect appears to be based on the simply cognitive phenomenon of recognizing rare, meaningful information (42). P300 amplitude is also associated with the amount of focused attention to stimuli (21). This suggests that since a financial incentive should increase attention [confirmed in the TOMM test here and in (1), with the finding of fewer error/deceptive trials in paid Ps], the incentive manipulation should also produce larger P300s to familiar stimuli. However, once attention is enough to assure recognition of probes within a memory type category, the resulting P300s consequently generated in a more all-or-none manner are no longer influenced by motivation. Apparently, in the present study as in Rosenfeld et al. (1), attention to stimuli was adequate to assure recognition, whose consequent P300s, were no longer modifiable by motivation.

Moreover, as noted above, paid Ps appeared more motivated to follow self-imposed instructions suggesting that the best way to convincingly appear head injured was to not miss *all* test items, but to try to balance honest and dishonest responses during the P300 test. However, if this was the case in the present paid Ps, they would be experiencing a greater workload during the CIT, tending typically to reduce P300 amplitudes and CIT responses—which we didn't observe here. There are thus many complexly organized psychological factors with many neural substrates interacting to yield the present effects, and it is clear that much more research will be required to fully account for the present lack of effects of financial incentive in the P300 CIT.

A critical remaining question is: Why do *uninstructed* malingerers behave as if they were instructed to approximate a 50% accuracy rate? Perhaps in the absence of specific instructions, the *default* response style is to not respond falsely on all trials. More likely, the present instructions could have inadvertently suggested or implied an accuracy rate closer to 50% than to 0%: Both BtNo and Bt $ groups were told, prior to the P300 CIT: "Your goal is to produce the disability in such a way that the examiner would not know you are faking or pretending." Prior to the TOMM, these same subjects were told, "your goal is to produce the symptoms of the disability, so we ask you to keep pretending that you are suffering memory loss and thus not able to recognize some of the pictures, and therefore to not press all the response buttons correctly." Such explicit instructions could easily have served to implicitly warn participants not to press all buttons *incorrectly* also. In contrast, as is next discussed, in the previous SCR-based forensic CITs, participants are directed to respond falsely to all probe-type trials.

We have noted here that the present head injury malingering scenario differs from the mock crime-forensic scenario. Perhaps this difference is the reason why financial incentive affects SCR-based forensic scenarios but not P300-based malingering scenarios. The previous SCR studies typically gave test-beating instructions emphasizing that Ps not respond to *any* crime-relevant probe stimuli, e.g., "You are about to take a polygraph test in which enhanced responses to the critical item would indicate guilt. Your task is to avoid being detected and if you beat the test and are classified as innocent, you will receive a cash reward of $10" [This was based on a review of the original submission of (1), by Gershon Ben Shakhar]. In contrast, as noted above, the incentivized participant in Rosenfeld et al. (1) was instructed to try to duplicate the behavior of actual head injured patients, who do *not* fail to respond to all critical probes, but to only about 50% of them. This is the typical strategy of instructed simulated malingerers in most of the numerous head injury malingering studies [see (43)], including our Ellwanger et al. (12) study, although as noted already, the specific malingering instructions were omitted in the present report. Nevertheless, the present participants behaved as if they were following such an instruction set, perhaps self-imposed. It therefore is not clear that results of this malingering strategy (of not making 100% errors) are strictly comparable to those strategies used earlier ("don't respond to *any* probes") to defeat a classical SCR-based CIT of the older ANS studies based on a mock theft scenario. Nevertheless, it is certainly clear from the present dataset and from Rosenfeld et al. (1) that financial incentive does not influence P300 in malingering performers. Moreover, we have now run a classical *mock theft scenario* using the CTP with participants motivated to beat the test, with one group paid and the other unpaid to beat the test (as in the present malingering study), and reported that (44) there was no effect of financial incentive on the P300 CIT in mock crime performance, just as with the present malingerers. Increasingly, the lack of effect of financial motivation on the P300 CIT effect becomes clearer.

As has been long argued [e.g., (45)], semantic information is stored more powerfully than incidentally acquired episodic information. The present results, along with the previous Rosenfeld et al. (1) results, are quite consistent with that notion. First, the probe-irrelevant differences and probe P300s are clearly larger with participant birthday stimuli (BD) than with experimenter name stimuli (NM). However, probe-irrelevant P300 differences with NM stimuli, however reduced, were seen here in contrast to RT effects, suggesting a greater sensitivity of P300 to weak memory traces, than of RT. Second, we did observe a significant effect of malingering instructions on BD-evoked but not NM-evoked P300s and RTs. Third, our bootstrap data showed expectedly higher detection rates for BD data than for NM data. Moreover, the effect of motivational and instructional incentives on BD detection rates were clearly different than on NM detection rates (**Figure 6**). This may be related to the greater variability seen for NM than BD bootstrap scores (**Table 2**), although, remarkably, not seen in the P300 data. It appears that participants are more uniformly detected with semantic than with episodic stimuli: Participants' detection scores cluster in a narrow range above 90% detection with semantic stimuli, but vary across a wide range with episodic stimuli. This implies that semantic stimuli are recognized on many more trials than are episodic stimuli.

It may seem surprising that financial incentive has no *incremental* effect after participants are instructed to defeat the test. This may be since our reward of $10 (US) for beating the test may be too inadequate to appeal to our mostly upper class undergraduates at a prominent private university. On the other hand, perhaps the intellectual challenge suggested by inviting participants to defeat the test may be more motivating than financial reward. This is an empirical question. Nevertheless, it is not unreasonable to conclude that the effect of financial reward is less in the P300-CIT [both forensic and malingering scenarios; (44, 46)] than in the autonomic CIT, since in the latter, similarly small rewards do in fact affect detection when added to instructions to beat the test (5, 47). This is important because it suggests that findings with student participants in university settings may well be applicable to field situations with higher stakes.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and the **Supplementary Files**.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Northwestern University Institutional Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Northwestern University Institutional Review Board.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00189/full#supplementary-material

REFERENCES

1. Rosenfeld JP, Labkovsky E, Davydova E, Ward A, Rosenfeld L. Financial incentive does not affect P300 (in response to certain episodic and semantic probe stimuli) in the Complex Trial Protocol (CTP) version of the Concealed Information Test (CIT) in detection of malingering. *Psychophysiology.* (2017) 54:764–72. doi: 10.1111/psyp.12835
2. Rosenfeld JP, Labkovsky E, Davydoya E, Ward AC. Financial incentive (motivation) has no effect on P300-based CTP performance. In: *28th Annual Meeting of American Psychological Society*, Chicago, IL (2016).
3. Lykken DT. The GSR in the detection of guilt. *J Appl Psychol.* (1959) 43:385. doi: 10.1037/h0046060
4. Verschuere B, Ben Shakhar G, Meijer E (eds). *Memory Detection: Theory and Application of the Concealed Information Test*. Cambridge: Cambridge University Press (2011). 63–89.
5. Meijer EH, Selle NK, Elber L, Ben-Shakhar G. Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology.* (2014) 51:879–904. doi: 10.1111/psyp.12239
6. Rosenfeld JP, Ben Shakhar G, Ganis G. Chapter 10 Physiologically based methods of concealed memory detection. In: Sinnott-Armstrong W, Schauer FD, Nadel L, editor. *Memory and Law*, Oxford: Oxford University Press (2012).
7. Rosenfeld JP. P300 in detecting concealed information. In: Verschuere B, Ben Shakhar G, Meijer E, editors. *Memory Detection: Theory and Application of the Concealed Information Test*. Cambridge: Cambridge University Press (2011). p. 63–89. doi: 10.1017/CBO9780511975196.005
8. Labkovsky E, Rosenfeld JP. A novel dual probe complex trial protocol for detection of concealed information: superiority of pictorial vs. verbal presentation. *Psychophysiology*. (2014) 51:1122–30. doi: 10.1111/psyp.12258
9. Kleinberg B, Verschuere B. The role of motivation to avoid detection in reaction time-based concealed information detection. *J Appl Res Memory Cogn.* (2016) 5:43–51. doi: 10.1016/j.jarmac.2015.11.004
10. Suchotzki K, Verschuere B, Crombez G, De Houwer J. Reaction time measures in deception research: comparing the effects of irrelevant and relevant stimulus–response compatibility. *Acta psychologica.* (2013) 144:224–31. doi: 10.1016/j.actpsy.2013.06.014
11. Suchotzki K, Verschuere B, Van Bockstaele B, Ben-Shakhar G, Crombez G. Lying takes time: a meta-analysis on reaction time measures of deception. *Psychol Bull.* (2017) 143:428–53. doi: 10.1037/bul0000087
12. Ellwanger J, Rosenfeld JP, Sweet JJ, Bhatt M. Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. International *J Psychophysiol.* (1996) 23:9–23. doi: 10.1016/0167-8760(96)00035-9
13. Rosenfeld JP, Ellwanger J, Sweet J. Detecting simulated amnesia with event-related brain potentials. *Int J Psychophys.* (1995) 19:1–11. doi: 10.1016/0167-8760(94)00057-L
14. van Hooff JC, Sargeant E, Foster JK, Schmand BA. Identifying deliberate attempts to fake memory impairment through the combined use of reaction time and event-related potential measures. *Int J Psychophys.* (2009) 73:246–56. doi: 10.1016/j.ijpsycho.2009.04.002
15. Rosenfeld JP, Ellwanger JW, Nolan K, Wu S, Bermann RG, Sweet J. P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *Int J Psychophysiol.* (1999) 33:3–19. doi: 10.1016/S0167-8760(99)00021-5
16. Rosenfeld JP, Sweet JJ, Chuang J, Ellwanger J, Song L. Detection of simulated malingering using forced choice recognition enhanced with event-related potential recording. *Clini Neuropsychol.* (1996) 10:163–79. doi: 10.1080/13854049608406678
17. Rosenfeld JP, Labkovsky E, Winograd M, Lui AM, Vandenboom C, et al. The Complex Trial Protocol (CTP): a new, countermeasure-resistant, accurate P300-based method for detection of concealed information. *Psychophysiology.* (2008) 45:906–19 doi: 10.1111/j.1469-8986.2008.00708.x
18. Tombaugh TN. *Test of Memory Malingering: TOMM*. New York, NY: MHS (1996).
19. Sweet JJ, Benson LM, Nelson NW, Moberg PJ. The American academy of clinical neuropsychology, national academy of neuropsychology, and society for clinical neuropsychology (APA Division 40) 2015 TCN professional practice and 'salary survey': professional practices, beliefs, and incomes of US neuropsychologists. *Clini Neuropsychol.* (2015) 29:1069–162. doi: 10.1080/13854046.2016.1140228
20. Martin PK, Schroeder RW, Odland AP. Neuropsychologists' validity testing beliefs and practices: a survey of North American professionals. *Clini Neuropsychol.* (2015) 29:741–76. doi: 10.1080/13854046.2015.1087597
21. Donchin E, Kramer A, Wickens C. Applications of brain event related potentials to problems in engineering psychology. In: Coles M, Porges S, Donchin E, editors. *Psychophysiology: Systems, Processes and Applications.* New York, NY: Guilford (1986). p. 702–10.
22. Polich J. Updating P300: an integrative theory of P3a and P3b. *Clini Neurophysiol.* (2007) 118:2128–48. doi: 10.1016/j.clinph.2007.04.019
23. Rees LM, Tombaugh TN, Gansler DA, Moczynski NP. Five validation experiments of the Test of Memory Malingering (TOMM). *Psychol Assess.* (1998) 10:10. doi: 10.1037/1040-3590.10.1.10
24. Weinborn M, Orr T, Woods SP, Conover E, Feix J. A validation of the test of memory malingering in a forensic psychiatric setting. *J Clini Exp Neuropsychol.* (2003) 25:979–90. doi: 10.1076/jcen.25.7.979.16481
25. Teichner G, Wagner MT. The Test of Memory Malingering (TOMM): normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Arch Clini Neuropsychol.* (2004) 19:455–64. doi: 10.1016/S0887-6177(03)00078-7
26. Hilsabeck RC, Gordon SN, Hietpas-Wilson T, Zartman AL. Use of trial 1 of the Test of Memory Malingering (TOMM) as a screening measure of effort: suggested discontinuation rules. *Clini Neuropsychol.* (2011) 25:1228–38. doi: 10.1080/13854046.2011.589409
27. Soskins M, Rosenfeld JP, Niendam T. The case for peak-to-peak measurement of P300 recorded at.3 hz high pass filter settings in detection ofdeception. *Int. J. Psychophysiol.* (2001) 40:173–80. doi: 10.1016/S0167-8760(00)00154-9
28. Rosenfeld JP, Ward A, Frigo V, Drapekin J, Labkovsky E. Evidence suggesting superiority of visual (verbal) vs. auditory test presentation modality in the P300-based, Complex Trial Protocol for concealed autobiographical memory detection. *Int J Psychophysiol.* (2015) 96:16–22. doi: 10.1016/j.ijpsycho.2015.02.026
29. Semlitsch HV, Anderer P, Schuster P, Presslich O. A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology.* (1986) 23:695–703. doi: 10.1111/j.1469-8986.1986.tb00696.x
30. Meijer EH, Smulders FT, Merckelbach HL, Wolf AG. The P300 is sensitive to concealed face recognition. *Int J Psychophysiol.* (2007) 66:231–7. doi: 10.1016/j.ijpsycho.2007.08.001
31. Keil A, Debener S, Gratton G, Junghöfer M, Kappenman ES, Luck SJ, et al. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*. (2014) 51:1–21. doi: 10.1111/psyp.12147
32. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Berlin: CRC press (1994).
33. Cohen J. *Statistical Power Analysis for the Behavioural Sciences.* New York, NY: Academic Press (1969).
34. Richardson JT. Eta squared and partial eta squared as measures of effect size in educational research. *Ed Res Rev.* (2011) 6:135–47. doi: 10.1016/j.edurev.2010.12.001
35. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev.* (2009) 16:225–37. doi: 10.3758/PBR.16.2.225
36. Jeffreys H. *The Theory of Probability*. Oxford: Oxford University Press (1961).
37. Schönbrodt FD, Wagenmakers EJ, Zehetleitner M, Perugini M. Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. *Psychol Methods*. (2017) 22:322. doi: 10.1037/met0000061
38. Kass RE, Raferty AE. Bayes factors. *J Am Stat Assoc.* (1995) 90:377–95. doi: 10.1080/01621459.1995.10476572
39. Seymour TL, Seifert CM, Shafto MG, Mosmann AL. Using response time measures to assess" guilty knowledge". *J Appl Psychol.* (2000) 85:30–7. doi: 10.1037/0021-9010.85.1.30

40. Rosenfeld JP, Hu X, Labkovsky E, Meixner J, Winograd M. Review of recent studies and issues regarding the P300-based, complex trial protocol for detection of concealed information. *Int J Psychophysiol.* (2013) 90:118–34. doi: 10.1016/j.ijpsycho.2013.08.012

41. klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Orienting versus inhibition in the concealed information test: different cognitive mechanisms drive different physiological measures. *Psychophysiology.* (2015) 53:579–90.doi: 10.1111/psyp.12583

42. Johnson R. The amplitude of the P300 component of the event-related potential: Review and synthesis. *Adv Psychophysiol.* (1988) 3:69–137.

43. Sweet JJ. Malingering. *Forensic Neuropsychology: Fundamentals and Practice.* Boca Raton, FL: CRC Press (1999). p. 255–73.

44. Rosenfeld JP, Sitar E, Wasserman JD, Ward AC. Moderate financial incentive does not appear to influence the P300 Concealed Information Test (CIT) effect in the Complex Trial Protocol (CTP) version of the CIT in a forensic scenario, while affecting P300 peak latencies and behavior. *Int J Psychophysiol*. (2018) 125:42–49. doi: 10.1016/j.ijpsycho.2018.02.006

45. Tulving E. Episodic and semantic memory. *Org Mem.* (1972) 1:381–403.

46. Rosenfeld JP, Ward C, Wasserman JD, Sitar E, Davydova L. Effects of motivational manipulations on the P300-based complex trial protocol for concealed information detection. In: Rosenfeld JP, editor. *Detecting Concealed Information and Deception: Recent Developments*. London: Academic Press/Elsevier (2018).

47. Elaad E, Ben-Shakhar G. Effects of motivation and verbal response type on psychophysiological detection of information. *Psychophysiology*. (1989) 26:442–51. doi: 10.1111/j.1469-8986.1989.tb01950.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fpsyt.2019.00389

Introducing Machine Learning to Detect Personality Faking-Good in a Male Sample: A New Model Based on Minnesota Multiphasic Personality Inventory-2 Restructured Form Scales and Reaction Times

*Cristina Mazza[1], Merylin Monaro[2], Graziella Orrù[3], Franco Burla[1], Marco Colasanti[1], Stefano Ferracuti[1] and Paolo Roma[1]**

[1] Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy, [2] Department of General Psychology, University of Padua, Padua, Italy, [3] Department of Surgical, Medical, Molecular & Critical Area Pathology, University of Pisa, Pisa, Italy

OPEN ACCESS

Edited by:
Cristina Scarpazza,
University of Padova, Italy

Reviewed by:
Marije E. Keulen-de Vos,
Forensic Psychiatric Center (FPC),
Netherlands
Ruth J. Tully,
Tully Forensic Psychology Ltd,
United Kingdom

****Correspondence:***
Paolo Roma
paolo.roma@uniroma1.it

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *13 March 2019*
Accepted: *16 May 2019*
Published: *06 June 2019*

Citation:
Mazza C, Monaro M, Orrù G, Burla F, Colasanti M, Ferracuti S and Roma P (2019) Introducing Machine Learning to Detect Personality Faking-Good in a Male Sample: A New Model Based on Minnesota Multiphasic Personality Inventory-2 Restructured Form Scales and Reaction Times. Front. Psychiatry 10:389. doi: 10.3389/fpsyt.2019.00389

Background and Purpose. The use of machine learning (ML) models in the detection of malingering has yielded encouraging results, showing promising accuracy levels. We investigated the possible application of this methodology when trained on behavioral features, such as response time (RT) and time pressure, to identify faking behavior in self-report personality questionnaires. To do so, we reintroduced the article of Roma et al. (2018), which highlighted that RTs and time pressure are useful variables in the detection of faking; we then extended the number of participants and applied an ML analysis.

Materials and Methods. The sample was composed of 175 subjects, of whom all were graduates (having completed at least 17 years of instruction), male, and Caucasian. Subjects were randomly assigned to four groups: honest speeded, faking-good speeded, honest unspeeded, and faking-good unspeeded. A software version of the Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF) was administered.

Results. Results indicated that ML algorithms reached very high accuracies (around 95%) in detecting malingerers when subjects are instructed to respond under time pressure. The classifiers' performance was lower when the subjects responded with no time restriction to the MMPI-2-RF items, with accuracies ranging from 75% to 85%. Further analysis demonstrated that *T*-scores of validity scales are ineffective to detect fakers when participants were not under temporal pressure (accuracies 55–65%), whereas temporal features resulted to be more useful (accuracies 70–75%). By contrast, temporal features and *T*-scores of validity scales are equally effective in detecting fakers when subjects are under time pressure (accuracies higher than 90%).

Discussion. To conclude, results demonstrated that ML techniques are extremely valuable and reach high performance in detecting fakers in self-report personality questionnaires over more the traditional psychometric techniques. Validity scales MMPI-2-RF manual criteria are very poor in identifying under-reported profiles. Moreover, temporal measures are useful tools in distinguishing honest from dishonest responders, especially in a no time pressure condition. Indeed, time pressure brings out malingerers in clearer way than does no time pressure condition.

Keywords: Minnesota Multiphasic Personality Inventory-2 Restructured Form, faking-good, machine learning, response latency, time pressure

INTRODUCTION

A crucial issue in medico-legal settings is the determination of whether a given symptom presentation or claimed cognitive impairment is credible, particularly when there is an external incentive, such as compensation or secondary gain (1). Recently, an increasing number of studies have addressed the phenomenon of malingering, which refers to an individual's deliberate choice to distort his/her mental or physical symptoms in order to achieve personal goals or advantages (2–4). More specifically, people can fake a clinical evaluation in two manners: faking-bad or faking-good. Faking-bad involves fabricating or exaggerating symptoms in an attempt to gain secondary benefits (e.g., financial compensation) (5). Faking-good, in contrast, involves presenting oneself in a more positive manner (6). Faking-good behaviors occur with alarming frequency in a variety of settings, from employee selection to forensic evaluation (7), making the prevention and identification of this phenomenon a field of great interest especially for practitioners and also for researchers. For instance, a candidate might lie about his/her personality during an employee selection process in order to secure a job that requires a particular profile. The problem of testing fit to work is crucial when a person is called to cover a position for which certain personality profiles are potentially dangerous; this could apply, for instance, to soldiers, police officers and intelligence staff, train drivers, and pilots. The prevalence of faking-good behaviors is unknown, but it likely exceeds malingering (8). Baer and Miller (9) estimated a dissimulation rate of 30% in job applicants; according to Donovan et al. (10), approximately 50% of applicants admit exaggerating qualities or characteristics of themselves, such as dependability or reliability, and over 60% of applicants de-emphasize their negative attributes. Again, the identification of faking-good subjects is critical in forensic settings, in which individuals can obtain some advantages by presenting themselves favorably (11). This is particularly true in forensic evaluations of parental skills (12) in the context of child custody hearings in which from 20% to as high as 74% of custody litigants (9) are prone to ménage a positive impression of themselves. A similar risk concerns psychological evaluations for obtaining gun or driving licenses. In a study involving offenders referred for impaired driving, Lapham et al. (13) found that about 30% of them lied about substance abuse. Thus, faking-good behavior is an important issue; however, to date, most studies have investigated faking-bad behavior [for a review, see Ref. (14)], and faking-good behavior remains underinvestigated (15).

The literature shows that faking is difficult to detect on the basis of a clinical interview only (16–18). For this reason, psychometric techniques have been proposed to evaluate systematic distortions concerning psychiatric symptoms. Validity scales of personality questionnaires are traditionally the main measure used to detect fakers by assessing the presence of responding biases (19) (i.e., the systematic tendency to answer items on a personality inventory in a manner inconsistent with accurate self-presentation). The validity scales of the Minnesota Multiphasic Personality Inventory (MMPI) (20, 21), also in its restructured form (MMPI-2-RF) (22), are the most suitable prototypes for this purpose. These scales allow researchers to measure response consistency, the presence of overexaggerating symptoms (23), and symptom minimization (underreporting) patterns (24). Specifically, the logic behind the underreporting Lie scale (L-r) and the Correction-Defensiveness scale (K-r) is that only people who want to provide a socially virtuous and well-adapted image of themselves will not answer genuinely those items that refer to common behaviors or small infractions that the majority of individuals are keen to admit to (e.g., "Sometimes I get angry"). However, validity scales are not always effective for detecting faking, as many items show high transparency; thus, test takers are often able to discern the constructs that the items are designed to measure and feign their answers towards the desired direction.

On the basis of this observation, many authors have searched for indirect measures of simulation. In 1972, for example, Dunn and colleagues (25) suggested that response times (RTs) to single items on a personality questionnaire could be used to distinguish malingerers from honest respondents, considering that the cognitive processes involved in lying are different from those involved when a person answers truthfully. As lying is a more complex mental operation than honesty, and because of the additional cognitive processing that is assumed to be involved with faking, simulators are exposed to a greater cognitive load than are truth tellers. Consequently, simulators are expected to obtain longer RTs than are controls and truth tellers (26, 27). A recent meta-analysis (6) on the relationship between RT and faking confirmed that honest respondents take less time to answer. The difference observed in RTs between faking and honest respondents is statistically significant only when test takers endorse items; it is not present when items are rejected. Similar evidence has been produced by researchers investigating the behavioral responses of honest and faking subjects using more complex measures, such as mouse tracking (28, 29) and keystroke dynamics (30, 31). Moreover, time pressure is a technique that has been shown to be effective in identifying malingering respondents (32). Research has shown that speeded tests, which impose time constraints by asking test takers to answer as quickly as possible, may increase accuracy in detecting fakers. In this context, Sutherland and Spilka (33) reported that time pressure accentuated a response style oriented to social desirability. Khorramdel and Kubinger (3) reported that the effect of time pressure on accentuating faking-good behavior is greater with a dichotomous response format. The rationale behind this phenomenon is that malingerers under time constraint pay less attention to the item selection and endorse more socially desirable items than they normally would, generating less believable profiles.

Roma et al. (34, in press) recently conducted a study of the faking-good personality profile, measuring RTs in a time pressure/no time pressure condition. In their experimental paradigm, participants were randomly assigned to one of four groups, each with different instructions based on the two manipulated factors (honest vs. faking-good; speeded vs. unspeeded). Interestingly, the authors found significant differences in terms of test fulfillment time and L-r/K-r scale completion time in both the time pressure and no time pressure conditions. The speeded condition increased *T*-scores in the L-r and K-r scales but decreased *T*-scores in some of the Restructured Clinical (RC) scales.

More recently, lie detection research machine learning (ML) models, which comprise "a category of algorithms that allow software applications to become more accurate in predicting outcomes without being explicitly programmed," have been used to distinguish between faking and honest respondents in many contexts, from the detection of false identities (35) to the detection of faked depression (5), with extremely promising accuracy levels. In the latter study, for instance, ML models were trained on behavioral features (e.g., number of symptoms of depression declared, mouse trajectory, and RTs) collected from depressed patients and malingerers; the resulting algorithms correctly identified malingerers with an accuracy approaching 96%. Indeed, ML has been demonstrated to outperform traditional statistical methods in terms of model complexity and classification accuracy in a wide variety of fields, including neuroimaging (36).

Here, we extend the results reported by Roma et al. (1) investigating whether the adoption of ML techniques may improve the detection of faking-good behavior, relative to traditional psychometric techniques.

MATERIALS AND METHODS

Participants and Research Design

Roma et al. (1) initially collected 140 young adult volunteers over a period of 2 months, from October to November 2017. These subjects participated in the study for a small reward (European breakfast in a café). To limit confounding variables, all subjects were aged 25 to 30 years (*M* = 26.64, *SD* = 1.88 years), healthy male (i.e., male without a diagnosed psychiatric disorder), Caucasian, and (non-psychology) graduates having completed at least 17 years of education. Participants were randomly assigned to one of four research groups, defined by a combination of the two manipulated factors relating to instruction (honest vs. faking-good) and time pressure (speeded vs. unspeeded): honest without time pressure (*n* = 33), faking-good without time pressure (*n* = 34), honest speeded (*n* = 35), and faking-good speeded (*n* = 33). In the unspeeded condition, participants were instructed to take all the time they needed to choose their answer, whereas in the speeded condition, participants were asked to answer as fast as they could, but no actual time limitation was imposed on them. Five subjects were excluded from data analysis for one or more of the following reasons: (a) failure to follow instructions as assessed by the final request (*n* = 2), (b) one or more changes in answers (*n* = 2), or (c) too brief a latency in one or more responses (*n* = 1, 3,000 ms). The final sample was composed of 135 subjects. No statistically significant differences were observed on age or level of education between groups.

Subsequently, from September to October 2018, we recruited an additional 45 young adult volunteers, whom we intended to use as an out-of-sample evaluation group for the models, built on the original sample collected by Roma et al. (1). All participants were rewarded with a breakfast ticket. Subjects were all tested in the morning and were randomly assigned to one of the four instruction groups listed above. Five subjects were excluded from the data analysis for one of the following reasons: (a) failure to follow instructions as assessed by the final request (*n* = 3) or (b) too brief a latency in one or more responses (*n* = 2, 3,000 ms). The remaining 40 persons were aged 23 to 32 years (*M* = 27.10, *SD* = 2.24 years), male, Caucasian, and (non-psychology) graduates. No statistically significant differences were observed on age or level of education. Overall, 175 young adult volunteers were recruited. Our samples were composed only of males both in an attempt to limit confounding variables and because researches indicated that men are more likely than women to use form of deception such as lying to obtain what they want (37) and to engage in harsher form of impression management (38, 39). Moreover, according to Volkema (40), women maintain higher ethical standards than do men. Such findings have been recently confirmed by Hogue et al. (41), which show that men have greater intentions than women to invent untrue personal information.

Experimental Procedure and Stimuli

The experimental procedure and stimuli were the same as those used and described in the research conducted by Roma et al. (34). In more detail, after filling in the demographic questionnaire and reading the instructions for the research task relative to their assigned group, participants responded to MMPI-2-RF items that were loaded onto the Microsoft Excel platform. Finally, their understanding of the instructions was checked. For more details on the materials and methods, please refer to Roma et al. (34). For each participant, 21 independent variables were collected. These independent variables included latencies (temporal features) and raw scores for each of the MMPI-2-RF scales (see **Table 1**).

Machine Learning Models: General Method

We performed two ML analyses: the first aimed at classifying participants under time pressure and the second aimed at classifying participants without time pressure. Analyses were run in WEKA 3.9 (42) following a best practice workflow: feature selection, model training, and then model testing in an

TABLE 1 | Features calculated for each participant.

	Feature	Description
Temporal performance	Total time (tt)	Time taken to compile the entire Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF) protocol
	1st part time (1t)	Time taken to compile the first part (items 1–112) of the MMPI-2-RF protocol
	2nd part time (2t)	Time taken to compile the second part (items 113–224) of the MMPI-2-RF protocol
	3rd part time (3t)	Time taken to compile the third part (items 225–338) of the MMPI-2-RF protocol
	L-r time (Lrt)	Time taken to respond to the L-r scale items
	K-r time (Krt)	Time taken to respond to the K-r scale items
	F-r time (Frt)	Time taken to respond to the F-r scale items
	Neutral time (Nt)	Time taken to respond to the 10 neutral questions
***T*-score**	L-r *T*-score (L-r)	*T*-score obtained in the L-r scale
	K-r *T*-score (K-r)	*T*-score obtained in the K-r scale
	F-r *T*-score (F-r)	*T*-score obtained in the F-r scale
	RCd *T*-score (RCd), RC1 *T*-score (RC1), RC2 *T*-score (RC2), RC3 *T*-score (RC3), RC4 *T*-score (RC4), RC6 *T*-score (RC6), RC7 *T*-score (RC7), RC8 *T*-score (RC8), RC9 *T*-score (RC9)	*T*-scores obtained in the RCd, RC1, RC2, RC3, RC4, RC6, RC7, RC8, and RC9 scales, respectively
	Total RC *T*-score (RCtot)	Sum of the *T*-scores obtained in all MMPI-2-RF scales

out-of-sample group (41). Following standard practice, given the high number of independent variables, the optimal subset was used in model building. Features selection is a widely used procedure in the construction of ML models (44), aimed at removing redundant and irrelevant features in order to increase the model generalization by reducing overfitting (45) and noise in the data. In this experiment, non-redundant features were extracted on the basis of their correlation with the outcome (faking vs. non-faking) and their mutual intercorrelation. In other words, we singled out the features that were more correlated with the predicting classification (faking vs. honest) and less correlated with one another. This procedure was performed using a correlation-based feature selector (CFS) (44), as implemented in WEKA 3.9 (42). The CFS algorithm, using a "greedy stepwise" search method, evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy with other predictors. Subsets of features that were highly correlated with the classification (the dependent variable) but with low intercorrelation were selected. For each selected predictor, we reported the point biserial correlation coefficient (r_{pb}), which related to the correlation with the outcome variable, and the correlation matrix with the other selected features.

The predictors resulting from the feature selection were fed as inputs to a number of ML models in order to evaluate the accuracy of the subjects' classification as faking or honest. Models were trained on the first sample of participants collected by Roma et al. (34), called the training set, following a 10-fold cross-validation procedure (46). K-fold cross-validation is a technique used to evaluate predictive models by repeatedly partitioning the original sample (e.g., 60 participants) into a training set to train the model, and a validation set to evaluate it. Specifically, in 10-fold cross-validation, the original sample is randomly partitioned into 10 equal-size sub-samples, or folds (e.g., 10 sub-samples of 6 participants each). Of the 10 sub-samples, a single sub-sample is retained as validation data to test the model, and the remaining 9 sub-samples are used as training data. The process is repeated 10 times, with each of the 10 folds used exactly once as validation data. The results from the 10 folds are then averaged to produce a single estimation of prediction accuracy.

In order to evaluate the model's capacity for generalization, it was tested on completely new data to reduce bias (47). Because classifiers are built to fit the data, it is important to know how an existing model fits unseen data. For this reason, we collected a new group of participants to evaluate the real performance of the classifiers. Data were collected by a different experimenter, and subjects were randomly assigned to the experimental conditions, in order to eliminate *a priori* knowledge of how the classifiers work during the test collection. The sample size of the test group was 40 subjects, corresponding to approximately 30% of the training sample—a percentage that is usually regarded as satisfactory (48). For each model, we reported accuracy, recall (sensitivity or true positive rate), and precision.

As stated above, we evaluated the accuracy of different ML classifiers in order to investigate whether the results were stable across classifiers and independent of specific model assumptions. In fact, the algorithms that we chose were representative of different underlying classification strategies, as follows:

- Logistic regression: measures the relationship between the categorical dependent variable and the independent variables by estimating probabilities using a logistic function (49).
- Support vector machine (SVM): a binary linear classifier that maps the space and divides the examples of separate categories by as large a margin as possible (50, 51).
- Naive Bayes: a probabilistic classifier based on Bayes' theorem, which assumes independence between features (52).
- Random forest: an ensemble learning method that operates by constructing a multitude of decision trees and combining their results (53).
- Logistic model tree (LMT): combines logistic regression and decision tree learning (54).

ML models, such as some of those reported above, are difficult to interpret. Often, the mechanics that yield the algorithm to identify a single participant as honest or faking-good is unclear.

For this reason, ML models are sometimes analyzed on the basis of decision rules such as a tree model called J48 (55). This is one of the simplest—if not the simplest—classifier in terms of the transparency of operations, and it highlights the classification logic (albeit not in the most efficient way) (56). In our research, it was helpful to use this method to explain the operations performed by the algorithm.

All algorithms were run using default parameters set by WEKA 3.9 (41). Therefore, there was no fine-tuning of the parameters to increase classification accuracy.

RESULTS

No Time Pressure Models

Sixty-seven participants (33 honest and 34 faking) from Roma et al. (34) were used to train the models, whereas the 40 new participants (10 honest and 10 faking) collected for this study were used to test the model. All participants completed the MMPI-2-RF without time pressure.

The feature selection, which was run as described above, identified the following predictors: first part time (1t), K-r time (Krt), RC4 *T*-score (RC4), and RC9 *T*-score (RC9). **Table 2** reports the correlation matrix between each selected feature and the outcome variable (faking vs. non-faking). The time taken by the subject to complete the first part of the MMPI-2-RF turned out to be the feature that best distinguished the two groups, as faking-good respondents were, on average, slower than honest respondents in responding to the first 112 MMPI-2-RF items (faking *M* = 11.59 min, *SD* = 1.28; honest *M* = 7.46 min, *SD* = 0.99; see **Figure 1**). Moreover, it is worth noting that the MMPI-2-RF validity scales (L-r, F-r, and K-r) did not contribute to the identification of faking behavior.

The results obtained by different ML algorithms in the training set and the test set are reported in **Table 3**. It is noticeable that all classifiers reached a very high accuracy (97–100%) in the training set. However, the accuracy in the test set dropped to 75%, with the logistic classifier outperforming other classifiers (logistic accuracy = 85%). These results indicate that out-of-sample accuracy was degraded, despite the errors being equally distributed amongst faking-good and faking-bad behavior. In **Figure 2**, the output of a J48 tree (used to facilitate understanding of a classification strategy) is reported. The algorithm achieved an accuracy of 95.9% (recall = 0.956, precision = 0.956) in the training set and 75% in the test set (recall = 0.750, precision = 0.753). It should be noted that J48 bases its outcome exclusively on the time spent by each participant in completing the first part of the MMPI-2-RF.

Time Pressure Models

Sixty-eight participants (35 honest and 33 faking) from Roma et al. (34) were used to train the models, whereas 20 new participants (10 honest and 10 faking) were used for out-of-sample testing. All participants performed the MMPI-2-RF under time pressure.

The CFS feature selector identified the following predictors: first part time (1t), third part time (3t), total time (tt), L-r time (Lrt), K-r time (Krt), L-r *T*-score (L-r), F-r *T*-score (F-r), and RC4 *T*-score (RC4). **Table 4** reports the correlation matrix between each feature and the dependent variable (faking vs. honest). Also, in this case, the time used to complete the first part of the MMPI-2-RF protocol was the variable that best discriminated between the two samples (faking vs. honest), with faking-good respondents taking longer than honest respondents (faking *M* = 8.09 min, *SD* = 1.25; honest *M* = 5.69 min, *SD* = 1.06).

Table 3 reports the results obtained by different ML algorithms in the 10-fold cross-validation and the test set. All ML models reached 95–100% accuracy in the training set, and similar results were achieved in the test set (95% for all classifiers). In this case, the trained classifiers showed good generalization when tested on a completely new sample. Errors were equally distributed across the two classes, with a similar rate of faking-good and faking-bad behavior.

Finally, **Figure 3** describes the output of the J48 algorithm. To classify subjects as honest or faking, the classification rule considers the time used to respond to L-r scale items, followed by K-r scale items. The algorithm achieved an accuracy of 92.53% (recall = 0.925, precision = 0.926) in the training set, which remained stable in the test set (accuracy = 90%, recall = 0.900,

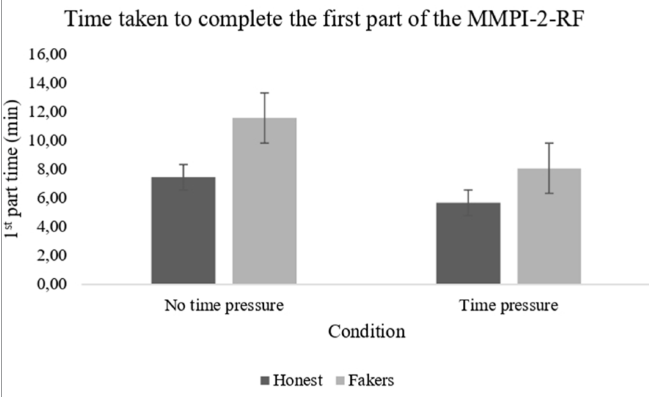


FIGURE 1 | The bar plots represent the time taken by participants in different experimental conditions to complete the first part of the MMPI-2-RF protocol.

TABLE 2 | The table reports the correlation matrix for the four features selected by the CFS algorithm in the group of participants under time pressure. The point biserial correlation (r_{pb}) between each selected feature and the dependent variable (faking vs. honest) is also reported.

	1t	Krt	RC4	RC9	Faking vs. honest
1t	1.00	0.31	–0.27	–0.15	0.88
Krt	0.31	1.00	–0.28	–0.42	0.42
RC4	–0.27	–0.28	1.00	0.02	–0.36
RC9	–0.15	–0.42	0.02	1.00	–0.31
Faking vs. honest	0.88	0.42	–0.36	–0.31	1.00

TABLE 3 | The table reports the accuracy, recall, and precision measures for each ML model. Results are reported for the 10-fold cross-validation (training) set and the test set, for both the time pressure and no time pressure groups.

	Training set (10-fold cross-validation)			Test set		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
No time pressure models						
Logistic	100%	1.00	1.00	85%	0.85	0.854
SVM	98.53%	0.985	0.986	75%	0.750	0.753
Naive Bayes	100%	1.00	1.00	75%	0.750	0.753
Random forest	98.53%	0.985	0.986	75%	0.750	0.753
LMT	97.06%	0.971	0.972	75%	0.750	0.775
Time pressure models						
Logistic	98.51%	1.00	0.986	95%	0.95	0.955
SVM	98.51%	0.985	0.986	95%	0.95	0.955
Naive Bayes	100%	1.00	1.00	95%	0.95	0.955
Random forest	97.01%	0.970	0.970	95%	0.95	0.955
LMT	95.52%	0.955	0.959	95%	0.95	0.955

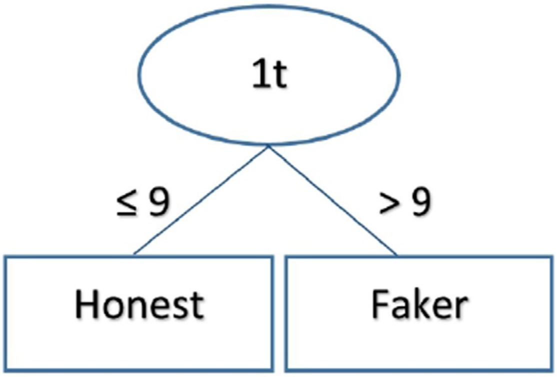


FIGURE 2 | The figure reports the rules that the J48 decision tree used to classify participants as faking-good or honest in the no time pressure sample. According to this algorithm, subjects who took fewer than 9 minutes to complete the first part of the questionnaire were classified as honest, whereas subjects who took more than 9 minutes were classified as faking.

precision = 0.917). Again, temporal features were sufficient to identify faking responders.

Models Based Only on the MMPI-2-RF Validity Scales

As reported above, in the no time pressure condition, the MMPI-2-RF validity scales were not included as predictors. This means that the time required to respond to the questionnaire may be sufficient to detect faking-good respondents with a level of accuracy that is greater than chance (75%). The same was shown by the J48 model built on time pressure data, which reached an accuracy of 90% based only on temporal predictors (Lrt and Krt).

In order to address the question "How useful are temporal features compared to traditional validity scales in the identification of faking-good respondents?," (i) we applied the MMPI-2-RF suggested criteria (see 22) to identify tests showing clear underreporting, and then (ii) we ran new ML models using only the *T*-scores of the validity scales (L-r, F-r, and K-r) as input. Similarly, we ran classifiers using only temporal features (tt, 1t, 2t, 3t, Lrt, Frt, Krt, and Nt) as input.

According to the MMPI-2-RF manual (22), a linear *T*-point score ≥ 80 in the L-r scale is a valid and reliable cut-off for identifying underreporting, as well as a *T*-score ≥ 70 in the K-r scale. Based on these suggested cut-offs, in the original sample of Roma et al. (34), only 12 out of 135 MMPI-2-RF protocols were surely invalid due to underreporting, generating an accuracy in detecting faking-good respondents of only 8.8%. Applying the same criteria to the 40 subjects in the validation study, we did not identify any invalid MMPI-2-RF protocols due to underreporting; that is, we did not detect any faking respondents, in either the time pressure or the no time pressure condition.

Results from new ML models using only *T*-scores of the validity scales or temporal features as input are reported in **Table 5**, for both time pressure and no time pressure conditions. With respect to the no time pressure condition, the *T*-scores of the validity scales were very poor in detecting faking-good behavior. Indeed, model accuracies ranged from 55% to 65%, just above chance. Considering only temporal features, model performance improved slightly (10–15%), reaching an accuracy of 70–75%. In regard to the temporal pressure condition, both validity scale scores and temporal features were good predictors of faking behavior when time pressure instructions were given. In this scenario, all models achieved greater than 90% accuracy.

DISCUSSION

Most cognitive and behavioral symptoms can be easily faked, even by naive, non-coached examinees; for this reason, psychometric tools are needed to objectively confirm whether test scores accurately reflect dysfunctions or whether respondents have attempted to simulate or overexaggerate difficulties (57). While malingering is a widely studied topic, there is a lack of research on methods and strategies to detect faking-good behavior (34, 58, 59). Most investigations have focused on techniques to spot faking-bad, rather

TABLE 4 | The table reports the correlation matrix for the eight features selected by the CFS algorithm in the no time pressure group. The point biserial correlation (r_{pb}) between each selected feature and the dependent variable (faking vs. honest) is also reported.

	1t	3t	tt	Lrt	Krt	L-r	F-r	RC4	Faking vs. honest
1t	1.00	0.05	0.72	0.67	0.68	0.67	−0.17	−0.27	0.72
3t	0.05	1.00	0.65	0.35	0.36	0.32	−0.14	−0.22	0.37
tt	0.72	0.65	1.00	0.74	0.77	0.73	−0.16	−0.33	0.83
Lrt	0.67	0.35	0.74	1.00	0.86	0.75	−0.15	−0.29	0.84
Krt	0.68	0.36	0.77	0.86	1.00	0.81	−0.22	−0.37	0.88
L-r	0.67	0.32	0.73	0.75	0.81	1.00	0.03	−0.33	0.83
F-r	−0.17	−0.14	−0.16	−0.15	−0.22	0.03	1.00	0.08	−0.28
RC4	−0.27	−0.22	−0.33	−0.29	−0.37	−0.33	0.08	1.00	−0.37
Faking vs. honest	0.72	0.37	0.83	0.84	0.88	0.83	−0.28	−0.37	1.00

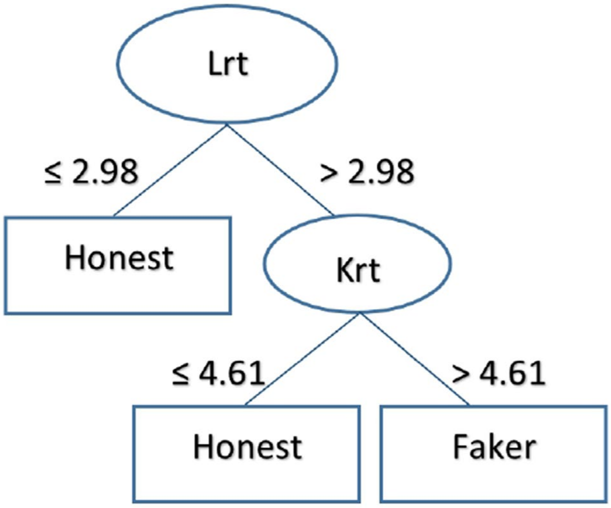


FIGURE 3 | The figure represents the classification logic of the J48 decision tree for the group under temporal pressure. According to the tree, subject who took fewer than 2.98 minutes to fill in the items of the L-r scale were classified as honest; subjects who took more than 2.98 minutes were classified as faking. Subjects who took fewer than 4.61 minutes to complete the K-r scale items were classified as honest; subjects who took more than 4.61 minutes were classified as faking.

than faking-good, behavior. However, in many legal conditions (e.g., child custody hearings), examinees are prone to faking-good. Recent advances in psychometric tools have indicated that ML techniques may boost classification accuracy, relative to standard statistical techniques. Accordingly, the goal of this research was to apply ML analysis in the identification of faking-good MMPI-2-RF test takers.

The results showed that ML algorithms achieved very high accuracy in detecting fakers when subjects were instructed to respond under time pressure (in fact, in the out-of-sample test set, all trained models showed an accuracy of 95%). However, the performance of classifiers was lower when subjects responded without time restriction to the MMPI-2-RF items, with accuracies ranging from 75% to 85% in the test set.

To demonstrate whether ML analysis can detect fakers more accurately than traditional validity scales, we detected invalid protocols for underreporting following the MMPI-2-RF suggested criteria. Using these criteria on the very same set of participants that we used to compute the algorithms accuracy resulted in no identification.

Moreover, to investigate whether validity scales are useful for the detection of faking behavior, we ran two sets of ML models: one using only the *T*-scores of the validity scales (L-r, F-r, and K-r) as features and the other using only temporal features (tt, 1t, 2t, 3t, Lrt, Frt, Krt, and Nt). The results showed that the *T*-scores of the validity scales were ineffective for detecting fakers when participants were not under time pressure (achieving only 55–65% accuracy), whereas temporal features were more useful (achieving 70–75% accuracy). By contrast, temporal features and the *T*-scores of the validity scales were equally effective in detecting faking behavior when subjects were under time pressure (achieving accuracies > 90%). Results indicate that time pressure increase faking-good respondents' descriptions of socially desirable behavior. This result is consistent with previous literature (3, 34, 60, 61) that show that time pressure prevents subjects to think deeply about the content of the questions and the possible lack of credibility of their responses. In other words, time limitations urge people to focus on responding faster, and this accentuates their fake behavior and prevents them from taking the time to consider whether their responses are exaggeratedly good, thus breaking the warning instruction ("your deception should not be detected").

To conclude, the results suggest that time—in the form of both RTs and time pressure—is a critical factor in the detection of faking behavior. Moreover, the use of ML is extremely valuable and offers the following advantages: first, it detects faking-good respondents on the MMPI-2-RF with significantly higher accuracy than do the validity scales criteria published in the manual; second, it works automatically, so it is more objective than human evaluation; third, it considers a variety of parameters, making it nearly impossible for fakers to successfully cheat; and finally, its predictions can be applied to completely new subjects, strengthening the replicability of the results. It can therefore be concluded that i) the MMPI-2-RF manual criteria with respect to the validity scales are very poor in identifying underreporting and ii) temporal measures are useful for distinguishing between honest and faking respondents, especially in a no time pressure condition.

Widely, our results found that time pressure revealed fakers more clearly than did a no time pressure condition. The ML models in the former condition were also more generalizable.

TABLE 5 | The table reports the results of the ML models using only *T*-scores of the validity scales (L-r, F-r, and K-r) as input. Results for the ML models using only temporal features (tt, 1t, 2t, 3t, Lrt, Frt, Krt, and Nt) as input are also reported. Results refer to accuracy, recall, and precision.

	Models based only on T-scores of the validity scales			Models based only on temporal features		
	Accuracy	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**
No time pressure models						
Logistic	60%	0.600	0.600	75%	0.750	0.753
SVM	60%	0.600	0.604	70%	0.700	0.738
Naive Bayes	55%	0.550	0.551	75%	0.750	0.753
Random forest	65%	0.650	0.700	70%	0.700	0.708
LMT	55%	0.550	0.551	75%	0.750	0.775
J48	55%	0.550	0.551	75%	0.750	0.753
Time pressure models						
Logistic	95%	0.95	0.955	90%	0.900	0.900
SVM	95%	0.95	0.955	95%	0.950	0.955
Naive Bayes	90%	0.900	0.900	95%	0.950	0.955
Random forest	95%	0.95	0.955	95%	0.950	0.955
LMT	95%	0.95	0.955	95%	0.950	0.955
J48	90%	0.900	0.917	90%	0.900	0.917

It is reasonable, therefore, to conclude that time pressure, which forces subjects to respond to a self-report questionnaire as quickly as possible, can effectively facilitate the detection of simulators. When it is not possible to instruct participants to respond with maximum speed, as is usually the case in forensic settings, the validity scales of the MMPI-2-RF are insufficient to accurately detect fakers; therefore, it is important to record RTs. To summarize, time pressure is the most reliable method to identify faking-good behavior. However, in the absence of time pressure, RTs are a more accurate measure than validity scales.

Despite that faking-good remains underinvestigated (15), it is a widespread behavior that commonly occurs in all that settings in which individuals are prone to ménage a positive impression of themselves. In employee selection, for instance, 30% of the candidates tend to provide an improved and socially adapted self-image in order to gain a job position. In forensic setting, furthermore, from 20% to as high as 74% of child custody litigants tend to deny or omit negative features of their personality in order to present themselves in a better light, to show more adaptive psychological and behavioral functioning, and to appear as responsible caregivers who will provide for the best interests of their child. A similar risk concerns psychological evaluations for obtaining gun or driving licenses. The present study adds useful insight to the debate over the methods that can be effectively used to detect faking-good behaviors. Based on findings described herein, personality assessment in personnel and forensic contexts could be improved, for example, by introducing time pressure asking subjects to fulfill self-report questionnaires (e.g., MMPI-2-RF) as soon as possible or again, using software that could record the reaction times to test item. To the best of our knowledge, this study was the first to have applied ML to bring out good-fakers.

STRENGTHS AND LIMITATIONS

The present study meant to overcome one of the limitations of the previous research conducted by Roma et al. (34) by expanding the sample size. At the same time, it also provides insight into the use of ML models for the detection of faking behavior. The main limitation of the study, however, is that the sample was selected for specificity (graduate males aged 23–32 years), and this reduces the generalizability of the findings. One important future direction would be to test the accuracy of the ML algorithms developed in this study on the forensic population. Future research could also analyze whether limiting the time available to fulfill a self-report personality questionnaire (rather than simply imposing time pressure) could lead to the same results, as such an approach could more easily be employed in forensic settings and personnel selection.

DATA AVAILABILITY STATEMENT

The dataset used and analyzed in the current study is available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study was carried out with written informed consent by all subjects, in accordance with the Declaration of Helsinki. It was approved by the local ethics committee (Board of the Department of Human Neuroscience, Faculty of Medicine and Dentistry, Sapienza University of Rome).

AUTHOR CONTRIBUTIONS

PR, SF and FB conceived the experiment. Data acquisition was done by PR, CM and MC. Data analysis was done by MM and CM. Data interpretation was done by MM, CM, PR, and GO. Drafting of the manuscript was done by MM, CM, PR, and GO. All authors revised the manuscript critically and gave final approval of the version to be published.

REFERENCES

1. Bush SS, Heilbronner RL, Ruff RM. Psychological assessment of symptom and performance validity, response bias, and malingering: official position of the Association for Scientific Advancement in Psychological Injury and Law. *Psychol Inj Law* (2014) 7(3):197–205. doi: 10.1007/s12207-014-9198-7
2. Karner T. The volunteer effect of answering personality questionnaires. *Psychol Beitr*(2002) 44(1):42–9.
3. Khorramdel L, Kubinger KD. The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychol Sci* (2006) 48(3):378–97.
4. Ziegler M, MacCann C, Roberts RD. *New perspectives on faking in personality assessment*. New York, NY, US: Oxford University Press (2012). doi: 10.1093/acprof:oso/9780195387476.001.0001
5. Monaro M, Toncini A, Ferracuti S, Tessari G, Vaccaro MG, De Fazio P, et al. The detection of malingering: a new tool to identify made-up depression. *Front Psychiatry* (2018c) 9:249. doi: 10.3389/fpsyt.2018.00249
6. Maricuțoiu LP, Sârbescu P. The relationship between faking and response latencies. *Eur J Psychol Assess* (2016) 35:1, 3–13. doi: 10.1027/1015-5759/a000361
7. Andrews P, Meyer RG. Marlowe–Crowne social desirability scale and short form C: forensic norms. *J Clin Psychol* (2003) 59:4:483–92. doi: 10.1002/jclp.10136
8. Rogers R. Detection strategies for malingering and defensiveness. In: Rogers R, Bender SD, editors. *Clinical assessment of malingering*, 4th ed. New York: Guilford Publications (2018). p. 31–2.
9. Baer RA, Miller J. Underreporting of psychopathology on the MMPI-2: a meta-analytic review. *Psychol Assess* (2002) 14:16–26. doi: 10.1037/1040-3590.14.1.16
10. Donovan JJ, Dwight SA, Hurtz GM. An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *HumPerform* (2003) 16(1):81–106. doi: 10.1207/S15327043HUP1601_4
11. Giacchetti N, Roma P, Pancheri C, Williams R, Meuti V, Aceti F. Personality traits in a sample of Italian filicide mothers. *Riv Psichiatr* (2019) 54(2):67–74. doi: 10.1708/3142.31247
12. Roma P, Ricci F, Kotzalidis GD, Abbate L, Lavadera AL, Versace G, et al. MMPI-2 in child custody litigation: a comparison between genders. *Eur J Psychol Assess* (2014) 30(2):110–6. doi: 10.1027/1015-5759/a000192
13. Lapham SC, C'de Baca J, Hunt WC, Berger RL. Are drunk-driving offenders referred for screening accurately reporting their drug use? *Drug Alcohol Depend* (2001) 66(3):243–53. doi: 10.1016/S0376-8716(02)00004-2
14. Sartori G, Orrù G, Zangrossi A. Detection of malingering in personal injury and damage ascertainment. In: Ferrara SD, Boscolo-Berto R, Viel G, editors. *Personal injury and damage ascertainment under civil law*. Switzerland: Springer-Cham (2016). doi: 10.1007/978-3-319-29812-2_29
15. Crighton AH, Marek RJ, Dragon WR, Ben-Porath YS. Utility of the MMPI-2-RF validity scales in detection of simulated underreporting: implications of incorporating a manipulation check. *Assessment* (2017) 24(7):853–64. doi: 10.1177/1073191115627011
16. Rosen J, Mulsant BH, Bruce ML, Mittal V, Fox D. Actors' portrayals of depression to test interrater reliability in clinical trials. *Am J Psychiatry* (2004) 161:1909–11. doi: 10.1176/ajp.161.10.1909
17. Rosenhan D. On being sane in insane places. *Science* (1973) 179:250–8. doi: 10.1126/science.179.4070.250
18. Roma P, Piccinni E, Ferracuti S. Using MMPI-2 in forensic assessment. *Rass Ital Criminol* (2016) 10(2):116–22.
19. Paulhus DL. Socially desirable responding: the evolution of a construct. In: Braun HI, Jackson DN, Wiley DE, editors. *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Erlbaum (2002). p. 49–69.
20. Hathaway SR, McKinley JC. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *J Psychol* (1940) 10(2):249–54. doi: 10.1080/00223980.1940.9917000
21. Hathaway SR, McKinley JC. *The Minnesota Multiphasic Personality Inventory*, Rev. ed., 2nd printing. Minneapolis, Minnesota: University of Minnesota Press (1943).
22. Ben-Porath YS, Tellegen A. *Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF)*. Minneapolis, Minnesota: University of Minnesota Press (2008). doi: 10.1037/t15121-000
23. Sellbom M, Bagby RM. Detection of overreported psychopathology with the MMPI-2-RF form validity scales. *Psychol Assess* (2010) 22(4):757–67. doi: 10.1037/a0020825
24. Jimenez-Gomez F, Sanchez-Crespo G, Ampudia-Rueda A. Is there a social desirability scale in the MMPI-2-RF? *Clin Salud* (2013) 24(3):161–8. doi: 10.1016/S1130-5274(13)70017-3
25. Dunn TG, Lushene RE, O'Neil HF. Complete automation of the MMPI and a study of its response latencies. *J Consult Clin Psychol* (1972) 39(3):381–7. doi: 10.1037/h0033855
26. Walczyk JJ, Schwartz JP, Clifton R, Barett A, Wei M, Zha P. Lying person to person about life events: a cognitive framework for lie detection. *Pers Psychol* (2005) 58(1):141–70. doi: 10.1111/j.1744-6570.2005.00484.x
27. Foerster A, Pfister R, Schmidts C, Dignath D, Kunde W. Honesty saves time (and justifications). *Front Psychol* (2013) 4:473. doi: 10.3389/fpsyg.2013.00473
28. Monaro M, Gamberini L, Sartori G. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS One* (2017b) 12(5):e0177851. doi: 10.1371/journal.pone.0177851
29. Monaro M, Fugazza FI, Gamberini L, Sartori G. How human–mouse interaction can accurately detect faked responses about identity. In: Gamberini L, Spagnolli A, Jacucci G, Blankertz B, Freeman J, editors. *Symbiotic Interaction. Symbiotic 2016. Lecture Notes in Computer Science*, vol. 9961, Cham: Springer (2017a). p. 115–24. doi: 10.1007/978-3-319-57753-1
30. Monaro M, Galante C, Spolaor R, Li Q. Covert lie detection using keyboard dynamics. *Sci Rep* (2018a) 8:1976. doi: 10.1038/s41598-018-20462-6
31. Monaro M, Businaro M, Spolaor R, Li QQ, Conti M, Gamberini Let al. The online identity detection *via* keyboard dynamics. In: Arai K, Bhatia R, Kapoor S, editors. *Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in Intelligent Systems and Computing 881*, vol. 2, Basingstoke, United Kingdom: Springer Nature (2019). p. 342–57. doi: 10.1007/978-3-030-02683-7_24
32. Degner J. On the (un-)controllability of affective priming: strategic manipulation is feasible but can possibly be prevented. *Cogn Emot* (2009) 23(2):327–54. doi: 10.1080/02699930801993924
33. Sutherland BV, Spilka B. Social desirability, item-response time, and item significance. *J Consult Psychol* (1964) 28(5):447–51. doi: 10.1037/h0047898
34. Roma P, Verrocchio MC, Mazza C, Marchetti D, Burla F, Cinti ME, et al. Could time detect a faking-good attitude? A study with the MMPI-2-RF. *Front Psychol* (2018) 9:1064. doi: 10.3389/fpsyg.2018.01064
35. Monaro M, Gamberini L, Zecchinato F, Sartori G. False identity detection using complex sentences. *Front Psychol* (2018b) 9:283. doi: 10.3389/fpsyg.2018.00283
36. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev* (2012) 36(4):1140–52. doi: 10.1016/j.neubiorev.2012.01.004
37. Dreber A, Johannesson M. Gender differences in deception. *Econ Lett* (2008) 99(1):197–9. doi: 10.1016/j.econlet.2007.06.027
38. Turnley WH, Bolino MC. Achieving desired images while avoiding undesired images: exploring the role of self-monitoring in impression management. *J Appl Psychol* (2001) 86(2):351. doi: 10.1037/0021-9010.86.2.351
39. Guadagno RE, Cialdini RB. Gender differences in impression management in organizations: a qualitative review. Sex Roles. *J Res* (2007) 56(7–8):483–94. doi: 10.1007/s11199-007-9187-3
40. Volkema RJ. Demographic, cultural, and economic predictors of perceived ethicality of negotiation behavior: a nine-country analysis. *J Bus Res* (2004) 57(1):69–78. doi: 10.1016/S0148-2963(02)00286-2
41. Hogue M, Levashina J, Hang H. "Will I fake it? The interplay of gender, Machiavellianism, and self-monitoring strategies for honesty in job interviews". *J Bus Ethics* (2013) 117(2):399–411. doi: 10.1007/s10551-012-1525-x
42. Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor* (2009) 11(1):10–8. doi: 10.1145/1656274.1656278
43. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*. Berlin, Germany: Springer-Verlag (2009). doi: 10.1007/978-0-387-84858-7
44. Hall MA. *Correlation-based feature selection for machine learning. Dissertation Thesis*. Hamilton, New Zealand University of Waikato (1999).

45. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C. Application of high dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep* (2015) 5(10312):1–12. doi: 10.1038/srep10312
46. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Kaufmann M, editor. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. California, USA: Stanford University Press, Palo Alto (1995). p. 1137–43.
47. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. The reusable holdout: preserving validity in adaptive data analysis. *Science* (2015) 349(6248):3–6. doi: 10.1126/science.aaa9375
48. Nelles O. *Nonlinear system identification from classical approaches to neural networks and fuzzy models*. Berlin, Germany: Springer-Verlag (2001). doi: 10.1007/978-3-662-04323-3_17
49. le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. *Appl Stat* (1992) 41(1):191–201. doi: 10.2307/2347628
50. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput* (2001) 13(3):637–49. doi: 10.1162/089976601300014493
51. Platt JC. Fast training of support vector machines using sequential minimal optimization. In: Burges CJC, Schölkopf B, Smola AJ, editors. *Advances in kernel methods*. Cambridge: MIT Press (1999).
52. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. San Francisco, California, USA: Morgan Kaufmann Publishers Inc. (1995) p. 338–45.
53. Breiman L. Random forest. *Mach Learn* 45(1):5–32 (2001). doi: 10.1023/A:1010933404324
54. Landwehr N, Hall M, Frank E. Logistic model trees. *Mach Learn* (2005) 59(1–2):161–205. doi: 10.1007/s10994-005-0466-3
55. Quinlan JS. *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers (1993).
56. Mitchell T. "Decision tree learning" In: Mitchell T, editor. *Machine Learning*. New York, USA: McGraw Hill (1997).
57. Sartori G, Zangrossi A, Orrù G, Monaro M. In: Ferrera S *Detection of malingering in psychic damage ascertainment, in P5 medicine and justice*. Switzerland: Springer-Cham Springer (2017), 330–41. doi: 10.1007/978-3-319-67092-8_21
58. Roma P, Mazza C, Mammarella S, Mantovani B, Mandarelli G, Ferracuti S. Faking-good behavior in self-favorable scales of the MMPI-2. *Eur J Psychol Assess* (2019a) 1–9. doi: 10.1027/1015-5759/a000511
59. Roma P, Mazza C, Ferracuti G, Cinti ME, Ferracuti S, Burla F. Drinking and driving relapse: data from BAC and MMPI-2. *PLoS One* (2019b) 14(1):e0209116. doi: 10.1371/journal.pone.0209116
60. Holden RR, Wood LL, Tomashewski L. Do response time limitations counteract the effect of faking on personality inventory validity. *J Pers Soc Psychol* (2001) 81:160–9. doi: 10.1037/0022-3514.81.1.160
61. Shalvi S, Eldar O, Bereby-Meyer Y. Honesty requires time (and lack of justifications). *Psychol Sci* (2012) 23(10):1264–70. doi: 10.1177/0956797612443835

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration, with one of the authors MM.

frontiers in Psychiatry

ORIGINAL RESEARCH
published: 17 June 2019
doi: 10.3389/fpsyt.2019.00424

The Detection of Malingered Amnesia: An Approach Involving Multiple Strategies in a Mock Crime

Stefano Zago[1]*, Emanuela Piacquadio[1], Merylin Monaro[2], Graziella Orrù[3], Erika Sampaolo[1,4], Teresa Difonzo[1], Andrea Toncini[2] and Eugenio Heinzl[5]

[1] U.O.C. Neurologia, IRCSS Fondazione Ospedale Maggiore Policlinico di Milano, Milano, Italy, [2] Department of General Psychology, University of Padova, Padova, Italy, [3] Department of Surgical, Medical, Molecular & Critical Area Pathology, University of Pisa, Pisa, Italy, [4] IMT School for Advanced Studies Lucca, Lucca, Italy, [5] Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Milano, Italy

The nature of amnesia in the context of crime has been the subject of a prolonged debate. It is not uncommon that after committing a violent crime, the offender either does not have any memory of the event or recalls it with some gaps in its recollection. A number of studies have been conducted in order to differentiate between simulated and genuine amnesia. The recognition of probable malingering requires several inferential methods. For instance, it typically involves the defendant's medical records, self-reports, the observed behavior, and the results of a comprehensive neuropsychological examination. In addition, a variety of procedures that may detect very specific malingered amnesia in crime have been developed. In this paper, we investigated the efficacy of three techniques, facial thermography, kinematic analysis, and symptom validity testing in detecting malingering of amnesia in crime. Participants were randomly assigned to two different experimental conditions: a group was instructed to simulate amnesia after a mock homicide, and a second group was simply asked to behave honestly after committing the mock homicide. The outcomes show that kinematic analysis and symptom validity testing achieve significant accuracy in detecting feigned amnesia, while thermal imaging does not provide converging evidence. Results are encouraging and may provide a first step towards the application of these procedures in a multimethod approach on crime-specific cases of amnesia.

Keywords: amnesia, crime, mock crime, malingering detection techniques, malingering

OPEN ACCESS

Edited by:
Cristina Scarpazza,
University of Padova, Italy

Reviewed by:
Antonietta Curci,
University of Bari Aldo Moro, Italy
Charles Scott,
University of California, Davis,
United States

****Correspondence:***
Stefano Zago
stefano.zago@unimi.it

Specialty section:
This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry

Received: *18 December 2018*
Accepted: *29 May 2019*
Published: *17 June 2019*

Citation:
Zago S, Piacquadio E, Monaro M, Orrù G, Sampaolo E, Difonzo T, Toncini A and Heinzl E (2019) The Detection of Malingered Amnesia: An Approach Involving Multiple Strategies in a Mock Crime. Front. Psychiatry 10:424. doi: 10.3389/fpsyt.2019.00424

INTRODUCTION

Crime-related amnesia is a controversial problem and the subject of a prolonged debate (1–5). It has been observed that offenders report total or partial amnesia regarding a violent homicide in a range of 10% to 70% of the cases, depending on the literature reviewed (5–8).

Tracing the history of the phenomenon, the interesting stories of crime-related amnesia of Rudolf Hess (9) and Guenther Podola (10) can be found, where amnesia seems easy to pretend and difficult to disprove, and it arises as part of a defense strategy using loss of memory as mental incompetency to stand trial (11).

Even in the recent past, some courts expressed the view that amnesia is an important point to consider when answering the question about whether or not the defendant can receive a fair trial (12).

Most clinicians, forensic experts, and judges though are skeptical about the development of such an authentic crime-related amnesia. Notwithstanding this marked skepticism, researchers demonstrate that, apart from malingering, some cases of crime-related amnesia are genuine and could be attributed to a range of temporary brain dysfunctions. Acute alcohol and drug intoxications (13, 14), sleep disorders (15, 16), psychotic episodes (17), or dissociative states in traumatic and/or under stressful events (6, 18) are some examples. In particular, it has been hypothesized that dissociative states relate to neurotransmission and neuroendocrine dysregulations, underlying an organic cause. However, as Pyszora and colleagues argued, crime-related amnesia has often a psychogenic origin (19, 20), a condition that could determine an adverse effect in attention and in the consolidation of memories related to crime (21). Yet, a partial or complete recovery of memories is possible (19).

According to many adversarial criminal systems (US, UK, and most European countries), amnesia for crime, as an isolated reported symptom, which is not clustered within another neurological or psychiatric disorder, cannot be the basis of any mental insanity or reduced capacity claim. However, when associated with a neurological or psychiatric disorder, amnesia may call for additional safeguards to guarantee a fair trial (11, 22, 23). According to the Italian Penal Code, this is the case when the defendant is suffering from a genuine amnesia (e.g., from a neurological illness), preventing his recollection of the real fact as it is unfolded during the crime itself. Consequently, the defendant's amnesia would give the prosecutor an improper advantage in the legal confrontation. This improper advantage would undermine the legal basis of the adversarial system, which requires equal opportunity of prosecution and defense in front of the judge. When loss of memory appears to be temporary, the trial could be deferred for a reasonable period of time to allow improvements of the defendant's amnesia. In Italy, for instance, these cases undergo periodic reassessments (usually every 6 months).

Given that fraudulent claims about amnesia are easy to be feigned, it is important to evaluate whether such amnesia is genuine or made up.

Forensic experts require a significant clinical and testing expertise in order to accurately evaluate amnesic disorders in criminal proceedings. Hence, a deep knowledge of neurobiological correlates of memory is needed. The same understanding is required for its processes, various amnesic syndromes, and their underlying organic or psychogenic causes (24). During the examination, it is important to have a thorough examination of offenders, reconstructing the history of their memory disorder, interviewing them, analyzing the circumstances of the crime in all its details, and inquiring into all the situations that preceded and followed the violent performance. Along with a neuropsychological standard evaluation including memory tests, a set of neuroimaging [Computerized axial tomography (CAT) and Magnetic resonance imaging (MRI) scans] and neurophysiological [Electroencephalography (EEG)] acquisitions may highlight lesions related to diseases or anomalous brain functioning. On a practical level, this preliminary knowledge and tools themselves could lead to a solution in some cases.

However, forensic specialists can also supply these traditional applications by using different tools specifically designed to detect lie and autobiographical memory veracity. Polygraph, event-related potentials (ERPs), functional magnetic resonance imaging (fMRI), facial expression analysis, thermal imaging, or neuropsychological procedures such as symptom validity testing (SVT), autobiographical Implicit Association Test (a-IAT), or kinematic technique have all been proposed as potential methods to detect genuine crime amnesia (25–27).

Despite the extended literature on laboratory studies regarding lie and memory detection, published single case reports on defendant's amnesia where these emerging techniques are applied are very few. Single-case application is, to our knowledge, limited to some studies with polygraph, ERPs, and SVT. For example, Jelicic (28) applied SVT in a case of a 29-year-old man who stabbed his girlfriend to death while claiming to have forgotten the details of the crime. Using SVT, the author argued that this was a circumstance where malingering occurred.

An alternative possibility to study feigned amnesia is the preliminary application of these techniques in mock crime experiments [e.g., Refs. (29–33)]. However, only in the study of Giger and colleagues was a more realistic homicide scenario built in which the participants had to hit with great force the victim. In the other studies, subjects were instructed to perform petty thefts of things or money. A limitation of the mock crime studies has been raised by Merckelbach et al. (29) that pointed out the little ecological validity of this kind of experimental design. For example, there is no doubt that, in cases of criminal amnesia, there are higher levels of emotional arousal (34), which are impossible to replicate in mock crime experimental conditions. It will therefore be necessary in the future, when feasible, to directly assess the compatibility of laboratory data (e.g., mock crime) and real-life data of offenders.

Currently, a clear line of demarcation between experimental analysis and real practical forensic application has yet to be defined. Nowadays, it is possible to see these techniques as a useful support to the clinical analysis of crime-related amnesia. Moreover, it is crucial to satisfy the Daubert Standard Criteria within such well-established practices. The U.S. Supreme Court, in Daubert v. Merrell Dow, outlined six criteria for the federal judge to consider when determining the admissibility of evidence (35). These criteria govern the acceptability of scientific tests based on the percentage of reliability of a technique, the publication of relevant studies in peer-reviewed journals, and the general consent among the scientific community.

The purpose of this paper is to evaluate the efficacy of three emerging techniques in evaluating crime-related amnesia, i.e., thermal imaging, kinematic analysis, and SVT, in a group of subjects invited to simulate, or not, amnesia following a mock homicide. The choice of these three techniques is motivated by the fact that they can be administered in a multi-method approach in a simple and non-interfering way. Thermal imaging is based on autonomic responses, while kinematic analysis and SVT are based on cognitive elaboration. A brief review of these techniques is reported below.

THREE EMERGING TECHNIQUES TO EVALUATE CRIME-RELATED AMNESIA

Thermal Facial Imaging

Thermal infrared imaging is a widely used technique to measure heat emission from the body, transformed into an infrared band of the electromagnetic spectrum (36). Body temperature, and in particular facial temperature, reflects the activity of the autonomic nervous system during the natural exposition to social interaction and communication (37). For this reason, psychophysiologists are interested in the measurement and recording of these bodily changes. An interesting application of thermal infrared imaging is in the lie detection field. In particular, researchers analyzed facial skin surface temperature (SST) in deceptive and non-deceptive participants while performing a Concealed Information Test (CIT) (38–41). During the arousal, an increase in SST in the periorbital region around the eye and the nose was found. This may suggest a plausible association with specific emotions. Generally, data showed that deceptive subjects had a higher temperature in these regions compared to non-deceptive ones. For example, in the study of Pavlidis et al. (38), 83% of the participants were correctly recognized as mentors (75%) or innocents (90%) by the analysis of thermographic images. For the same subjects analyzed with the polygraph, the accuracy was lowered up to 70%.

A simple objection of thermal imaging application is that an increase in blood flow in the periorbital zone is also associated with prolonged stress, and a stressed person could be wrongly judged as guilty.

To our knowledge, there are no studies analyzing thermal imaging results on crime-related amnesia.

Kinematic Technique

A recent technique, also referred to as *kinematic technique*, has been introduced by Monaro et al. (42, 43) to detect fake responses regarding identity. It is based on recording motor response of subjects involved in a computer task while using a mouse. The mouse movement analysis may be used as an implicit measure to investigate the cognitive processes underlying a task (44), including the cognitive processes underlying the deception production (45). Indeed, lying is more cognitive demanding than truth-telling, and this challenging cognitive process reflects itself in the human behavior, like reaction times (RTs) (43, 46) or mouse responses.

During this activity, participants are asked to answer truthfully or untruthfully to phrases shown on the monitor, using the mouse to click one between two alternative responses ("yes" or "no") that appear on the screen. The analysis of the mouse trajectory highlights how false responses can be distinguished from the true ones. This statement is based on temporal and spatial dynamic parameters, such as the time to compute the response and the width of the mouse trajectory, as well as other kinematic parameters like speed and acceleration (47). Indeed, liars show wider and more erratic trajectories; they make more errors and take more time to compute their responses. On the other hand, truth-tellers are more rapid; they make fewer errors, and they are characterized by mouse trajectories straight to the responses.

The kinematic technique has been recently applied also to the detection of psychiatric disorders simulation. Monaro et al. (47) proposed to apply the kinematic analysis to detect the simulation of depression, catching mouse movements while the patient is engaged in responding to double-choice questions about depressive symptoms. The authors analyzed the difference in mouse trajectories between depressed patients and participants who were instructed to simulate a depressive disorder in order to gain a financial reward. Results demonstrate that this technique is able to detect feigned depression with an accuracy up to 96%.

Currently, there are no studies on crime-related amnesia in which kinematic technique was applied.

SVT Procedures

One additional strategy to detect malingered amnesia in crimes consists of using forced-choice recognition memory tests, such as SVT (33, 48, 49). This is a well-known procedure used in civil courts to detect malingering, especially in mild traumatic brain injury. Its logic is as follows: if a patient is genuine, with an unfeigned impairment, he will not be able to choose the correct answer between the two stimuli; in this case, he should perform at chance level over many trials. On the contrary, malingerers usually select the wrong response deliberately and thus they perform significantly below chance. The most likely explanation for this performance is that the examinee knows the correct answer but decides not to choose it (49, 50).

The SVT procedure was also adapted to assess criminal defendants who claim to suffer from amnesia. The offender is asked to answer a series of questions based on facts or details linked to the crime deriving from police reports or third-party testimonies. Each question has at least two possible answers, one correct and the other incorrect but plausible. Generally, this information is presented orally or in written form on a computer screen, but alternatively, when visual material obtained during police inspections or images of the crime scene are available, it is possible to set up the test with such material [see, for example, Ref. (33)]. The visual presentation of images seems, in our opinion, preferable due to a reduced mnemonic load in terms of working memory.

Brandt et al. (48) proposed a first application of SVT in crime-related amnesia. They examined LG, a 64-year-old man charged with the murder of his wife claiming complete amnesia of the event. He was asked to freely recall a 20-item word list and then to attempt two-alternative forced-choice recognitions of each of the 20 words. LG freely recalled only four target words, and in a forced-choice recognition, he correctly selected only three of the target words. This was a performance worse than chance indicating that, at some level, he knew most of the 20-item words. It was suggested that he was feigning his anterograde memory deficit for violent crime.

Similarly, Denney (49) used SVT to evaluate crime-related amnesia in three cases of homicide. He collected a series of autobiographical information and data concerning the criminal events in order to create a questionnaire. Subjects were presented with written sentences where 50% of the cases were referred to real events occurring before the homicide. The remaining 50% of the phrases described a similar, but unreal, event. The task consisted of reading the sentences and saying whether they were true or false. All the subjects responded below the chance level, a result that indicates a voluntary strategy to avoid correct answers pretending memory loss of the criminal event. Recently, Jelicic (28) described the case of Randy, a 29-year-old man, accused of his girlfriend's homicide, who claimed a complete amnesia of the murder. Reconstructing the crime scene, an SVT with 20 forced-choice questions with correct and incorrect but, plausible and similar, answers was created. Randy's amnesia resulted in 14 incorrect answers out of the 20 items. According to binomial statistics, the probability that his response pattern was based on random guessing was <6%, indicating that there was a <6% chance that his amnesia was genuine. Those two elements led to converging evidence that Randy had feigned his amnesia for the stabbing. The court also found his amnesia claim not credible (28).

Again, in a multi-method approach study, Giger et al. (33) applied a forced-choice SVT in a mock homicide and found out only a low sensitivity of the procedure. The authors argued that the results are probably due to a few utilized numbers of items.

METHOD

Participants

Forty volunteers (20 female, 20 males; mean age = 24.5, *sd* = 8.27, range = 19–60 years) were recruited from the staff of the Bicocca University of Milan and the IRCCS Fondazione Caa Granda Ospedale Maggiore Policlinico of Milan. All participants had normal visual acuity and were screened for a history of psychiatric, neurological, or medical illnesses. The Ethics Committee of the Bicocca University of Milan and the Istituti di Ricovero e Cura a Carattere Scientifico (IRCCS) Fondazione Cá Granda Ospedale Maggiore Policlinico of Milan exempted us from initiating the practice of approval, considering the study as an observational type without drug use. Each subject participated in the study voluntarily, without remuneration. Before the experiment, all participants signed a disclaimer form in order to take part in the study in accordance with the guidelines of the University Committee. Two random groups, balanced for gender and age, were used.

Procedure

The experiment took place in a single session and lasted between 30 and 45 min. It was designed in four steps: a) baseline thermography; b) mock crime; c) thermal imaging during kinematic test, and d) Symptom Validity Test procedure. At first, each subject entered a room and a baseline thermographic image was taken. The areas behind the eyes and to the sides of the nose corresponding approximately to the tear ducts were explored in detail. Lacrimal caruncle temperature (°C) was recorded by a certified technician (EH) using an infrared camera (NEC Avio TVS500; Nippon Avionics Co., Ltd, Tokyo, Japan). It was not possible to regulate room temperature and humidity, but they were relatively stable across all situations (minimum = 18°C, maximum = 22°C; mean = 20°C). Before every session, to define the radiance emission and to nullify the effect of surface reflections on tested participants, the same image of a Lambert surface was taken. Only images perfectly on focus were used. Grayess IRT Analyzer 6.0 software (51) was used to calculate the maximum temperature (°C) of a circular area traced around the caruncle area and of the body surface; this value was used for subsequent analysis.

Figure 1 reports some examples of images obtained at baseline thermal imaging.

Afterwards, each subject was instructed with the following orders:

> "You have to enter the room and pick up the knife on the table. Don't worry, it is a fake knife which can harm no one! So, now go into the room. You will see a girl sitting at a desk with her back to you. This is a mannequin even if it seems real. There is a big box on the table, inside there may or may not be some money. Stab the girl violently in the back and check whether the box contained money or not. If it does, take it and run back to the room where you found the knife."

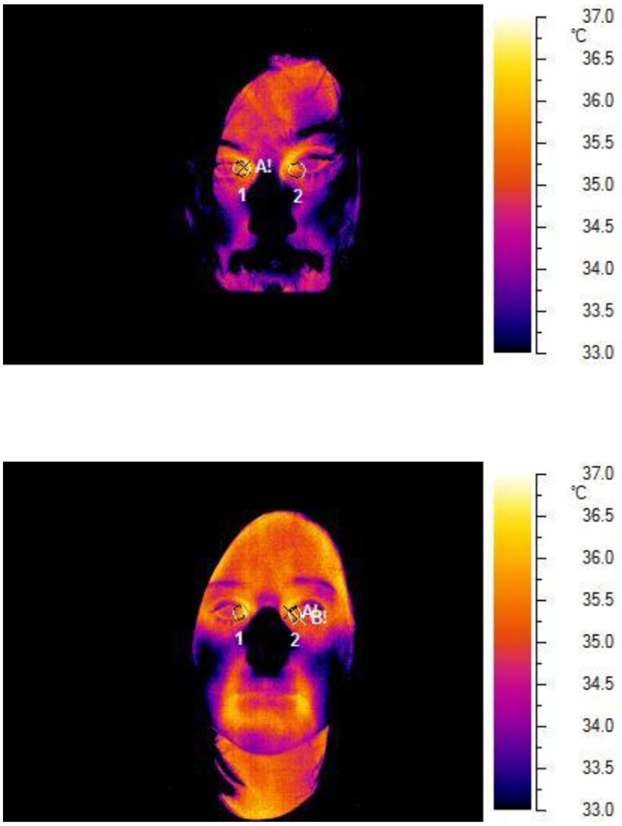


FIGURE 1 | Thermal imaging.

The mannequin was wearing a pink cap, black sweater with a white motif, a white lace skirt, and black boots. It also had sunglasses, earrings, a black necklace, and a yellow watch. Furthermore, the crime scene was composed of the following objects: two chairs, a desk, two red apples, a red rose, a fork, a black bag, a box containing jewels, and a computer. See **Figure 2** for a full representation of the crime scene.

Once subjects returned to the original room, they were assigned to two different experimental groups. The first group (*honest; n* = 20) was instructed to be honest and to perform accordingly in all the experimental phases. The second group (*naïve malingerers; n* = 20) was instructed to simulate a crime-related amnesia to avoid any criminal responsibility.

Immediately after, participants were asked to sit in front of a computer and to carry out a kinematic test, which analyzes mouse dynamics to detect deceptive responses. The details of the procedure have been described in-depth by Monaro and colleagues, in their paper regarding malingered depression (47). In this study, the mouse dynamics test was adapted to the analysis of crime-related amnesia. For example, test instructions for the honest group of our study were the following:

> "The following questions concern the actual moment and the simulated homicide in question. Please answer all the questions honestly. If you are undecided about a question, mark the answer which you think is more correct. To answer, click 'yes' on top right or 'no' on top left of the screen. To see each question, click on 'start' at the center bottom of the screen. Some questions are composed of two phrases. To answer these questions, you should click 'yes' only if you agree with both phrases. To start the experiment press 'shift' on the keyboard."

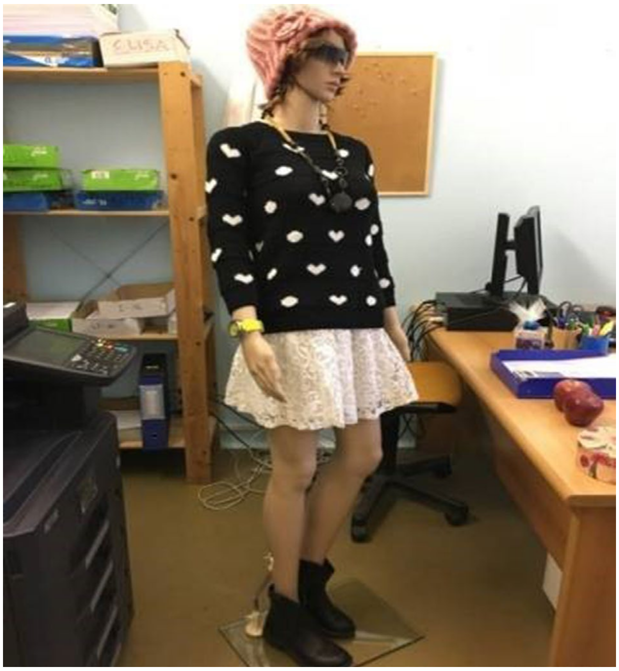


FIGURE 2 | Mannequin used in the mock crime scene.

The task was programmed using *MouseTracker* software (52). Seventy-one sentences randomly appeared on the upper part of the computer screen and presented to the subjects. Participants were instructed to respond to each question by clicking on one of the two alternative responses ("*yes*" on the upper left or "*no*" on the upper right). The 71 stimuli included 16 types of sentences according to the complexity of the sentence (simple vs. complex sentences), to the required response (yes vs. no), and to the sentence topic (memory of mock crime vs. crime scene vs. test setting). Simple sentences (*n* = 15) contained only one piece of information related to the crime scene, the test setting or the amnesia symptoms (e.g., "*Do you remember the face of the mannequin?*"). Complex sentences (*n* = 56) were those containing two or more pieces of information—about the crime scene, the test setting, or amnesia symptoms—in the same phrase (e.g., "*Do you remember the face of the mannequin and are you wearing shoes right now?*"). Each piece of information in the phrase could be true or false, so a complex question required a "yes" response when both parts were true, whereas it requires a "no" response when at least one of the two was false (53). Simple and complex sentences regarded the memory of the mock crime (e.g., *"Do you remember what happened in the room?"*), the crime scene (e.g., "*Do you remember an apple and a bag in the room?"*), or the test setting (e.g., "*Are you wearing shoes right now?*"). In the **Online Supplementary Information**, the list of the sentences presented to the subjects is reported, including the information about the type of sentence and the expected response for each experimental condition.

Complex questions have been proved to be an accurate strategy to increase liars' cognitive load and, as a consequence, to spot them. Indeed, responding to complex questions, the subject has to monitor the plausibility of more than one information and retain it in working memory to finally decide if the entire sentence is true or false. While truth-tellers can speedily carry out this sequence of mental operations, liars need more time to match the plausibility of each information with the lie they told (54). Responding to complex questions, liars have been demonstrated to have slower RTs and worst accuracy than truth-tellers (53).

The *MouseTracker* software recorded the spatial and temporal features of the mouse trajectory while the subject was responding (see **Table 1**). After computing the average value of all stimuli for each participant, the kinematic spatial and temporal features were used to compute statistical analysis. Finally, for each participant, we also calculated the average value of each feature for the 16 types of sentences. Then, these data were entered in machine learning (ML) models to predict whether a subject was honest or a naïve malingerer.

During the kinematic test, a second infrared thermographic image was taken for a comparison with the baseline image previously made. At the end of kinematic session, a self-filling computerized two forced-choice task (SVT) was administered

TABLE 1 | Spatial and temporal features recorded by Mousetracker sotware.

	Feature	Description
Temporal features	Initiation time (IT)	The time between the appearance of the question and the beginning of the mouse movement.
	Reaction time (RT)	The time from the appearance of the question to the click on the response box.
	Maximum deviation time (MD-time)	The time to reach the point of maximum deviation.
	Velocity on *x*- and *y*-axis	The speed of movement of the mouse on *x*- and *y*-axis during the response.
	Acceleration on *x*- and *y*-axis	The movement acceleration of the mouse on *x*- and *y*-axis during the response.
Spatial features	Maximum deviation (MD)	The largest perpendicular distance between the actual trajectory and the ideal trajectory.
	Area under the curve (AUC)	The geometric area between the actual trajectory and the ideal trajectory.
	x-flip	The number changes in direction along the *x*-axis.
	y-flip	The number changes in of direction along the *y*-axis.

to the subjects. It was composed of 25 questions concerning the mock crime scene. As reported in the introduction, SVT is one of the most extensively investigated measures for the detection of memory malingering and has been used in some studies to evaluate memory in a criminal forensic setting. A forced-choice SVT is based on the binomial theorem. It predicts whether, when an individual is asked questions with only two possible answers of equal probability, test results fall within a predictably random range and distribution. In particular, below-chance performance alone would be predicted by binomial values.

In our procedure, to make the SVT more sensitive, we modeled the task on the *Free and Cued Selective Reminding Test* (*FCRST*) (55). FCSRT is a measure of memory under conditions that control attention and cognitive processing. The aim is to obtain an assessment of memory unaffected by normal age-related changes in cognition. Differently from other memory tests, the FCSRT requires a study phase designed to control attention and cognitive processing in order to identify memory impairment, not secondary to other cognitive deficits. Subjects identify pictured items (e.g., grapes, vest) in response to category cues (fruit, clothing). In the test phase, subjects are asked to recall the items they learned (free recall). The category cues are used for a prompt recall of items not retrieved during the free recall to generate a score termed cued recall. The sum of free and cued recall is called total recall. Originally, this was composed of 12 figures of both living and non-living things. In our test, the 12 original images were replaced by six images of objects present in the crime scene and six distractors. All the images were subdivided into three cards with four items on each. The six objects in the crime scene were a fork, a pink rose, red apples, a necklace, a sweater, and a pink hat. The four images were placed in front of the participant who had to name all of them.

Participants were then asked to remember the 12 items. For the images they did not remember, a semantic cue was given (e.g., there was a flower). The procedure was carried out three times. Then, an interference task lasting for about 20 min was presented to the subject. As interference test, we used the "Deux Barrages Test" (56), which only implies attentional capabilities without overloading or stimulating memory recall. If naïve malingering subjects report a score below chance level, it is possible to state with good probability that they are malingerers.

RESULTS

All the participants followed the instructions and committed the mock crime. Data from the experiment were processed with IBM SPSS (version 24) and WEKA software (57).

Thermal Imaging

We carried out two *t tests* on an independent sample to compare temperatures within the groups, one on the baseline condition and the other on the experimental condition. With regard to the baseline condition, the result demonstrates no difference between the two experimental groups (t = 1.675, df = 20.908, p = .109; see **Table 2**). Surprisingly, in the experimental condition, the lacrimal caruncle temperature decreases in the deceptive group compared to the honest one (p = .003). These data contrast with results obtained in previous studies [e.g., Refs. (38, 40)] where an increase in facial temperature was found in deceptive participants.

Kinematic Technique

The kinematic results compared the responses of malingerers with honest subjects by averaging the responses to all stimuli across individuals. The analysis of kinematic spatial features, relative to the average of all stimuli to which subjects responded, shows that honest trace wider trajectory compared to malingerers [average honest maximum deviation (MD) = 0.69, sd = 0.21, area under the curve (AUC) = 1.46, sd = 0.72; average malingerers MD = 0.6, sd = 0.24, AUC = 1.35, sd = 1.09]. The average trajectories of both malingerers and honest are represented in **Figure 3**.

An independent sample t test was carried out on the 11 kinematic features [initiation time (IT), RT, MD, maximum deviation time (MD-time), AUC, x-flip, y-flip, velocity, and acceleration on x- and y-axis] obtained by averaging the 71 stimuli for each subject. To avoid the multiple testing problem, we applied a Bonferroni correction and the p value was set to .0045. Results showed a significant statistical difference between the two groups only for *MD-time* [$t_{(36)} = -3.27$, $p <$ .0045, $sd = -1.04$].

Then, we analyzed the same 11 features by averaging the subjects' responses to each of the 12 types of stimuli. Using a correlation-based feature selector (CFS), as implemented in WEKA software (57), we identified the features that are highly

TABLE 2 | Thermal imaging analysis with Levene's test.

		Baseline		Experimental group	
		Equal variances assumed	**Equal variances not assumed**	**Equal variances assumed**	**Equal variances not assumed**
Levene's Test for Equality of Variances					
F		6.765		1.967	
Sig.		.014		.170	
T		1.741	1.675	3.159	3.210
df		32	20.908	32	31.326
***t* Test for Equality of Means**					
Sig. (two-tailed)		.091	.109	.003	.003*
Mean difference		.27986	.27986	.37727	.37727
s.e. difference		.16078	.16706	.11943	.11753
95% confidence interval	Lower	–.04764	–.06765	.13400	.13766
of the difference	Upper	.60736	.62737	.62054	.61687

F, test statistic of Levene's test; Sig., p value corresponding to Levene's test; t, computed test statistic; df, degrees of freedom; Sig. (two-tailed), –value corresponding to the given test statistic and degrees of freedom; s.e., standard error.

*t test was used to compare temperatures within the baseline condition and the experimental condition. In the experimental condition, a statistically significant difference between the two groups was demonstrated (*p < .001, p = .003).*

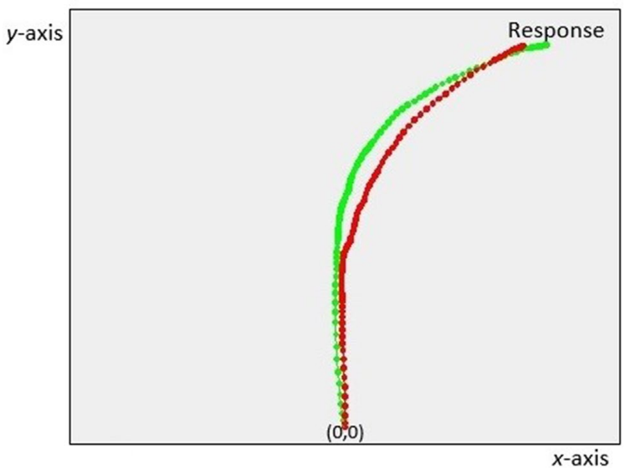


FIGURE 3 | Average trajectories for liars and truth-tellers. The figure shows the average trajectories between the subjects to the expected YES and unexpected YES questions, respectively, for honest (in red) and for malingerers (in green). Expected and unexpected questions that require a YES response are those to which the malingerers lied. The values of the maximum deviation (MD), area under the curve (AUC), *x*-flip, and *y*-flip parameters for the two groups are reported. The gray area represents the difference in the AUC parameter between the malingerers and honest.

correlated with the dependent variable (honest vs. malingerers) while having low inter-correlation. Seven variables were selected and included as predictors within different ML models. The selected features are summarized in **Table 4**.

The trained ML algorithms were the following: Naïve Bayes, Random Forest, SVM, and K-nearest neighbours classifier (IBk). All the classifiers were trained using a 10-fold cross-validation procedure and reached an accuracy between 80% and 90% in distinguishing honest from malingerers. The accuracy for each classifier is reported in **Tables 3** and **4**.

Symptom Validity Test

Finally, in the *Symptom Validity Test*, a *t* test for independent samples showed a statistical difference between the two groups ($t = 17.7$; $df = 31.22$; $p < .001$). In addition, the results demonstrate that malingerers scored significantly below the chance level ($t = -8.\ 159$, $df = 19$, $p < .001$; $Z = -1.84$).

DISCUSSION AND CONCLUSIONS

One of the main goals in crime-related amnesia is to find methods to detect malingering. Techniques of investigation are aimed to assist the court in evaluating the reliability of declarative proof that has been devised and perfected over a century. An increasing number of researches involving new lie detectors such as modern polygraphs, ERPs, thermal imaging, fMRI, kinematic analysis, facial analysis, or neuropsychological measures are applied today. Overall, studies have resulted in many promising findings. However, most of them highlighted the need of advances in the field with the consolidation of new methods driven by technical improvements.

The purpose of the present study was to investigate crime-related amnesia through the comparison of three new emerging methods (facial infrared thermography, kinematic analysis, and SVT) in a group of subjects invited to simulate, or not, an amnesia following a mock homicide. The results showed that kinematic analysis and SVT acquired significant accuracy in distinguishing honest from malingerers. However, thermal imaging results do not appear in line with those studies that reported more heat absorbed around the eyes when people lie.

With regard to SVT, the results of the present study clearly show better significance levels than those obtained by Giger et al. (33), who designed one of the first realistic mock crime experiments. Moreover, our data seem to be in line with earlier studies on real offenders (28, 49). It should be noted that in our SVT procedure, visual stimuli were used, along with controlling for the correct coding of the stimuli. In our opinion, this procedure

TABLE 3 | Description of the seven variables selected by the correlation-based feature selector (CFS) and entered in the machine learning (ML) models and their correlation with the dependent variable.

Feature	r_{pb}
x-flip of simple sentences about the testing situation	0.44
MD-time of complex sentences about the memory of the mock crime and the testing situation requiring a no response	0.40
Velocity on *x*-axis of complex sentences about the memory of the mock crime requiring a yes response	0.44
Velocity on *x*-axis of complex sentences about the crime scene requiring a yes response	0.53
Acceleration on *y*-axis of complex sentences about the crime scene requiring a yes response	0.16
Velocity on *x*-axis of complex sentences about the crime scene and the testing situation requiring a yes response	0.63
MD-time of complex sentences about the crime scene and the testing situation requiring a no response	0.39

TABLE 4 | Accuracy in distinguishing malingerers and honests obtained by four different ML classifiers using a 10-fold cross-validation procedure. Precision, recall, and *F* measure are also reported.

Classifier	Accuracy in 10-fold cross-validation	Precision	Recall	*F* measure
Naïve Bayes	89.7%	0.902	0.897	0.897
SVM	84.6%	0.862	0.846	0.845
Random Forest	89.7%	0.902	0.897	0.897
IBk	92.3%	0.924	0.923	0.923

IBk, K-nearest neighbours classifier; SVM, support-vector machine.

and, above all, the implementation of visual material offer greater guarantee than the verbal version of the SVT in determining the veracity of crime-related amnesia.

To our knowledge, this is the first study to apply a kinematic analysis on an experiment involving crime-related amnesia. The results demonstrate the efficacy of this technique in detecting feigned amnesia, but they need to be further verified by additional studies.

Regarding infrared thermal imaging, we found that malingerers were slightly cooler than the honest subjects. A possible interpretation of this unexpected result is that such experimental conditions do not elicit a real emotional state. It should be noted that this is a measure of sympathetic nervous system and it differs from the other two techniques in which cognitive aspects, such as memory recall, are more prevalent. It has the advantage of being a contactless and non-invasive device able to record the spontaneous thermal irradiation of the face. We analyzed a specific region, the lacrimal caruncle, differently from most of the studies in the literature. Indeed, previous studies took into account the analysis of more distributed areas such as periorbital, supraorbital, and maxillary regions without focusing only on the lacrimal caruncle (36).

Animal studies suggest a relationship between the temperature of this area and the sympathetic nervous system (58). We examined this region to find a correspondence in humans. Our results show a little and non-significant decrease in the lacrimal caruncle temperature of the malingered group. Recently, Huggins and Rakobowchuk (59) applied a cold pressor test (CPT) and a muscle chemoreflex (MCR) to healthy subjects in order to activate the autonomic nervous system. No significant alteration in the temperature of the lacrimal caruncle was found. As the authors claimed, it is likely that changes in this region are more difficult to be detected using the infrared thermal imaging. Another possible explanation is that the human response is different compared to animals. The results of this study did not show an increase in the eye temperature between the baseline and the experimental condition. Since our aim was to find a very subtle variation when people lie, it is possible that the used infrared vision camera was not sensitive enough to detect such a change. A plausible interpretation of discordant results in literature is probably related to the complexity of the sympathetic nervous system in the lacrimal caruncle. It may be possible that, in this region, there is a different pattern of activation compared to the periorbital or supraorbital areas. Additional studies, with more refined thermal imaging approaches, are needed to clarify the activity of the autonomic nervous system through temperature changes in the human lacrimal caruncle. Moreover, the potential of this technique as a lie detector should be assessed more precisely.

In conclusion, the results of this preliminary study clearly highlighted the role of new lie detection methods in empirically supporting forensic professionals when discriminating between genuine and malingering crime-related amnesia. A multi-technique approach seems desirable and will be crucial in the translation of mock experimental to real single criminal case evaluation. In particular, future work, with defendant's amnesia, will allow a more informed use of the three methods we have studied here.

ETHICS STATEMENT

The Ethics Committee of the Bicocca University of Milan and the IRCCS Fondazione Ca Granda Ospedale Maggiore Policlinico of Milan exempted us from initiating the practice of approval, considering the study as an observational type without drug use. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

Conceived the experiment: SZ and EP. Designed the experimental task: SZ, EP, MM, and AT. Data acquisition: EP, SZ and EH. Data analysis: SZ and MM. Data interpretation: SZ and MM. Drafting of the manuscript: SZ, MM, ES, TD and GO. All the authors revised the manuscript critically and gave the final approval of the version to be published.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00424/full#supplementary-material. The complete list of stimuli presented to participants during the mouse tracking task.

REFERENCES

1. Hopwood JS, Snell HK. Amnesia in relation to crime. *J Ment Sci* (1933) 79:27–41. doi: 10.1192/bjp.79.324.27
2. Schacter DL. Amnesia and crime: how much do we really know? *Am Psychol* (1986) 41:286–95. doi: 10.1037/0003-066X.41.3.286
3. Kopelman MD. Crime and amnesia: a review. *Behav Sci Law* (1987) 5:323–42. doi: 10.1002/bsl.2370050307
4. Herman JL. Crime and memory. *J Am Acad Psychiatry Law Online* (1995) 23:5–17.
5. Bourget D, Whitehurst L. Amnesia and crime. *J Am Acad Psychiatry Law Online* (2007) 35:469–80.
6. Porter S, Birt AR, Yuille JC, Hervé HF. Memory for murder: a psychological perspective on dissociative amnesia in legal contexts. *Int J Law Psychiatry* (2001) 24:23–42. doi: 10.1016/S0160-2527(00)00066-2
7. Cima M, Nijman H, Merckelbach H, Kremer K, Hollnack S. Claims of crime-related amnesia in forensic patients. *Int J Law Psychiatry* (2004) 27:215–21. doi: 10.1016/j.ijlp.2004.03.007
8. Woodworth M, Porter S, ten Brinke L, Doucette NL, Peace K. Campbell MA. A comparison of memory for homicide, non-homicidal violence, and positive life experiences. *Int J Law Psychiatry* (2009) 32:329–34. doi: 10.1016/j.ijlp.2009.06.008
9. Rees JR. *The Case of Rudolph Hess*. New York: WW Norton & Company, Incorporated (1948).
10. Furneaux R. *Crime Documentary NO.1. Guenther Podola*. London: Stevens & Sons Limited (1960).
11. Hoctor S. Amnesia and criminal responsibility. *S Afr J Crim Just* (2000) 13:273.
12. Cocklin K. Amnesia: the forgotten justification for finding an accused incompetent to stand trial. *Washburn Law J* (1980) 20:289.
13. Van Oorsouw KIM, Merckelbach H, Ravelli D, Nijman H, Mekking-Pompen I. Alcoholic blackout for criminally relevant behavior. *J Am Acad Psychiatry Law Online* (2004) 32:364–70.
14. Granacher RP. Commentary: alcoholic blackout and allegation of amnesia during criminal acts. *J Am Acad Psychiatry Law Online* (2004) 32:371–74.
15. Podolsky E. Somnambulistic homicide. *Dis Nerv Syst* (1959) 20:534.
16. Broughton R, Billings R, Cartwright R, Doucette D, Edmeads J, Edwardh M, et al. Homicidal somnambulism: a case report. *Sleep* (1994) 17:253–64. doi: 10.1093/sleep/17.3.253
17. Eronen M, Tiihonen J, Hakola P. Schizophrenia and homicidal behavior. *Schizophr Bull* (1996) 22:83–9. doi: 10.1093/schbul/22.1.83
18. Bradford JMW, Smith SM. Amnesia and homicide: the padola case and a study of thirty cases. *J Am Acad Psychiatry Law Online* (1979) 7:219–31.
19. Pyszora N, Fahy T, Kopelman M. Amnesia for violent offenses: factors underlying memory loss and recovery. *J Am Acad Psychiatry Law* (2014) 42:202–13.
20. Pyszora NM, Barker AF, Kopelman MD. Amnesia for criminal offences: a study of life sentence prisoners. *J Forens Psychiatry Psychol* (2003) 14:475–90. doi: 10.1080/14789940310001599785
21. Bourget D, Gagné P, Wood S. Dissociation: defining the concept in criminal forensic psychiatry. *J Am Acad Psychiatry Law* (2017) 45:147–60.
22. Roesch R, Golding SL. Amnesia and competency to stand trial: a review of legal and clinical issues. *Behav Sci Law* (1986) 4:87–97. doi: 10.1002/bsl.2370040107
23. Go G. Amnesia and criminal responsibility. *J Law Biosci* (2017) 4:194–204. doi: 10.1093/jlb/lsx003
24. Markowitsch HJ, Staniloiu A. Amnesic disorders. *Lancet* (2012) 380:1429–40. doi: 10.1016/S0140-6736(11)61304-4
25. Peters MJV, van Oorsouw KIM, Jelicic M, Merckelbach H. Let's use those tests! Evaluations of crime-related amnesia claims. *Memory* (2013) 21:599–607. doi: 10.1080/09658211.2013.771672
26. Zago S, Fumagalli M, Inglese S, Rossetti I, Sartori G, Priori A, et al. Remembering and lying in relation to crime: clinical and research implications. In: *Criminal Behaviors. Impacts, Tools and Social Networks*. Mantova: FDE Institute Press (2014).
27. Sartori G, Zangrossi A, Monaro M. Deception detection with behavioral methods: the autobiographical implicit association test, concealed information test–reaction time, mouse dynamics, and keystroke dynamics. In: Rosenfeld JP, editor. *Detecting Concealed Information and Deception*. London: Academic Press (2018). p. 215–41. doi: 10.1016/B978-0-12-812729-2.00010-0
28. Jelicic M. Testing Claims of crime-related amnesia. *Front Psychiatry* (2018) 9:617. doi: 10.3389/fpsyt.2018.00617
29. Merckelbach H, Hauer B, Rassin E. Symptom validity testing of feigned dissociative amnesia: a simulation study. *Psychol Crime Law* (2002) 8:311–8. doi: 10.1080/10683160208401822
30. Jelicic M, Merckelbach H, Bergen S. Symptom validity testing of feigned amnesia for a mock crime. *Arch Clin Neuropsychol* (2004) 19:525–31. doi: 10.1016/j.acn.2003.07.004
31. Jelicic M, Merckelbach H, van Bergen S. Symptom validity testing of feigned crime-related amnesia: a simulation study. *J Credibility Assess Witness Psychol* (2004) 5:1–8.
32. Van Oorsouw K, Merckelbach H. Simulating amnesia and memories of a mock crime. *Psychol Crime Law* (2006) 12:261–71. doi: 10.1080/10683160500224477
33. Giger P, Merten T, Merckelbach H, Oswald M. Detection of feigned crime-related amnesia: a multi-method approach. *J Forensic Psychol Pract* (2010) 10:440–63. doi: 10.1080/15228932.2010.489875
34. Swihart G, Yuille J, Porter S. The role of state-dependent memory in "red-outs." *Int J Law Psychiatry* (1999) 22:199–212. doi: 10.1016/S0160-2527(99)00005-9
35. Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
36. Gołaszewski M, Zajac P, Widacki J. Thermal vision as a method of detection of deception: a review of experiences**. *Eur Polygr* (2015) 9:5–24. doi: 10.1515/ep-2015-0001
37. Shastri D, Merla A, Tsiamyrtzis P, Pavlidis I. Imaging facial signs of neurophysiological responses. *IEEE Trans Biomed Eng* (2009) 56:477–84. doi: 10.1109/TBME.2008.2003265
38. Pavlidis I, Eberhardt NL, Levine JA. Seeing through the face of deception. *Nature* (2002) 415:35–35. doi: 10.1038/415035a
39. Dery GM. Lying eyes: constitutional implications of new thermal imaging lie detection technology. *Am J Crim L* (2003) 31:217.
40. Pollina DA, Dollins AB, Senter SM, Brown TE, Pavlidis I, Levine JA, et al. Facial skin surface temperature changes during a "concealed information" test. *Ann Biomed Eng* (2006) 34:1182–9. doi: 10.1007/s10439-006-9143-3
41. Zhu Z, Tsiamyrtzis P, Pavlidis I. Forehead thermal signature extraction in lie detection. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2007) 243–6. doi: 10.1109/IEMBS.2007.4352269
42. Monaro M, Gamberini L, Sartori G. Spotting faked identities via mouse dynamics using complex questions. *in Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI* 2018*)* (2018). doi: 10.14236/ewic/HCI2018.8
43. Monaro M, Gamberini L, Sartori G. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS One* (2017) 12:e0177851. doi: 10.1371/journal.pone.0177851
44. Freeman JB, Dale R, Farmer TA. Hand in motion reveals mind in motion. *Front Psychol* (2011) 59:1–6. doi: 10.3389/fpsyg.2011.00059
45. Duran ND, Dale R, McNamara DS. The action dynamics of overcoming the truth. *Psychon Bull Rev* (2010) 17:486–91. doi: 10.3758/PBR.17.4.486
46. Mazza C, Monaro M, Orrù G, Burla F, Colasanti M, Ferracuti S, et al. Introducing machine learning to detect personality faking-good: a new model based on MMPI-2-RF scales and reaction times. *Front Psychiatry* (2019) 10:389.
47. Monaro M, Toncini A, Ferracuti S, Tessari G, Vaccaro MG, De Fazio P, et al. The detection of malingering: a new tool to identify made-up depression. *Front Psychiatry* (2018) 249:1–12. doi: 10.3389/fpsyt.2018.00249
48. Brandt J, Rubinsky E, Lassen G. Uncovering malingered amnesia. *Ann N Y Acad Sci* (1985) 444:502–3. doi: 10.1111/j.1749-6632.1985.tb37625.x
49. Denney R. Symptom validity testing of remote memory in a criminal forensic setting. *Arch Clin Neuropsychol* (1996) 11:589–603. doi: 10.1016/0887-6177(95)00042-9
50. Sartori G, Zangrossi A, Orrù G, Monaro M. Detection of malingering in psychic damage ascertainment. In: *P5 Medicine and Justice*. Springer, Cham (2017). p. 330–41. doi: 10.1007/978-3-319-67092-8_21
51. Grayess. *IRT Analyser Users' Manual*. Bradenton, USA: GRAYESS, Inc. (2007).

52. Freeman JB, Ambady N. MouseTracker: software for studying real-time mouse-tracking method. *Behav Res Methods* (2010) 42:226–41. doi: 10.3758/BRM.42.1.226
53. Monaro M, Gamberini L, Zecchinato F, Sartori G. False identity detection using complex sentences. *Front Psychol* (2018) 283:1–10. doi: 10.3389/fpsyg.2018.00283
54. Gombos VA. The cognition of deception: the role of executive processes in producing lies. *Genet Soc Gen Psychol Monogr* (2006) 132:197–214. doi: 10.3200/MONO.132.3.197-214
55. Grober E, Buschke H. Genuine memory deficits in dementia. *Dev Neuropsychol* (1987) 3:13–36. doi: 10.1080/87565648709540361
56. Zazzo R. Test des deux barrages. In: *Actualités pédagogiques et psychologiques*. Neuchatel, Switzerland: Delachaux et Nestle (1974).
57. Hall MA. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand (1998).
58. Travain T, Colombo ES, Heinzl E, Bellucci D, Prato Previde E, Valsecchi P. Hot dogs: thermography in the assessment of stress in dogs (Canis familiaris)—A pilot study. *J Vet Behav* (2015) 10:17–23. doi: 10.1016/j.jveb.2014.11.003
59. Huggins J, Rakobowchuk M. Utility of lacrimal caruncle infrared thermography when monitoring alterations in autonomic activity in healthy humans. *Eur J Appl Physiol* (2019) 119:531–8. doi: 10.1007/s00421-018-4041-6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a shared affiliation, though no other collaboration, with several authors MM, GO, AT at the time of review.

ORIGINAL RESEARCH
published: 11 September 2019
doi: 10.3389/fpsyt.2019.00636

Evaluative Observation in a Concealed Information Test

Wolfgang Ambach[1]*, *Birthe Assmann*[2], *Blanda Wielandt*[1] *and Dieter Vaitl*[3,4]

[1] Department Clinical and Physiological Psychology, Institute for Frontier Areas of Psychology and Mental Health, Freiburg, Germany, [2] Lower Saxony Institute of Early Childhood Education and Development, Osnabrück, Germany, [3] Institute for Frontier Areas of Psychology and Mental Health, Freiburg, Germany, [4] Bender Institute of Neuroimaging, University of Giessen, Giessen, Germany

The Concealed Information Test (CIT) is a valid method to detect hidden knowledge by means of psychophysiological measures. Concealing information is always a social behavior; yet, the role of social aspects has barely been investigated in recent CIT research favoring standardized, computer-based experiments. Evaluative observation is known to influence social behavior as well as physiological measures; examining the impact of evaluative observation on physiological responding in a CIT is the aim of this study. Sixty-three students completed a mock-crime and then underwent a CIT. In a between-subjects manipulation, half of the participants were observed through a camera and were faced with the real-time video of the experimenter watching them while completing the CIT. The other half completed the CIT without observation and video. Electrodermal activity, respiration line length, phasic heart rate, and finger pulse waveform length were registered. A specific questionnaire captured the individual fear of negative evaluation. Typical differential CIT responses occurred in both groups and with each measure. Contrary to expectations, differential CIT responses did not differ between groups. No modulatory influence of the fear questionnaire score on physiological responding was found. A ceiling effect, involving high attention and high motivation to avoid detection as well as high arousal in both groups due to the CIT procedure per se is discussed as explanation for these results, while the independence of the orienting reflex of social and motivational influence appears less likely in the light of previous literature.

Keywords: Concealed Information Test, deception, mock crime, social stimuli, evaluative observation, orienting

OPEN ACCESS

Edited by:
Giuseppe Sartori,
University of Padova, Italy

Reviewed by:
Pasquale Capuozzo,
University of Padova, Italy
George Visu-Petra,
Babeş-Bolyai University, Romania

****Correspondence:***
Wolfgang Ambach
ambach@igpp.de

Specialty section:
This article was submitted to Forensic Psychiatry, a section of the journal Frontiers in Psychiatry

Received: *25 October 2018*
Accepted: *06 August 2019*
Published: *11 September 2019*

Citation:
Ambach W, Assmann B, Wielandt B and Vaitl D (2019) Evaluative Observation in a Concealed Information Test Front. Psychiatry 10:636. doi: 10.3389/fpsyt.2019.00636

INTRODUCTION

The Concealed Information Test

Among the manifold manifestations of deception, the concealment of information is a common type of deceptive behavior. For example, a culprit may exhibit this specific social behavior to appear uninvolved in a specific criminal act. The Concealed Information Test (CIT) is a scientific psychophysiological method to detect such intentionally hidden information. A systematic interrogation is hereby combined with a multi-channel physiological measurement. The CIT relies on the assumption that, if the examined subject is guilty, his or her physiological responses will differ between crime-related and crime-irrelevant information (1; for an overview, see 2), whereas an innocent examinee will not differentiate between crime-relevant and irrelevant details. The CIT consists of several multiple-choice questions each referring to another detail of the crime under investigation. Typically, there are four to five answer alternatives to each question; only one of these alternatives, the probe, refers to the critical detail. For example, if an envelope was stolen out of an office, a typical CIT question could be: "An office requisite has been stolen. Is this the stolen object?" This question is combined with a sequence of five pictures representing the respective answer alternatives, e.g., a picture of a) a pencil sharpener, b) an envelope, c) a highlighter, d) a stapler and e) a writing pad. In this example, the picture of the envelope (b) is the probe item; the other items are referred to as irrelevant. It is assumed that only guilty (knowledgeable) subjects will exhibit a different physiological response to the probe item. Unknowledgeable subjects, in contrast, will not exhibit different physiological responses to probe vs. irrelevant items; their response pattern will be unsystematic. The high validity of the CIT in the differentiation between guilty and innocent subjects was proven by a multitude of studies, as summarized by Ben-Shakhar and Elaad (3) or Meijer et al. (4).

The theory of the CIT strongly relies on cognitive aspects such as the orienting response (5, 6). Besides the orienting response, influences of motivation and emotion are discussed to play only a moderating role in the CIT when conducted in the laboratory. These influences might be greater when the CIT is applied in the field (2). Until now, only few studies focused on the influence of social factors, such as attention, intention, motivation, or emotion, on the CIT performed in the laboratory or in the field. Different authors have shown that motivation, intention, and emotion can affect response differences in the CIT (2). As an example, physiological response differences in the CIT have been shown to be enhanced by demanding a deceptive answer from the examinee, rather than demanding a truthful answer or no answer at all [e.g. Refs. (7, 8)], on the motivation to inhibit one's own physiological arousal (9), and on the subject's belief in the effectiveness of the physiological detection (10). Social aspects in turn are likely to have an impact on these factors. It seems worthwhile to study social and motivational influences on the CIT, as well as possible mediators of these influences, with more intensity and with a stronger focus on the effective mechanisms.

Social Aspects and the CIT

Physiological functions and physiological responses are always influenced by the social context and by social stimuli. Zajonc (11) showed in his 'social facilitation theory', that the mere presence of another person enhances the physiological level of arousal. Perception of social gaze, which is evolutionary meaningful (12) is accompanied by specific subjective sensations and neurophysiological reactions (13). With respect to the CIT, which relies on physiological responding to specific stimuli, social influence has barely been investigated. Presumably, due to the desire to standardize CIT experiments as far as possible and also driven by the increasing use of computers in experiments, social aspects have played only a minor role in past CIT research. However, concealing knowledge from an interrogator is always a social act. Some decades ago, the social influence on the CIT was investigated in a small number of studies (14, 15). For instance, ethnic differences between subject and investigator, which were known to influence physiological parameters in general (16), enhanced physiological response differences in the CIT (15). Orne (14) broached the issue of possible differences between a friendly and an antagonistic investigator. However, these studies did not report on systematic experimental manipulations of social stimuli, social interaction, or social roles (17).

Particularly in real-life CIT examinations, as applied at a large scale and on a daily basis in Japan, social influences are inevitable and extensive (18). Emotion and motivation are supposed to be intense in an interrogation referring to a real crime. Elements of social interaction between examiner and examinee preceding and during the CIT could have an additive impact on a suspect's motivation to remain undetected, on the intention to conceal, and on emotions like fear during the CIT. Notably, the contact between suspect and examiner in the CIT includes a wide spread of social elements: Social presence, eye contact, speech, sight, gestures, verbal interaction, and observation are just some examples. In the real-life CIT, these elements always occur and co-act in varying and hard to specify combinations. This makes it difficult to investigate them element by element. Experimentally varying single components of social influence in the laboratory is the best way to identify the components actually effective. Interestingly, a real-life interaction between examiner and examinee may vary in its positive vs. negative emotional impact. The same may hold for the positive vs. negative aspects of being observed and evaluated in real-life CIT examinations.

The influence of a first, specific set of social stimuli on physiological responding in the CIT was investigated in an earlier study (19). Employing the "voice of an interrogating person" asking the CIT questions, combined with presenting the image of the "face of an interrogating person" during the questions, lead to increased response differences. It remained open whether it was the acoustic questioning or the presented face, or their combination, that impacted the examinee's physiological responses. Further, if a presented face in fact co-determines physiological responding in the CIT, then the specific connotation of that face for the subject becomes a central question. The impact of facial emotional expressions representing a virtual investigator on reaction times in a CIT was examined by Varga et al. (20). Interestingly, the mere presence of a virtual investigator's neutral face led to an increase in overall but not differential reaction times, whereas emotional expression in this face was found to differentially increase reaction times to probe items. Most likely, a presented face or even a presented pair of eyes (21, 22) induces a feeling of being watched, controlled, or judged by another person (23), which should facilitate socially approved behavior while disapproved behavior, like deception, should become more difficult. For the present study, we focused on varying the "watching," more precisely the "evaluative observation" component of social interaction.

Evaluative Observation

Evaluative observation in a social context denotes a situation in which one person, while watching another, evaluates the behavior and performance of the other. Chapman (24) showed that the awareness of being observed evaluatively enhances arousal and raises the muscular tone. Cottrell et al. (25) showed that performing a task in front of an audience increases a person's physiological arousal. Additionally, the presence of an audience enhanced dominant responses but the mere presence of others did not, which is contrary to Zajonc's social facilitation theory (11).

Other studies (26, 27) showed that anticipated evaluation of performance facilitated dominant responses but evaluation without awareness of being observed evaluatively did not. The finding that evaluative observation (with awareness) exerted social impact independently from social presence illustrates the importance and the possibility of experimentally separating individual components of social influence.

The same was hypothesized with respect to the CIT: We aimed to investigate the impact of evaluative observation on physiological responding in a CIT independently of the presence of another person. To create an environment, in which the subjects were observed evaluatively without the presence of another person, we decided to use a video camera. In this way, the experimenter could observe the subjects without being present in the experimental cabin. To ensure that the subjects were fully aware of being watched and critically evaluated, we installed, in addition to the according instructions, a second camera and presented a live video of the observing experimenter situated outside, on the subject's screen.

Building upon the motivational impairment hypothesis (see 3, 28), we supposed that an examinee would show greater physiological response differences in the CIT when observed evaluatively. If this holds true, it might, on the one hand, help to enhance detection accuracy, and on the other hand, it might contribute to CIT theory by shedding light on the interplay of social influences, emotional-motivational factors, and physiological responding in the CIT.

To explore whether the impact of evaluative observation varies between subjects according to specific traits, we included the German version of the FNE (Fear of Negative Evaluation Scale) (29), called SANB (Skala Angst vor Negativer Bewertung; 30), which captures an individual's fear of negative evaluation that is commonly understood as the cognitive component of social phobia. Following the finding of increased heart rate (HR) and palmar sweating in socially relevant situations (31), and also in view of earlier studies on trait influences on skin conductance (32) and reaction times (33) in the CIT, we supposed that people exhibiting a greater fear of negative evaluation would show not only increased overall responses but, due to a motivational impairment effect, also greater response differences between probe and irrelevant item types in the CIT. Additionally, we expected that people with a high fear of negative evaluation would also worry more about being detected, which in turn would facilitate detectability (34).

Aim of the Present Study

1. This study focused on the influence of evaluative observation on physiological responding in a CIT. Two variants of the CIT, one condition, "with observation" and a second condition "without observation," were manipulated between-subjects. Differential physiological responses were compared between groups. Greater differential responding was expected in the condition with evaluative observation for all physiological measures.
2. The study further investigated whether the differential physiological responses in the CIT are moderated by the individual fear of negative evaluation; therefore, the SANB questionnaire was included. With higher SANB scores, an enhancing influence of evaluative observation on physiological response differences was expected.

MATERIALS AND METHODS

Subjects

Sixty-three healthy students (31 males, 32 females; mean age, 22.8 ± 2.4 years) voluntarily participated in the study. They were paid 12 Euros, with an additional incentive of 3 Euros. Data from one subject were discarded from evaluation because of a technical failure. The ethics committee of the German Psychological Society (DGPs) confirmed that the study met all ethical requirements (ID: WA122013).

Procedure and Design

The experiment consisted of two parts, a mock-crime in an "office room" and a detection procedure in the "laboratory," each guided by a different experimenter. The first experimenter welcomed the participant, explained the procedure, and accepted written informed consent. First, participants were told they had to perform a "special task" in an "office room," for which they were asked to choose one out of five rolled-up instruction documents with different instructions. In fact, all documents contained the same mock-crime instruction. After the mock crime, participants walked over to the "laboratory" where the second experimenter expected them. This experimenter, in fact blind with respect to the mock-crime objects a particular participant had handled in the first part, was introduced as the person responsible for "detecting whether the subjects had stolen something in the office room or not." After completing the CIT and a subsequent memory test, subjects filled in the SANB questionnaire (referring to trait anxiety of negative evaluation) before they were debriefed and released. Payment included the incentive of 3 Euros, regardless of a participant's responding in the CIT.

Subjects were randomly assigned to either of two groups: Half of the subjects (i.e., the observation group; 31 valid data sets) underwent a CIT with particular emphasis on the fact that the experimenter was evaluatively observing them throughout the CIT; the other half (i.e., the no observation group; 31 valid data sets) underwent a CIT without evaluative observation.

For the observation group, a conspicuous camera was placed on top of the participant's monitor, in addition to the inconspicuous camera generally surveying the experimental room from a corner. Written instructions stated that the experimenter's aim was to find out "by means of precise observation *via* cameras, and by physiological measurement" whether the participant had stolen items from the office room or not. In three instances in advance of running the experiment, the experimenter himself explicitly mentioned these cameras and the importance of observation. To further direct attention to evaluative observation, the monitor-placed camera was adjusted again immediately before starting the CIT. Moreover, participants in the observation group viewed—between two subsequent item presentations—a full-screen live video of the experimenter critically watching them from outside and making written notices. The experimenter, while being watched *via* camera by the subjects, behaved in a predefined manner which emphasized attentive observation while excluding talking, laughing, gazing straight into the camera lens, as well as direct responding to the subject's behavior. In fact, the experimenter filled in a score sheet continuously during the CIT, according to his or her conjectures about the items stolen by the individual participant.

In the no observation group, only the inconspicuous camera in the corner was installed, which was indispensable for conducting the experiment according to ethical standards; this camera was only briefly mentioned to the subjects as warranting they were in good hands. The experimenter's aim was explained as "finding out by means of physiological measurement" whether the participant had stolen items from the office room or not.

Mock-Crime Scenario

Alone and unwatched in an office room of the institute, subjects unrolled the document they had obtained from the first experimenter. By instruction, they had to remove ("steal") nine objects from this room after having extensively viewed each of them. The choice of the nine objects, one from each category, was randomized and balanced across subjects. The nine object categories, each comprising five objects, were: key pendants, kitchen objects, boxes, office materials, cosmetics, wooden toy fruits, drink packages, playing cards, and plastic flowers.

Subjects were advised to collect all nine items in a suitcase, which they should keep closely to themselves throughout the remaining experiment. An amount of 3 Euros was hidden in one of the stolen objects (a box); later, this served as an incentive to "remain undetected" in the subsequent CIT.

Concealed Information Test

The second experimenter conducted the CIT in the laboratory. For the so-called "physiological investigation," recording devices were attached first. The CIT consisted of nine blocks referring to the nine item categories (e.g., key pendants, cosmetics). Each block comprised one question with five answer alternatives: the probe ("stolen") item of each category and four corresponding irrelevant items, which were all unknown to the subjects.

CIT questions were presented acoustically with a pre-recorded male voice *via* speakers; "voice" questioning had turned out earlier as the more efficient variant (19). Different from the typical CIT wording, an active questioning format was chosen, which had also shown itself more efficient (35): Questions were, e.g., "Did you steal this cosmetic product from the administration room?" Each question was presented five times in sequence, each time directly followed by a different picture of one of the five answer alternatives.

The first item presented for each question served as buffer item; the according trials were discarded from analysis. Preceding each block, two neutral items were presented as distractors. The according questions referred to everyday objects that had to be identified (e.g., "Is this a slide projector?"). The two questions had to be answered correctly, one with "yes" and the other with "no" (in a pseudorandomized sequence) to keep participants alert and prevent them from answering automatically with "no" throughout the CIT. Responses to these neutral questions were not evaluated. Together with the two neutral questions preceding each category, the entire procedure resulted in a total of 63 item presentations. The main run was preceded by a training run consisting of two blocks, each with five neutral items.

Figure 1 shows the intra-trial development of the screen for both groups. Acoustic question presentations were accompanied by 3.5 s of blank screen in both groups. Item pictures were presented for 10 s foveally on a 19-inch monitor at a distance of 90 cm, followed by the equally distributed 4.5 to 6.5 s display of either a live video showing the observing experimenter (in the observation group), or a blank screen (no observation group).

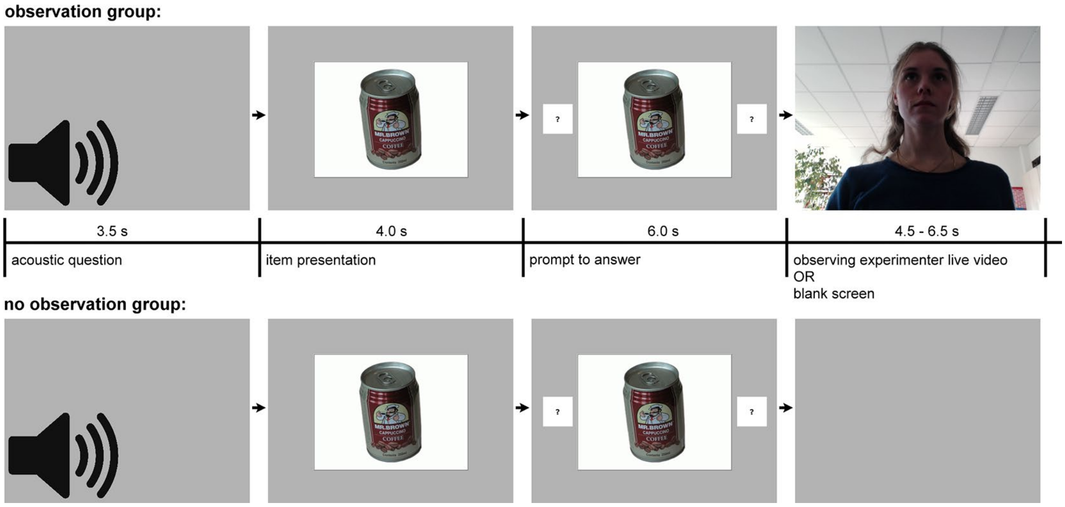


FIGURE 1 | Temporal course of question and item presentation in the Concealed Information Test in the observation group and the no observation group. The acoustic question was presented first, CIT items appeared 3.5 s later, and fields with question marks succeeded 4.0 s thereafter. After the key press, a "yes" or "no" text (reflecting the subject's answer) replaced the question marks.

Picture size was 14.3° by 10.7° of visual angle for the CIT items. Four seconds after an item was presented, two indication fields containing question marks appeared on either side of the item picture, this prompted the subjects to answer. Answers had to be given as quickly as possible by pressing one of the two response keys and by vocally responding with "yes" or "no." Key assignment was balanced across subjects. Following the answer, the given "yes" or "no" replaced the question marks and remained visible on the screen as long as the item was presented. Subjects were told to hide their knowledge about the objects that had been stolen from the administration room, i.e., to deny all knowledge about probe items.

After subjects were disconnected from the leads, they underwent a memory test: All five pictures of each category were presented on the screen simultaneously, one item category after the other; subjects were asked to identify the item they had stolen within each category.

Physiological Measures

The physiological recordings took place in a dimly lit and electrically and acoustically shielded experimental chamber (Industrial Acoustics GmbH, Niederkrüchten, Germany). Subjects sat in an upright position so that they could comfortably see the monitor and reach the keyboard. Temperature in the cabin was set to 21°C at the beginning of the first run, with an increase of maximum 2°C throughout the course of the experiment.

Skin conductance, respiratory activity, electrocardiogram (ECG), and finger plethysmogram were registered. Physiological measures were A/D-converted and logged by the Physiological Data System I 410-BCS manufactured by J&J engineering (Poulsbo, Washington). The A/D-converting resolution was 14 bit, allowing skin conductance to be measured with a resolution of 0.01 µS. All data were sampled with 510 Hz. Triggers indicating question onsets were registered with the same sampling frequency.

For skin-conductance recordings, standard Ag/AgCl electrodes (Hellige; diameter 0.8 cm), electrode paste of 0.5% saline in a neutral base (TD 246 Skin Resistance, Mansfield R&D, St. Albans, Vermont), and a constant voltage of 0.5 V were used. The electrodes were fixed at thenar and hypothenar sites of the nondominant hand. For registration of respiratory activity, two PS-2 biofeedback respiration sensor belts (KarmaMatters, Berkeley, California) with a built-in length-dependent electrical resistance were used. They were fixed at the upper thorax and the abdomen. ECG was measured with Hellige electrodes (diameter, 1.3 cm) according to Einthoven II. Finger pulse signal was transmitted by an infrared system in a cuff around the middle finger of the nondominant hand.

Behavioral Measures

Subjects responded verbally as well as by pressing a key. Key presses indicating "yes" or "no" answers were time-logged, synchronized with the physiological measures and stored on the stimulus-presenting computer. Importantly, answers were delayed by 4 s in this study. After this delay, most stimulus processing and answer preparation can be assumed to be completed; in addition, it is rather easy to perform strategic manipulations by voluntarily controlling reaction speed after the delay. Therefore, behavioral data were not analyzed. CIT questions with at least one item answered incorrectly were discarded from the analysis, which resulted in a loss of 1.4% of the data.

Questionnaire

As the last part of the experiment, participants filled in the SANB questionnaire. It comprises 20 items to assess the individual fear of negative evaluation as a trait variable. The sum scale was calculated from the raw data according to Vormbrock and Neuser (30).

Data Processing

Skin conductance data from four subjects (two from the observation group, two from the no observation group) had to be discarded from the analysis because of electrodermal non-responding. Skin conductance reactions were assessed by a computerized method [see Refs. (7, 19, 36)] based on the decomposition of overlapping reactions as proposed by Lim et al. (37). This method was chosen, because two subsequent physiological reactions occurred with a short delay, due to the delay of 4 s between a question and the prompt to answer. With short interstimulus intervals, conventional trough-to-peak evaluation is inadequate (38) because the first of two reactions causes a diminishing bias in the estimation of the second one. The size of this bias is determined by the size of the first reaction and by the time interval between both reactions. Decomposition aims at overcoming this problem of overlapping electrodermal responses.

After optimizing model coefficients for each subject, all trials were evaluated by decomposing electrodermal activity (EDA) by use of each subject's individual model coefficients. Then, magnitudes of all EDA responses that were elicited within a time window of 0.5 to 4.5 s after item presentation were additively combined to a first response (EDA_1). Magnitudes of EDA responses, which began between 4.5 and 8.5 s after item presentation, i.e., between 0.5 and 4.5 s after the subjects were prompted to answer, were additively combined to a second response (EDA_2). In addition, a combined response measure (EDA_sum) was calculated by adding both components per trial. For each time window, the decomposed responses were transformed into their equivalent amplitudes in µS according to each subject's individual electrodermal response template.

Respiratory data were low-pass filtered (10 dB at 2.8 Hz); respiration line length (RLL) was automatically computed over a time interval of 15 s after trial onset. The RLL measure integrates information about frequency and depth of respiration. The method was derived from Timm (39) and modified by Kircher and Raskin (40). Respiratory data from nine subjects (four from the observation group, five from the no observation group) were discarded due to sensor problems. For analysis, raw scores from both respiratory channels were averaged.

ECG data obtained from one subject (from the observation group) had to be excluded from analysis because of technical failure. After notch filtering at 50 Hz, R-wave peaks were automatically detected and visually controlled. The R-R intervals were transformed into HR and real-time scaled (41). The HR during the last second before trial onset served as pre-stimulus baseline. The phasic HR (pHR) was calculated by subtracting this baseline value from each second-per-second poststimulus value. To extract the trial-wise information of the phasic HR, the mean change in HR within 15 s after trial onset, compared to the prestimulus baseline, was calculated [see Refs. (42, 43)].

Finger pulse waveform length (FPWL) data from four subjects (three from the observation group, one from the no observation group) had to be discarded from analysis because of insufficient signal quality. The FPWL within the first 15 s after trial onset was calculated from the finger pulse waveform and then subjected to further analyses (44). It comprises information about both HR and pulse amplitude.

To compare indicators of arousal between groups, we additionally computed the individual averages of non-standardized skin conductance level (SCL) and HR at trial onsets. The SCL and HR data were averaged over the last second before the onset of a CIT question, i.e., 3.5 to 4.5 s before item onset.

A within-subject standardization of measured values has been proposed by Lykken and Venables (45). Here, according to Ben-Shakhar (46), Gamer et al. (47), and Gronau et al. (48), the physiological measures are z-transformed for each subject and for each data channel. All probe and irrelevant trials (but neither neutral trials nor the first trials of each stimulus category) were used to calculate individual means and standard deviations. The z-transformed values were used in subsequent statistical analyses.

Statistical Analysis

Statistical analyses were performed with SYSTAT, Version 13 (SYSTAT Software, Inc., Monte Carlo).

For each physiological measure, mean responses to probe vs. irrelevant items were compared using one-tailed t-tests (matched samples) separately for observation and no observation group. An additional t-test (two-tailed, independent samples) was performed to test whether the probe-minus-irrelevant response differences differed between groups. Cohen's d was calculated as estimate of effect size (49, 50). To test for group effects on tonic physiological measures of arousal, means of SCL and HR were determined in the second preceding the acoustic question presentations, i.e., from 4.5 to 3.5 s before item onsets. To test between groups, two-tailed independent-samples t-tests were conducted on the basis of raw values. Significance level of all analyses was set to 0.05.

For identifying the fear of negative evaluation as a moderator of differential physiological responding in the CIT, correlation coefficients were calculated for the individual SANB sum scores and the individual standardized probe-minus-irrelevant response differences for each physiological data channel. Testing whether the individual SANB score is moderating the influence of evaluative observation on differential responding in the CIT was later dropped from analysis, after the influence of evaluative observation, *per se*, turned out insignificant.

RESULTS

Memory Test

In the memory test, 98.6% of the probe items were identified correctly (97.8% in the observation and 99.3% in the no observation group). Categories with false identification of the probe item were entirely discarded from evaluation.

Overview of Psychophysiological Measures

Preceding data standardization and test statistics, descriptive statistics based on raw scores are presented. **Table 1** summarizes means and standard errors of means of raw scores for each data channel separately for both groups.

Figure 2 illustrates the differential responses to probe vs. irrelevant items for both groups. Response differences (z-scores) between probe and irrelevant trials are depicted for each of the physiological measures.

Skin Conductance

Figure 3 shows the averaged intra-trial course of skin conductance depicting grand means for trials with probe and irrelevant items separately for both groups. The grand means show two strong EDA response components with an onset and peak asynchrony of 4 s, which is in accordance with the 4-s delay between item onset and prompt to answer. Response amplitudes to probe items exceeded those to irrelevant items by far in both groups, with no apparent difference between groups. The additional EDA response, which was observed 3.5 s before the response to item onset, can be ascribed to the onset of the acoustic question presentation (which was the same for all items of a category).

EDA_1 responses were greater to probe than to irrelevant items in the observation group (t_{26} = 7.41; $p < 0.001$; d = 1.43) as well as in the no observation group (t_{27} = 9.85; $p < 0.001$; d = 1.86). The between-groups t-test for probe-minus-irrelevant EDA_1 response differences did not reveal a group difference (t_{53} = −1.41; $p > 0.1$).

EDA_2 responses were greater to probe than to irrelevant items in the observation group (t_{26} = 4.19; $p < 0.001$; d = 0.81) as well as in the no observation group (t_{27} = 5.07; $p < 0.001$; d = 0.96). The between-groups t-test for probe-minus-irrelevant EDA_2 response differences did not reveal a group difference (t_{53} = 0.07; $p > 0.1$).

TABLE 1 | Means and standard errors of means (SEM) of raw scores for each data channel. Responses to *probe* and *irrelevant* items are listed separately for *observation* and *no observation* group.

	***Observation* group**				***No observation* group**			
	***Probe* items**		***Irrelevant* items**		***Probe* items**		***Irrelevant* items**	
	Mean	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**
EDA_1 [nS]	262	47	139	24	298	60	132	23
EDA_2 (nS]	261	52	178	35	295	54	209	31
pHR [1/min]	−4.13	0.59	−1.58	0.36	−3.38	0.57	−1.59	0.29
RLL [arb. units]	2218	212	2517	229	2278	199	2645	204
FPWL [arb. units]	1416	149	1603	158	1702	190	1947	210

The physiological measures were first electrodermal response component (EDA_1), second electrodermal response component (EDA_2), phasic heart rate (pHR), respiration line length (RLL), and finger pulse waveform length (FPWL).

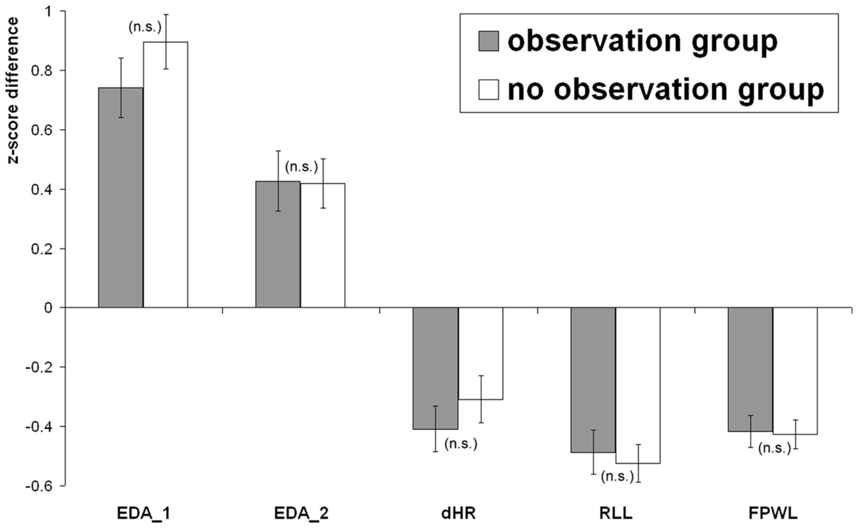


FIGURE 2 | Differential responses (z-scores) to probe vs. irrelevant items: For the observation and the no observation group, standardized response differences are depicted for first electrodermal reaction (EDA_1), second electrodermal reaction (EDA_2), phasic heart rate (pHR), respiration line length (RLL), and finger pulse waveform length (FPWL). Error bars represent the standard error of the mean; the level of significance of the group difference is indicated by "n.s." (not significant; p > 0.05).

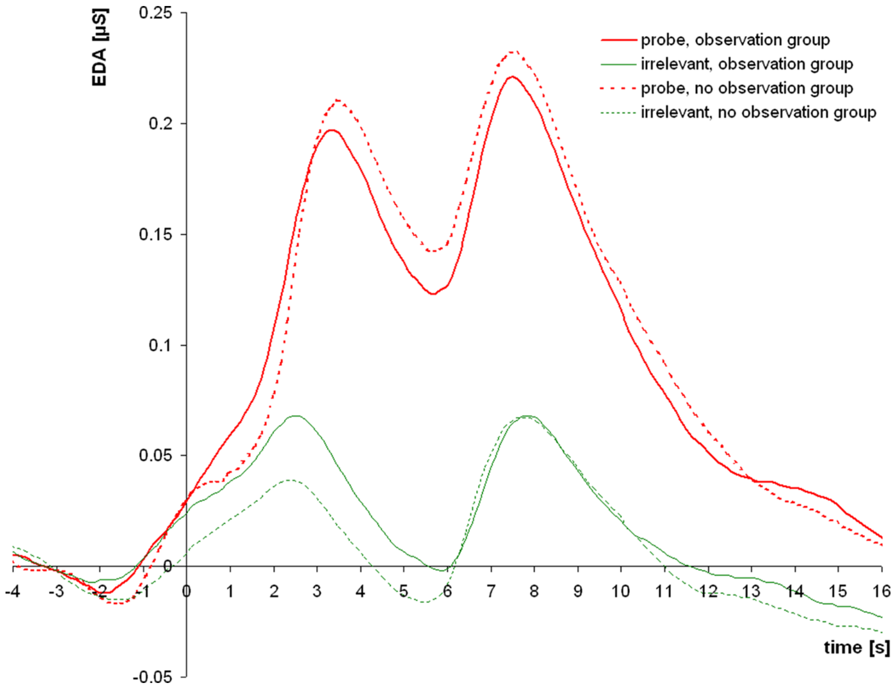


FIGURE 3 | Grand means of skin conductance responses to probe and irrelevant items for the observation group and the no observation group. After a small initial response to the acoustic question presentation, two subsequent electrodermal responses of interest (EDA_1 and EDA_2) follow the item presentation and the prompt to answer, respectively.

Both EDA components were then additively combined in a single measure: EDA_sum. EDA_sum responses were also greater to probe than to irrelevant items in the observation group ($t_{26} = 6.77$; $p < 0.001$; $d = 1.30$) as well as in the no observation group ($t_{27} = 8.38$; $p < 0.001$; $d = 1.58$). Probe-minus-irrelevant response differences for EDA_sum did not differ between groups ($t_{53} = -1.00$; $p > 0.1$).

Respiration

RLL values were smaller after probe than after irrelevant items in the observation group ($t_{26} = -6.45$; $p < 0.001$; $d = -1.24$) as well as in the no observation group ($t_{25} = -8.12$; $p < 0.001$; $d = -1.59$). Probe-minus-irrelevant response differences for RLL did not differ between observation and no observation group ($t_{51} = 0.38$; $p > 0.1$).

HR

HR decelerations were more pronounced after probe than after irrelevant items in the observation group ($t_{29} = -5.25$; $p < 0.001$; $d = -0.96$) as well as in the no observation group ($t_{30} = -3.84$; $p < 0.001$; $d = -0.69$). Probe-minus-irrelevant response differences for pHR did not differ between observation and no observation group ($t_{59} = -0.90$; $p > 0.1$).

Finger Pulse

FPWL values were smaller after probe than after irrelevant items in the observation group ($t_{27} = -7.69$; $p < 0.001$; $d = -1.45$) as well as in the no observation group ($t_{29} = -8.68$; $p < 0.001$; $d = -1.58$). The between-groups t-test for probe-minus-irrelevant FPWL differences did not reveal a difference between groups ($t_{56} = 0.15$; $p > 0.1$).

Tonic Measures of Arousal

As indicators of arousal, SCL and HR at trial onsets were analyzed. **Figure 4** depicts grand means of SCL (top) and HR (bottom) in the course of the experiment; data were collapsed over trials within categories; the first two data points of each subplot correspond to the two categories of the training run.

SCL appeared higher in the no observation group (4.64 ± 1.59 µS) than in the observation group (4.28 ± 1.66 µS).

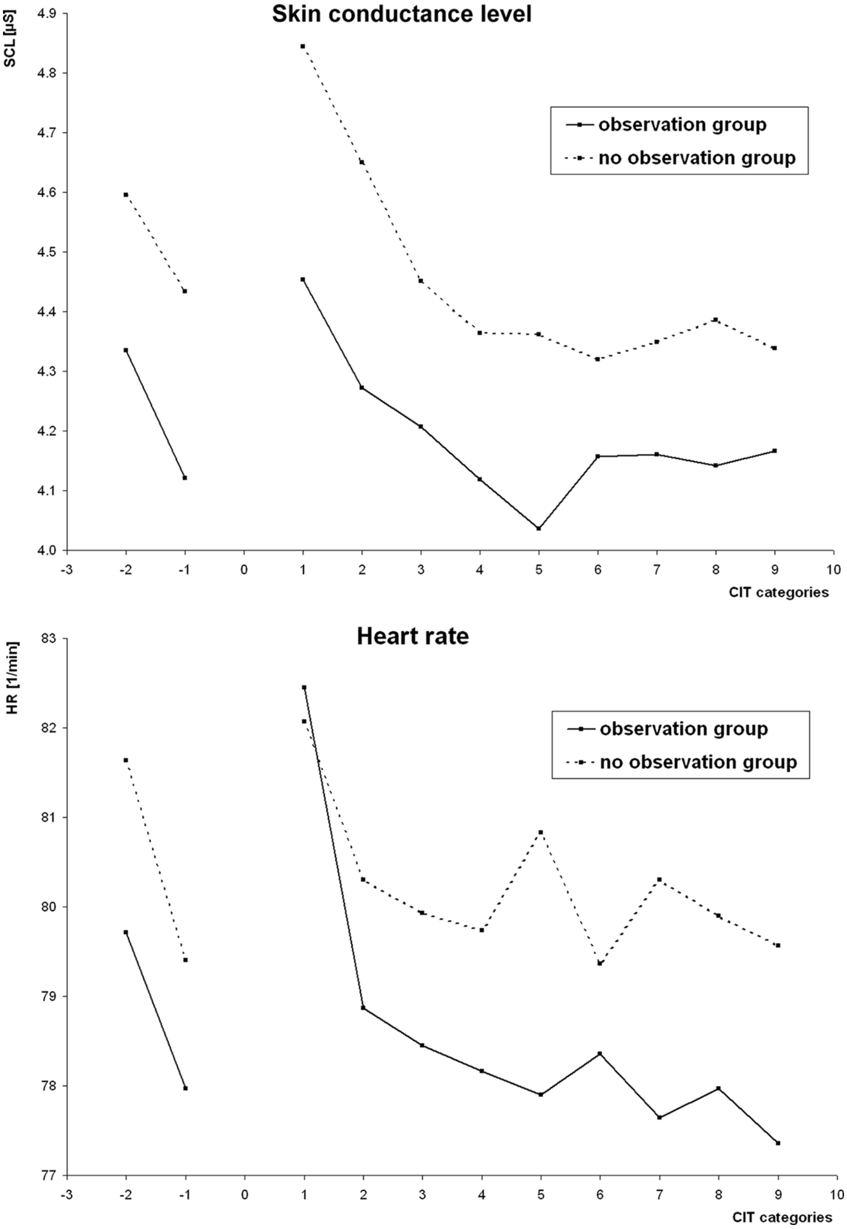


FIGURE 4 | Skin conductance level (SCL) and heart rate (HR) as tonic measures of arousal. Grand means of raw values in the course of the experiment are depicted for the observation group and the no observation group. Data were collapsed within item categories; the first two data points of each plot reflect the training run.

This was contrary to the expectation; yet, this difference was not statistically significant (t_{53} = 0.83, p > 0.1). Inspection of the raw data, see **Figure 4** (left), indicated that this result was due to an initially higher EDA level in the no observation group that was preserved throughout the entire examination. HR appeared higher in the no observation group (80.01 ± 9.14 bpm) than in the observation group (78.79 ± 11.24 bpm); yet, this difference was not statistically significant (t_{59} = 0.47, p > 0.1). The visual impression of a larger HR decrease over the experiment in the no observation group, see **Figure 4** (right), was not confirmed in a 2 (groups) × 9 (categories) ANOVA (F8 = 1.05; p > 0.1).

SANB Questionnaire

The individual SANB sum scores were calculated. SANB data from the one participant precluded from physiological analysis were treated as missing data.

SANB sum scores were 45.68 ± 10.15 (median, 46) for female participants (n = 31) and 41.55 ± 9.58 (median: 41) for male participants (n = 31). This difference was not statistically significant (t_{60} = −1.65, p > 0.1). Scores for males as well as females markedly exceeded those reported originally for a student sample (median for females: 37, for males: 36; standard deviation unreported) (30). Data from males and females were then collapsed for further analyses.

SANB scores did not differ between observation and no observation group (43.84 ± 9.83 and 43.39 ± 10.34, respectively; t_{60} = −0.176, p > 0.1).

Correlation coefficients between individual SANB sum scores and probe-minus-irrelevant response differences for each physiological measure (EDA_sum, pHR, FPWL, and RLL) were calculated across groups as well as separately for the no observation and the observation group. Correlation coefficients for RLL were 0.31 across groups, 0.03 for the no observation, and 0.51 for the observation group. Here, the correlation across groups did significantly differ from zero before but not after Bonferroni correction for multiple testing (p = 0.024 and p > 0.1, respectively), whereas the correlation for the observation group did significantly differ from zero before as well as after Bonferroni correction (p = 0.003 and p = 0.036, respectively). Thus, differential respiratory responding in the CIT was found to be moderated by SANB sum scores in the observation group. For the other physiological measures, none of the corresponding correlations exceeded ±0.15; correspondingly, none of the correlations did significantly differ from zero (all p > 0.1, uncorrected).

DISCUSSION

The present study followed the idea that being observed evaluatively by an examiner during a CIT might enhance the examinee's differential physiological responsiveness to probe vs. irrelevant items. A CIT condition "with observation," in which subjects were observed *via* a conspicuous camera and presented with a live video of the experimenter watching them, was contrasted with a CIT "without observation." A questionnaire on fear of negative evaluation was administered to explore a specific trait as possible moderator variable.

Differential Responding in the CIT

Response differences between probe and irrelevant items were found for the electrodermal, the respiratory, and the cardiovascular measure. The observed effect sizes were in line with the large physiological response differences characteristic for the CIT (4). As an additional detail adopted from earlier studies, two components of electrodermal responding were separated, one after item presentation and one after the prompt to answer. Both components, with the first component exceeding the second with respect to effect size (cf. 7), mirrored the typical CIT response pattern. In line with a number of earlier studies (19, 44, 51), FPWL yielded larger effect sizes than pHR and RLL, but did not exceed those of EDA in this study. Yet, it has to be noted that effect sizes obtained after a restriction of recorded data sets to "valid" data sets, e.g., due to electrodermal non-responding, frequent extrasystoly, or insufficient pulse signal quality, should be interpreted with caution.

The Impact of Evaluative Observation

The two CIT conditions with and without evaluative observation were administered in a between-subjects manipulation. Features differing critically between experimental conditions were the written CIT instructions, verbal instructions, and non-verbal behavior by the experimenter, presence of a conspicuous camera on top of the monitor, and presentation of a live videostream of the experimenter watching.

Contrary to the *a priori* expectation, none of the physiological measures exhibited a statistically significant enhancement of probe-vs.-irrelevant response differences in the observation group. The observed, insignificant group differences in the mean differential responses did, with the exception of pHR, not even meet the predicted direction. Hence, the main alternative hypothesis for this study, i.e., the assumption of an enhanced CIT effect under evaluative observation, was rejected.

Explanations, Implications, and Limitations

At a first glance, one might suspect that the experimental manipulation was insufficient. The participants' motivation to cope with the test and their prospect of success, known to affect differential responding in the CIT (52, 53), and also their attention during the CIT, might not have differed sufficiently between conditions. Tonic physiological measures at stimulus onsets, i.e., tonic HR and SCL, were analyzed to test for differences in arousal between conditions. The lack of a statistically significant difference in these measures between conditions principally supports the idea of an insufficient experimental manipulation. Also, the video of the experimenter was shown to participants only in between CIT questions, so that the time window of its impact may be discussed. On the other side, subjects' verbal reports after the experiment (gathered unsystematically during debriefing) underlined that the experimental manipulation was visually impressive and psychologically effective. Particularly the real-time view of the experimenter watching was described by participants as challenging, reminding of the presence of an opponent, and thereby enhancing the motivation to hide the critical knowledge "without letting anything show." The influence of arousal on the CIT effect, which gave rise to the analysis of the two tonic measures in this study, was recently summarized by Klein Selle et al. (9). Given that no arousal difference between groups was found in this study, also the interpretation of other group effects deserves caution.

The psychological difference between conditions might also have been diminished by the fact that there was always one camera present in the room, even in the condition without observation. This camera, which was indispensable for the ethical correctness of the experiment, was not conspicuous, not positioned centrally, and mentioned only briefly by the experimenter as "necessary to make sure you are alright." Nevertheless, this camera might have been sufficient to make participants even in the no observation group feel that they were watched critically throughout the test. Although conceivable, this possible explanation did not find support in participants' later verbal reports. In addition, the difference between groups in the written instructions might have led to instruction-induced effects differing between groups. Given that the interaction between participant and examiner is always complex, the different instructions might have confounded possible group effects.

Next, subjects in both experimental groups were motivated to pass the test without being detected. The incentive of three Euros probably strengthened this aim. Yet, experimental participants generally tend to exhibit such motivation, rather than absolving the CIT incuriously. Thus, participants of either group were highly motivated not to be detected in the CIT. It can then be questioned to what extent such high motivation can be even further enhanced by additional observation and evaluation. In case of a general high level of motivation, a ceiling effect is well conceivable, which might contribute to explaining the lack of a difference between conditions in differential physiological responding.

A similar kind of ceiling effect might be caused by the physiological recording. An examinee might fear that his feelings will be uncovered by this recording, independently of his control. The rather uncommon experience of being attached to a physiological recording device while being questioned might lead subjects to feel like "being watched" and "being evaluated" intensively, even without camera. If so, then it can be questioned to what extent the feeling of being observed evaluatively can still be enhanced by an additional, "visual" observation.

In addition, it has to be noted that CIT questions were presented acoustically in both conditions. In contrast to the text variant of question presentation, acoustic presentation was thought to entail greater physiological response differences (19), presumably by inducing a more "social" experience, perhaps emphasizing the need to actively conceal knowledge and increasing the difficulty of doing so without letting anything show. Thus, the acoustic variant of question presentation that was employed in both conditions used the more social and presumably more efficient stimuli. This might have contributed to a ceiling effect: Voice as a social stimulus might have augmented physiological responding even in the no observation (control) condition, so that further augmentation by additional social stimuli approached a limit.

The study did not include groups of innocent (unknowledgeable) participants. This design was chosen in favor of comparing differential physiological responding between the two experimental groups with a maximum test power in relation to the number of participants. As a consequence, classification statistics, which would have been based on simulated data only, were omitted in this study.

Summarizing, whatever mediator variables are joining social impact and physiological responding, these mediators might perhaps already be augmented to a near-maximum degree in the CIT variant that was used as control condition in this study. Then, additional social influence, which was thought to heighten these mediator variables, would hardly be able to induce further enhancement. It can be seen as one limitation of the present study that no "low arousal," "low social impact," or "low motivation" condition was included which might have left more space for enhancement of differential physiological responding between conditions.

A more theoretical attempt to explain the lacking influence of evaluative observation refers to the orienting reflex and its modulation. Bradley (54) suggested to regard the orienting response as embedded in motivational and attentional systems that are active and fluctuating within an individual. This view gave rise to the assumption that social stimuli and social interaction influence the orienting response to stimuli; it was hypothesized that also the differential response to stimuli of different subjective significance is sensitive to social influence. Perceiving orienting from a classical perspective [cf. Ref. (6)] and focusing on the correlation between features of the individual stimulus (novelty, intensity, and significance) and the corresponding physiological response, one might possibly start to doubt that psychophysiological responding in the CIT depends on social factors at all. The fact that results of this study did not just miss statistical significance but did not even exhibit a clear direction, might be interpreted as support for this viewpoint. However, given the above alternative explanations and given that this study was the first attempt to manipulate evaluative observation in a CIT, the primary implication of negative results cannot refer to theory before clearing out the conjectured limitations of this study.

A more conceptual limitation of this study refers to the process of evaluative observation, which according to Chapman (24) was initially regarded as one elementary component of social interaction to be distinguished, e.g., from mere presence (55). In the aftermath, observation, evaluation, and the way in which both were realized and combined in this study, are thought to have entailed a set of social components more complex than intended. Although experimenter instructions were meant to help standardize the interaction between subject and experimenter, this interaction presumably had remained complex and multi-faceted.

Trait Aspects: The Fear of Negative Evaluation and the CIT

For male as well as for female participants, SANB average scores exceeded those of a student sample reported earlier (30). Due to the small number of samples reported in the literature, the source of this difference cannot be pinpointed. Temporal change is conceivable, but also a biased sample cannot be ruled out in this study. Higher scores, however, would be expected to lead to greater rather than lower sensitivity of participants to the manipulation of evaluative observation, so that a biased sample is unlikely responsible for the negative results.

Overall, differential physiological responding was not correlated with SANB scores and SANB scores did not interact with the experimental manipulation. Possibly, the trait fear of negative evaluation, which the SANB captures, is of limited relevance when information is concealed from a social counterpart. Being observed with the assumed aim to detect bodily and behavioral indicators of deception might be critically different from being evaluated by observation with respect to performance, correctness, adequateness, or other. The one specific exception to that, namely, the interactive influence of trait fear and evaluative observation on the respiration measure, tentatively points toward a specific sensitivity of fearful examinees to evaluative observation; yet, this finding deserves replication before conclusions can be drawn. Data collection in this study did not include a state measure of fear which might have been fruitfully related to the experimental manipulation and which might have helped to sharpen possible effects of trait fear of evaluative observation.

Suggestions for Future Studies

Continuing the study of social influence on physiological responding in the CIT is necessary. Recent attempts to resume this earlier line of research [see, e.g., Ref. (15)] abandoned in favor of standardization of experiments revealed specific difficulties. Problems arise from the complexity of social interaction, per se, and the severity of disentangling and "isolating" the individual components of social stimuli and social action.

One line of future research might split the focus into two: Studies might investigate the direct influence of those mediator variables that are assumed to link social influence to CIT responding, while other studies might investigate how the manipulation of social stimuli and interaction affects these mediator variables.

From a CIT application perspective, however, research manipulating the social context, in which the CIT is performed, directly in CIT studies is presumably indispensable. The experimental manipulations of social conditions in these CIT studies should then become less complex. For example, evaluative observation might be decomposed into two components, observation and evaluation, which should then be manipulated independently. Taking our experimental setup as an example, the live video showing the experimenter that was employed to remind the participants of being watched might be replaced by a more uniform implementation of the manipulation pursuing the same objective, i.e., induce awareness of being observed during the CIT. Furthermore, replacing acoustic question presentation by a textual interrogation, but perhaps also replacing the active wording of CIT questions by a passive wording, might help to avoid a ceiling effect and thus allow for greater response differences between conditions. Some caution will be needed, however, to prevent such tailored experimental manipulations to become too artificial for an applied setting. With respect to the applied CIT, it has to be borne in mind that in present field applications of the CIT, e.g., in Japan, the examiner is usually present within the same room as the examinee. In this light, laboratory studies like the present one, in which details of the social context in the CIT are manipulated, might contribute first to our theoretical understanding of basic socio-psychophysiological correlations, and perhaps second to practical implications.

CONCLUSIONS

Being observed and evaluated during the CIT with awareness but in the absence of a physical examinator did not enhance an examinee's differential physiological responding in the test. Furthermore, the individual fear of negative evaluation by others did not moderate physiological CIT responses. Standardization of experiments and investigation of social action are essentially conflicting aims even today. To further examine influences of the social situation in which the CIT is performed on physiological responding, it is suggested to manipulate social stimuli and elements of social action at an even more elementary level in future studies.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of "the German Science Foundation and the German Society for Psychology (DGPs)," with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the "ethics committee of the German Society for Psychology (DGPs)."

AUTHOR CONTRIBUTIONS

WA, BA, and DV were in charge of planning the study. BW was in charge of preparing stimulus material and recruiting participants. BW and WA were in charge of conducting the experiments. WA was in charge of data analysis. WA, BA, BW, and DV were in charge of writing the paper.

REFERENCES

1. Lykken DT. The GSR in the detection of guilt. *J Appl Psychol* (1959) 43:385–8. doi: 10.1037/h0046060
2. Verschuere B, Ben-Shakhar G. Theory of the Concealed Information Test. In: Verschuere B, Ben-Shakhar G, Meijer EH, editors. *Memory detection*. Cambridge University Press (2011). p. 128–48. doi: 10.1017/CBO9780511975196.008
3. Ben-Shakhar G, Elaad E. The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. *J Appl Psychol* (2003) 88(1):131–51. doi: 10.1037/0021-9010.88.1.131
4. Meijer EH, Klein Selle N, Elber L, Ben-Shakhar G. Memory detection with the concealed information test: a meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology* (2014) 51:879–904. doi: 10.1111/psyp.12239
5. Lykken DT. Psychology and the lie detector industry. *Am Psychol* (1974) 29:725–39. doi: 10.1037/h0037441
6. Sokolov EN. Higher nervous functions: the orienting reflex. *Annu Rev Physiol* (1963) 25:545–80. doi: 10.1146/annurev.ph.25.030163.002553
7. Ambach W, Stark R, Peper M, Vaitl D. Separating deceptive and orienting components in a concealed information test. *Int J Psychophysiol* (2008) 70:95–104. doi: 10.1016/j.ijpsycho.2008.07.002

8. Furedy JJ, Ben-Shakhar G. The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology* (1991) 28:163–71. doi: 10.1111/j.1469-8986.1991.tb00407.x
9. Klein Selle N, Verschuere B, Kindt M, Meijer E, Ben-Shakhar G. Orienting versus inhibition in the Concealed Information Test: different cognitive processes drive different physiological measures. *Psychophysiology* (2016) 53(4):579–90. doi: 10.1111/psyp.12583
10. Stern RM, Breen JP, Watanabe T, Perry BS. Effect of feedback of physiological information on responses to innocent associations and guilty knowledge. *J Appl Psychol* (1981) 66:677–81. doi: 10.1037//0021-9010.66.6.677
11. Zajonc RB. Social facilitation. *Science* (1965) 149:269–74. doi: 10.1126/science.149.3681.269
12. Emery NJ. The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci Biobehav Rev* (2000) 24:581–604. doi: 10.1016/S0149-7634(00)00025-7
13. Schilbach L, Wohlschläger AM, Krämer NC, Newen A, Shah JN, Fink GR, et al. Being with virtual others: neural correlates of social interaction. *Neuropsychologia* (2006) 44:718–30. doi: 10.1016/j.neuropsychologia.2005.07.017
14. Orne MT. Implications of laboratory research for the detection of deception. *Polygraph* (1975) 2:169–99.
15. Waid WM, Orne MT. Cognitive, social, and personality processes in the physiological detection of deception. *Adv Exp Soc Psychol* (1981) 14:61–106. doi: 10.1016/S0065-2601(08)60370-6
16. Rankin RE, Campbell DT. Galvanic skin response to Negro and white experimenters. *J Abnormal Psychol* (1955) 51:30–3. doi: 10.1037/h0041539
17. Iacono WG. The detection of deception. In: Cacioppo JT, Tassinary LG, Berntson GG, editors. *Handbook of psychophysiology*. Cambridge University Press (2000). p. 772–93.
18. Osugi A. Daily application of the Concealed Information Test: Japan. In: Verschuere B, Ben-Shakhar G, Meijer EH, editors. *Memory Detection*. Cambridge University Press (2011). p. 253–76. doi: 10.1017/CBO9780511975196.015
19. Ambach W, Assmann B, Krieg B, Vaitl D. Face and voice as social stimuli enhance differential physiological responding in a Concealed Information Test. *Front Psychol* (2012) 3:510. doi: 10.3389/fpsyg.2012.00510
20. Varga M, Visu-Petra G, Miclea M, Visu-Petra L. 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* (2015) 10(2): e0116087. doi: 10.1371/journal.pone.0116087
21. Haley KJ, Fessler DMT. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evol Hum Behavior* (2005) 26:245–56. doi: 10.1016/j.evolhumbehav.2005.01.002
22. Jones ME, Nettle D, Bateson M. Effects of eye images on everyday cooperative behavior: a field experiment. *Evol Hum Behavior.* (2011) 32:172–8. doi: 10.1016/j.evolhumbehav.2010.10.006
23. Sproull L, Subramani M, Kiesler S, Walker J, Waters K. When the interface is a face. *Hum-Comp Interact* (1996) 11:97–124. doi: 10.1207/s15327051hci1102_1
24. Chapman AJ. An electromyographic study of apprehension about evaluation. *Psychol Rep* (1973) 33:811–4. doi: 10.2466/pr0.1973.33.3.811
25. Cottrell NB, Wack DL, Sekerak GJ, Rittle RH. Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *J Person Soc Psychol* (1968) 9:245–50. doi: 10.1037/h0025902
26. Paulus PB, Murdoch P. Anticipated evaluation and audience presence in the enhancement of dominant responses. *J Exp Soc Psychol* (1971) 7:280–91. doi: 10.1016/0022-1031(71)90028-X
27. Henchy T, Glass D. Evaluation apprehension and the social facilitation of dominant and subordinate responses. *J Pers Soc Psychol* (1968) 10:446–54. doi: 10.1037/h0026814
28. Gustafson LA, Orne MT. Effects of heightened motivation on the detection of deception. *J Appl Psychol* (1963) 47:408–11. doi: 10.1037/h0041899
29. Watson F, Friend R. Measurement of social-evaluative anxiety. *J Consult Clin Psychol* (1969) 33:448–57. doi: 10.1037/h0027806
30. Vormbrock F, Neuser J. Konstruktion zweier spezifischer Fragebögen zur Erfassung von Angst in sozialen Situationen (SANB und SVSS). *Diagnostica* (1983) 292:165–82.
31. Edelmann RJ, Baker SR. Self-reported and actual physiological responses in social phobia. *Br J Clin Psychol* (2002) 41:1–14. doi: 10.1348/014466502163732
32. Giesen M, Rollison MA. Guilty knowledge versus innocent associations: effects of trait anxiety and stimulus context on skin conductance. *J Res Pers* (1980) 14:1–11. doi: 10.1016/0092-6566(80)90035-5
33. Visu-Petra G, Miclea M, Visu-Petra L. Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Appl Cogn Psychol* (2011) 26(3): 342–51. doi: 10.1037/e669802012-248
34. Stern RM, Ray WJ, Quigley KS. *Psychophysiological recording*. 2nd ed. Oxford: University Press (2001). doi: 10.1093/acprof:oso/9780195113594.001.0001
35. Ambach W, Dummel S, Lüer T, Vaitl D. Physiological responses in a concealed information test are determined interactively by encoding procedure and questioning format. *Int J Psychophysiol* (2011) 81(3):275–82. doi: 10.1016/j.ijpsycho.2011.07.010
36. Ambach W, Bursch S, Stark R, Vaitl D. A Concealed Information Test with multimodal measurement. *Int J Psychophysiol* (2010) 75:258–67. doi: 10.1016/j.ijpsycho.2009.12.007
37. Lim CL, Rennie C, Barry R, Bahramali H, Lazzaro I, Manor B, et al. Decomposing skin conductance into tonic and phasic components. *Int J Psychophysiol* (1997) 25:97–109. doi: 10.1016/S0167-8760(96)00713-1
38. Lim CL, Gordon E, Rennie C, Wright JJ, Bahramali H, Li WM, et al. Dynamics of SCR, EEG, and ERP activity in an oddball paradigm with short interstimulus intervals. *Psychophysiology* (1999) 36:543–51. doi: 10.1111/1469-8986.3650543
39. Timm HW. Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *J Appl Psychol* (1982) 67:391–400. doi: 10.1037/0021-9010.67.4.391
40. Kircher JC, Raskin DC. *The computerized polygraph system II (Software Version 4.01)*. Salt Lake City, Utah (USA): Scientific Assessment Technologies (2003).
41. Velden M, Wölk C. Depicting cardiac activity over real time: a proposal for standardization. *J Psychophysiol* (1987) 1:173–5.
42. Gamer M, Verschuere B, Crombez G, Vossel G. Combining physiological measures in the detection of concealed information. *Physiol Behav* (2008) 5:333–40. doi: 10.1016/j.physbeh.2008.06.011
43. Verschuere B, Crombez G, Koster EHW, De Clercq A. Antisociality, underarousal and the validity of the Concealed Information Polygraph Test. *Biol Psychol* (2007) 74:309–18. doi: 10.1016/j.biopsycho.2006.08.002
44. Elaad E, Ben-Shakhar G. Finger pulse waveform length in the detection of concealed information. *Int J Psychophysiol* (2006) 61:226–34. doi: 10.1016/j.ijpsycho.2005.10.005
45. Lykken DT, Venables PH. Direct measurement of skin conductance: a proposal for standardization. *Psychophysiology* (1971) 8:656–72. doi: 10.1111/j.1469-8986.1971.tb00501.x
46. Ben-Shakhar G. Standardization within individuals: a simple method to neutralize individual differences in skin conductance. *Psychophysiology* (1985) 22:292–9. doi: 10.1111/j.1469-8986.1985.tb01603.x
47. Gamer M, Rill HG, Vossel G, Gödert HW. Psychophysiological and vocal measures in the detection of guilty knowledge. *Int J Psychophysiol* (2006) 60:76–87. doi: 10.1016/j.ijpsycho.2005.05.006
48. Gronau N, Ben-Shakhar G, Cohen A. Behavioral and physiological measures in the detection of concealed information. *J Appl Psychol* (2005) 90:147–58. doi: 10.1037/0021-9010.90.1.147
49. Cohen J. *Statistical power analysis for the behavioral sciences*. San Diego, CA: McGraw-Hill (1988).
50. Rosnow RL, Rosenthal R. Computing contrasts, effect sizes, and counternulls on other people's published data: general procedures for research consumers. *Psychol Methods* (1996) 1:331–40. doi: 10.1037//1082-989X.1.4.331
51. Vandenbosch K, Verschuere B, Crombez G, De Clercq A. The validity of finger pulse line length for the detection of concealed information. *Int J Psychophysiol* (2009) 71:118–23. doi: 10.1016/j.ijpsycho.2008.07.015
52. Zvi L, Nachson I, Elaad E. Effects of coping and cooperative instructions on guilty and informed innocents' physiological responses to concealed

information. International. *J Psychophysiol* (2012) 84:140–8. doi: 10.1016/j.ijpsycho.2012.01.022

53. Zvi L, Nachson I, Elaad E. Effects of perceived efficacy and prospect of success on detection in the Guilty Actions Test. *Int J Psychophysiol* (2015) 95:35–45. doi: 10.1016/j.ijpsycho.2014.12.010
54. Bradley MM. Natural selective attention: orienting and emotion. *Psychophysiology* (2009) 46:1–11. doi: 10.1111/j.1469-8986.2008.00702.x
55. Chapman AJ. An electromyographic study of social facilitation: a test of the 'mere presence' hypothesis. *Br J Psychol* (1974) 65:123–8 doi: 10.1111/j.2044-8295.1974.tb02777.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychiatry

ORIGINAL RESEARCH
published: 16 October 2019
doi: 10.3389/fpsyt.2019.00725

MMPI-2-RF Profiles in Child Custody Litigants

*Cristina Mazza[1], Franco Burla[1], Maria Cristina Verrocchio[2], Daniela Marchetti[2], Alberto Di Domenico[2], Stefano Ferracuti[1] and Paolo Roma[1]**

[1] Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy, [2] Department of Psychological, Health, and Territorial Sciences, G. d'Annunzio University of Chieti-Pescara, Chieti, Italy

Background and Purpose: A psychological assessment of parents in post-divorce child custody disputes highlighted parents' motivation to appear as adaptive and responsible caregivers. The study hypothesized that personality self-report measures completed by child custody litigants (CCLs) during a parental skills assessment would show underreporting, rendering the measures worthless. The study also analyzed gender differences in a CCL sample, general CCL profiles, and the implicit structure of the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) in the CCL sample.

Materials and Methods: The sample comprised 400 CCLs undergoing personality evaluation as part of a parenting skills assessment. The mean age of the 204 mothers was 41.31 years (*SD* = 6.6), with an overall range of 24–59 years. Mothers had a mean educational level of 14.48 years (*SD* = 3.2). The 196 fathers were aged 20–59 years (*M* = 42.31; *SD* = 7.8), with an average of 14.48 years (*SD* = 3.9) of education. The MMPI-2-RF was administered. To test the hypotheses, multivariate analyses of variance (MANOVAs) and two-step cluster analyses were run.

Results: CCL subjects reported higher scores in underreporting (L-r and K-r) and lower scores in overreporting [F-r, Fp-r, Fs-r, and response bias scale (RBS)] validity scales and restructured clinical (RC) scales, with the exception of RC2 and RC8. RC6 (Ideas of Persecution) was the most elevated. Intercorrelations within the RC scales significantly differed between CCL and normative samples. Women appeared deeply motivated to display a faking-good defensive profile, together with lower levels of cynicism and antisocial behaviors, compared to CCL men. Two-step cluster analyses identified three female CCL profiles and two male CCL profiles. Approximately 44% of the MMPI-2-RF profiles were deemed possibly underreporting and, for this reason, considered worthless.

Discussion: The present study adds useful insight about which instruments are effective for assessing the personality characteristics of parents undergoing a parental skills assessment in the context of a child custody dispute. The results show that almost half of the MMPI-2-RF protocols in the CCL sample were worthless due to their demonstration of an underreporting attitude. This highlights the necessity to interpret CCL profiles in light of normative data collected specifically in a forensic setting and the need for new and promising methods of mainstreaming and administering the MMPI-2-RF.

Keywords: Minnesota Multiphasic Personality Inventory-2-Restructured Form, custody litigants, parenting skills, personality, forensic evaluation

OPEN ACCESS

Edited by:
*Giuseppe Sartori,
University of Padova, Italy*

Reviewed by:
*Nicola Carone,
University of Pavia, Italy
Claudio Imperatori,
Università Europea di Roma, Italy*

****Correspondence:***
*Paolo Roma
paolo.roma@uniroma1.it*

Specialty section:
*This article was submitted to
Forensic Psychiatry,
a section of the journal
Frontiers in Psychiatry*

Received: *13 March 2019*
Accepted: *09 September 2019*
Published: *16 October 2019*

Citation:
*Mazza C, Burla F, Verrocchio MC,
Marchetti D, Di Domenico A,
Ferracuti S and Roma P (2019)
MMPI-2-RF Profiles in
Child Custody Litigants.
Front. Psychiatry 10:725.
doi: 10.3389/fpsyt.2019.00725*

INTRODUCTION

In any child custody evaluation, parental adequacy must be assessed in order to guarantee the best interests of the child. Among all couples who request a separation in Italy, 15–20% are subjected to psychological evaluation as part of a parental skills assessment; this percentage was released by the Supreme Court of Appeal of Rome on December 4, 2018, at the "New questions in parental competency on child custody" congress. When assessing parental fitness, examiners evaluate factors such as the social context, the child's condition, the relationship between each parent and the child, and the personality characteristics of the child custody litigants (CCLs). Parental couples are among the more problematic in the judicial setting, as they are often in litigation over economic issues and may be less amenable to mediation agreements aimed at securing the best interests of their child. CCLs are also often characterized by impaired psychological functioning, poor coping strategies, and unrealistic ideas of themselves and others (1, 2), despite their tendency to present themselves as psychologically stable and responsible (3–7).

An overwhelming proportion of child custody evaluations involve psychometric measures, which are predominantly used to assess the personality characteristics and functioning of the litigants. One such measure, the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) (8), is a well-established psychological instrument that is frequently used in forensic assessment (9–14). Ackerman and Pritzl (15), in their 20-year follow-up survey of practice and methods in child custody evaluations, highlighted that, in 97.2% of all cases, clinicians use the MMPI-2 when evaluating parents. This finding is consistent with data presented in other studies (16–18). Due to the wide use of the MMPI-2 in child custody evaluations, there is a considerable literature regarding the MMPI-2 psychometric characteristics of CCLs (4, 5, 11, 19–22, 23). This literature indicates that, overall, subjects undergoing a parental skills evaluation obtain, on some scales, significantly different scores relative to non-CCL subjects and the normative population. In more detail, CCL respondents tend to deny or omit negative features of their personality in order to present themselves in a better light, to show more adaptive psychological and behavioral functioning, and to appear as responsible caregivers who will provide for the best interests of their child. This underreporting attempt—stemming from a faking-good profile and usually combined with elevated scores on the MMPI-2 clinical scales of Hysteria (Hy), Psychopathic Deviate (Pd), and Paranoia (Pa)—is thought to be an effect of the legal environment (6, 11, 24, 25).

Recently, a restructured and shortened version of the MMPI-2, the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF), was introduced (26, 27). The MMPI-2-RF is composed of items extracted from the MMPI-2 (338 vs. 567 items), arranged into 51 scales (vs. the 118 scales of the MMPI-2) (8, 28). Compared to the MMPI-2, the MMPI-2-RF has some advantages: it is shorter, it requires less time to administer, and it comprises a limited set of scales. Because it is easy to score and interpret, it reduces the potential for mistakes to be made in the assessment process, which is critically important in forensic contexts. To the best of our knowledge, only three studies have addressed the use of the MMPI-2-RF in CCL samples. In the first study, Sellbom and Bagby (7) focused on the MMPI-2-RF validity scales L-r and K-r in a group of 109 CCLs (56 men, 53 women), compared to a group of 140 university students. The students were split into two groups, with one group instructed to underreport and the other instructed to follow the standard MMPI-2 protocol. The results indicated that the CCL sample produced higher mean T-scores in the L-r and K-r scales, relative to the underreporting students; this finding underlies the role of these scales in discriminating between honest and faking-good respondents. Additionally, the authors found substantial consistency between the L-r and K-r scales, suggesting that test administrators could benefit from analyzing these scales in conjunction when making decisions about underreporting.

In the second study, Archer et al. (3) studied all MMPI-2-RF scales in a sample of 344 North American CCLs (172 men, 172 women). The authors found two major differences between this group and the general population: higher scores in underreporting validity scales (L-r and K-r) and a lower cumulative percentage frequency of restructured clinical (RC) scales with T-scores > 65. Specifically, the most commonly elevated RC scale (as shown by 15.1% of men and 18% of women) was RC6 (Ideas of Persecution). Among men, RC4 (Antisocial Behavior) was the second most commonly elevated scale, whereas RC1 (Somatic Complaints) was the second most frequently elevated scale among women. The study also examined the alpha coefficients of the H-O, Somatic/Cognitive, Internalizing, Externalizing, Interpersonal, and PSY-5 scales of the MMPI-2-RF for men, women, and the combined sample and found consistency between the internal reliabilities of these scales and those reported in the MMPI-2-RF manual. Nevertheless, the scale intercorrelation patterns were found to be very similar to those reported for other populations. The study did not examine the association between MMPI-2-RF scores and other relevant factors of individual parenting ability, and the researchers underlined that it was not possible to reach a causal inference of parents' psychopathology on the eventual emotional disturbance of their children. Finally, Kauffman et al. (6) examined the MMPI-2-RF performance of a sample of 49 CCLs (25 men, 24 women). The results were similar to those of the previous studies of Sellbom and Bagby (7) and Archer et al. (3), indicating elevated scores in the scales of L-r (with 67% of the sample showing T-scores ≥ 55) and K-r (with 80% of the sample showing T-scores ≥ 55), in comparison to the other validity scales, which showed mean T-scores of 59.78 and 59.49, respectively. In addition, only RC6 (Ideas of Persecution) achieved a mean T-score > 50. Specifically, 43% of the sample demonstrated elevated T-scores ≥ 55, and 14% showed elevated scores in the clinical range (T-scores > 65). These results suggest that CCLs had the tendency to experiment with high levels of suspiciousness and mistrust, relative to the normative sample, and to present themselves as responsible and socially desirable.

The research of Sellbom and Bagby (7) considered only two (out of 51) MMPI-2-RF scales and acknowledged the necessity for future research to enlarge the sample for the purposes of cross validation. The results of Archer et al. (3), which considered all MMPI-2-RF scales, also require confirmation by further research. Finally, Kauffman et al. (6) findings, despite contributing to the analysis of CCL personalities, were based on a relatively small sample, which limits the generalizability of their results. Furthermore, all of the aforementioned studies administered the MMPI-2 and only retrospectively generated and scored each individual's MMPI-2-RF, with a high risk of noisy factors (e.g., subject fatigue and overworking caused by responding to a scale of almost twice the length of the MMPI-2-RF, with item redundancy). Lastly, as reported in the literature, CCLs have specific attributes of personality and psychological functioning; thus, their MMPI-2-RF profiles should be interpreted in light of normative data collected in a forensic setting. It is also questionable whether the use of the MMPI-2-RF is altogether worthwhile, considering the different psychological characteristics of CCLs relative to the normative population (against whom data are standardized) and their common underreporting profiles, which cloud the test's ability to discriminate by reducing values on the clinical scales. Thus, building on the research of Sellbom and Bagby (7), Archer et al. (3), and Kauffman et al. (6), the present study used the MMPI-2-RF 338-item protocol to test the following hypotheses in a large CCL sample:

H1. CCL subjects would report higher scores in underreporting validity scales (L-r and K-r) and lower scores in overreporting validity scales (F-r, Fp-r, Fs-r, and RBS), relative to the normative sample;

H2. CCL subjects would report lower scores in RC scales compared to the normative sample, and RC6 would be most elevated among CCLs;

H3. the MMPI-2-RF profiles of CCL women would differ from those of CCL men;

H4. CCL MMPI-2-RF profiles would demonstrate intercorrelations between scales that do not significantly differ from those of the normal/non-forensic population.

Furthermore,

H5. As mean MMPI-2-RF profile scores are limited in their ability to accurately characterize individuals (because low and high scores may cancel each other out), the study tested for the presence of typical CCL personality profiles through a cluster analysis of the MMPI-2-RF scores. While this approach is not widely used in the field, it has generated important results in other settings (e.g., with respect to studies of driving under the influence of alcohol subjects and filicide).

Finally, the study sought to investigate

H6. The percentage of underreporting MMPI-2-RF protocols in the CCL sample, expecting this to be very high.

Overall, the study aimed at testing the utility of the MMPI-2-RF in forensic settings, analyzing the percentage of useless protocols, implicit structural differences, and typical CCL profiles (in both women and men), compared to a normative sample. Given the high percentage of useless protocols due to the well-documented underreporting attitude of CCL subjects, the study was considered useful to clinicians in a position of choosing whether or not to administer this test to couples undergoing a parental skills assessment.

MATERIALS AND METHODS

Participants

At first, the subjects were 451 parents undergoing a psychological evaluation of personality and parenting ability, as prescribed by judges in the context of a child custody dispute. Each parent agreed to participate in the study for research purposes. Thirty-six subjects compiled the MMPI-2-RF but did not give informed consent to the research, mainly because they didn't willingly accept the CCLs assessment (consequently they refused the consent to research purpose).

In more detail, the sample comprised 196 couples plus 8 mothers whose ex-partners did not complete the MMPI-2-RF in a valid and reliable way. The 196 fathers were aged 20–59 years (*M* = 42.31; *SD* = 7.8), with an average of 14.48 (*SD* = 3.9) years of education. The 204 mothers were aged 24–59 years (*M* = 41.31; *SD* = 6.6), with a mean educational level of 14.48 years (*SD* = 3.2). No statistically significant differences were observed across genders in age and years of education, and these measures were also sufficiently aligned with the data provided for Italian divorced couples by the Italian National Institute of Statistics (29, 30). According to these latter statistics, in 2015, the majority of Italian divorced women (20.3%) were aged 40–44 years, with an average of 45 years for the entire sample; most Italian divorced men (19.7%) were aged 45–49 years, with an average of 48 years. Within this normative sample, 44.3% of women and 41% of men had a mean educational level of 13 years. The study sample was collected between 2015 and 2017 from five regional courts throughout Italy, with the collaboration of local experts in psychology who were called to evaluate parents and administer the MMPI-2-RF protocol during assessments of parental fitness. Fifteen cases were excluded, as they contained 15 or more items that were unanswered and because the Variable Response Inconsistency (VRIN-r) or True Response Inconsistency (TRIN-r) scale T-scores were ≥80. All 400 cases were court ordered, and data were only collected from child custody dispute cases; no data were collected from other child protection matters, as the literature suggests that there is a difference between these specific judicial contexts. On the one hand, child custody disputes are civil cases concerning disagreements between parents about legal and/or physical custody; on the other hand, in evaluations of parental competency, criminal charges (e.g., allegations of abuse, neglect, etc.) may co-occur, forcing the involvement of government agencies with the purpose of protecting the children involved (31).

Materials

MMPI-2-RF

The full Italian version of the MMPI-2-RF (32) was used. the MMPI-2-RF (33) is a 51-scale measure of personality and psychopathology with 338 items, selected from the 567 items of the MMPI-2 (26, 34). The MMPI-2-RF has the following: nine validity scales, most of which are revised versions of MMPI-2 validity scales; nine RC scales, which were developed by Tellegen et al. and released in 2003; three higher order (HO) scales, which were derived from factor analyses to identify the basic domains of affect, thought, and behavior; 23 specific problem (SP) scales, which highlight important characteristics associated with particular RC scales; and revised versions of the personality psychopathology five (PSY-5) scales, which link the MMPI-2-RF to a five-factor model of personality pathology (26). All of the raw scores of the MMPI-2-RF scales, with the exception of the validity and interest scales, register uniform T-scores, as developed for the MMPI-2 by Tellegen and Ben-Porath (35). For these scales, a uniform T-score of 65 corresponds to the 92nd percentile and indicates the minimal level of elevation required for the interpretive recommendations. The MMPI-2-RF validity and interest scales, however, register linear T-scores, as the scales have distinct distributions, dissimilar to the composite uniform distribution. for this scale, the T-score interpretation is variable: for the TRIN-R and VRIN-R scales, T-scores > 79 could measure inconsistency; for the L scale, T-scores > 64 Could demonstrate possible underreporting; for the K scale, T-scores > 59 could show possible underreporting; and for the F "family" scales, T-scores > 79 could represent possible overreporting (relative to T-scores > 80 for Fs-R, RBS, and FBS-r).

Statistical Analyses

To test H1 and H2, the frequency of elevation (in terms of percentile score) was studied for the seven validity scales and the nine RC scales. For the purposes of verifying H3, a multivariate analysis of variance (MANOVA) was run using gender as the independent variable and MMPI-2-RF validity and RC scale T-scores as dependent measures. The Bonferroni correction was applied for multiple comparisons. The effect sizes of the score differences between groups were recorded, with values of 0.02, 0.13, and 0.26 considered indicative of small, medium, and large effects, respectively (36). The intercorrelation for the nine RC scales in the CCL sample was compared to that of the normative sample through a z-score analysis (37), in order to verify H4. H5 was tested using a two-step cluster analysis in which the BIC criterion was used to define the profiles of female and male CCLs, respectively. This method first identified groupings using a quick cluster algorithm (pre-clustering) and then ran hierarchical cluster models in the second step. MMPI-2-RF validity and RC scales were used in the cluster model. In order to achieve natural clustering, the number of clusters was set to automatic (38). MANOVAs were also performed between gender clusters using the cluster as the independent variable and MMPI-2-RF validity and RC scale T-scores as dependent measures. Scheffé (39) method was used to assess *post hoc* pair differences ($p < 0.05$). Finally, the frequency of underreporting elevation (in terms of percentage) was also inspected for the L and K validity scales to test H6. Invalid MMPI-2-RF protocols were not included in the statistical analyses. The SPSS-18 statistical package (SPSS Inc., Chicago, IL) was used for all analyses.

RESULTS

Differences Between Normative and CCL Samples

Table 1 provides data on the frequency of elevations in the MMPI-2-RF validity and RC scales, both collapsed across genders and in the combined sample. According to the technical manual (31), in the normative sample, 10% of subjects achieved a linear T-score ≥ 65 in the validity scales, while in the RC scales, uniform T-scores of 65 fell in the 8th percentile. **Table 1** reveals that, in the underreporting scales (L-r and K-r), the percentage of CCL subjects who achieved a linear T-score ≥ 65 was almost twice the expected proportion. In the overreporting scales (F-r, Fp-r, Fs-r, FBS, and RBS), however, the percentage of CCL subjects demonstrating a linear T-score ≥ 65 was lower than the 8% expected. In relation to the RC scales, only three (out of nine) scales (RC1, RC2, and RC6) had more than 8% of CCL subjects achieving uniform T-scores ≥ 65.

To evaluate whether the relationship between scales differed between the CCL and normative samples, correlation values were compared. **Table 2** shows the raw score intercorrelations between the nine RC scales, with findings for men presented in the upper diagonal and values for women presented in the lower diagonal. In the same table, the intercorrelation values reported in the Italian technical manual of the MMPI-2-RF (29) are displayed. No gender differences emerged in the correlations. Out of 36 correlations, 5 were significantly different for men, while 15 were significantly different for women. RC1 intercorrelations in both CCL women and CCL men showed the greatest differences relative to the normative intercorrelations reported in the technical manual (31). For women, most other differences were found in the RC8 scale. The great number of meaningful differences suggests that the implicit structure of the MMPI-2-RF was significantly different in the CCL sample.

TABLE 1 | Frequency of elevations ≥65 for men and women on the MMPI-2-RF Validity and RC scales in the CCL sample.

	Scale	Combined (%)	Male (%)	Female (%)
Validity scales	L-r (Uncommon Virtues)	18.3	14.8	21.6
	K-r (Adjustment Validity)	20	16.8	23
	F-r (Infrequent Responses)	2.8	2.6	2.5
	Fp-r (Infrequent Psychopathology Responses)	2	2.6	1.5
	Fs (Infrequent Somatic Responses)	1.3	2	0.5
	FBS-r (Symptom Validity)	6	4.6	7.4
	RBS (Response Bias Scale)	2.5	2.6	2.5
Restructured clinical scales	RCd (Demoralization)	1.3	2.6	0
	RC1 (Somatic Complaints)	10.3	10.2	10.3
	RC2 (Low Positive Emotions)	9.8	12.2	7.4
	RC3 (Cynicism)	7.5	13.8	1.5
	RC4 (Antisocial Behavior)	5.3	8.2	2.5
	RC6 (Ideas of Persecution)	14.3	14.8	13.7
	RC7 (Dysfunctional Negative Emotions)	1.5	2	1
	RC8 (Aberrant Experiences)	2.3	2	2.5
	RC9 (Hypomanic Activation)	5.3	7.1	3.4

TABLE 2 | Raw score intercorrelation table for MMPI-2-RF RC scales presented separately by gender.

		RCd	RC1	RC2	RC3	RC4	RC6	RC7	RC8	RC9
RCd	**CCL**	-	0,58	0,47	0,46	0,58	0,56	0,81	0,55	0,52
	Normative	-	0,55	0,51	0,39	0,52	0,45	0,79	0,63	0,42
RC1	**CCL**	.53	-	0,60	0,21	0,53	0,53	0,43	0,46	0,20
	Normative	.59	-	0,46	0,24	0,52	0,61	0,61	0,56	0,36
RC2	**CCL**	.32	.57	-	-0,02	0,30	0,39	0,25	0,33	-.08
	Normative	.56	.31	-	-0,01	0,34	0,30	0,36	0,28	-.07
RC3	**CCL**	.39	.17	.12	-	0,41	0,43	0,53	0,33	0,62
	Normative	.48	.38	.18	-	0,40	0,28	0,46	0,36	0,54
RC4	**CCL**	.50	.56	.32	.14	-	0,48	0,51	0,51	0,53
	Normative	.38	.32	0,19	.32	-	0,48	0,55	0,53	0,49
RC6	**CCL**	.40	.30	.07	.35	.33	-	0,48	0,61	0,35
	Normative	.50	.52	.18	.47	.39	-	0,56	0,64	0,36
RC7	**CCL**	.70	.41	.13	.48	.34	.50	-	0,47	0,55
	Normative	0,77	.57	.39	.56	0.7	.51	-	0,71	0,55
RC8	**CCL**	.34	.28	.05	.22	.31	.39	.37	-	0,41
	Normative	.51	.57	.15	.43	.39	.59	0,57	-	0,51
RC9	**CCL**	.41	.22	-.19	.42	.31	.36	.49	.42	-
	Normative	.34	.36	.12	.47	.45	.44	.52	.51	-

Men are represented in the upper right part, women in the lower left part. CCL correlations are from the present study data; Normative correlations are from the Italian normative data. Red circles indicate differences in correlations higher than 0.15.

In the CCL sample, approximately 44% of the MMPI-2-RF profiles could be deemed possibly underreporting and, for this reason, worthless. This estimation was based on the percentage of protocols with both linear T-scores ≥ 65 in the L-r scale and T-scores ≥ 60 in the K-r scale, in line with the cutoffs for underreporting in the technical manual (31).

Gender Differences in the MMPI-2-RF Validity and RC Scales

A 2 × 7 MANOVA (gender × MMPI-2-RF validity scales) showed a significant gender effect on the MMPI-2-RF validity scales, $V = 0.11$, F (6, 393) = 8.12, p < 0.001, $\text{par}\eta^2 = 0.110$. Separate univariate ANOVAs on the outcome variables revealed a significant gender effect on the following validity scales: L-r [F(1, 398) = 5.74, $p = 0.017$, $\text{par}\eta^2 = 0.014$], K-r [F(1, 398) = 6.82, $p = 0.009$, $\text{par}\eta^2 = 0.017$], and FBS [F(1, 398) = 29.38, p = 0.001, $\text{par}\eta^2 = 0.069$].

With respect to the RC scales, a 2 × 9 MANOVA (gender × MMPI-2-RF RC scales) showed a significant overall gender effect, $V = 0.22$, F (9, 390) = 12.32, p < 0.001, $\text{par}\eta^2 = 0.221$. Separate univariate ANOVAs on the outcome variables revealed a significant gender effect on the following RC scales: RC1 [F(1, 398) = 6.21, $p = 0.013$, $\text{par}\eta^2 = 0.015$], RC3 [F(1, 398) = 35.22, p = 0.001, $\text{par}\eta^2 = 0.081$], RC4 [F(1, 398) = 12.25, $p = 0.001$, $\text{par}\eta^2 = 0.030$], and RC9 [F(1, 398) = 12.65, p = 0.001, $\text{par}\eta^2 = 0.031$]. **Table 3** shows the descriptive values of the two groups (men vs. women) for all outcome variables. Compared to men, women scored higher on all significant MMPI-2-RF validity scales (L-r, K-r, and FBS) and the RC1 scale. Men had higher scores on the RC3, RC4, and RC9 scales.

Cluster Analysis

The two-step cluster analysis of the 204 female CCL subjects revealed three clusters with significant differences in mean score profiles (see **Table 4**). A 3 × 16 MANOVA showed a significant clustering effect (cluster 1 vs. cluster 2 vs. cluster 3) on the MMPI-2-RF validity and RC scales, $V = 1.21$, F (30, 376) = 19.24, $p < 0.001$, parη2 = 0.606. In more detail, separate univariate ANOVAs on the outcome variables revealed a significant clustering effect in all MMPI-2-RF scales except for the L-r scale [F(2, 201) = 1.74, $p = 0.179$, $\text{par}\eta^2 = 0.017$]. Characteristics of the CCL women in each cluster were as follows:

- Cluster 1 ($N = 18$) women had very high scores (T-scores ≥ 66) in the RC1, RC6, and RC2 scales; the FBS, F-r, RBS, F-s, Fp-r, L-r, RC8, RC4, and RC9 scales showed moderately high scores (T-scores = 55–60). All other MMPI-2-RF scales showed T-scores < 55.
- Cluster 2 ($N = 110$) women scored moderately high (T-scores > 55) to high (T-scores > 60) in the L-r scale. All other MMPI-2-RF scales showed T-scores < 55.
- Cluster 3 ($N = 76$) women scored high (T-scores ≥ 60) in the K-r scale and moderately high (T-scores = 55–60) in the L-r scale. All other MMPI-2-RF scales showed T-scores < 55.

The two-step cluster analysis of the 196 male CCL subjects revealed two clusters with significant differences in mean score profiles (see

TABLE 3 | Mean T-scores and standard deviations for women and men for the MMPI-2-RF validity and RC scales with associated univariate F values and effect sizes.

MMPI-2-RF		Total sample *N* = 400 *M* (SD)	Women *N* = 204 *M* (SD)	Men *N* = 196 *M* (SD)	*F*	parη^2
Validity scales						
	F-r	47.84 (7.08)	48.09 (6.50)	47.59 (7.66)	0.51	0.001
	Fs	45.76 (6.85)	45.97 (6.10)	45.55 (7.55)	0.37	0.001
	FBS	51.96 (8.39)	54.11 (7.90)	49.72 (8.31)	29.38***	0.069
	L-r	55.83 (9.15)	56.89 (8.75)	54.71 (9.44)	5.74*	0.014
	K-r	54.57 (8.98)	55.71 (8.26)	53.38 (9.55)	6.82**	0.017
	Fp-r	47.79 (6.92)	47.79 (6.66)	47.79 (7.21)	0.00	0.000
	RBS	44.84 (7.08)	45.09 (6.50)	44.59 (7.66)	0.51	0.001
RC scales						
	RCD	45.36 (7.60)	45.06 (6.56)	45.66 (8.56)	0.62	0.002
	RC1	49.71 (9.65)	50.88 (8.61)	48.49 (10.50)	6.21*	0.015
	RC2	50.27 (11.19)	50.15 (10.24)	50.40 (12.14)	0.05	0.000
	RC3	46.17 (10.29)	43.30 (7.95)	49.16 (11.55)	35.22***	0.081
	RC4	47.94 (8.99)	46.42 (7.78)	49.52 (9.87)	12.25***	0.030
	RC6	53.00 (10.48)	53.98 (10.43)	51.99 (10.47)	3.61	0.009
	RC7	44.34 (7.37)	44.02 (6.57)	44.67 (8.13)	0.77	0.002
	RC8	50.86 (6.97)	50.44 (7.13)	51.30 (6.79)	1.53	0.004
	RC9	43.75 (11.44)	41.79 (10.78)	45.80 (11.77)	12.65***	0.031

**p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001.*

Table 5). A 2 × 16 MANOVA showed a significant clustering effect (cluster 1 vs. cluster 2) on the MMPI-2-RF validity and RC scales, $V = 0.73$, $F(15, 180) = 40.97$, $p < 0.001$, parη$^2 = 0.773$. In more detail, separate univariate ANOVAs on the outcome variables revealed a significant clustering effect in all MMPI-2-RF scales. Characteristics of the CCL men in each cluster are summarized below. CCL men in cluster 2 scored higher in all MMPI-2-RF scales compared to CCL men in cluster 1, save for the L-r and K-r scales.

- Cluster 1 (N = 151) men scored moderately high (T-scores = 55–60) in the K-r and L-r validity scales. All other MMPI-2-RF scales showed T-scores < 55.
- Cluster 2 (N = 45) men scored high (T-scores ≥ 60) in the RC6, RC2, RC1, and RC4 scales; and moderately high (T-scores > 55) to high (T-scores > 60) in the F-r, FBS, Fs-r, Fp-r, and RBS validity scales and the RC3, RC8, RCD, and RC9 scales.

DISCUSSION

The main purpose of the research was to investigate if use of the MMPI-2-RF, as it is currently administered, could successfully increase our knowledge of the personality features of CCL

TABLE 4 | T-scores for the validity and RC scales of the MMPI-2-RF for Women-1, Women-2, and Women-3 clusters.

	MMPI-2-RF	Cluster 1 *N* = 18 *M* (SD)	Cluster 2 *N* = 110 *M* (SD)	Cluster 3 *N* = 76 M (SD)	*F*	parη^2
Validity scales						
	F-r	62.94 (6.25)^{a}	49.20 (3.70)^{b}	42.97 (2.05)^{c}	247.82***	.711
	Fs	59.50 (4.79)^{a}	47.20 (3.70)^{b}	40.97 (2.05)^{c}	245.01***	0.709
	FBS	63.17 (8.44)^{a}	54.85 (8.38)^{b}	50.89 (4.46)^{c}	22.56***	0.183
	L-r	59.39 (8.46)	55.93 (8.71)	57.70 (8.78)	1.74	0.017
	K-r	47.17 (9.36)^{a}	52.95 (6.97)^{b}	61.72 (5.36)^{c}	55.02***	0.354
	Fp-r	59.50 (5.13)^{a}	48.31 (5.92)^{b}	44.26 (4.15)^{c}	62.25***	0.382
	RBS	59.94 (6.25)^{a}	46.20 (3.70)^{b}	39.97 (2.05)^{c}	247.82***	0.711
RC scales						
	RCD	53.78 (5.36)^{a}	47.54 (5.08)^{b}	39.42 (3.58)^{c}	105.03***	0.511
	RC1	67.00 (8.77)^{a}	52.71 (5.74)^{b}	44.42 (5.10)^{c}	120.59***	0.545
	RC2	65.83 (19.65)^{a}	50.52 (7.23)^{b}	45.89 (6.53)^{b}	37.83***	0.273
	RC3	47.89 (9.87)^{a}	45.01 (7.61)^{a}	39.74 (6.53)^{b}	15.00***	0.130
	RC4	56.50 (7.63)^{a}	48.72 (6.91)^{b}	40.70 (4.00)^{c}	66.65***	0.399
	RC6	66.44 (10.58)^{a}	57.13 (8.58)^{b}	46.46 (7.27)^{c}	59.19***	0.371
	RC7	50.50 (9.18)^{a}	46.49 (4.91)^{b}	38.91 (4.15)^{c}	64.12***	0.390
	RC8	59.17 (10.22)^{a}	50.10 (7.12)^{b}	48.87 (4.36)^{b}	18.08***	0.152
	RC9	56.44 (19.53)^{a}	42.55 (8.64)^{b}	37.22 (6.81)^{c}	30.63***	0.234

****p ≤ 0.001. For each line, different letters indicate differences between columns.*

TABLE 5 | T-scores for the validity and RC scales of the MMPI-2-RF for Men-1 and Men-2 clusters.

	MMPI-2-RF	Cluster 1 *N* = 151 *M* (SD)	Cluster 2 *N* = 45 *M* (SD)	*F*	parη^2
Validity scales					
	F-r	44.28 (3.82)	58.69 (6.82)	330.20^{***}	0.630
	Fs	42.28 (3.82)	56.53 (6.56)	335.19^{***}	0.633
	FBS	47.46 (6.00)	57.31 (10.32)	64.68^{***}	0.250
	L-r	55.72 (9.47)	51.36 (8.61)	7.65^{**}	0.038
	K-r	56.80 (6.76)	41.91 (8.57)	147.94^{***}	0.433
	Fp-r	45.38 (4.87)	55.87 (7.91)	117.30^{***}	0.377
	RBS	41.28 (3.82)	55.69 (6.82)	330.20^{***}	0.630
RC scales					
	RCD	42.44 (5.20)	56.49 (8.78)	178.44^{***}	0.479
	RC1	44.60 (6.78)	61.56 (10.26)	167.77^{***}	0.464
	RC2	46.84 (7.69)	62.33 (16.19)	79.15^{***}	0.290
	RC3	46.32 (9,78)	58.71 (12.01)	49.91^{***}	0.205
	RC4	46.13 (7.26)	60.91 (8.95)	128.65^{***}	0.399
	RC6	48.72 (8.35)	62.96 (9.42)	94.98^{***}	0.329
	RC7	42.01 (5.71)	53.58 (8.73)	109.07^{***}	0.360
	RC8	49.34 (5.05)	57.87 (7.75)	75.51^{***}	0.280
	RC9	42.76 (9.50)	56.00 (12.99)	56.243^{***}	0.225

$^{**}p \le 0.01$; $^{***}p \le 0.001$.

subjects undergoing a psychological evaluation of parental fitness. The primary aim was to test the hypothesis that CCL personality profiles, as measured by the MMPI-2-RF, differ from normative profiles. This hypothesis was based on the underreporting tendencies of CCL subjects reported in the literature, characterized by elevated L-r, K-r, and RC6 scales, suggesting the motivation of these subjects to present themselves in a positive light. Furthermore, the study differentiated between CCL women and men in order to determine whether there are specific MMPI-2-RF profiles among each gender.

First, it was assumed that CCL subjects would report higher scores in the underreporting validity scales (L-r and K-r) and lower scores in the overreporting validity scales (F-r, Fp-r, Fs-r, and RBS), compared to a normative sample (H1). The results confirmed this hypothesis, in line with the aforementioned literature (3, 6, 7). CCLs showed underreporting MMPI-2-RF profiles with elevated L-r and K-r linear T-scores approximately five points higher than the medium value of the normative data. In more detail, women's scores were almost seven points higher in the L-r scale and approximately six points higher in the K-r scale, relative to the normative sample. Men, in contrast, demonstrated an elevation of almost five points in the L-r scale and approximately three points in the K-r scale. These results aligned with the findings of Sellbom and Bagby (7), Archer et al. (3), and Kauffman et al. (6), though the latter two studies reported even higher mean T-scores for the combined sample, relative to the subjects in the present study. The CCL subjects in the present study showed elevated validity scales (L-r and K-r), with 18% demonstrating an elevated L-r scale and 20% demonstrating an elevated K-r scale at or above a T-score of 65—almost twice the 10% expected according to the standardized data. These results were especially salient for CCL women, who represented themselves as more adapted and unusually virtuous compared to normative subjects. CCL MMPI-2-RF profiles were also characterized by lower linear T-scores (ranging from two to five points) in the overreporting validity scales (F-r, Fp-r, Fs-r, and RBS), compared to the standard average. Furthermore, data on the frequencies of elevations in MMPI-2-RF validity scales reveal that such elevations should only be expected in 8% of the sample; however, a lower percentage of CCL subjects in the present study produced T-scores > 65, confirming that caregivers in child custody disputes are prone to describing themselves as more righteous, healthy, and vigorous than they effectively are. The findings with respect to the underreporting and overreporting validity scales are also consistent with other MMPI-2 research (8), which has shown CCL subjects to be more psychologically defensive than other groups, as reflected in their responses to MMPI-2 validity scales relating to defensiveness (4, 5, 19, 21, 22).

With respect to the RC scales (H2), CCL subjects in the present study scored lower than the normative sample on all but RC2 (Depressive Symptoms) and RC8 (Thinking Disorders), which showed scores in the average range. RC6 (Ideas of Persecution) was the most elevated of the RC scales, as also shown in previous studies (3, 6, 7). Elevations in the clinical range occurred most frequency in RC1 (10.3%), RC2 (9.9%), and RC6 (14.3%). Elevations above a 65 T-score in RC6 were highlighted by Kauffman et al. (6), Archer et al. (3), and Sellbom and Bagby (7). In the other RC scales, the percentage of subjects showing elevated T-scores was lower than the expected 8%, based on the normative sample. Overall, the results suggest that CCL subjects have a greater propensity to present themselves in a socially desirable way, together with higher levels of suspiciousness and mistrust and fewer displayed symptoms and feelings of negativity.

The findings support the hypothesis that there are gender differences in the MMPI-2-RF profiles of CCL subjects undergoing clinical assessment (H3), as previously highlighted by Archer et al. (3) with the MMPI-2-RF and Roma et al. (11) with the MMPI-2. In more detail, women appeared deeply motivated to display a faking-good defensive profile, together with lower levels of cynicism and antisocial behaviors, compared to CCL men.

This trend could be explained by several reasons: women may have a stronger desire to gain custody of their children in order to avoid the social stigma of being judged as unsuitable mothers; mothers are generally considered the leading figures in operative caregiving, due to a rigid and conservative view of feminine roles that leads them to deny psychological imperfections; women are frequently in a weaker economical position relative to men, and this may lead them to develop a defensive attitude.

According to the fourth hypothesis (H4), it was expected that the MMPI-2-RF of the CCL sample would demonstrate a comparable implicit structure to that of a normal, non-forensic population. The findings did not bear out this assumption: rather, in contrast to the findings of Archer et al. (3), the intercorrelations reported among the nine RC scales in the CCL sample differed from those reported in the technical manual. This was true especially for women, whose RC scales showed 15 (out of 36) significantly different intercorrelations compared to women in the normative sample. This result suggests a different implicit structure of the MMPI-2-RF and highlights the need to interpret CCL profiles in the context of normative data collected specifically in a forensic setting.

In order to determine whether the MMPI-2-RF could be used to more deeply classify CCL subjects, both with and without recourse to gender, the validity and RC scales were used to define CCL typologies based on the psychological characteristics CCL subjects were aware of or wished to communicate (H5). Two-step cluster analyses showed three typical female CCL profiles and two typical male CCL profiles. Women in cluster 1 (8.8%) complained of problems related to health, cognitive symptoms, low positive emotions, and suspiciousness. In cluster 2, which comprised 53.9% of female CCLs, subjects showed a mixed profile characterized by a constricted range of feeling with limited emotional responsiveness across a wide spectrum. They also complained of medical symptomatology and unusual thoughts. Women in cluster 3 (37.3%) tended to show more adaptive psychological functioning and attempted to deny, rationalize, and limit self-disclosure, probably due to the evaluative/forensic setting. It is interesting to note that the three clusters did not differ in their communication of uncommon virtues (L-r) and thus their attitude to underreporting. Among CCL men, 77% fell in cluster 1, demonstrating underreporting profiles that masked other personality characteristics. Men in cluster 2 (23%), however, showed more problematic profiles with low positive emotion, mistrust, somatic complaints, and difficulties with people in a position of authority.

Finally, among the entire CCL sample, approximately 44% of the MMPI-2-RF profiles showed possible underreporting and, for this reason, could be considered worthless (H6). To the best of our knowledge, this was the first study to have included this kind of evaluation, digging up an overwhelming percentage of worthless protocols and calling researchers and forensic experts to join together to develop more effective methods of measuring CCL personality characteristics.

STRENGTHS AND LIMITATIONS

One limitation of the research design is that the sample was not classified according to participant age; however, this lack of stratification was consistent with the normative group. The present study adds useful insight to the debate over the instruments that can be effectively used in forensic settings to assess the psychopathology and personality characteristics of parents undergoing a parental skills assessment. To the best of our knowledge, this study was the first to have administered the MMPI-2-RF in its own form and not to instead interpret scores that have been extracted and converted from the MMPI-2 (a similar but longer test). Moreover, the study analyzed the MMPI-2-RF protocols of men and women involved in a real forensic parenting skills evaluation, avoiding an experimental paradigm. On the basis of the results, many issues arise for researchers and practitioners. Most notably, the worthlessness of approximately half of all MMPI-2-RF protocols, due to the underreporting attitude of CCL respondents, requires the test to be administered in combination with a clinical interview and other measures (e.g., projective methods) that are less subject to simulation. This alarming finding is comparable with the results of previous studies of the MMPI-2-RF and MMPI-2 in forensic settings (40) with subjects who have driven under the influence of alcohol (13) and mothers who have committed filicide (41, 42), as well as studies on malingering (12, 14, 41). The worrying percentage of pointless protocols highlights the need to mainstream and administer the MMPI-2-RF more effectively with new and promising methods and strategies, drawing on, for instance, reaction time, machine learning, and mouse tracking (12, 43). Future studies could investigate the personality profile of CCL subjects, comparing the MMPI-2-RF with other personality assessment instruments; research could also examine whether differences exist within the personality profiles of CCLs involved in child protection matters for neglect, violence or abuse, relative to a normative population.

DATA AVAILABILITY STATEMENT

The dataset used and analyzed during the current study is available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study was carried out with written informed consent by all subjects, in accordance with the Declaration of Helsinki. It was approved by the local ethics committee (Board of the Department of Human Neuroscience, Faculty of Medicine and Dentistry, Sapienza University of Rome).

AUTHOR CONTRIBUTIONS

All authors helped to conceive and plan the study and prepared and approved the final manuscript. PR conducted the data collection and produced the first draft of the final manuscript. SF, MCV, and DM supervised the data collection. PR and CM conducted the analyses and wrote the manuscript. MCV, DM, and AD carefully read the final version of the manuscript and revised it.

REFERENCES

1. Bonieskie LM. An examination of personality characteristics of child custody litigants on the Rorschach. *Diss Abstr Int* (2000) 61(6-B):3271.
2. Kennelly JJ (2002). Rorschach responding and response sets in child custody evaluations. *Dissertation Abstracts International: B. The Sciences and Engineering,* 63(6-B): 3034.
3. Archer EM, Hagan L, Mason J, Handel R, Archer RP. MMPI-2-RF characteristics of custody evaluation litigants. *Assessment* (2012) 19(1): 14–20. doi: 10.1177/1073191110397469
4. Bathurst K, Gottfried AW, Gottfried AE. Normative data for the MMPI–2 in child custody litigation. *Psychol Assess* (1997) 9(3):205–11. doi: 10.1037/1040-3590.9.3.205
5. Bagby RM, Nicholson RA, Buis T, Radovanovic H, Fidler BJ. Defensive responding on the MMPI-2 in family custody and access evaluations. *Psychol Assess* (1999) 11(1): 24–8. doi: 10.1037/1040-3590.11.1.24
6. Kauffman CM, Stolberg R, Madero J. An examination of the MMPI-2-RF (Restructured Form) with the MMPI-2 and MCMI-III of child custody litigants. *J Child Custody* (2015) 12(2): 129–51. doi: 10.1080/15379418.2015.1057354
7. Sellbom M, Bagby RM. Validity of the MMPI-2-RF (Restructured Form) L-r and K-r scales in detecting underreporting in clinical and nonclinical samples. *Psychol Assess* (2008) 20(4): 370–6. doi: 10.1037/a0012952
8. Butcher JN, Graham JR, Ben-Porath YS, Tellegen A, Dahlstrom WG, Kaemmer B. *MMPI-2 (Minnesota Multiphasic Personality Inventory 2): manual for administration, scoring, and interpretation, revised edition*. Minneapolis, MN: University of Minnesota Press (2001). doi: 10.1037/t15120-000
9. Archer RP, Buffington-Vollum JK, Stredny RV, Handel RW. A survey of psychological test use patterns among forensic psychologists. *J Pers Assess* (2006) 87(1): 84–94. doi: 10.1207/s15327752jpa8701_07
10. Otto R. Use of the MMPI-2 in forensic settings. *J Forensic Psychol Pract* (2002) 2(3): 71–92. doi: 10.1300/J158v02n03_05
11. Roma P, Ricci F, Kotzalidis GD, Abbate L, Lavadera AL, Versace G, et al. MMPI-2 in child custody litigation: a comparison between genders. *Eur J Psychol Assess* (2014) 30(2):110–6. doi: 10.1027/1015-5759/a000192
12. Roma P, Verrocchio MC, Mazza C, Marchetti D, Burla F, Cinti ME, et al. Could time detect a faking-good attitude? A study with the MMPI-2-RF. *Front Psychol* (2018) 9:1064. doi: 10.3389/fpsyg.2018.01064
13. Roma P, Mazza C, Ferracuti G, Cinti ME, Ferracuti S, Burla F. Drinking and driving relapse: data from BAC and MMPI-2. *PLoS One* (2019a) 14(1):e0209116. doi: 10.1371/journal.pone.0209116
14. Roma P, Mazza C, Mammarella S, Mantovani B, Mandarelli G, Ferracuti S. Faking-good behavior in self-favorable scales of the MMPI-2: a study with time pressure. *Eur J Psychol Assess* (2019b) 1–9. doi: 10.1027/1015-5759/a000511
15. Ackerman MJ, Pritzl TB. Child custody evaluation practices: a 20-year follow-up. *Family Court Rev* (2011) 49(3):618–28. doi: 10.1111/j.1744-1617.2011.01397.x
16. Ackerman MJ, Ackerman MC. Custody evaluation practices: a survey of experienced professionals (revisited). *Prof Psychol: Res Pract* (1997) 28(2):137–45. doi: 10.1037/0735-7028.28.2.137
17. Bow J, Flens J, Gould J. MMPI-2 and MCMI-III in forensic evaluations: a survey of psychologists. *J Forensic Psychol Pract* (2010) 10(1) :37–52. doi: 10.1080/15228930903173021
18. Quinnell FA, Bow JN. Psychological tests used in child custody evaluations. *Behav Sci Law* (2001) 19(4):491–501. doi: 10.1002/bsl.452
19. Carr GD, Moretti MM, Cue BJH. Evaluating parenting capacity: validity problems with the MMPI-2, PAI, CAPI, and ratings of child adjustment. *Prof Psychol: Res Pract* (2005) 36(2): 188–96. doi: 10.1037/0735-7028.36.2.188
20. Fariña F, Redondo L, Sejio D, Novo M, Arce R. A meta-analytic review of the MMPI validity scales and indexed to detect defensiveness in custody evaluations. *Int J Clin Health Psychol* (2017) 17(2):128–38. doi: 10.1016/j.ijchp.2017.02.002
21. Siegel JC. Traditional MMPI-2 validity indicators and initial presentation in custody evaluations. *Am J Forensic Psychol* (1996) 14(3):55–63.
22. Strong DR, Greene RL, Hoppe C, Johnston T, Olesen T. Taxometric analysis of impression management and self-deception on the MMPI-2 in child-custody litigants. *J Pers Assess* (1999) 73(1):1–18. doi: 10.1207/S15327752JPA730101
23. Roma P, Piccini E, Ferracuti S. Using MMPI-2 in forensic assessment. *Rassegna Italiana di Criminologia* (2016) 10(2): 116–122.
24. Caldwell A. Symposium conducted at the annual convention of the American Psychological Association. Interpreting MMPI data custody evaluations: a clinical perspective. In S. Podrygula (Chair), MMPI use in child custody evaluations: integrating the data (1995).
25. Caldwell A. How can the MMPI-2 help child custody examiners? *J Child Custody: Res Issues and Pract* (2005) 2(1/2):83–117. doi: 10.1300/J190v02n01_06
26. Ben-Porath YS, Tellegen A. Empirical correlates of the MMPI-2 restructured clinical (RC) scales in mental health, forensic and nonclinical settings: an introduction. *J Pers Assess* (2008) 90(2): 119–21. doi: 10.1080/00223890701845120
27. Tellegen A, Ben-Porath YS. *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2-Restructured Form): technical manual*. Minneapolis, MN: University of Minnesota Press (2008). doi: 10.1037/t15121-000
28. Butcher JN, Dahlstrom WG, Graham JR, Tellegen A, Kaemmer B, (1989). *MMPI-2: manual for administration and scoring*. Minneapolis: University of Minnesota Press.
29. ISTAT. *Matrimoni, separazioni e divorzi*. Rome: ISTAT (2015). https://www.istat.it/it/files//2016/11/matrimoni-separazioni-divorzi-2015.pdf.
30. ISTAT. *Rapporto annuale 2012—La situazione del Paese*. Rome: ISTAT (2014). https://www.istat.it/it/files/2012/05/Rapporto-annuale-2012.pdf.
31. Resendes J, Lecci L. Comparing the MMPI-2 scale scores of parents involved in parental competency and child custody assessments. *Psychol Assess* (2012) 24(4):1054. doi: 10.1037/a0028585
32. Sirigatti S, Faravelli C. *MMPI-2-RF*. Firenze: Giunti OS (2012).
33. Tellegen A, Ben-Porath YS. *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2-Restructured Form): technical manual*. Minneapolis, MN: University of Minnesota Press (2011).
34. Tellegen A, Ben-Porath YS, McNulty JL, Arbisi PA, Graham JL, Kaemmer B. *The MMPI-2 Restructured Clinical (RC) scales: development, validation and interpretation*. Minneapolis, MN: University of Minnesota Press (2003).
35. Tellegen A, Ben-Porath YS. The new uniform T-scores for the MMPI-2: rationale, derivation, and appraisal. *Psychol Assess* (1992) 4(2): 145. doi: 10.1037/1040-3590.4.2.145
36. Pierce CA, Block RA, Aguinis H. Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educ Psychol Meas* (2004) 64(6): 916–24. doi: 10.1177/0013164404264848
37. Field A. *Discovering statistics using IBM SPSS statistics*. SAGE (2013). London: Sage Publication.
38. Wendler T, Gröttrup S. Cluster analysis. In: *Data mining with SPSS Modeler*. Springer, Cham (2016), 587–712. doi: 10.1007/978-3-319-28709-6_7
39. Scheffé H. *The analysis of variance*. New York, NY: John Wiley & Sons (1959), 351–8.
40. Roma P, Pazzelli F, Pompili M, Girardi P, Ferracuti S (2013). Shibari: double hanging during consensual sexual asphyxia. *Archi Sex Behav*, (5): 895–900.
41. Giacchetti N, Roma P, Pancheri C, Williams R, Meuti V, Aceti F. Personality traits in a sample of Italian filicide mothers. *Rivista di Psichiatria* (2019) 54(2):67–74. doi: 10.1708/3142.31247
42. Mazza C, Monaro M, Orrù G, Colasanti M, Ferracuti S, Burla F, et al. Introducing machine learning to detect personality faking-good in a male sample: a new model based on Minnesota multiphasic personality inventory-2 restructured form scales and reaction times. *Front Psychiatry* (2019) 10:389. doi: 10.3389/fpsyt.2019.00389
43. Monaro M, Gamberini L, Sartori G. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS One* (2017b) 12(5):e0177851. doi: 10.1371/journal.pone.0177851

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article’s readership