Self-initiated versus instructed cheating in the physiological Concealed Information Test

Geven, L.M.; klein Selle, N.; Ben-Shakhar, G.; Kindt, M.; Verschuere, B.

DOI
10.1016/j.biopsycho.2018.09.005

Publication date
2018

Document Version
Final published version

Published in
Biological Psychology

Biological Psychology 138 (2018) 146–155

Contents lists available at ScienceDirect

Biological Psychology

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

Self-initiated versus instructed cheating in the physiological Concealed Information Test

Linda Marjoleine Geven[a,b,*], Nathalie klein Selle[b], Gershon Ben-Shakhar[b], Merel Kindt[a], Bruno Verschuere[a]

[a] Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018WS Amsterdam, the Netherlands
[b] Department of Psychology, Hebrew University of Jerusalem, Israel Mount Scopus, 91905, Jerusalem, Israel

ARTICLE INFO

Keywords:
Concealed Information Test (CIT)
Memory detection
Deception
External validity
Cheating
Dishonesty

ABSTRACT

The validity of the Concealed Information Test (CIT) to detect recognition of critical details has been demonstrated in hundreds of laboratory studies. These studies, however, lack the factor of deliberate intent to deceive. This disparity between research and practice may affect the generalizability of laboratory based CIT findings.

In the current study, 65 out of 174 participants cheated on their own initiative in a trivia quiz. These self-initiated cheaters were compared to 68 participants who were explicitly requested to cheat. Skin conductance, heart rate, and respiration were found to detect concealed information related to cheating. No significant differences emerged between self-initiated and instructed cheaters, supported by Bayesian statistics showing substantial evidence for the null hypothesis. The data demonstrate that the validity of the CIT is not restricted to instructed deception. This finding is encouraging from an ecological validity perspective and may pave the way for further field implementation of memory detection.

1. Introduction

The Concealed Information Test (CIT; initially called the Guilty Knowledge Test; Lykken, 1959) seems to offer a valid method of detecting whether examinees recognize critical details. Yet, there are notable differences between laboratory and real-world applications. The current study examines whether the validity of the CIT for self-initiated cheating (characteristic of real-world applications) differs from the validity observed when examinees are instructed to cheat (typical of laboratory studies).

The CIT is designed to detect concealed knowledge rather than deception and is constructed like a multiple-choice test. Specifically, it comprises several questions, each having one critical (e.g., crime-related) and several equally plausible irrelevant items. The CIT assesses whether the examinee recognizes the critical item, assumed to be known only to individuals involved in a crime (i.e., the perpetrator and the investigative team). For example, in case of a homicide, the CIT might include questions concerning the murder weapon or the location of the victim in the crime scene (e.g., 'Where was the victim found? a) bathroom, b) kitchen, c) bedroom, d) garden, e) living room').

During the presentation of the CIT items, physiological responses (e.g., Skin Conductance, Heart Rate, and Respiration) are assessed. Distinct responses to the correct item compared to responses to the plausible, but incorrect alternatives, indicate recognition of the critical items. This differential response to the crime-related information is known as the CIT effect. Innocent examinees who are unaware of the location of the victim, in this example, are expected to respond similarly to all presented options.

Traditionally, the CIT effect has been accounted for by Orienting Response (OR) theory (see Lykken, 1974). As ORs are elicited by salient stimuli (Sokolov, 1963), it was assumed that enhanced responses to the critical CIT items reflect an OR. More specifically, the critical detail holds an increased significance to the perpetrator, but not to an innocent examinee. While OR theory generally fits the data (e.g., habituation effects upon repetition; Verschuere, Crombez, De Clercq, & Koster, 2004), it has recently been demonstrated that this theory does not explain the CIT effect with all physiological measures.

Specifically, new insights (Klein Selle, Verschuere, Kindt, Meijer, & Ben-Shakhar, 2016; klein Selle, Verschuere, Kindt, Meijer, & Ben-Shakhar, 2017) suggest that the OR account of the CIT effect particularly holds for the Skin Conductance Response (SCR), while Heart Rate (HR) and Respiration Line Length (RLL) may reflect attempts to inhibit the arousal associated with the presentation of the critical items. In order to remain undiscovered, knowledgeable individuals need to

* Corresponding author at: Faculty of Social and Behavioural Sciences, Department of Clinical Psychology, University of Amsterdam, Amsterdam, the Netherlands.
E-mail address: L.M.Geven@uva.nl (L.M. Geven).

https://doi.org/10.1016/j.biopsycho.2018.09.005
Received 4 April 2018; Received in revised form 10 September 2018; Accepted 12 September 2018
Available online 17 September 2018
0301-0511/

suppress this increased arousal which accompanies the presentation of crime-related details. Yet, this effort comes at a cost (Pennebaker & Chew, 1985), paradoxically reflected by an even stronger physiological response to the critical items.

1.1. Laboratory research: current findings and limitations

The majority of the existing data confirming the validity of the CIT is derived from laboratory studies. There exist several paradigms, with the mock crime procedure being amongst the most popular (Meijer, Klein Selle, Elber, & Ben-Shakhar, 2014). In the mock crime procedure, participants in the "guilty" condition are instructed to enact an antisocial act for the sake of the experiment (e.g., theft of an envelope with money). The details of this crime (e.g., the amount of money stolen) subsequently serve as critical items in the CIT. Meta-analyses of CIT studies have demonstrated large effect sizes for differentiating between knowledgeable and unknowledgeable individuals (d = 1.55, 95% CI [1.44; 1.66] for SCR, d = 1.11, 95% CI [1.00; 1.22] for RLL, and d = 0.89, 95% CI [0.80; 0.99] for HR; Meijer et al., 2014).

Yet, mock crime procedures differ from real life in many aspects, including characteristics of the examinees (Lykken, 1959; Verschuere, Crombez, Koster, & De Clercq, 2007), the stakes of the test outcome, and the (emotional) nature of the questions (Klein Selle, Verschuere, Kindt, Meijer, Nahari et al., 2017). Therefore, questions can be raised about the suitability of mock-crime studies to assess the validity of the CIT. A notable difference is that in real life, people deliberately engage in the antisocial act under investigation, whereas laboratory participants merely follow the experimental instructions in a mock-crime (paradoxically making the instructed antisocial act actually a prosocial act). Thus, while laboratory research is important for establishing a controlled environment in which particular variables can be disentangled, it has limited external validity. This gap between research and practice may affect the generalizability of laboratory based CIT findings (see also Ben-Shakhar & Nahari, 2018).

1.2. Bridging the gap

In recent years, researchers have attempted to tackle the external validity problem by systematically manipulating various factors that differ between the laboratory and real world setting. As deception is commonly defined as a voluntary act (see Vrij, 2004), this study focuses on the factor of intent when investigating the validity of the CIT. As far as we know, only two studies have looked at the influence of deliberate deception. Nahari, Breska, Elber, Klein Selle, and Ben-Shakhar (2017) let participants freely choose to perform either a mock-crime or an innocent computer task and compared those participants to an 'instructed' condition in which the subjects were explicitly ordered by the experimenter to do either of the tasks. The study revealed similar CIT detection efficiency for participants who made an active decision to commit the mock crime and those who committed the crime upon instruction. A downside of their design was that participants choosing to commit the mock crime in fact still obediently complied with the researchers' instructions.

To address this limitation, we developed a new design that provides participants with an opportunity and an incentive to cheat, allowing a more unambiguous comparison of self-initiated and instructed cheating. In this design, participants are randomly assigned to either a condition in which they were ordered to cheat on a 10-item trivia quiz by looking up the answers online (mimicking the typical laboratory set-up) or to a condition in which they were provided with the opportunity and incentive to cheat, yet without explicit instructions to do so. Unbeknownst to the participants, the trivia quiz was constructed in a way that the first eight questions were fairly easy, but earning the bonus would require also correctly answering the last two questions that were piloted extensively to be nearly impossible to solve without cheating.

A first study with this design, conducted in an online setting and relying on response time as the dependent measure in the CIT, found that 175 out of 259 (67.6%) participants cheated on their own initiative (Geven, Ben-Shakhar, Kindt, & Verschuere, 2018). The results revealed that the RT-CIT was a valid measure of detecting concealed information, while its validity was unaffected by self-initiated versus instructed cheating.

In the present study, we use our novel design to study the impact of spontaneous cheating on autonomic responding to concealed information. This extension is important for at least two reasons. First, applied usage in criminal proceedings currently relies solely on autonomic nervous system measures (Osugi, 2018). Second, new insights reveal response fractionation: different CIT response measures tap into different processes. The RT-CIT (Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017), presumed to rely on response inhibition, may not be affected by putative process variables such as item saliency (klein Selle, Verschuere, Kindt, Meijer, & Ben-Shakhar, 2017) or increased motivation to avoid detection (Kleinberg & Verschuere, 2016) associated with self-initiated cheating. While HR and RLL mainly depend on response inhibition, SCR is reasoned to primarily reflect the orienting response. As self-initiated cheating might increase item saliency, it may affect the SCR CIT effect.

As a secondary research question, our design allowed to examine whether personality traits predict self-initiated cheating. Previous research has shown that low Honesty-Humility scores - reflecting the inclination to break rules in order to obtain material or financial gains - were predictive of cheating behavior (Hilbig & Zettler, 2015). Following the findings from Hilbig and Zettler (2015), it was expected that participants who deliberately cheat in the current paradigm show lower Honesty-Humility scores than fair players. It remains unclear to date whether such differences are driven by the personality of the cheater (i.e., particularly low HH scores) and/or by that of fair players (i.e., particularly high HH scores). Besides investigating whether indeed participants who deliberately cheat show lower Honesty-Humility scores than fair players, the current paradigm allows for an additional comparison with the randomly assigned group of instructed cheaters. The instructed cheating group can serve as a norm group to reveal whether the effect of personality on cheating is driven by significantly low Honesty-Humility of cheaters, or by high Honesty-Humility scores of fair players.

2. Method

Ethical approval was obtained from the Ethics Review Board of the University of Amsterdam (2016-CP-6440). All participants provided written consent before taking part in the study. All materials, data, and scripts are publicly available on https://osf.io/gkj5w/.

2.1. Participants

The sample consisted of 260 individuals (75% female), who were recruited through a university portal or through advertisements on social media and received a monetary compensation. Their average age was 24.33 years old (SD = 7.68, range from 18 to 67). Participants were randomly allocated to the instructed cheating versus the possibility to cheat condition, with a 1:3 ratio. The latter condition was subsequently split in self-initiated cheaters versus fair players, depending on the participants' performance on the trivia quiz. This randomization resulted in non-significant differences in age and gender between participants instructed to cheat and participants given the possibility to cheat, $t(240) = 1.14$, $p = .254$, and $X^2(1) = 0.20$, $p = .655$, respectively.

2.1.1. Instructed cheaters, self-initiated cheaters and fair players

Based upon the instructions and their performance on the trivia quiz, participants were classified as instructed cheaters (i.e., those given

explicit instructions to cheat), self-initiated cheaters (i.e., those who had the opportunity to cheat and answered the two cheating-evoking questions correctly) or fair players (i.e., those who had the opportunity to cheat and answered neither of the two cheating-evoking questions correctly).

Eight individuals, who either reported a wrong answer or failed to enter the answer they did look up, were excluded from further analyses. Ten participants were not included in the data analyses because they did not follow the instructions for the trivia quiz (i.e., they did not look up the correct answers, although explicitly instructed to do so) or due to technical problems. Thus, the final sample consisted of 242 participants (75.6% female) with an average age of 24.28 years (SD = 7.37).

Sixty-five out of 174 participants (37.4%) in the opportunity to cheat condition cheated on their own initiative, and were labeled as self-initiated cheaters. The other 109 participants were labeled fair players. The condition in which cheating was explicitly instructed, consisted of 68 participants. Group characteristics and comparisons in age, sex, and personality measures are described in the Results.

2.2. Procedure

To conceal the true purpose of the experiment, the study was phrased as an English language proficiency and knowledge test, rather than a lie detection experiment. Since the goal was to compare deliberate, self-initiated deceptive behavior with instructed cheating, it was crucial that participants did not have prior expectations or knowledge on the real aim of the study. Participants were informed in advance about the use of psychophysiological measures while doing a task on the computer, but were told this would assess their physical reactions to recognition of English words rather than to detect cheating.

Upon arrival, all participants were asked to wash their hands, read the information brochure and sign the informed consent. Next, they completed the Dutch 100-item version of the HEXACO Personality Inventory Revised (De Vries, Lee, & Ashton, 2008). Then participants engaged in two seemingly relevant filler tasks in separate browser tabs on Qualtrics assessing their English grammar and verbal proficiency, followed by a trivia quiz with ten open (eight easy and two difficult, cheating-evoking) questions in English.

Before beginning these tasks and being left alone in the room, participants were told that answering all questions correctly would entitle them to a €5 bonus in addition to the standard payment for participation in the experiment. An additional instruction was given to participants in the instructed to cheat condition only, explicitly mentioning that they could use Google to find any answers they did not know. All participants were unaware of the fact that the quiz was constructed in a way that it would be almost impossible to earn a bonus without looking up the correct answers to the last two questions. As a result, participants claiming the trivia bonus for answering all ten questions correctly, were assumed to be cheaters. In addition to relying solely on the participants' answers, the experimenter could remotely access the computer screen (through split screen) and observe cheating when participants used the desktop computer to look up the answer.

Various tactics were used to give participants sufficient opportunity to engage in self-initiated cheating. First, all participants were left alone in a room to complete the two filler tasks and the trivia quiz on the computer. They were told that the experimenter would not be back for at least 20 minutes, thereby reflecting low supervision and low risk of getting caught cheating (Mazar, Amir, & Ariely, 2008; Nagin & Paternoster, 1993). Moreover, cheating would have been quick and almost effortless as it was ensured that the correct answer would appear within the first three results in the search engine Google (see Domnich et al., 2015). As a monetary incentive has been shown to increase deceptive behavior (Gino & Pierce, 2009), participants received a bonus upon answering all questions correctly. Additionally, the two cheating-evoking questions were intentionally placed at the bottom of the questionnaire as participants in a study by Effron, Bryan, and Murnighan (2015) were more likely to cheat and lie when facing the last chance to obtain a monetary reward. Furthermore, participants were given ten minutes time to finish the trivia quiz and decide whether or not to cheat, which could increase the rate of deceptive behavior. Lastly, in addition to the three tabs in the web browser reflecting the two different filler tasks and the quiz, a fourth tab was already opened on the Google search engine page. These four tabs were visible during the entire first phase of the experiment.

An instructed cheating condition was included, using the same procedure, except that participants in this condition were asked to actively look up the correct answers on the trivia questions. This condition was included to mimic the group of deceptive participants in laboratory experiments who are instructed to cheat or conceal information. This design thereby allows for a comparison of the psychophysiological responses of instructed and self-initiated cheaters. Participants in all conditions were not informed that they would undergo a lie detection test during the experiment. By doing so, encoding of the critical information reflected rather spontaneous behavior that is supposed to be externally more valid compared to typically used overlearned stimuli in mock-crime procedures (see also Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Meixner & Rosenfeld, 2014).

2.2.1. CIT

Only participants who cheated continued to the second part of the experiment. Since the aim of the current study is to compare self-initiated cheaters with those who were instructed to cheat, the participants who did not cheat were debriefed and did not continue to the CIT. A second experimenter, who was blind regarding the cheating condition (hence self-initiated or instructed) of the participant, attached the RLL belts as well as the SCR and HR electrodes and conducted the full CIT procedure.

Participants were told that they were suspected of cheating during the trivia quiz. Moreover, they were instructed to prove their innocence in the subsequent polygraph test by concealing their knowledge of the correct answers to the quiz questions. To ensure motivation to avoid detection, participants were offered a €5 incentive for successful concealment of their knowledge of the correct answers.[1]

Following an initial rest period of two minutes after attachment of the electrodes, participants were presented with the CIT questions. In addition to the two questions the participants cheated on during the trivia quiz, the CIT included two manipulation checks: an easy question from the trivia quiz, serving as a baseline for recognition (i.e., guilt check) and a new difficult question to which the participants did not know the correct answer to. This question served as a baseline for non-recognition (i.e., innocence check). This procedure allows to confirm basic CIT effects upon (non-)recognition, before investigating the influence of self-initiated versus instructed cheating on the outcomes of the CIT. The order of the four questions and their answering alternatives were randomly determined. In total, all four questions were repeated twice, with a short break between the two blocks to maintain participant's attention.

After the CIT, all cheating participants were debriefed on the true purpose of the experiment and were asked to answer truthfully several questions assessing their memory for the relevant items, using both recall and recognition formats. These memory tests were administered to examine whether participants remembered the relevant items correctly and could distinguish them from the irrelevant alternatives. Next, they were asked to complete a questionnaire regarding their motivation, memory and countermeasure attempts. Finally, participants were

[1] The bonus was paid out when the average SCR Z-score to the correct answers of the two cheating-evoking questions and the guilt-check question was below 0, which reflects no recognition of the critical items. The innocence check question was not included in this calculation, since all participants were unknowledgeable of its correct answer.

informed about their performance in the CIT and compensated for their participation in the experiment.

2.3. Material

2.3.1. HEXACO

Personality traits were assessed with the 100-item Dutch version of the HEXACO Personality Inventory Revised (Lee & Ashton, 2016). This inventory measures the six major dimensions of personality: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness and Openness to experience. Participants indicated the extent to which they agreed with each statement on a 5-point Likert scale ranging from 1 (*Completely disagree)* to 5 (*Completely agree)*. For the current sample, each scale revealed a good reliability using Cronbach's alpha (Cronbach, 1951): Honesty-Humility ($\alpha = .78$), Emotionality ($\alpha = .82$), Extraversion ($\alpha = .85$), Agreeableness ($\alpha = .84$), Conscientiousness ($\alpha = .82$), and Openness to experience ($\alpha = .78$).

2.3.2. Trivia quiz

Participants took part in a trivia quiz with ten open questions and were offered a monetary bonus of €5 for correctly answering all ten questions. This bonus was paid in cash after participants completed the first part of the experiment (i.e., the filler tasks and the trivia quiz, yet before commencing the CIT). They were, however, unaware of the fact that the quiz was constructed in a way that it would be almost impossible to earn the bonus without looking up the correct answers to the last two questions. In a second room, the experimenter could remotely access the computer screen (through split screen) and observe cheating, at least when participants used the desktop computer, rather than their personal smartphone. For the purpose of reaching external validity, the current paradigm involved active, self-initiated cheating as opposed to temptation resistance paradigms, for example, in which participants merely need to peek in an answer key that was carelessly left in the room by the experimenter (see DeAndrea, Carpenter, Shulman, & Levine, 2009).

The eight easy questions were correctly solved by 88–100% of the participants in extensive pilot studies. The last two questions, which were correctly answered by 0–6% of the pilot's participants, were classified as cheating-evoking.[2] Consequently, answering all questions correctly without cheating would be highly unlikely. Participants claiming the trivia bonus for answering all ten questions correctly, including the two cheating-evoking questions, were assumed to be cheaters.

2.3.3. CIT

Besides the two questions the participants cheated on during the trivia quiz (i.e., 'Who coined the term dinosaur?' as well as 'Who wrote the autobiographical book Prairie Tale: A Memoir?' as a replacement for 'Who wrote the autobiographical book Wishful Drinking'[3]), the CIT included one easy question from the trivia quiz (i.e., 'In which city is Buckingham Palace, the symbol and home of the British monarchy, located?') serving as a baseline for recognition. This 'guilt check' was added to show the usual CIT effect on a question that the participants definitely knew the correct answer to. Additionally, the CIT included a new question that had not been in the previous trivia quiz (i.e., 'Who assassinated American president James Garfield in 1881?'), but to which all participants in the pilot questionnaire did not know the correct answer. This 'innocence check' therefore served as a baseline for non-recognition, showing that participants would not show enhanced responses to unfamiliar items. Together, these additional questions allowed for a comparison between mere recognition and the influence of cheating on the outcome of the CIT (see Table 1).

The order of the four questions and their answering alternatives were randomly determined. In total, the four questions were repeated twice, with a short break between the two blocks to maintain participant's attention. Each question and alternative was presented both verbally through headphones and visually on the computer screen. The audio files were pre-recorded by a third party who was blind to the procedure. The question remained on the screen for 10 seconds, followed by the answering alternatives that were presented for 5 seconds each, with a mean inter-stimulus interval of 18 seconds (range 16–20). Between stimulus presentation, a fixation-cross appeared on the screen to maintain the attention of the participant.

The first answering alternative following the question was always a buffer-item, designed to absorb the initial OR. Subsequently, the critical item, four irrelevant items and a single catch item were presented in a random order. Catch items were included in the CIT to ensure the participants' attention to all presented items. Upon identifying the catch item amongst the presented alternatives, participants were instructed to repeat this specific item verbally. In the current CIT, the catch items consisted of random numbers ranging from 1 to 10, written in words. Whenever the answering alternatives consisted of two words (for example a first and last name), the catch item consisted of two numbers (e.g., three seven) in order to maintain a homogeneous sample of alternatives. Besides from repeating the catch items, participants were instructed to respond to all other items with a verbal "no". Altogether, participants were presented with two blocks of four questions, each consisting of seven items (i.e., 1 buffer, 1 critical, 4 irrelevant and 1 catch item), totaling 56 items.

2.3.4. Recall and recognition

Memory for the correct answers of the trivia quiz was assessed with a free recall followed by a recognition test. Firstly, the questions from the trivia quiz were presented on the screen with a text box in which participants were asked to freely recall and enter the correct answer to the guilt check, innocence check and the two cheating-evoking questions that were used in the CIT. For the recognition test, the questions from the trivia quiz were presented on the screen, each with five alternative options (the correct answer and four irrelevant options). Participants were asked to select the answer they deemed correct.

For free recall, answers were coded as either correct (1) or incorrect (0), leading to a total score per item type (i.e., 0–1 for the guilt and the innocence check and 0–2 for the cheating-evoking questions). Using arbitrary criteria, for the cheating-evoking questions and the innocence check answers were coded as correct if participants recalled both the first and last name correctly (e.g., Melissa Gilbert as the author of the novel Prairie Tale: A Memoir) or when they only recalled the last name correctly (e.g., Gilbert). When an incorrect first name was entered in combination with a correct last name, or only the first name was mentioned, the recall was coded as incorrect. For the guilt-check, the item consisted of a single word (e.g., London) that had to be entered correctly. For recognition, items were scored as either correct (1) or incorrect (0), leading to a total score per item type (i.e., 0–1 for the guilt and the innocence check and 0–2 for the cheating-evoking questions).

2.3.5. Follow-up questionnaire

Participants rated six questions designed to assess their motivational state on a 5-point Likert scale. This questionnaire measured how well participants were able to focus on the screen during the CIT, how involved they were in the study and how much they tried to avoid

[2] The cheating-evoking questions were correctly answered by 0% (Who coined the term dinosaur?), 3% (Who wrote the autobiographical book Wishful Drinking?) and 6% (Who wrote the autobiographical book Prairie Tale: A Memoir?) of the pilot sample. Thus, the likelihood of answering two of these questions correctly is close to 0.

[3] One of the two questions of the trivia quiz inserted to evoke possible cheating was on the author of the novel 'Wishful drinking'. However, halfway during data collection the author died, which evoked multiple news items occasionally mentioning her novel. Because of possible familiarity with the probe, we replaced this question with an equally difficult question (i.e., Who wrote the autobiographical book Prairie Tale: A Memoir?).

Table 1
Question types and experimental goals.

Question-Type	Question [Answer]	Experimental goal
Cheating-evoking	Who coined the term dinosaur? [Richard Owen] Who wrote the autobiographical book 'Prairie Tale: A Memoir'? [Melissa Gilbert]	Compare physiological response to answer for self-initiated and instructed cheating
Guilt-check	In which city is Buckingham Palace, symbol and home of the British monarchy, located? [London]	Manipulation check for recognition
Innocence-check	Who assassinated American president James Garfield in 1881? [Charles Guiteau]	Manipulation check for non-recognition

detection and appear innocent on the CIT. Then, participants rated their effort to suppress or raise their physiological responses during the test and freely elaborated on whether they had used strategies to avoid detection.

2.4. Data acquisition and reduction

The experiment was conducted in an air-conditioned laboratory. Stimulus presentation was performed using Presentation® software (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). Psychophysiological responses were measured and recorded with Vsrrp89 software, developed by the Technical Support Social and Behavioral Sciences at the University of Amsterdam.

Electrodermal activity was recorded with an amplifier using a 50 Hz, sine-shaped excitation voltage with an amplitude of 1Vpp. Two curved-shape sintered silver–silver chloride (Ag/AgCl) electrodes (20 × 16 mm) were connected to the palmar surface of the distal phalanges of the left index and left ring finger with adhesive tape. The SCR was measured from 1 second to 5 seconds after stimulus onset and defined as the maximal increase in conductance during this time window.

The ECG measure was acquired by placing a set of three Ag/AgCl electrodes (3M™ Red Dot™ disposables, type 2249-50) in a standard Einthoven lead-II configuration: one electrode attached near the distal end of the left collarbone, one electrode placed near the distal end of the right collarbone, and one electrode placed on the left lateral base of the chest. Prior to analysis, the inter-beat intervals were converted to HR in beats per minute (bpm) per real-time epoch (1 second). The 15 second-by-second post-stimulus HR values were baseline-corrected by subtracting the average pre-stimulus baseline HR value (mean HR in the three seconds preceding stimulus onset), resulting in 15 post-stimulus difference scores (ΔHR). The average of these 15 scores was used as the HR deceleration dependent measure.

Respiration was measured with two Braebon Respiratory Effort Sensors, type REF 0522. The thoracic respiration belt was attached around the upper chest of the participant, just below the armpits. The second belt was placed around the participants' abdomen, slightly below the rib cage. Respiration responses were defined on the basis of the total RLL during a 15-second interval following stimulus onset, across the two belts. The RLL measure is a combination of the participants' depth of breathing (respiratory amplitude) as well as the rate of breathing (respiratory cycle). Each response is measured using ten 15-second windows, each beginning 100 milliseconds later than the preceding time window. The RLL (following Elaad, Ginton, & Jungman, 1992) was established by calculating the mean of these 10 length measures (ranging from 0.1 s after stimulus onset through 15.1 s, from 0.2 s through 15.2 s after stimulus onset, etc.).

2.4.1. Exclusion criteria

On participant level, individuals whose standard deviation of the raw SCR scores was below 0.01 throughout the entire CIT procedure ($n = 2$) were considered to be skin conductance non-responders and their data were eliminated from all SCR analyses. In addition, three participants were excluded from SCR analyses due to technical errors with the electrodes. Data from the HR measure were eliminated from analyses when anomalies ($n = 1$) or technical errors occurred ($n = 2$). RLL data from the thoracic respiration belt were excluded due to technical errors ($n = 4$).

For exclusions on response level, standard scores per item were computed (buffer and catch items were not included in the standardization procedure; see Ben-Shakhar & Elaad, 2002; Elaad & Ben-Shakhar, 1997). These within-subject Z-scores reflect the mean response to the irrelevant items subtracted from the mean response to the critical item, divided by the respective standard deviation, across both repetitions of the questions.

Further exclusions on response level were performed when participants showed a standard deviation of the raw SCR scores below 0.01 during the presentation of a question and the subsequent alternatives. In these cases, all SCR measurements regarding that specific question were discarded from further analyses due to strong habituation. Moreover, for each of the three dependent measures, item specific responses were removed if the standardized score was smaller than -5 or larger than 5, reflecting outliers. When a movement coincided with a positive standardized score (for SCR) or a positive standardized score larger than 2 or lower than -2 (for HR and RLL), the item was discarded from analyses (see also klein Selle et al., 2016; klein Selle, Verschuere, Kindt, Meijer, Ben-Shakhar, 2017). Following these exclusion criteria, 85.2% of the SCR data was included in the analyses. For the HR and RLL measures, 98.1% and 97.7% of the total number of responses, respectively were included in analyses.

3. Results

All analyses used an alpha level of 0.05. Effect sizes for the ANOVA are reported using Cohen's *f*. For follow-up contrasts Cohen's *d* is used[4]. Cohen's *d* for within-subject and between-subject comparisons are annotated as d_{within} and $d_{between}$. As a rule of thumb, Cohen (1992) proposed 0.20, 0.50 and 0.80 as thresholds for "small", "moderate" and "large" effects, respectively, for *d* values and 0.10, 0.25 and 0.40 as thresholds for "small", "moderate" and "large" effects, respectively for *f* values.

In addition, JZS Bayes factors (BF) were computed using JASP software version 0.8.4, representing numerical values quantifying the odds ratio between the null and the alternative hypothesis given the data. BF_{01} annotates how much more likely the data are under the null as compared to the alternative hypothesis, and BF_{10} annotates how much more likely the data are under the alternative as compared to the null hypothesis. For one-tailed testing, Bayes factors are reported as either predicting the null (BF_{0+}) or the alternative hypothesis (BF_{+0}). Default JZS prior with scaling factor $r = 0.707$ was used for the alternative hypothesis (see Rouder, Speckman, Sun, Morey, & Iverson, 2009). Using Jeffreys (1961) criteria, a Bayes factor (BF) larger than 3, 10, and 100 is taken as substantial, strong and decisive evidence for the respective hypothesis. It should be noted that values close to 1 fail to support either hypothesis.

[4] The probe-irrelevant within-subject contrast was calculated as $d_{within} = M_{(Response_{(probes)} - Response_{(irrelevants)})} / \sqrt{(SD_{(probes)}^2 + SD_{(irrelevants)}^2 - 2*r*SD_{(probes)}*SD_{(irrelevants)}}$, where *r* is the Pearson correlation between $Response_{(probes)}$ and $Response_{(irrelevants)}$. The between-subject contrast was calculated as $d_{between} = (M_{Response(probe\text{-}irrelevant\ difference\ group\ 1)} - M_{Response(probe\text{-}irrelevant\ difference\ group\ 2)}) / \sqrt{(((n_{(group\ 1}-1)*SD_{(probe\text{-}irrelevant\ difference\ group\ 1)}^2 + (n_{(group\ 2}-1)*SD_{(probe\text{-}irrelevant\ difference\ group\ 2)}^2)/n_{group1} + n_{group2} - 2)}$, see also Lakens (2013) and Suchotzki et al. (2017).

Table 2
Mean raw scores on the innocence check averaged for self-initiated cheaters and instructed cheaters.

Measure	Item	*M(SD)*	*t-test*	*p*-value	d_{within} *(95% CI)*	*BF*
SCR (n = 106)	Probe	0.36 (0.59)	$t(105) = 1.11$	.271	0.11 [−0.08; 0.30]	$BF_{01} = 5.13$
	Irrelevant	0.31 (0.44)				
HR (n = 130)	Probe	1.47 (3.25)	$t(129) = 1.55$	.125	0.14 [−0.04; 0.31]	$BF_{01} = 3.22$
	Irrelevant	0.98 (1.78)				
RLL (n = 133)	Probe	66.10 (19.42)	$t(132) = 0.02$	.984	0.00 [−0.17; 0.17]	$BF_{01} = 10.37$
	Irrelevant	66.11 (18.05)				

SCR in μS, HR change (from 3 s pre to 15 s post stimulus onset) in bpm and RLL in arbitrary units.

3.1. Physiological measures

As a first manipulation check, three two-tailed paired-samples t-tests were conducted comparing the raw responses of each physiological measure to the critical (i.e., *probe*) and irrelevant items of the innocence check question. Since participants did not know the correct answer, no significant difference should emerge in any of these comparisons.

As a second manipulation check, three one-tailed paired-samples t-tests were conducted to compare the raw responses to the probe and irrelevant items of the guilt check question, separately for the SCR, HR and RLL. Since participants did know the correct answer, a significant result is expected in each of these comparisons. Specifically, an increase in SCR and a decrease in HR and RLL are expected for the probe compared to the irrelevant options.

For the main analysis, a 2 (Condition: self-initiated cheaters vs. instructed cheaters, between-subjects) by 2 (Stimulus: probe vs. irrelevant, within-subjects) mixed ANOVA was conducted on the raw responses of each physiological measure for the cheating-evoking questions. We expect a main effect of Stimulus (i.e., a significant difference in responsivity to the probe compared to the irrelevant items, hence a CIT effect) and an interaction between Stimulus and Condition, particularly for SCR (i.e., the CIT effect being larger for self-initiated cheaters than for instructed cheaters).

3.1.1. Innocence check

Table 2 shows the averaged SCR, HR, and RLL results for probe and irrelevant items in self-initiated and instructed cheaters. For all three physiological measures, the difference between responses to the probe and irrelevant options was not significant, with the 95% confidence interval of the effect size including zero. Moreover, for all three measures, the Bayes factors showed substantial to strong evidence for the null hypothesis. These results reveal that no knowledge was indicated.

3.1.2. Guilt check

Table 3 shows the averaged SCR, HR, and RLL results for probe and irrelevant items in self-initiated and instructed cheaters. The probe-irrelevant difference was significant for all three physiological measures. Effect sizes were small for RLL, and moderate for the SCR and HR. The Bayes factors indicated that there was decisive evidence that the data were more likely under the alternative hypothesis for SCR and HR. For RLL this evidence was anecdotal.

3.1.3. Cheating-evoking questions

Table 4 shows the SCR, HR, and RLL results for probe and irrelevant items separately for self-initiated and instructed cheaters.

For the SCR, the ANOVA revealed a significant main effect of Stimulus, $F(1, 126) = 76.58$, $p < .001$, $f = 0.78$, $BF_{10} = 5.32e+11$, no significant main effect of Condition, $F(1, 126) = 0.02$, $p = .889$, $f = 0.00$, $BF_{01} = 4.55$, and no significant interaction between Condition and Stimulus $F(1, 126) = 0.02$, $p = .887$, $f = 0.00$, $BF_{01} = 5.34$. This indicates that there was a decisive probe-irrelevant difference in SCR, with substantial evidence that the SCR-CIT effect does not differ between self-initiated and instructed cheaters.

For the HR, the ANOVA revealed a significant main effect of Stimulus, $F(1, 128) = 58.70$, $p < .001$, $f = 0.68$, $BF_{10} = 1.35e+11$, no significant main effect of Condition, $F(1, 128) = 0.46$, $p = .498$, $f = 0.06$, $BF_{01} = 5.17$, and no significant interaction between Condition and Stimulus $F(1, 128) = 0.63$, $p = .428$, $f = 0.05$, $BF_{01} = 4.16$. This indicates that there was a decisive probe-irrelevant difference in HR, with substantial evidence that the HR-CIT effect does not differ between the self-initiated and instructed cheaters.

For the RLL, the ANOVA revealed a significant main effect of Stimulus, $F(1, 131) = 11.13$, $p < .001$, $f = 0.29$, $BF_{10} = 21.24$, no significant main effect of Condition, $F(1, 131) = 0.12$, $p = .725$, $f = 0.03$, $BF_{01} = 2.26$, and no significant interaction between Condition and Stimulus $F(1, 131) = 0.18$, $p = .674$, $f = 0.03$, $BF_{01} = 4.85$. This indicates that there was a strong probe-irrelevant difference across the conditions ($BF_{10} = 21.24$), with substantial evidence that the RLL-CIT effect does not differ between the self-initiated and instructed cheaters ($BF_{01} = 4.85$).

3.2. Who Cheats? Age, gender and personality

For all 242 participants who succesfully completed the HEXACO personality measure, no difference emerged in gender distribution between the three conditions (i.e., self-initiated cheaters, fair players, and instructed cheaters, $X^2(2) = 3.78$, $p = .151$). However, significant differences were revealed in the mean age of the participants, $F(2, 239) = 4.80$, $p = .009$, $f = 0.20$, $BF_{10} = 3.07$. This effect was driven by eight outliers of participants that differed more than three standard deviations in age from the mean. Table 5 shows the mean age, gender proportions and HEXACO factor scores for the self-initiated cheaters, fair players and the instructed cheaters after the exclusion of participants with outlying age (n = 234).[5]

A one-way ANOVA of condition on the HEXACO Honesty-Humility scores, revealed a statistically significant effect of condition, $F(2, 231) = 3.51$, $p = .032$, $f = 0.17$, $BF_{10} = 1.00$. For sake of completion, the groups were contrasted on their Honesty-Humility score. A planned one-tailed t-test showed a significant difference between self-initiated cheaters ($M = 3.28$, $SD = 0.54$) and fair players ($M = 3.48$, $SD = 0.52$, $t(165) = 2.29$, $p = .012$, $d = 0.37$, $BF_{10} = 3.78$). Post-hoc comparisons with Bonferroni correction revealed that self-initiated cheaters as well as fair players did not differ from the control group (i.e., instructed cheaters; $p = .067$, $d = 0.15$, $BF_{10} = 2.41$, and $p = 1.00$, $d = 0.01$, $BF_{01} = 5.90$, respectively). We also mention that the ANOVAs on the

[5] Analysis without exclusions based on age reached the following results (n = 242). A one-way ANOVA of condition on the HEXACO Honesty-Humility factors, revealed no statistically significant effect of condition on Honesty-Humility, $F(2, 239) = 2.63$, $p = .074$, $f = 0.15$, $BF_{01} = 2.21$. For completion sake, the groups were contrasted on their Honesty-Humility score. A planned one-tailed t-test showed a significant difference between self-initiated cheaters ($M = 3.32$, $SD = 0.56$) and fair players ($M = 3.48$, $SD = 0.52$, $t(172) = 1.90$, $p = .030$, $d = 0.30$, $BF_{10} = 1.72$). Post-hoc comparisons with Bonferroni correction revealed that self-initiated cheaters as well as fair players did not differ from the control group of instructed cheaters ($M = 3.50$, $SD = 0.44$), $p = .121$, $d = 0.13$, $BF_{10} = 1.35$, and $p = 1.00$, $d = 0.02$, $BF_{01} = 5.73$, respectively).

Table 3
Mean raw scores on the guilt check averaged for self-initiated cheaters and instructed cheaters.

Measure	Item	*M(SD)*	*t-test*	*p*-value	d_{within}(95% CI)	*BF*
SCR (n = 118)	Probe	1.02 (1.11)	$t(117) = 8.48$	< .001	0.78 [0.57;0.99]	$BF_{+0} = 1.79e+11$
	Irrelevant	0.41 (0.49)				
HR (n = 130)	Probe	0.08 (3.37)	$t(129) = 4.41$	< .001	−0.39 [−∞; −0.24]	$BF_{-0} = 1406$
	Irrelevant	1.52 (1.82)				
RLL (n = 133)	Probe	64.70 (18.35)	$t(132) = 2.35$	.010	−0.20 [−∞;0.06]	$BF_{-0} = 2.71$
	Irrelevant	66.37 (17.55)				

SCR in μS, HR change (from 3 s pre to 15 s post stimulus onset) in bpm and RLL in arbitrary units.

Table 4
Mean raw scores and main effect of Stimulus (in Cohens' *F*) on the cheating-evoking questions.

Measure	Item	*M(SD)*	*M(SD)*	$d_{between}$	*BF*
		Self-Initiated Cheaters (n = 61)	Instructed Cheaters (n = 67)		
SCR	Probe	0.89 (1.05)	0.90 (1.00)	0.78	$BF_{10} = 5.32e+11$
	Irrelevant	0.35 (0.44)	0.38 (0.47)		
		Self-Initiated Cheaters (n = 64)	Instructed Cheaters (n = 66)		
HR	Probe	−0.63 (3.01)	−0.61 (2.51)	0.68	$BF_{10} = 1.35e+11$
	Irrelevant	1.60 (1.23)	1.20 (1.50)		
		Self-Initiated Cheaters (n = 65)	Instructed Cheater (n = 68)		
RLL	Probe	63.58 (17.64)	64.94 (17.11)	0.29	$BF_{10} = 21.24$
	Irrelevant	66.25 (17.69)	67.01 (18.67)		

SCR in μS, HR change (from 3 s pre to 15 s post stimulus onset) in bpm and RLL in arbitrary units.

other HEXACO factors showed no effects for Emotionality, Extraversion, Agreeableness, Conscientiousness, or Openness to Experience.

3.3. Follow-up questionnaire

For each question in the post-CIT motivation questionnaire, an independent-samples *t*-test was conducted to evaluate whether the ratings of participants on their reported focus, involvement, motivation and use of countermeasures in the self-initiated cheating condition (n = 65) differed significantly from participants who were instructed to cheat (n = 68).

Analysis revealed no statistically significant difference between the two cheating conditions on reported focus, $t(131) = 0.19$, $p = .850$, $d_{between} = 0.03$, $BF_{01} = 5.30$, involvement, $t(118.66) = 0.39$, $p = .698$, $d_{between} = 0.07$, $BF_{01} = 5.03$, motivation to avoid detection, $t(131) = -1.70$, $p = .092$, $d_{between} = 0.30$, $BF_{01} = 1.45$, or to either suppress, $t(131) = 0.25$, $p = .800$, $d_{between} = 0.05$, $BF_{01} = 5.24$ or enhance physiological reactions, $t(131) = 0.80$, $p = .428$, $d_{between} = 0.14$, $BF_{01} = 4.02$.

3.4. Memory

For all memory data, two-tailed independent-samples *t*-tests were conducted to evaluate whether the memory of participants in the self-initiated cheating condition (n = 65) differed significantly from participants who were instructed to cheat (n = 68).

3.4.1. Recall

The easy (guilt-check) question of the trivia quiz was correctly answered by all but one instructed cheater and correctly answered by all self-initiated cheaters. Consequently, no difference occurred between the two conditions, $t(67) = 1.00$, $p = .321$, $d_{between} = 0.17$, with substantial evidence for the null hypothesis ($BF_{01} = 3.42$).

The difficult (innocence-check) question was correctly answered by only two self-initiated cheaters (3.1%) and one instructed cheater (1.5%). No statistically significant difference in recall ability was revealed between the two conditions, $t(131) = 0.62$, $p = .536$, $d_{between} = 0.11$, with substantial evidence for the null hypothesis ($BF_{01} = 4.52$).

There was no significant difference in recall rate for the two cheating-evoking questions between the instructed cheaters (M = 1.37, SD = 0.77) and the self-initiated cheaters (M = 1.52, SD = 0.71), $t(131) = 1.21$, $p = .229$, $d_{between} = 0.20$, with anecdotal evidence for the null hypothesis ($BF_{01} = 2.77$).

3.4.2. Recognition

The easy (guilt-check) question was correctly answered by all participants. The difficult (innocence-check) question was correctly chosen by 16 self-initiated cheaters (24.6%) and 9 instructed cheaters (13.2%). These recognition rates represents chance level and no significant differences between the conditions emerged, $t(121.48) = 1.68$, $p = .096$, $d_{between} = 0.29$, with anecdotal evidence for the null hypothesis ($BF_{01} = 2.05$).

Moreover, there was no significant difference in recognition rate for the two cheating-evoking questions, between the instructed cheaters (M = 1.99, SD = 0.12) and the self-initiated cheaters (M = 1.92,

Table 5
Mean age, gender proportions and HEXACO factor scores.

Measure	*M(SD)* Self-Initiated Cheaters (n = 59)	*M(SD)* Fair Players (n = 108)	*M(SD)* Instructed Cheaters (n = 67)	*p*-value	Effect size *f*	BF_{01}
Age	23.93 (5.43)	22.99 (4.35)	22.94 (4.07)	.378	0.09	9.07
Proportion Female	67.8%	80.6%	74.6%	.181	$\varphi_c = 0.12$	3.57
Honesty-Humility	3.28 (0.54)	3.48 (0.52)	3.49 (0.44)	.032	0.17	1.00
Emotionality	3.11 (0.60)	3.25 (0.57)	3.23 (0.56)	.286	0.11	7.04
Extraversion	3.44 (0.60)	3.57 (0.52)	3.58 (0.62)	.309	0.10	7.60
Agreeableness	2.87 (0.53)	3.03 (0.56)	3.00 (0.61)	.187	0.12	4.82
Conscientiousness	3.42 (0.53)	3.55 (0.54)	3.49 (0.55)	.275	0.11	6.69
Openness to Experience	3.67 (0.53)	3.51 (0.53)	3.58 (0.55)	.165	0.12	4.96

SD = 0.32), $t(81.16) = 1.46$, $p = .147$, $d_{between} = 0.29$, with anecdotal evidence for the null hypothesis ($BF_{01} = 2.77$).

4. Discussion

The present study seeks to bridge the gap between laboratory studies on the CIT and real-life applications. Contrasting the psychophysiological responses of self-initiated and instructed cheaters allowed us to examine whether the CIT is affected by the voluntary decision to cheat. We found that 37.4% of the participants who had the opportunity to maximize self-profit by behaving dishonestly, cheated on their own initiative. The guilt- and innocence check confirmed the validity of our CIT to pick up both recognition and lack of recognition. More importantly, we found that the CIT was similarly effective in detecting knowledge of instructed and spontaneous cheating with all three physiological measures.

4.1. External validity of the Concealed Information Test

Meta–analytic results based on laboratory paradigms have demonstrated the validity of the CIT in detecting the presence or absence of crime-related knowledge in an interviewees' memory (Meijer et al., 2014), with large effect sizes for SCR, RLL and HR measures (Cohen's d = 1.55, 1.11, and 0.89, respectively). Yet, for implementation in the field it is crucial to investigate the external validity of the CIT.

Research findings are generally expected to overestimate the effect that would be observed in the field. The current paradigm was created in an attempt to more closely mimic real-life dishonest behavior within a controlled environment and thereby verify whether this influences the validity of the autonomic CIT. Similar to the results obtained using response latency as a dependent measure (Geven et al., 2018), mainly targeting the inhibition component, the current study using SCR, HR and RLL did not reveal any detrimental effects of self-initiated cheating on the validity of the Concealed Information Test. In fact, Bayesian statistics showed that there was no apparent difference between self-initiated cheaters and the traditionally used condition in which rule-breaking behavior was explicitly instructed. Interestingly, it seems that the act of spontaneous cheating did not increase item saliency, reflected by equivalent levels of orienting, or arousal inhibition.

Moreover, other studies have examined boundary conditions that might limit the external validity of CIT laboratory studies. For example, Verschuere, Meijer, and de Clercq (2011) conducted a study in which actual suspects under police investigation had to conceal which card they had picked from a deck of cards. While these suspects revealed a higher baseline heart rate compared to laboratory participants, their physiological responses significantly changed upon the presentation of the picked card as opposed to irrelevant options. Osugi and Ohira (2018) investigated the effect of arousal, expected to be present in actual perpetrators committing a crime. Before engaging in a mock-murder, in which participants stabbed a mannequin with a sharp tool, participants viewed highly arousing pictures. Compared to the group who encoded the crime in a neutral state, an even larger CIT effect was observed for participants who committed the crime under arousal.

Clearly, field validity requires further investigation as it is typically lower than laboratory validity (National Research Council, 2003). Meanwhile, our findings along with other laboratory studies that have examined the role of factors that differentiate the lab from the field, show that such factors associated with more realistic set-ups do not necessarily harm the validity of the CIT.

4.2. Detecting knowledge of naturally encoded information

Another important difference between laboratory and realistic settings is related to the form by which critical items are encoded. Mock-crime items are typically encoded under optimal conditions: a highly controlled setting in which the pre-tested details are rehearsed until remembered perfectly (Carmel et al., 2003; Honts, Raskin, & Kircher, 2002). Moreover, participants are often aware of the fact that a deception detection test will be administered subsequently. These factors affecting memory might be an important limitation when studying the external validity of the CIT. First, in the field it is not known to the investigator whether the perpetrator actually paid attention to the details of the crime-scene and secondly, it cannot be assured that the culprit retains the critical items in memory and retrieves them during the administration of the CIT. In an attempt to address this important limitation, Meixner and Rosenfeld (2014) investigated the sensitivity of the CIT for incidentally acquired memory traces. On the first day of the experiment, participants walked around with a body camera during four hours. On the next day, participants were tested on their capacity to recognize and distinguish words related to the events recorded on the previous day from to irrelevant events. The findings showed good discrimination between the twelve knowledgeable participants and a control group of individuals who were tested on irrelevant items only.

The current results add to this literature, by revealing high detection and memory accuracy for both cheating groups, although the critical items were not perfectly rehearsed, but merely searched on Google. Since participants did not know that these items would be used later in the study, let alone in a memory detection test, both encoding as well as memory retention reflected natural processes.

Yet, while in the typical experiments the CIT is administered immediately after participants are exposed to critical items, realistic tests are mostly administered several days, weeks, or even months after the crime occurred (Ben-Shakhar & Furedy, 1990). Since memory naturally declines after a delay, this could be a serious pitfall for memory detection. In realistic scenarios, more reliable responses can be expected for questions targeting items that are directly associated with the crime, such as the murder weapon.

4.3. Who cheats? Age, gender, and personality

In addition to the main question we investigated whether personality affects the tendency to cheat and behave dishonestly. In line with previous research (Hilbig & Zettler, 2015), the deliberate decision to cheat or not (i.e., using Google to answer trivia questions, resulting in self-initiated cheaters and fair players) was related to Honesty-Humility scores (Ashton & Lee, 2005), although the effect was small (d = 0.37).

Now, are self-initiated cheaters especially dishonest or are fair players rather especially honest? To explore this question, we also compared these two groups to the instructed cheaters. The Honesty-Humility scores of fair players and instructed cheaters were near identical. Numerically, it were the scores of the self-initiated cheaters that differed from the other two groups, but this comparison failed to reach significance. While the findings are suggestive of dishonest players rather than honest players differing in personality from the control condition, follow-up research with larger samples is needed.

Interestingly, Honesty-Humility did not predict cheating in our first study using the current paradigm (Geven et al., 2018). There may be at least two possible explanations. First, the predictive effect of Honesty-Humility on cheating in the present study was small, implying it requires substantial power to be picked up. Second, an important difference between our first study and the present study is that the former was conducted online and the latter in the laboratory. The more anonymous online setting may have promoted cheating. Indeed, in the online setting, the majority of participants cheated (67.6%). In an experimental manipulation that reflects an unambiguous 'strong situation', such as the possibility to anonymously cheat without consequences, the response variety across participants will be rather minimal (see Lissek, Pine, & Grillon, 2006). However, the laboratory environment in the current study (with a corresponding lower cheating rate of 37.4%) provided a more ambiguous situation in which the possible influence of personality traits on the decision to cheat could be investigated in a more valid manner.

4.4. Limitations and suggestions for future research

This study is not without its limitations. First, cheating is not directly observed but rather inferred from participants answering both cheating-evoking questions correctly, with pilot testing showing it would statistically be highly unlikely to answer both questions correctly. Although the experimenter was able to watch the duplicated screen in the laboratory, some participants admitted to use their smartphone to look up the correct answer and avoid possible detection. For future research our paradigm could be adapted to establish full ground truth. A suggestion would be to create an fictitous question with the correct answer only available on a web page designed by the experimenter.

Second, it cannot be fully confirmed whether participants felt they were actually cheating when looking up the correct answers online. In order to create a situation in which cheating behavior would be evoked spontaneously, we did not explicitly mention that it was not allowed to use Google. Thus, we cannot rule out that some participants considered cheating a viable option and might have assumed that it was not necessarily prohibited. Note however that many real-life situations are also ambiguous, and prominent psychological theories of cheating (e.g., Mazar et al., 2008) actually argue this is why people cheat: an ambiguous situation allows people to justify their cheating. It is precisely the inherent ambiguity that allows self-serving justifications as a result of which many people will cheat a little (see also Peer, Acquisti, & Shalvi, 2014).

Similar to the most popular cheating paradigms and to many real-life situations, our paradigm created a context in which the rules can be bent to maximize self-profit. Again, by not explicitly prohibiting participants to gain extra money in the quiz by looking up the answer, it is possible that some participants thought they did not engage in rule-breaking. Yet, the number of participants who decided to look up the correct answers to the cheating-evoking questions ($n = 65$) in comparison to the number of participants who had the opportunity to cheat ($n = 174$, hence 37%) and the fact that these participants reported lower Honesty-Humility scores than fair players, seems to indicate that we created a situation in which looking up the answer indeed constituted cheating.

Third, cheating on a trivia quiz for a monetary reward obviously does not completely reflect the scope of lies that could be expected in a high stakes criminal situation. However, contrary to previous research, in which participants were explicitly asked to lie, in the current experiment participants might in fact have felt guilty about deceiving the experimenter. Still, the consequences of cheating were negligible. Although the monetary reward upon a positive CIT outcome could have encouraged participants' motivation to avoid detection, no punishment was involved for a negative outcome.

5. Conclusions

The data reveal that even realistic deception elicits the typically observed response pattern of recognition (i.e., an increased SCR followed by a decrease in HR and RLL). Upon comparison with instructed deception, these findings imply that experimental findings in the laboratory do in fact resemble those expected in field settings. This result is encouraging from an ecological validity perspective and may pave the way for further successful field implementation of memory detection.

Acknowledgements

This research was funded by a grant, no. 238/15, from the Israel Science Foundation to Gershon Ben-Shakhar and also supported by the Psychology Research Institute (PsycRes) of the University of Amsterdam. We cordially thank Bert Molenkamp for programming the experiment, and Anouk Bercht, Zhu Han Chang, Elisa Cordesius and Jeannique Zimmerman for their assistance during data collection.

References

Ashton, M. C., & Lee, K. (2005). Honesty-humility, the big five, and the five-factor model. *Journal of Personality, 73*(5), 1321–1354. https://doi.org/10.1111/j.1467-6494.2005.00351.x.

Ben-Shakhar, G., & Elaad, E. (2002). Effects of questions' repetition and variation on the efficiency of the Guilty Knowledge Test: A reexamination. *Journal of Applied Psychology, 87*(5), 972–977. https://doi.org/10.1037/0021-9010.87.5.972.

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective.* New York, NY: Springer.

Ben-Shakhar, G., & Nahari, T. (2018). The external validity of studies examining the detection of concealed knowledge using the Concealed Information Test. In J. P. Rosenfeld (Ed.). *Detecting concealed information and deception* (pp. 59–76). London, UK: Elsevier. https://doi.org/10.1016/B978-0-12-812729-2.00003-3.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied, 9*(4), 261–269. https://doi.org/10.1037/1076-898X.9.4.261.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555.

De Vries, R. E., Lee, K., & Ashton, M. C. (2008). The Dutch HEXACO Personality Inventory: Psychometric properties, self-other agreement, and relations with psychopathy among low and high acquaintanceship dyads. *Journal of Personality Assessment, 90*(2), 142–151. https://doi.org/10.1080/00223890701845195.

DeAndrea, D. C., Carpenter, C., Shulman, H., & Levine, T. R. (2009). The relationship between cheating behavior and sensation-seeking. *Personality and Individual Differences, 47*(8), 944–947. https://doi.org/10.1016/J.PAID.2009.07.021.

Domnich, A., Panatto, D., Signori, A., Bragazzi, N. L., Cristina, M. L., Amicizia, D., et al. (2015). Uncontrolled web-based administration of surveys on factual health-related knowledge: A randomized study of untimed versus timed quizzing. *Journal of Medical Internet Research, 17*(4), e94. https://doi.org/10.2196/jmir.3734.

Effron, D. A., Bryan, C. J., & Murnighan, J. K. (2015). Cheating at the end to avoid regret. *Journal of Personality and Social Psychology, 109*(3), 395–414. https://doi.org/10.1037/pspa0000026.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology, 34*(5), 587–596. https://doi.org/10.1111/j.1469-8986.1997.tb01745.x.

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology, 77*(5), 757–767.

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12353.

Gino, F., & Pierce, L. (2009). The abundance effect: Unethical behavior in the presence of wealth. *Organizational Behavior and Human Decision Processes, 109*, 142–155. https://doi.org/10.1016/j.obhdp.2009.03.003.

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality, 57*, 72–88. https://doi.org/10.1016/j.jrp.2015.04.003.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Vol. Eds.), *Modern scientific evidence: The law and science of expert testimony: Vol. 2*, (pp. 446–483). St. Paul, MN: West Publishing.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2016). Orienting versus inhibition in the Concealed Information Test: Different cognitive processes drive different physiological measures. *Psychophysiology, 53*(4), 579–590. https://doi.org/10.1111/psyp.12583.

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the Concealed Information Test. *Psychophysiology, 54*(4), 628–639. https://doi.org/10.1111/psyp.12825.

Klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., Nahari, T., & Ben-Shakhar, G. (2017). Memory detection: The effects of emotional stimuli. *Biological Psychology, 129*, 25–35. https://doi.org/10.1016/j.biopsycho.2017.07.021.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863.

Lee, K., & Ashton, M. C. (2016). *Psychometric properties of the HEXACO-100. Assessment.* https://doi.org/10.1177/1073191116659134.

Lissek, S., Pine, D. S., & Grillon, C. (2006). The strong situation: A potential impediment to studying the psychobiology and pharmacology of anxiety disorders. *Biological Psychology, 72*(3), 265–270. https://doi.org/10.1016/J.BIOPSYCHO.2005.11.004.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*(6), 385–388. https://doi.org/10.1037/h0046060.

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist, 29*(10), 725–739. https://doi.org/10.1037/h0037441.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633.

Meijer, E. H., klein Selle, N., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879–904. https://doi.org/10.1111/psyp.12239.

Meixner, J. B., & Rosenfeld, J. P. (2014). Detecting knowledge of incidentally acquired, real-world memories using a P300-based Concealed-Information Test. *Psychological Science, 25*(11), 1994–2005. https://doi.org/10.1177/0956797614547278.

Nagin, D. S., & Paternoster, R. (1993). Enduring individual differences and rational choice theories of crime. *Law and Society Review, 27*(3), 467. https://doi.org/10.2307/3054102.

Nahari, T., Breska, A., Elber, L., Klein Selle, N., & Ben-Shakhar, G. (2017). The external validity of the Concealed Information Test: The effect of choosing to commit a mock crime. *Applied Cognitive Psychology, 31*(1), 81–90. https://doi.org/10.1002/acp.3304.

National Research Council (2003). *The polygraph and lie detection.* Washington, DC: The National Academies Press.

Osugi, A. (2018). Field findings from the Concealed Information Test in Japan. In J. P. Rosenfeld (Ed.). *Detecting concealed information and deception* (pp. 97–121). London, UK: Elsevier. https://doi.org/10.1016/B978-0-12-812729-2.00005-7.

Osugi, A., & Ohira, H. (2018). Emotional arousal at memory encoding enhanced P300 in the Concealed Information Test. *Frontiers in Psychology, 8*, 2334. https://doi.org/10.3389/fpsyg.2017.02334.

Peer, E., Acquisti, A., & Shalvi, S. (2014). "I cheated, but only a little": Partial confessions to unethical behavior. *Journal of Personality and Social Psychology, 106*(2), 202–217. https://doi.org/10.1037/a0035392.

Pennebaker, J. W., & Chew, C. H. (1985). Behavioral inhibition and electrodermal activity during deception. *Journal of Personality and Social Psychology, 49*(5), 1427–1433. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4078683.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Sokolov, E. N. (1963). *Perception and the conditioned reflex.* New York, NY: Macmillan.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428–453. https://doi.org/10.1037/bul0000087.

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. W. (2004). Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology, 41*(3), 461–466. https://doi.org/10.1111/j.1469-8986.00167.x.

Verschuere, B., Crombez, G., Koster, E. H., & De Clercq, A. (2007). Antisociality, underarousal and the validity of the Concealed Information Polygraph Test. *Biological Psychology, 74*(3), 309–318. https://doi.org/10.1016/j.biopsycho.2006.08.002.

Verschuere, B., Meijer, E., & de Clercq, A. (2011). Concealed information under stress: A test of the orienting theory in real-life police interrogations. *Legal and Criminological Psychology, 16*(2), 348–356. https://doi.org/10.1348/135532510X521755.

Vrij, A. (2004). Guidelines to catch a liar. In P. Granhag, & L. Stromwall (Eds.). *The detection of deception in forensic contexts* (pp. 287–314). Cambridge, UK: University Press. https://doi.org/10.1017/CBO9780511490071.013.