Comparing the effectiveness of different test paradigms in the detection of concealed information

Alexander A. Sommerkamp-Homann

University of Groningen

Author Note

Alexander A. Sommerkamp-Homann, Department of Neuroscience, University of Groningen.

I am grateful for the supervision of this research by Dr. Mark Nieuwenstein and Dr. Jacob Jolij and for the technical support by Dr. Mark Span. Special thanks to Dr. Berry Wijers for his help regarding data processing, to Timothy Sondej and Marco Wirthlin for their contributions to the statistical analyses and to Dr. Sabine Sommerkamp-Homann and Inka Frommhold for proofreading this thesis.

Correspondence concerning this article should be addressed to Alexander A. Sommerkamp-Homann, Department of Neuroscience, University of Groningen, 9712 CP Groningen. Contact: a.a.sommerkamp-homann@student.rug.nl

Abstract

The concealed information test is a powerful forensic instrument used to detect incriminating knowledge in crime suspects. Recently, researchers have increasingly based detection on a combination of event-related brain potentials and galvanic skin response yielding incremental validity. The optimal conditions for these two measurements, however, are thought to differ, particularly regarding inter-stimulus-interval length, complicating the construction of tests with simultaneous recordings. Furthermore, in standard test paradigms verbal or pictorial representations of the respective crime details (i.e. targets) are used, and presented singularly between distractor stimuli. Also, it is unclear whether distractors accidentally resembling details of the crime scene undermine the discriminatory potential of the procedure and how stable individual response differences are over time, relevant to the exclusion of prospective inadequate responders. Nineteen students participated in a pseudo-randomized study with a within-subjects design, investigating the use of the objects encountered during a mock crime as targets and the use of all-target sequences, and comparing different inter-stimulus interval lengths as well as distractors with resemblance to the mock crime scene to distractors without. While partial support for the overall detection of concealed information and for the stability of individual response differences over time was found, data quality issues constrained the extend of the analyses and no other hypotheses were statistically confirmed. Several factors possibly underlying the limited significance of the findings and scope of the analyses are discussed.

Keywords: concealed information, guilty knowledge, inter-stimulus-interval, new paradigm, object-based, all-target sequence

Comparing the effectiveness of different test paradigms in the detection of concealed information

As long as lies and deception have existed, the detection and exposure of such has been of vital interest. For the longest time, however, assessments of trustworthiness have relied on little more than intuition and dubious procedures, such as burning a suspect's skin with boiling water to investigate the degree of damage inflicted ("Ordeal of Boiling Water, 12th or 13th Century," 1996). It was not until 1908 that Hugo Münsterberg described a method which allowed for a scientific determination of deception in forensic contexts. His method utilizes the difference in physiological response when a guilty suspect is presented with crime-relevant vs. crime-irrelevant but similar information. If both types of information are otherwise indistinguishable, such a response difference indicates the possession of incriminating knowledge and may thus reflect a suspect's involvement in a crime. This principle of knowledge-detection has been implemented in a large variety of different test paradigms commonly known as the concealed information test (CIT) or guilty knowledge test, which is used by the police thousands of times each year (Matsuda, Nittono, & Allen, 2012). During a concealed information test, suspects are usually asked a question like "Which of the following items did you steal?" before different pieces of information are presented in sequential order and physiological and/or behavioral responses of the suspect are recorded. Some of the information pieces have particular relevance to the crime in question, such as a wallet which had been stolen, and are commonly referred to as *probes* or *targets*. Other pieces of information, such as a piece of jewelry, a pair of sunglasses or a mobile phone, are similar to the targets in that they are approximately equally plausible answers to the question asked, but are not relevant to the respective crime, and are called *distractors* or *non-targets*. These targets and non-targets are presented in random or pseudo-random order, either visually as written words or pictures, or aurally as spoken words. The suspect´s responses are compared between targets and non-targets and if the response difference between the two stimulus types is considered sufficiently large, it is interpreted as the possession of incriminating knowledge indicating a suspect´s involvement in the respective crime. Traditionally, the CIT has been conducted using autonomic measures, with galvanic skin response (GSR) yielding the best results (Meijer, Selle, Elber, & Ben-Shakhar, 2014). The observed change in the activity of the autonomic nervous system in response to targets is commonly interpreted as an *orienting response*, which is "a complex of physiological and behavioral reactions that reflect attentional processes and are evoked by any novel stimulus or by any change in stimulation" and can be enhanced by "stimuli that are significant for the individual" (Meijer et al., 2014). If perceived as crime-relevant information, target stimuli bear significance for the individual and represent a change in stimulation, as they are commonly presented less frequently than non-targets, often in a ratio of 1:5 (Meijer et al., 2014). In the late 1980s, CIT studies began to utilize event-related brain potentials (ERPs) as an alternative measure, primarily the P300 component (Rosenfeld, Nasman, Whalen, Cantwell, & Mazzeri, 1987). P300 amplitude is known to increase with decreasing presentation frequency (Donchin & Coles, 1988) and with the significance of stimuli, as constituted by task relevance (Castro & Diaz, 2001) or emotional value (Johnston, Miller, & Burleson, 1986), for example. The observed response differences between targets and non-targets may thus at least partially derive from the same perceptual processes as those related to autonomic activity, even though the extent to which the P300 can be seen as an orienting response is a matter of debate (Donchinet al., 1984). One of the most recent advances in CIT research has been the combined use of ERPs and autonomic measures. Typically, a separate analysis of responses is conducted with each type of measure employed, providing additional opportunities to observe a marked response difference between targets and non-targets. Meijer (2008), who was the first to take such an approach, did not observe incremental validity, but subsequent studies found significant increases in discriminatory potential by combined measurement (Ambach, Bursch, Stark, & Vaitl, 2010; Gamer & Berti, 2010; Matsuda, Nittono, & Ogawa, 2011). Yet, even though numerous lab studies support the validity of the CIT and the procedural methodology has repeatedly been advanced, detection accuracy is still far from perfect (Matsuda et al., 2012; Meijer et al., 2014). Correct classification rates roughly approximate around values like 80%, even though they should be interpreted with care as they are commonly derived from experiments in artificial settings. This study aims at improving the discriminatory potential of the CIT by investigating the effects of inter-stimulus interval (ISI) length and of stimuli resembling crime details accidentally being used as non-target stimuli, as well as by exploring the use of objects as test stimuli and of all-target sequences, and by investigating the stability of individual response differences over time. In the following sections, these issues will be explained in further detail.

Inter-stimulus interval length

Even though the combined use of ERPs and GSR in CITs has shown promising results, the simultaneous application of both techniques appears to complicate the design of tests, as the measures are thought to require different ISI lengths for optimal results. However, studies investigating the validity of this assumption have resulted in mixed findings, which this paragraph will briefly outline. On the one hand, in GSR-based CITs, stimuli are usually presented with long inter-stimulus intervals (ISIs) of 20 s or more in order to allow recovery of the slow response. This practice is supported by the Benedek and Kaernbach (2010), who observed GSR strength to positively correlate with ISI duration. On the other hand, ERP-based CITs usually use short ISIs of 1.5 - 3 s, as the fast recovery of components allows for numerous stimulus repetitions and thus for a large number of responses to be recorded in relatively short time. The use of short ISIs is supported by the finding that P300 amplitude differences between frequent and infrequent stimuli in an oddball task using ISIs between 2 - 4 s could no longer be observed with longer ISIs of 4 - 10 s (Polich, 1990). Matsuda, Nittono, Hirota, Ogawa, and Takasawa (2009) attributed this result to the long target-to-target interval (TTI) inducing a ceiling effect on the P300. Accordingly, using ISIs of 22 s in CITs with combined measurement, they and Matsuda, Nittono, and Ogawa (2011) found GSR but no P300 response differences. They did, however, also observe ERP response differences regarding the N200 (Matsuda et al., 2009) and the late positive component (LPC; Matsuda et al., 2009, 2011). Furthermore, significant GSR and other autonomic effects have been found in studies with short ISIs (Ambach et al., 2010; Meijer, 2008). Intermediate ISIs of 7.1 - 9 s were used by Gamer and Berti (2010) who found response differences in N200 and GSR but no significant P300 effect. In 2012, Gamer and Berti used random ISIs ranging from 2.4 to 16 s (M = 5 s) and observed significant differences in P300 and GSR but not in N200. Interestingly, ISI duration was not found to moderate either effect. Taken together, it is still unclear how ISI duration affects the discriminatory power of ERPs and autonomic responses in a CIT and thus which ISI duration could be considered optimal for a CIT with combined measurement. Theoretically, an estimation of optimal ISI duration could be derived from comparing respective effect sizes across studies, but as the number of published studies with combined measurement is small and the experimental as well as analytical differences between them are quite substantial, such an estimation would be virtually meaningless. Therefore, this study aims to obtain data from multiple CITs with combined measurement varying only in ISI duration.

The use of objects as test stimuli

One way to enlarge the response difference between targets and non-targets may be the use of target stimuli which provide a quantitative and qualitative increase in retrieval cues to the items "stolen" during the mock crime. To our knowledge, stimulus presentation in CIT research has been strictly unimodal and limited to the visual and aural domain, as targets and non-targets are presented either as written words, pictures or spoken words. Details of a crime such as weapons or stolen goods, however, which serve to elicit deviant responses during a CIT, are usually encountered multimodally by the perpetrator, not as words or pictures but as three-dimensional objects with complex visual and haptic properties. Between the initial exposure to crime details and their perception during the test thus exists a modal discrepancy. A considerable body of evidence shows that memory performance is improved when encoding and retrieval mode are identical, e.g. regarding haptic vs. visual information (Nabeta & Kawahara, 2006), and aural vs. visual information (for a review see Rummer, Schweppe, & Martin, 2013). This congruency effect was also observed by Stenberg, Radeborg, and Hedman (1995) and Stenberg (2006) for pictures vs. written words. As the CIT can at least partially be seen as a memory-based technique (Rosenfeld, Ward, Thai, & Labkovsky, 2015), a modal discrepancy between encoding and retrieval of crime details may thus weaken the desired response differences and impair detection accuracy. However, neither Ben-Shakhar, Frost, Gati, and Kresh (1996) nor Rosenfeld et al. (2015) who manipulated initial exposure and test presentation modality (written words vs. pictures) in a GSR-based and P300-based CIT, respectively, observed a congruency effect. Instead, they found that the use of pictures increased response differences, even when words were used either during the encoding or retrieval phase. Rosenfeld et al. (2015) argued that this picture superiority effect "overshadowed" the congruency effect. It is thus unclear whether the congruency effect can be observed in CIT paradigms. However, the unimodal presentation of stimuli in CITs also causes a reduction of stimulus detail. A written or spoken word or even a good photograph can arguably never provide the full detail and perceptual complexity of an actual object. Indeed, Ben-Shakhar and Gati (1987, 1990) showed in GSR-based CITs that responsivity depends on the level of crime stimulus detail present in the test stimuli: responsivity was lowest when the stimuli only resembled general properties of the crime detail and highest when resemblance was perfect. A reduction of the perceptual complexity of crime details to words or pictures may thus impair detection accuracy of the CIT. Furthermore, studies on intersensory facilitation show that stimulus identification, classification and recognition is improved by simultaneous exposure to additional stimulus details in a different sensory modality (Colonius & Diederich, 2012). For example, Giard and Peronnet (1999) observed that stimuli were identified faster and more accurately when both aural and visual features were presented compared to unimodal stimulation alone. Also, in studies with agnostic patients, stimulus identification was found to be facilitated when objects rather than line drawings were used, an observation referred to as *real-object advantage* (Farah, 1990; Ratcliff & Newcombe, 1982). Taken together, these findings suggest that an increase in stimulus detail also beyond the boundaries of one sensory modality may increase responsivity in the CIT. As arguably no representation of an item can provide a higher level of stimulus detail than the item itself, this study aims to utilize the objects encountered during a mock crime as targets in the subsequent CIT.

Non-targets resembling crime details

The experimental environment in CIT studies is usually controlled in that regard that non-targets do not closely resemble stimuli present at the mock crime scene or in an alternative target stimulus encoding procedure prior to the test. This is done to prevent non-targets accidentally being perceived as targets during the test. In a real-world setting, however, investigators may possess incomplete knowledge about the details of a crime. Even when sufficient information about the central components of a crime such as stolen objects is gathered, peripheral aspects such as other objects present at the crime scene may remain unrecognized. In such a case, apparent non-targets used in a CIT could accidentally resemble crime details. These stimuli may thus unintentionally elicit responses similar to targets, impairing detection accuracy. Responses to peripheral crime details have been investigated in several studies (Gamer & Berti, 2012; Nahari & Ben-Shakhar, 2010; Peth, Vossel, & Gamer, 2011) and were indeed found to resemble those to central crime details rather than to true non-targets. However, the peripheral details were used as targets, not as non-targets, and CIT questions analogous to “Did the poster in the room depict a …?” were thus aimed at those details specifically. In order to investigate the effect of a non-target accidentally resembling an 'unknown' crime detail, however, questions should not specifically address these but rather other details which the CIT was intentionally based on. Therefore, this study aims to compare responses to true non-targets with responses to putative non-target items which had been present at the mock crime scene, referred to as *semi-targets* hereafter.

The stability of individual response differences over time

Detection accuracy of the CIT can be increased by the exclusion of individuals indicative of not responding differently to targets and non-targets. This approach is taken by the Japanese police who conduct a screening prior to every CIT: the suspect is presented a series of different numbers of which one had to be memorized beforehand. If this number cannot be identified based on the difference in autonomic response, the suspect is not admitted to the CIT (Matsuda et al., 2012). Using a combination of GSR, heart rate, respiratory line length and normalized pulse volume, Matsuda, Ogawa, Tsuneoka, and Verschuere (2014) investigated the effectiveness of this method and found that screening increased the area under the receiver operating characteristic curve from 0.92 to 0.97. This means an increase in classification accuracy from 92% to 97%. Noordraven and Verschuere (2013) found support for screening procedures in a CIT based on reaction time. Subjects were instructed to respond to questions analogous to “Was the stolen item … a ring?” with “no” whenever the target or a non-target was included and to respond with “yes” to a special stimulus memorized before the test. The discriminatory potential of reaction time-based CITs relies on the increased latency of responses to targets compared to non-targets. Noordraven and Verschuere (2013) observed that non-increased reaction times to the special stimulus were predictive of non-increased reaction times to the equally infrequently presented target. By excluding the respective subjects, the area under the receiver operating characteristic curve increased from 0.87 to 0.95. Despite these promising findings, however, screening procedures for ERP-based CITs have not been investigated yet, to our knowledge. The observed correlations between ERPs and GSR in CIT studies with combined measurement are low (Gamer & Berti, 2010, 2012; Meijer, 2008) and therefore impede an unhesitant generalization of findings from one measure to the other. Also, the incremental validity found by combining reaction times and ERPs in order to detect faked memory impairments (Hooff, Sargeant, Foster, & Schmand, 2009) suggests that these two measures are not highly correlated, hindering a generalization in the same way. However, individual differences in ERPs are known to exist (Polich, 1997, 2007) and their utilization by Bayesian classification has shown a high discriminatory potential in the detection of memories (Allen, Iacono, & Danielson, 1992). Therefore, it may be promising to investigate the potential of screening procedures for ERP-based CITs. In order to confirm and explore the value of GSR-based and ERP-based screenings, respectively, this study aims to compare individual CIT response differences between two different test sessions.

The use of all-target sequences

Another way to increase the discriminatory potential of the CIT may be a facilitation of response differences between targets and non-targets by a modification of the stimulus presentation order. As GSR is known to accumulate with overlapping responses (Benedek & Kaernbach, 2010), the presentation of several targets in direct succession and with short ISIs may result in a composite response, stronger than responses to single targets in stimulus sequences with random order. Furthermore, following the reasoning of Matsuda et al. (2009) regarding a ceiling effect of the P300 with longer TTIs, a reduction of TTI length may also result in increased P300 response differences. However, as an all-target sequence holds an internal target probability of 1 and P300 response differences are known to correlate negatively with target probability in oddball-tasks (Pritchard, 1981), it may be that the presentation of several targets in direct succession counteracts the possible increase in response differences by shorter TTIs. This study aims to investigate whether the use of all-target sequences in a CIT increases the discriminatory potential of ERP and GSR measures.

Hypotheses of this study

Taken together, it is hypothesized that (1) responses differ between targets and non-targets, (2) between semi-targets and non-targets, and that (3) a combination of GSR and ERP measures provides incremental validity. It is further hypothesized that (4) the response difference between targets and non-targets differs between short and long ISIs, (5) is larger when objects instead of pictures are used test stimuli, and (6) is also larger when all-target sequences instead of singular targets are used in a CIT with short ISIs. Lastly, it is hypothesized that (7) GSR and (8) ERP response differences correlate between different test sessions.

Method

Participants

Nineteen first-year psychology students (eight female, eleven male) voluntarily participated in the study. The age ranged from 19 to 34 ($M = 21.26$, $SD = 3.69$). Ten of the participants were German, four were Dutch, and five had other nationalities. Inclusion criteria were: (corrected) normal vision and hearing, and fluency in English. Exclusion criteria were: dreadlocks, cornrows or similar hair styles, brain implants, history in epilepsy or recent panic attacks, and aversion towards physical interaction with people. Participants received course credit in compensation and, dependent on their performance, also had the chance to win a bike. All participants provided informed consent.

Materials, apparatus and data acquisition

Fifty objects and fifty digital pictures were used as experimental stimuli; each of the pictures was a photograph of one of the objects on a white background, see Appendix. The objects and the corresponding pictures belonged to ten different categories, each containing five different but similar items: bottle caps, data storage devices, dice, lighters, pens/pencils, tools, glasses, screws, toys, and electronic devices. A bike painted in the university colors was used as performance incentive in the study. A personal computer running Windows XP with a monitor refresh rate of 60 Hz was used to present experimental tasks which were programmed using OpenSesame (Mathôt, Schreij, & Theeuwes, 2012) and to record electroencephalographic (EEG) and electrodermal data using OpenVibe (Renard et al., 2010). EEG was recorded with tin electrodes which were attached to an electrocap (Electro-Cap International Inc., Eaton, Ohio, USA) at the Fz, Cz, Pz and Oz sites according to the international 10–20 system. The amplifier used was a PORTI7 (Twente-Medical Systems, Enschede, The Netherlands). An average reference was used and the sampling frequency was 250 Hz. An electrode placed at the left collarbone served as the subject´s ground. Four electrodes, placed at the left and right lateral canthi and above and below the left eye, were used to measure the electroocuologram. Impedance was kept <10 kOhms at all electrodes for all subjects except one for which impedance was kept <20 kOhms as a further reduction appeared challenging without harming the subject's skin. Electrodermal activity was recorded via two electrodes, one attached to the inside of the intermediate phalanx of the left middle finger, the other to the same position on the left ring finger, which were connected to the amplifier on an auxiliary channel through a GSR-4 module. For one subject, the ring finger electrode was instead attached to the middle finger as the ring finger had been numbed by traumatic nerve injury.

Procedure and Design

This study utilized a repeated measures design. Subjects went through two different experimental sessions, A and B, each containing two or more repetitions of several different CITs. The sessions were separated by exactly one week, taking place at the same weekly day and the same daily time. Session order, i.e. whether the subject went through session A or B first, was determined at random.

The study was advertised online for first-year psychology students at the University of Groningen with the following title: “Do you think you can beat a lie-detector?” Potential participants were provided information about the fact that the experiment would involve committing two mock crimes, that the detection of concealed information would be based on EEG and electrodermal recordings, and that a bike would be given as a reward to someone who successfully deceives the test. It was also mentioned that the CIT is used by the police and has an impressive detection accuracy.

In the beginning of the first session the subject was informed again about the mock crime procedure, the recording of EEG and electrodermal activity (EDA) and about the administration of different CITs. The experimenter also informed the subject that his assistant would be guiding the subject to the room for the mock crime, that the assistant would remain outside of that room in case of any questions and that his task was to assist the subject in the procedure. Furthermore, the experimenter informed the subject that he himself would be conducting the lie-detection procedure, the evaluation of the results, and that he would be blind to the subject's instructions for the mock crime. The subject was instructed to try best to remain innocent throughout the CITs, was then shown the bike and informed again that it would be given to someone who deceived the CIT successfully. In the beginning of the second session, the subject saw the bike again and was informed that the procedure would closely resemble the first session except for the CITs.

Mock crime. The subject was guided to a room in another building by the assistant and handed a bag and written instructions for the mock crime procedure. Inside the room, three objects from each of five of the categories were located clearly visible. The subject was instructed to pick up and thoroughly examine one specific object of each category and to place it into the bag, thereby “stealing” it. The other ten objects were merely present at the mock crime scene and did not require any interaction with the subject. Which objects were located in the room and which of them were chosen as targets was determined previously in a pseudo-random manner in such that all objects were chosen comparably often as targets, non-targets and semi-targets, respectively. When all five objects were placed inside the bag, the subject was required to hand over the bag to the assistant who was waiting outside of the room and to walk back to the testing room where the experimenter was located. The experimenter and the subject then went to a preparation room to attach the electrodes for the EEG and electrodermal measurements. In the meantime, the assistant collected all remaining objects from the crime scene room and brought these, together with the “stolen” objects, back to the testing room. Afterwards, the experimenter and the subject went back into the testing room, where the electrodes were connected to the recording apparatus. The stimulus categories utilized in session A were the bottle caps, data storage devices, dice, lighters, and pens/pencils, and those used in session B were the tools, glasses, screws, toys, and electronic devices.

CITs. Four different CIT paradigms were utilized in this study: a computer-based test in which stimulus categories were presented in succession and with long inter-stimulus intervals (ISIs), referred to as CSL hereafter, a computer-based test also with successively presented stimulus categories but with short ISIs, subsequently abbreviated as CSS, a computer-based test in which the presentation of stimuli was mixed between categories but also with short ISIs, called CMS in the following, and an object-based test in which stimuli categories were presented in succession and with long ISIs, referred to as OSL hereafter.

Object-based CIT: OSL (successively presented stimuli categories, long ISIs). The subject was seated in a comfortable chair at an empty table, with the right hand openly resting on it. At the opposite side of the table, at an approximate distance of 80 cm, the experimenter was sitting, facing the subject. The 25 experimental objects were located in close proximity to the experimenter, on a chair to his right side slightly below the level of the table, not in sight of the subject. While looking the subject straight into the eyes, the experimenter asked “Which of the following objects did you steal?” after which he gave a randomly selected object from a randomly selected category into the right hand of the subject. The subject was instructed to look at the object and was allowed to move it around in the hand. After 7000 ms, a 130.81 Hz tone appeared for 2000 ms. This tone served as the signal to hand the object back to the experimenter who put it back out of sight of the subject. After a pause of 18000 ms, this sequence, beginning with the presentation of the object, was repeated for the remaining four objects of the selected category. Then, the whole procedure, beginning with asking the question, was repeated for the other four categories. One complete run through all 25 objects took 685 s. The five objects "stolen" during the mock crime served as targets and the other ten objects present at the mock crime scene served as semi-targets. The remaining thirty-five objects served as non-targets.

Computer-based CITs. The subject was seated in a comfortable chair at a table with a desktop computer. The screen had a diameter of 22 inches and was located in an approximate distance of 50 cm to the subject's eyes. Also, a keyboard and two small speakers were positioned on the table in front of the subject and to the subject's left, respectively. All computer-based CITs utilized grey (#888888) as background color of the computer screen and black (#000000) as font color for all written content and as color for fixation dots. Furthermore, all respective content was presented centrally on the screen. Pictures of the five objects "stolen" during the mock crime served as targets and pictures of the other ten objects present at the mock crime scene served as semi-targets. The remaining thirty-five pictures served as non-targets.

CSS (successively presented stimuli categories, short ISIs). First, the question "Which of the following items did you steal?" was presented for 4000 ms. Then, after a blank screen for 1000 ms and a fixation dot for 500 ms, a randomly selected picture of a randomly selected category was shown for 800 ms, succeeded by another fixation dot for 500 ms. This sequence, beginning with the blank screen, was repeated for the four remaining pictures of the respective category. Then, after a blank screen for 1500 ms, the whole aforementioned procedure, beginning with the presentation of the question, was repeated for the remaining four categories. One complete run through all 25 stimuli took 97.5 s. The subject was instructed to look at the fixation dot when it appeared and to avoid major eye movements and any eye blinks during the presentation of the picture and until the second fixation dot disappeared. The subject was explicitly allowed to blink when the screen was blank.

CSL (successively presented stimuli categories, long ISIs). After a presentation of the question "Which of the following items did you steal?" for 2000 ms, a fixation dot was shown for 2500 ms, followed by a randomly selected picture from a randomly selected category for 9000 ms. During the last 2000 ms of the picture presentation, a 130.81 Hz tone was simultaneously presented to increase comparability with the OSL test. Then, a blank screen was shown for 15500 ms before the sequence, beginning with the fixation dot, was repeated for all other pictures of the selected category. The entire sequence, beginning with the presentation of the question, was subsequently repeated for the other four categories. One complete run through all 25 stimuli took 685 s. The subject was instructed to look at the fixation dot when it appeared and to avoid major eye movements and any eye blinks during the first one or two seconds of the presentation of the picture. The subject was explicitly allowed to blink and look elsewhere in the meantime.

CMS (mixed presentation of stimuli across categories, short ISIs). In the beginning, the question "Did you steal the following items?" was shown for 4000 ms. After a blank screen for 1000 ms and a fixation dot for 500 ms, a randomly selected picture from a randomly selected category was presented for 800 ms, followed by another fixation dot for 500 ms. Subsequently, this sequence, beginning with the blank screen, was repeated for one picture from each of the four remaining categories. If the sequence started with the presentation of a target, all other items in the sequence were also targets. If the sequence started with a semi-target or non-target, all other items in the sequence were also semi-targets or non-targets, respectively. Then, after a blank screen for 1500 ms, the entire procedure, beginning with the presentation of the question, was repeated until all 25 pictures were shown. One such run took 97.5 s. The subject was instructed to look at the fixation dot when it appeared and to avoid major eye movements and any eye blinks during the presentation of the picture and until the second fixation dot disappeared. The subject was explicitly allowed to blink when the screen was blank.

Test battery composition. Table 1 shows the order and the number of runs of different CITs conducted in session A and B. The order in which the different tests were conducted also differed between two conditions: paradigm order 0 and 1. One of these conditions was pseudo-randomly assigned to the subject in a way that counterbalanced the allocation of session order across subjects. The order of the different CITs was partially chosen to reduce the overall influence of GSR habituation, which is a decrease in responses to targets when the frequency of target presentation is increased (Barry, Feldmann, Gordon, Cocker, & Rennie, 1993). The CSL and OSL, which both feature relatively few stimulus repetitions, were therefore included early in the sessions. Even though the CMS features comparatively many repetitions, it was also included early in order to improve comparability with the

CSL, for exploratory reasons. In order to improve the comparability of responses across time, the two standard paradigms, CSS and CSL, were part of both sessions. The numerous repetitions of the CSS and the comparatively few repetitions of the CSL were chosen to orient at the usual implementation of ERP-based and GSR-based test paradigms, respectively (Meijer et al., 2014). In order to improve the comparability between the OSL and CSL, the number of repetitions of both tests was identical, and the same was true for the CSS and CMS. Prior to the first presentation of each paradigm, the subject was given a few practice trials to ensure the instructions were correctly understood. During these practice trials, other stimuli were used than those relevant to this study. Between the different CITs, the subject was allowed to rest for about one minute.

Table 1

Order, number of runs and duration of the CITs in the test batteries for both sessions and paradigm orders

Paradigm order 0						Paradigm order 1					
Session A			Session B			Session A			Session B		
CIT	Runs	Duration	CIT	Runs	Duration	CIT	Runs	Duration	CIT	Runs	Duration
CSL	1	685 s (~11min)	CSL	1	685 s (~11min)	OSL	1	685 s (~11min)	CMS	1	292.5 s (~5 min)
OSL	1	685 s (~11min)	CMS	3	292.5 s (~5 min)	CSL	1	685 s (~11min)	CSL	3	685 s (~11min)
CSL	1	685 s (~11min)	CSL	1	685 s (~11min)	OSL	1	685 s (~11min)	CMS	1	292.5 s (~5 min)
OSL	1	685 s (~11min)	CMS	3	292.5 s (~5 min)	CSL	1	685 s (~11min)	CSL	3	685 s (~11min)
CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CMS	3	292.5 s (~5 min)
CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CMS	3	292.5 s (~5 min)
CSS	3	292.5 s (~5 min)	CMS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)

CSS	3	292.5 s (~5 min)	CMS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)	CSS	3	292.5 s (~5 min)
			CSS	3	292.5 s (~5 min)				CMS	3	292.5 s (~5 min)
			CSS	3	292.5 s (~5 min)				CMS	3	292.5 s (~5 min)
			CMS	3	292.5 s (~5 min)				CSS	3	292.5 s (~5 min)
			CMS	3	292.5 s (~5 min)				CSS	3	292.5 s (~5 min)

After the testing procedure, the EEG and electrodermal recording apparatus was detached from the subject. A memory test assessed which objects the subject remembered to have "stolen" during the mock crime. The subject also filled out a questionnaire; all questions had to be answered on Likert scales ranging from 1 (not at all) to 5 (very much) unless indicated otherwise:

1. How much do you like the bike which you can obtain as a reward?

2. How motivated were you to fool the test, that is to remain innocent?

[During session A] 3. You underwent three different test procedures, one in which pictures were presented on a screen with a long interval between them (1), one in which they were presented at a shorter interval (2), and one in which the objects were given into your hand by the experimenter (3). How powerful or accurate did you perceive these tests to be? [Three Likert scales]

[During session B] 3. You underwent three different test procedures, one in which different categories of pictures were presented in succession with a long interval between pictures (1), one in which the categories they were presented in succession at a shorter interval (2), and one in which objects from different categories were presented in the same sequence with each sequence containing one object from each category (3). How powerful or accurate did you perceive these tests to be? [Three Likert scales]

4. How successful do you think you were at fooling the three different tests, that is to remain innocent? [Three Likert scales]

5. What was your technique in order to remain innocent? Please explain in detail. [Open question]

6. How good do you generally consider yourself at lying?

7. All in all, do you think that you successfully fooled the test procedure, that is to remain innocent? [Answer options: yes, no]

8. Do you have any additional remarks, criticism, feedback? [Open question]

The data obtained by the questionnaire was not further analyzed. The subject received an evaluation of the CIT performance after the end of the study.

Data processing

Part of the EEG and electrodermal data regarding the CSL were deleted for two subjects because of wrongly encoded event markers and part of the electrodermal data regarding the OSL was deleted for another subject for the same reason. The entire electrodermal data regarding the OSL of one subject was deleted because of missing event markers. Due to interruptions of the experiment, part of the recordings regarding the OSL, the CSL and the CMS were removed for two subjects, one subject and one subject, respectively. Part of the electrodermal data regarding the CLS was deleted for two subjects and regarding the CMS for one subject either because the hand to which the electrodes were attached was not kept still or because the electrodes were touched with the other hand during the recordings. The entire EEG and electrodermal data of one session of one subject was removed because the recorded values did neither seem to correspond with the experimental stimuli nor to reflect adequate physiological activity. Six subjects either selected a wrong object during the mock crime or remembered it incorrectly, and one subject selected or remembered only four objects instead of five; the corresponding item categories were excluded from the analysis of the respective subjects' data. In the CMS of one subject, a target item was wrongly coded as an irrelevant and vice versa; the item category containing both stimuli was excluded from the analysis of the respective subject's CMS data.

EEG was processed with BrainVision Analyzer 2 (Brain Products GmbH, Gilching, Germany). After a notch filter at 50 Hz was applied, the data was cut into segments of 1200 ms beginning 200 ms before stimulus presentation. Segments with a voltage gradient > 40 µV, with low activity (difference between minimal and maximal voltage < 0,5 µV across 100 ms) and with high activity (difference between minimal and maximal voltage > 100 µV across 1000ms) were removed. Ocular activity was corrected using the method of Gratton, Coles, and Donchin (1983). The remaining segments were averaged for targets, semi-targets and non-targets per subject for the CSS, CSL and CMS, respectively, and baseline corrected using a pre-stimulus interval of 200 ms. Averages containing less than five segments were not computed and tests with a missing average were excluded from further processing for the respective subject. Regarding the CMS, EEG data of 15 of the 17 subjects who underwent the test remained included; regarding the CSL and CSS, of the 19 subjects who underwent the tests 14 and 18 remained included, respectively. For the analysis of P300 components, the microvoltage mean amplitudes of the averages at the Pz site were computed from 250 ms to 400 ms after stimulus presentation. Because of the ambiguity regarding adequate interval choices in EEG analysis and in order to assess the stability of the results across chosen intervals, the microvoltage mean amplitudes were alternatively computed from 250 ms to 500 ms after stimulus presentation. For the analysis of LPCs, the microvoltage mean amplitudes of the averages at the Pz site were computed from 400 ms to 1000 ms and alternatively from 500 ms to 1000 ms after stimulus presentation.

EDA was also processed with BrainVision Analyzer 2. A notch filter at 50 Hz was applied to the data. Regarding the CSL, CMS and OLS, segments of 21500 ms were formed, beginning at 1000 ms before stimulus presentation. Note that only the first event marker in each series of the CMS was used for this segmentation in order to prevent overlap between segments. Also, regarding the CSS, CSL and CMS, electrodermal data was cut into segments of 4800 ms, beginning at 1000 ms before stimulus presentation. The segments were averaged for targets, semi-targets and non-targets per subject for the CSS, CSL, CMS and OSL, respectively, and baseline corrected using a pre-stimulus range of 200 ms. All 19 subjects who underwent the CSS and CSL remained included in the data. Of the 17 subjects who underwent the CMS, 16 remained included and of the 15 subjects who underwent the OSL, 13 remained included. The microvoltage mean amplitudes of the longer averages were then computed for an interval of 19500 ms and of the shorter averages for an interval of 2800 ms, both beginning 1000 ms after stimulus presentation.

Results

The waveforms of EEG and electrodermal grand averages across all subjects are displayed for all four tests in Figure 1. Due to technical issues that resulted in uncertainty about the quality of the EDA data, these data were not included in analysis. As a result, we could only test the following hypotheses: (1) Responses differ between targets and non-targets, and (2) between semi-targets and non-targets. (3) The response difference between targets and non-targets differs between short and long ISIs, and (4) is larger when all-target sequences instead of singular targets are used in a CIT with short ISIs. (5) ERP response differences correlate between different test sessions.

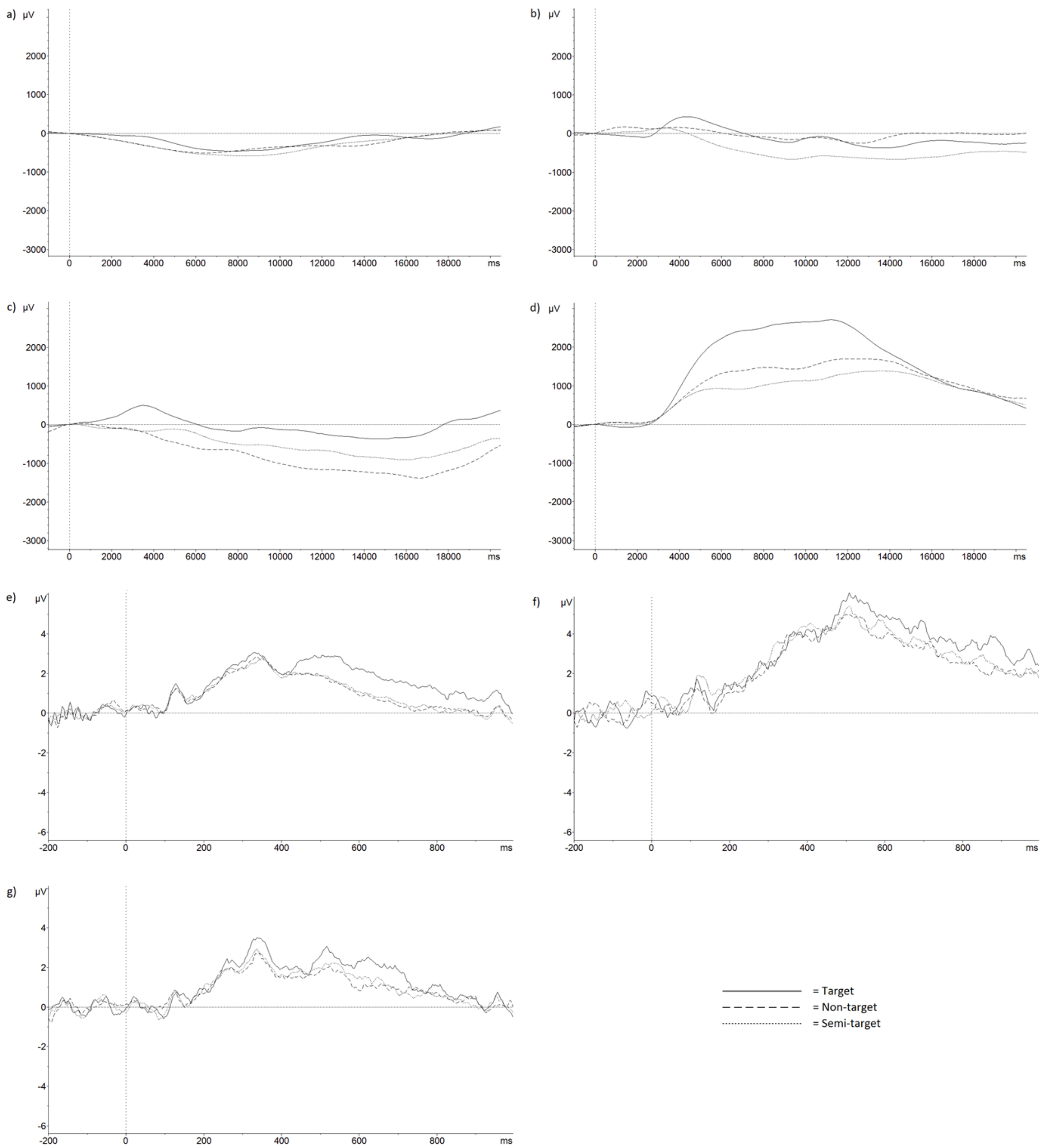

Figure 1. Grand averages across all subjects. EDA is displayed for a) CSS, b) CSL, c) CMS, and d) OSL. For illustrative purposes, equivalent intervals are displayed for all tests. Note, however, that regarding the CSS, responses potentially begin to overlap at 3800 ms. Regarding the CMS, only segments based on the first event marker in each series were selected and averaged. EEG at the Pz site is displayed for e) CSS, f) CSL, and g) CMS.

All analyses including EEG data were conducted twice, once including microvoltage mean amplitudes based on P300 averages ranging from 250 ms to 400 ms post-stimulus and based on LPC averages ranging from 400 ms to 1000 ms post-stimulus, and once with the corresponding ranges of 250 ms to 500 ms and of 500 ms to 1000 ms, respectively. No difference in outcome were found, and the statistics reported in this section are limited to results based on the former ranges.

In order to test hypotheses 1 to 4, two separate omnibus models were used to conduct analyses of variance regarding the P300 and LPC, respectively. Both models included the responses to targets, semi-targets and non-targets for the CSS, CSL and CMS tests. Due to missing values in the dataset, only 12 of the 19 subjects were included. Regarding the LPC, a main effect was observed for CIT type, $F(1.12, 12.33) = 18.01$, $p = .001$, and a trend-level main effect was observed for target type, $F(2, 22) = 2.84$, $p = .08$. No significant interaction effect was found, $F(4, 44) = 1.17$, $p = .34$. Regarding CIT type, also a quadratic effect was found, $F(1, 11) = 19.26$, $p = .001$. Pairwise comparisons, adjusted by Bonferroni correction, showed a significant difference between the CSS and the CSL, $p = .002$, and between the CMS and the CSL, $p = .005$, and no significant difference between the CSS and the CMS, $p = 1.00$. Regarding target type, a linear effect was found, $F(1, 11) = 8.16$, $p = .02$. Pairwise comparisons, adjusted by Bonferroni correction, showed a significant difference between targets and non-targets, $p = .047$, and no significant difference between semi-targets and non-targets, $p = 1.00$, and between semi-targets and targets, $p = .42$. Regarding the P300, no significant main effects were observed for CIT type, $F(2, 22) = 0.43$, $p = .66$, and for target type, $F(1.24, 13.66) = 0.93$, $p = .37$. Also, no significant interaction effect was found, $F(2.23, 24.57) = 1.14$, $p = .34$.

In order to test hypothesis 5, difference scores (target - non-target) of the P300 and LPC were computed and compared between the two experimental sessions for the CSS and CSL tests. Note that the CMS test was not included as it was only part of one of the sessions. Due to missing values in the dataset, only 11 and 8 of the 19 subjects were included for the CSS and CSL, respectively. Regarding the CSS, a positive correlation was found in P300s, $r(11) = .70$, $p = .017$, and in LPCs, $r(11) = .72$, $p = .013$. Regarding the CSL, a negative correlation was found in P300s, $r(8) = -.81$, $p = .013$, and in LPCs, $r(8) = -.73$, $p = .041$.

Discussion

Although CIT research has provided ample support for the validity of the test procedure, estimates of the overall accuracy have still indicated room for improvement. The goal of this study was to examine the influence of ISI length on the discriminatory potential of combined GSR and ERP measurement and the efficacy of the use of objects instead of pictures as test stimuli and of all-target sequences instead of singular targets in CITs with short ISIs. Furthermore, the goal was to assess the stability of response differences across test sessions and the effect of stimuli resembling crime details accidentally being used as non-target stimuli in a CIT. Due to the questionable quality of GSR measurements, analyses were limited to ERP data. Consequently, combined measurement and the use of objects as test stimuli was not further investigated.

The results showed significant response differences between targets and non-targets across CITs regarding the LPC. However, no such differences were observed regarding the P300. Interestingly, across target type, the LPC magnitude differed between CITs and was more pronounced in the CSL than in the CSS and CMS. No response differences between semi-targets and non-targets and between semi-targets and targets were found. Response differences were not found to be affected by ISI length or by the use of all-target sequences. The correlation of response differences between test sessions was positive for the CSS but negative for the CSL.

Regarding the LPC, the significant response differences between targets and non-targets across different tests stand in line with the corresponding literature (Matsuda et al., 2009, 2011). However, the non-significance of these differences regarding the P300 contrasts with the vast body of research conducted on CITs (for a review, see Meijer et al., 2014). A possible explanation for the difference in significance between the LPC and P300 may be that the former is less affected by variance. Whereas the P300 is a relatively transient potential, the LPC typically spans across a comparatively long time interval. The ability to observe the LPC over a longer post-stimulus period may reflect in a greater robustness of the potential against noise in the data. The LPC magnitude difference between the CSL and both the CSS and the CMS might be attributed to a difference in habituation. Similar to the GSR, the LPC has been shown to decrease with repeated stimulus exposure, at least to a certain degree (Codispoti, Ferrari, & Bradley, 2006; Paul, Kathmann, & Riesel, 2016). Featuring fewer stimulus repetitions than the CSS and CMS and longer ISIs, which translate to a decreased stimulus presentation frequency, the CSL may thus have elicited less habituation of the LPC than the other two tests. The non-significance of response differences between semi-targets and non-targets may suggest that it is not necessary to generally prevent the use of non-targets resembling details of crime scenes in respective CITs. However, given the overall limited significance of our findings and that the significant LPC response difference between targets and non-targets could not be observed between targets and semi-targets, such conclusions should not be readily drawn. The positive correlation of individual ERP difference scores observed between test sessions for the CSS seems to support the generalizability of GSR-based screenings, which are supported by the findings of Matsuda et al. (2014), to ERPs. On the other hand, the negative correlation observed for the CLS seems to oppose the possibility of effective ERP-based screening procedures. An inverse relationship between measurements of almost identical tests, however, appears illogical and suggests methodological errors to be of influence. Still, a conclusion about the stability of individual differences across CITs and about their potential utilization in screenings should not be readily drawn from our findings. Further research is needed to clarify these subject matters.

Our hypotheses regarding modifications of the standard CIT paradigms may indeed be false, contrary to what might be expected following the underlying line of empirical reasoning. However, multiple factors potentially contributed to the limited significance of findings and scope of the analyses. This study was originally designed to include 30 subjects. Due to complications regarding the availability of adequate testing facilities, however, the study was terminated after it was run with 19 subjects and of these subjects, 6 had not participated in the second experimental session yet. One subject also missed the second experimental session through absence. Failures of the recording equipment resulted in signal loss and in missing or unusable event markers. External interruptions during the recordings, deviations from task instructions and other procedural disturbances caused additional data to be discarded. Upon visual inspection, the raw EEG data appeared to be heavily confounded and artifact removal lead to numerous deletions of segments. The resulting individual averages, comprised of correspondingly few segments, appeared to carry much noise and comparably little signal strength. Regarding the GSR, unresolved questions regarding the quality of the recordings ultimately led to the exclusion of the entire corresponding data from the analyses. One contributing factor may have been the number of stimulus repetitions during each session, which was rather high compared to other GSR-based CIT studies, in which each target stimulus usually is presented only once or a few times (Meijer et al., 2014). In this study, targets were repeated 16 times during session A and 32 times during session B. Over the course of the experimental sessions, the frequent target exposure may have led to increased GSR habituation and thus to a reduction of response differences. However, several GSR-based CIT studies have also been conducted with 12 target repetitions and without an indication for a reduction in effect size (Meijer et al., 2014). Whether GSR indeed progressively decreased during the experimental sessions may be assessed by a re-analysis of the data comparing GSR between different runs of the same tests. Regarding the CMS and particularly the CSS, the short ISIs may have also resulted in GSR habituation within the tests. Furthermore, the low number of stimulus repetitions in the CSL as well as the long ISIs in the test may have prevented statistically detectable ERP differences. These potential influences, however, are related to the intrinsic properties of the CIT paradigms and thus do not reflect methodological shortcomings or external impediments of the study. Recording and processing of the GSR in this study can be seen as somewhat unconventional as usually a baseline period of several minutes is recorded prior to the presentation of the experimental stimuli and the slow, tonic component of the signal is subtracted from the fast, phasic activity generally following specific events such as stimulus presentations (Braithwaite, Watson, Jones, & Rowe, 2015). Also, GSR segments are usually assessed individually regarding the plausibility of a genuine event-related response and, accordingly, either included in or discarded from further processing and analysis. The general averaging and baselining of GSR segments which has been derived from EEG data processing and applied in this study may have thus caused variations other than the direct responses to the experimental stimuli to be included in the analyses. Furthermore, regarding both ERPs and GSR, the averaging of individual segments into composite scores on which the analyses are based discards a great deal of information. Although the use of averages is not uncommon for reasons of convenience, the power of statistical analyses is reduced hereby. Another limitation of this study is the lack of statistical investigations of pilot testing data. Steps in data processing may have been differently, and perhaps more adequately, chosen with prior knowledge of the effects. For example, a peak-to-peak amplitude measurement regarding the P300 or a tighter interval for GSR analyses regarding the CSL and OSL may have identified response differences more accurately. Furthermore, the subjects' motivation to deceive the different tests may have been unusually high. In CIT studies, small monetary rewards or some study credits are usually given to subjects for their participation. In this study, however, a bike was promised as a reward explicitly for a successful deception. The possibly much higher incentive-related motivation to conceal knowledge may have resulted in an increased use of countermeasures which are known to be effective in both ERP- and GSR-based CITs (Ben-Shakhar, 2011), thereby reducing response differences between targets and non-targets. Whether a higher motivation to conceal is indeed related to reduced response differences in the CIT or may even undermine detection accuracy remains subject to future research and may be of considerable relevance, especially to field applications of the CIT. Taken together, the limited significance of our findings and scope of the analyses may have been contributed to by the low amount of data on which the analyses were performed, particularly caused by the premature termination of the study and by the deletion and averaging of segments, a low signal-to-noise ratio in the EEG, GSR habituation, procedural constrains inherently related to properties of the test paradigms, poor data processing choices due to a lack of analyses on pilot testing data and a high reward for successful deception.

Taken together, partial support for the overall detection of concealed information and for the stability of individual response differences over time was found. Several factors were identified potentially underlying the limited significance of findings and scope of the analyses. The effectiveness of both new CIT paradigms proposed in this study as well the validity of our other hypotheses are yet to be determined by future research.

References

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The Identification of Concealed Memories Using the Event-Related Potential and Implicit Behavioral Measures: A Methodology for Prediction in the Face of Individual Differences. *Psychophysiology, 29*(5), 504-522. doi:10.1111/j.1469-8986.1992.tb02024.x

Ambach, W., Bursch, S., Stark, R., & Vaitl, D. (2010). A Concealed Information Test with multimodal measurement. *International Journal of Psychophysiology, 75*(3), 258-267. doi:10.1016/j.ijpsycho.2009.12.007

Barry, R. J., Feldmann, S., Gordon, E., Cocker, K. I., & Rennie, C. (1993). Elicitation and Habituation of the Electrodermal Orienting Response in a Short Interstimulus Interval Paradigm. *International Journal of Psychophysiology, 15*(3), 247-253. doi:10.1016/0167-8760(93)90008-d.

Ben-Shakhar, G. (2011). Countermeasures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory Detection: Theory and Application of the Concealed Information Test* (pp. 200-214). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511975196.012

Ben-Shakhar, G., Frost, R., Gati, I., & Kresh, Y. (1996). Is an apple a fruit? Semantic relatedness as reflected by psychophysiological responsivity. *Psychophysiology, 33*(6), 671-679. doi:10.1111/j.1469-8986.1996.tb02363.x

Ben-Shakhar, G., & Gati, I. (1987). Common and distinctive features of verbal and pictorial stimuli as determinants of psychophysiological responsivity.

Journal of Experimental Psychology: General, 116(2), 91-105. doi:10.1037//0096-3445.116.2.91

Ben-Shakhar, G., & Gati, I. (1990). Novelty and significance in orientation and habituation: A feature matching approach. *Journal of Experimental Psychology: General, 119*(3), 251-263.

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*. *47*(4), 647-58. doi:10.1111/j.1469-8986.2009.00972.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments (Revised version: 2.0). University of Birmingham. UK: Selective Attention & Awareness Laboratory (SAAL), Behavioural Brain Sciences Centre, School of Psychology.

Castro, A., & Díaz, F. (2001). Effect of the relevance and position of the target stimuli on P300 and reaction time. *International Journal of Psychophysiology, 41*(1), 43-52. doi:10.1016/s0167-8760(00)00182-3

Codispoti, M., Ferrari, V., & Bradley, M. (2006). Repetitive picture processing: Autonomic and cortical correlates. *Brain Research, 1068*(1), 213-220. doi:10.1016/j.brainres.2005.11.009

Colonius H., & Diederich A. (2012). Intersensory Facilitation. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer. doi.org/10.1007/978-1-4419-1428-6

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11*(3), 357. doi:10.1017/s0140525x00058027

Donchin, E., Heffley, E., Hillyard, S. A., Loveless, N., Maltzman, I., Öhman, A., Rösler, F., Ruchkin, D., & Siddle, D. (1984). Cognition and Event-Related Potentials II. The Orienting Reflex and P300. *Annals of the New York Academy of Sciences, 425*(1), 39-57. doi:10.1111/j.1749-6632.1984.tb23522.x

Farah, M. J. (1990). *Issues in the biology of language and cognition. Visual agnosia: Disorders of object recognition and what they tell us about normal vision.* Cambridge, MA, US: The MIT Press.

Gamer, M., & Berti, S. (2010). Task relevance and recognition of concealed information have different influences on electrodermal activity and event-related brain potentials. *Psychophysiology, 47*(2), 355-364. doi:10.1111/j.1469-8986.2009.00933.x

Gamer, M., & Berti, S. (2012). P300 amplitudes in the concealed information test are less affected by depth of processing than electrodermal responses. *Frontiers in Human Neuroscience, 6,* 308. doi:10.3389/fnhum.2012.00308

Giard, M. H., & Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of Cognitive Neuroscience, 11*(5), 473-490. doi:10.1162/089892999563544

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*(4), 468-484. doi:10.1016/0013-4694(83)90135-9

Hooff, J. C., Sargeant, E., Foster, J. K., & Schmand, B. A. (2009). Identifying deliberate attempts to fake memory impairment through the combined use of reaction time and event-related potential measures. *International Journal of Psychophysiology, 73*(3), 246-256. doi:10.1016/j.ijpsycho.2009.04.002

Johnston, V. S., Miller, D. R., & Burleson, M. H. (1986). Multiple P3s to Emotional Stimuli and Their Theoretical Significance. *Psychophysiology, 23*(6), 684-694. doi:10.1111/j.1469-8986.1986.tb00694.x

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324. doi:10.3758/s13428-011-0168-7

Matsuda, I., Nittono, H., & Allen, J. J. (2012). The Current and Future Status of the Concealed Information Test for Field Use. *Frontiers in Psychology, 3,* 532. doi:10.3389/fpsyg.2012.00532

Matsuda, I., Nittono, H., Hirota, A., Ogawa, T., & Takasawa, N. (2009). Event-related brain potentials during the standard autonomic-based concealed information test. *International Journal of Psychophysiology, 74*(1), 58-68. doi:10.1016/j.ijpsycho.2009.07.004

Matsuda, I., Nittono, H., & Ogawa, T. (2011). Event-related potentials increase the discrimination performance of the autonomic-based concealed information test. *Psychophysiology, 48*(12), 1701-1710. doi:10.1111/j.1469-8986.2011.01266.x

Matsuda, I., Ogawa, T., Tsuneoka, M., & Verschuere, B. (2014). Using pretest data to screen low-reactivity individuals in the autonomic-based concealed information test. *Psychophysiology, 52*(3), 436-439. doi:10.1111/psyp.12328

Meijer, E. H. (2008). *Psychophysiology and the detection of deception: promises and perils.* Maastricht, The Netherlands: Maastricht University Press.

Meijer, E. H., Selle, N. K., 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. doi:10.1111/psyp.12239

Münsterberg, H. (1908). *On the witness stand*. New York, NY: McClure Company.

Nabeta, T., & Kawahara, J. (2006). Congruency effect of presentation modality on false recognition of haptic and visual objects. *Memory, 14*(3), 307-315. doi:10.1080/09658210500277398

Nahari, G., & Ben-Shakhar, G. (2010). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology, 48*(6), 733-744. doi:10.1111/j.1469-8986.2010.01148.x

Noordraven, E., & Verschuere, B. (2013). Predicting the Sensitivity of the Reaction Time-based Concealed Information Test. *Applied Cognitive Psychology, 27*(3), 328-335. doi:10.1002/acp.2910

Ordeal of Boiling Water, 12th or 13th Century. (1996, January). In P. Halsall (Ed.), *Internet Medieval Sourcebook.* Fordham University. Retrieved from https://sourcebooks.fordham.edu/source/water-ordeal.asp

Paul, S., Kathmann, N., & Riesel, A. (2016). The costs of distraction: The effect of distraction during repeated picture processing on the LPP. *Biological Psychology, 117*, 225-234. doi:10.1016/j.biopsycho.2016.04.002

Peth, J., Vossel, G., & Gamer, M. (2011). Emotional arousal modulates the encoding of crime-related details and corresponding physiological responses in the Concealed Information Test. *Psychophysiology, 49*(3), 381-390. doi:10.1111/j.1469-8986.2011.01313.x

Polich, J. (1990). P300, Probability, and Interstimulus Interval. *Psychophysiology, 27*(4), 396-403. doi:10.1111/j.1469-8986.1990.tb02333.x

Polich, J. (1997). On the relationship between EEG and P300: Individual differences, aging, and ultradian rhythms. *International Journal of Psychophysiology, 26*(1-3), 299-317. doi:10.1016/s0167-8760(97)00772-1

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148. doi:10.1016/j.clinph.2007.04.019

Pritchard, W. S. (1981). Psychophysiology of P300. *Psychological Bulletin, 89*(3), 506-540. doi:10.1037/0033-2909.89.3.506

Ratcliff, G., & Newcombe, F. (1982). Object recognition: Some deductions from the clinical evidence. In A. W. Ellis (Ed.), *Normality and pathology in cognitive function*. London: Academic Press.

Renard, Y., Lotte, F., Gibert, G., Congedo, M., Maby, E., Delannoy, V., Bertrand, O., & Lécuyer, A. (2010). OpenViBE: An Open-Source Software Platform to Design, Test, and Use Brain–Computer Interfaces in Real and Virtual Environments. *Presence: Teleoperators and Virtual Environments, 19*(1), 35-53. doi:10.1162/pres.19.1.35

Rosenfeld, J. P., Nasman, V. T., Whalen, R., Cantwell, B., & Mazzeri, L. (1987). Late Vertex Positivity in Event-Related Potentials as a Guilty Knowledge Indicator: A New Method of Lie Detection. *International Journal of Neuroscience, 34*(1-2), 125-129. doi:10.3109/00207458708985947

Rosenfeld, J. P., Ward, A., Thai, M., & Labkovsky, E. (2015). Superiority of Pictorial Versus Verbal Presentation and Initial Exposure in the P300-Based, Complex Trial Protocol for Concealed Memory Detection. *Applied Psychophysiology and Biofeedback, 40*(2), 61-73. doi:10.1007/s10484-015-9275-z

Rummer, R., Schweppe, J., & Martin, R. C. (2013). Two modality effects in verbal short-term memory: Evidence from sentence recall. *Journal of Cognitive Psychology, 25*(3), 231-247. doi:10.1080/20445911.2013.769953

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18*(6), 813-847. doi:10.1080/09541440500412361

Stenberg, G., Radeborg, K., & Hedman, L. R. (1995). The picture superiority effect in a cross-modality recognition task. *Memory & Cognition, 23*(4), 425-441. doi:10.3758/bf03197244

Appendix