Psychophysiology, 44 (2007), 814–822. Blackwell Publishing Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2007.00543.x

Combining skin conductance and forced choice in the detection of concealed information

EWOUT H. MEIJER, FREN T. Y. SMULDERS, JAMES E. JOHNSTON, AND HARALD L.G.J. MERCKELBACH

Department of Experimental Psychology, Faculty of Psychology, Maastricht University, Maastricht, The Netherlands

Abstract

An advantage of the concealed information polygraph test (CIT) is that its false positive rate is determined on statistical grounds, and can be set a priori at arbitrary low levels (i.e., few innocents declared guilty). This criterion, however, inevitably leads to a loss of sensitivity (i.e., more guilty suspects declared innocent). We explored whether the sensitivity of a CIT procedure could be increased by adding an independent measure that is based on an entirely different psychological mechanism. In two experiments, we explored whether the accuracy of a CIT procedure could be increased by adding Symptom Validity Testing (SVT), a relatively simple, forced-choice, self-report procedure that has previously been used to detect malingering in various contexts. Results of a feigned amnesia experiment but not from a mock crime experiment showed that a combination measure of both tests yielded better detection than either test alone.

Descriptors: Polygraph testing, Guilty Knowledge Test, Concealed Information Test, Symptom Validity Test

The use of the polygraph in criminal investigations has been heavily criticized in the scientific literature (e.g., Ben Shakhar, 2002; Fiedler, Schmid, & Stahl, 2002; Lykken, 1998; National Research Council, 2003). This critique primarily concerns the Control Question Test (CQT), the technique most widely used in police investigations. During a CQT, physiological recordings to questions directly related to the incident under investigation (e.g., "Did you stab John Doe?") are compared to emotionally provocative control questions (e.g., "During the first 20 years of your life, did you ever hurt someone physically?"). Stronger physiological responding to the control questions is taken as an indication of innocence, whereas stronger physiological responding to the questions directly related to the crime is taken as an indication of deception. The CQT has been criticized for its lack of theoretical framework, its lack of standardization, and the fact that it relies on improper controls, resulting in a high percentage of false positives (i.e., innocent examinees tested guilty; see Ben-Shakhar, 2002; Lykken, 1998).

A different interrogation technique, first described by Münsterberg (1908), and later named the Guilty Knowledge Test (GKT; Lykken, 1959, 1960) or Concealed Information Test (CIT) is used less frequently. In fact, the only country where it is employed on a large scale is Japan (Hira & Furumitsu, 2002; Nakayama, 2002). In a CIT, physiological measures are similar to the CQT, but questions are presented in a multiple choice format (e.g., "Was the amount of money stolen \$10,000, \$20,000, \$30,000, \$40,000, or \$50,000?"). All questions concern intimate details of the crime, of which only the investigative authorities and the perpetrator are presumed knowledgeable. Consistent stronger physiological responding to the correct answers reflects knowledge of these details and thus involvement in the crime.

Unlike the CQT, the CIT is highly standardized and has a sound theoretical underpinning in orienting theory, as recently shown by Verschuere, Crombez, De Clercq, and Koster (2004). It has also been shown to be a robust tool for discriminating between guilty and innocent participants. A recent meta-analysis on 80 studies revealed an average effect size (d) of 1.55, and this effect size was even larger (3.12) under optimal conditions (i.e., motivational instructions, deceptive verbal response, at least five questions; Ben Shakhar & Elaad, 2003). Furthermore, the CIT provides adequate safeguards for the innocent in that the probability of a false positive outcome can be determined a priori. This probability depends on the number of questions, the number of alternative answers per question, and the criterion for guilt. For example, when a CIT contains four questions with four answers each, the probability for an innocent examinee to systematically show the strongest physiological response to the correct alternative is $(1/4)^4 = .004$. If the criterion for guilt is set at "respond maximally to at least three out of the four questions," this probability becomes .05.[1] Thus, for a CIT, the probability of

Jim Johnston is now at Best Evidence Inc., Tampa, Florida.

The authors thank Gershon Ben-Shakhar and two anonymous reviewers for their helpful comments. We also thank Nieke Elbers for her assistance in data collection.

Address reprint requests to: E.H. Meijer, Department of Experimental Psychology, Faculty of Psychology, Maastricht University. PO Box 616, Maastricht, The Netherlands. E-mail: eh.meijer@psychology.unimaas.nl

[1]This probability is composed of the nonmutually exclusive events of responding maximally to all four questions (A) and responding maximally to three out of the four questions (B), and can be calculated as

a false positive outcome can be set a priori by including sufficient questions and alternatives and by selecting a proper detection criterion. For use in criminal justice, detection criteria that result in a low number of false positive outcomes (i.e., high specificity) are important, because it adheres to the legal doctrine in most countries, abbreviated in the so-called Blackstone Maxim: "Better that ten guilty persons escape than that one innocent suffer" (Blackstone, 1882; Volokh, 1997).

However, an inevitable consequence of setting the detection criteria at levels corresponding with high specificity is that the percentage of false negative outcomes (i.e., a guilty examinee tested innocent) becomes larger. Several recent studies have estimated this percentage to range between 14% and 24% (Ben Shakhar & Elaad, 2003; Elaad, 1998). These studies primarily relied on mock crime scenarios to determine detection efficiency. The false negative rates in field studies on detection efficiency of the CIT are even higher. For example, Elaad (1990) and Elaad, Ginton, and Jungman (1992) found that the CIT missed approximately half of the guilty suspects, when using skin resistance response as the detection measure and confessions as an index of guilt. A more recent mock crime study by Carmel, Dayan, Naveh, Raveh, and Ben-Shakhar (2003) showed that under realistic conditions, the percentage of false negative outcomes was as high as 48%. The problem of false negative outcomes associated with the CIT cannot simply be explained by a perpetrator's failure to remember pertinent details. Elaad (1990), for example, presented the items from a selection of the records to independent judges. These judges were asked whether these items had at least an 80% likelihood of being recognized by the guilty subject. This was the case in the majority of the items, leading Elaad to conclude that the high proportion of false negatives was due to the low number of questions, rather than due to the fact that items were not encoded in memory. Likewise, in the study by Carmel et al. (2003), even though the false negative rate was only 10% under optimal conditions (i.e., a CIT containing only questions on central details, performed immediately following a realistic mock crime), this rate still ranged from 20% to 45% when only correctly recalled items were analyzed.

One approach to reduce the number of false negatives without sacrificing specificity is combining several highly specific measures. So far, this approach has primarily focused on the addition of new psychophysiological indices to the standard measurement of skin conductance. For example, a specific respiration parameter (respiration line length; RLL) has been shown to increase the sensitivity (i.e., proportion of correctly classified guilty examinees) of a skin conductance based CIT (e.g., Ben-Shakhar & Dolev, 1996; Ben-Shakhar & Elaad, 2002; Ben Shakhar, Gronau, & Elaad, 1999; Elaad et al., 1992; Timm, 1982). More recently, measures such as Finger Pulse Waveform Length (Elaad & Ben-Shakhar, 2006) and Normalized Pulse Volume (Hirota et al., 2003) have been proposed as useful additional CIT parameters. In general, the addition of extra physiological variables is useful if the false negative outcomes of the CIT are due to noise in the measurement of the underlying psychophysiological mechanism (i.e., the orienting response). In that case, increasing the number of physiological indices that are manifestations of this underlying mechanism will reduce the noise and hence the percentage of false negatives.

If the false negative outcomes are not due to measurement noise, but simply result from the absence of response in the underlying mechanism, adding more psychophysiological indices might be less fruitful. In this case, simply adding more physiological indices will not increase sensitivity. It might then be more fruitful to combine the CIT with highly specific measures that are independent of the psychophysiological mechanism tapped by CIT (see also Nies & Sweet, 1994). An interesting candidate measure that might be used in this way is the Symptom Validity Test (SVT). This test was developed to detect malingering and has been used in a wide variety of fields, including detection of malingering of perceptual deficits (Brady & Lind, 1961), short-term memory deficits (Hiscock & Hiscock, 1989), amnesia for specific events (Frederick, Carter, & Powel, 1995), and cognitive deficits attributed to chronic pain (Meyers & Diep, 2000) or posttraumatic stress disorder (Rosen & Powel, 2003).

The rationale of the SVT lies in the notion that the performance of honest individuals (i.e., individuals with genuine perceptual or memory deficits or without intimate knowledge of a crime) on a forced choice test will be at chance level. Take, for example, an individual with genuine color blindness. This person is presented with a series of events, for example red or blue illuminating bulbs. After each event, the individual is asked to name the color of the light bulb that lit up. These questions have a forced-choice format with answers of equal probability (e.g., 1. blue, 2. red). For this person, test performance will be at chance level. An individual with intact perception will perform above chance level. Deception (i.e., malingering of a perceptual deficits) is inferred when performance falls significantly below chance level as the person apparently has the ability to systematically avoid the correct answers and select the incorrect answer more often than predicted by chance (Denney, 1996). Even when aware of this rationale, dishonest people may still fail the test due to humans' incapacity to generate random series of responses (see also Haughton, Lewsley, Wilson, & Williams, 1979; Wagenaar, 1972). Because the distribution of the number of correct answers in individuals with no true ability is known (binomial), the SVT, like the CIT, allows for the computation of the probability of a false positive outcome at any chosen detection criteria.

Denney (1996) adapted the SVT for use in a forensic setting (see also Lieblich & Ninio, 1972; Lieblich, Shaham, & Ninio, 1976). He described three cases where the defendant claimed amnesia for his or her crime. In all cases, defendants performed well below chance level at an SVT consisting of questions concerning intimate details of the crime (e.g., "How did the perpetrator leave the bank? 1. walking, 2. running"). More recent research on the accuracy of the SVT in detecting feigned amnesia for mock crimes shows that, at specificity levels of 95%, its sensitivity ranges from 40% to 60% (Jelicic, Merckelbach, & van Bergen, 2004a, 2004b; Merckelbach, Hauer, & Rassin, 2002).

Because the SVT measures a different psychological mechanism than the CIT (i.e., limitations of the cognitive system in producing randomlike responses versus an orienting response), combining it with a CIT may decrease the number of false negative outcomes while maintaining high levels of specificity. The aim of the current experiments was twofold: First, we explored whether the SVT can be used as an indicator of guilt and deception. Second, we investigated whether the addition of SVT has an incremental value beyond that of a skin conductance based CIT. In the first experiment, participants performed a mock crime, mimicking the application of the CIT and SVT in a typical forensic application. In the second experiment, participants were

follows: P(A)+P(B) − P(A and B), and equals $.25^4 + 4 \times .25^3 - 4 \times .25^4 = .0508$.

instructed to feign complete amnesia of their identity, after which they were given a CIT and SVT containing biographical data. This test mimics deception in a somewhat different context, for example, to judge the veracity of amnesia claims or when someone's identity is the topic of investigation.

EXPERIMENT 1

Method

Participants

Participants were 65 undergraduates who received either course credits or a small financial compensation. Five participants were excluded from data analyses because they either failed to follow instructions or equipment malfunctioned. Thus, the remaining sample consisted of 60 students (20 men) with a mean age of 21 years ($SD = 2.8$). All participants read and signed a letter of informed consent before participating. The experiment was approved by the Faculty's ethical committee.

Physiological Measures

Skin conductance was measured using a 24-bit DC 0.5-V system. Two Beckmann Ag/AgCl electrodes (5 mm in diameter) were placed on the medial phalanges of the first and second fingers of the participants' nondominant hand. Electrodes were filled with isotonic electrode paste (0.9% NaCl). Respiration was measured using a strain gauge attached around the thorax. All data were acquired using Contact Precision Instruments bioamplifiers with a sample rate of 200 Hz.

Procedure

Upon arrival in the laboratory, the participant was given written instructions to carry out a scenario. For half of the participants ($N = 30$), these instructions entailed the guilty scenario. The other half received the innocent scenario. The guilty scenario consisted of stealing 20 euros and a mobile telephone hidden away in a jacket in a café located inside the university building. To gain access to this café, participants had to collect a key, which was located in a drawer of a kitchen unit. In the innocent condition, the task involved collecting a dirty cup from a kitchen and washing it elsewhere. Both innocent and guilty instructions were concluded by telling the participant to wait for further instructions in a waiting room. After 15 min the examiner entered the room and informed participants that "a crime has been committed and you are one of the suspects. If you are guilty, try to lie effectively during the lie detection test so that you will be declared innocent." Following this, the experimenter escorted the participants back to the laboratory where testing commenced.

The CIT consisted of one example question and six genuine questions. Questions were presented on a 15-in. monitor. Each question was followed by a set of six items, among which was the correct answer (critical item). The first item was never the critical item and served to absorb novelty orienting responses. The six questions of the CIT addressed both central and peripheral details of the crime. Each question was displayed for 10 s. Then, a blank screen followed for 3 s, after which the first item was displayed for 3 s. Next, another blank screen followed for 10 s. This cycle was repeated for each of the six items, creating a 26-s interstimulus interval. The critical item was always positioned at either the third, fourth, or fifth place. The order of the questions was determined by a balanced latin square. Participants had to respond to the presentation of each item with a verbal "no" answer. A participant-terminated break was given after completion of an entire question.

Upon completion of the CIT, participants were given the SVT. The SVT consisted of 12 questions, each with two equally plausible alternatives. These items were checked using a Doob and Kirschenbaum pilot procedure to ensure they were all equally plausible (Doob & Kirschenbaum, 1973). For this procedure, 10 naïve participants were given all questions and asked to pick the most plausible item. Any item for which the probability was below .3 or above .7 was discarded. Six of the 12 SVT questions resembled those of the CIT. The additional six questions concerned specific details of the café where the mock crime took place. The SVT was administered in the form of a booklet, containing only one question per page with the following instructions: "Complete this questionnaire by circling one of the answers to each question. You must always choose one option. If you do not know the answer, just guess. You must answer the questions in the order they are presented. Do not turn to the next page unless the question has been answered. Do not turn back the page under any circumstance." The thickness of the booklet was increased by adding 12 empty pages at the end. This was done so as to obscure the true length of the test, making it difficult for participants to calibrate their performance in accordance with chance. To prevent participants from deriving correct CIT answers from the SVT answers, the CIT was always administered first.

All testing took place in a dimly lit, sound-proof, air-conditioned laboratory. Participants were monitored from a control room by means of a video surveillance camera and a microphone.

Response Scoring and Data Analysis

The maximal positive deflection in skin conductance during the 1 s to 5 s interval after stimulus onset was defined as the SCR. To eliminate individual differences in responsivity, within-question standardized scores were computed by subtracting the mean of all five responses from the response to the critical item and dividing that by the standard deviation of all five responses (Ben-Shakhar, 1985). These standardized scores were then averaged over questions in order to produce a single detection score for the CIT.

Siegel's (1956) formula was used to calculate the z-score for the SVT: $z = ((x \pm 0.5) - NP)/\sqrt{(NP(1-P))}$. Here, z is the test statistic, x is the number of correct responses, N is the total number of questions (i.e., 12), and P is the probability of a correct discrimination given no true ability (i.e., 0.5). Due to the fact that the binomial distribution involves a discrete variable, a correction for continuity was made: adding 0.5 when $x < NP$ and subtracting 0.5 when $x > NP$.

Results and Discussion

The SVT and CIT scores within guilty participants were uncorrelated ($r = -.11$; $p = .56$). To derive accuracy rates, cutoff points for the detection measures were set at a z-score < -1.65 for the SVT (corresponding to a specificity of 95%), whereas for the CIT, the Lykken score was used. With the latter, each question is assigned 2 points if the response to the crime relevant item is the largest of all responses, 1 point is assigned if it is the second largest, and 0 points are assigned in all other cases. All points are then added, and a score of 6 or more is taken as a guilty test outcome. Based on the binomial theorem, this cutoff point

Table 1. *Means and Standard Deviations of the Concealed Information Test (CIT), Symptom Validity Test (SVT), and Their Combination for the Guilty and Innocent Conditions*[a]

Measure	Mean *z* guilty	Standard deviation guilty	Mean *z* innocent	Standard deviation innocent	*d*	*a*	95% CI of *a*
Experiment 1							
CIT	0.78	0.61	− 0.06	0.41	1.62	.86	.77–.96
SVT	0.73	1.41	− 0.15	0.71	0.79	.70	.56–.83
SVT & CIT	1.51	1.47	− 0.21	0.88	1.42	.84	.74–.95
Experiment 2							
CIT	0.95	0.60	0.10	0.39	1.68	.88	.81–.94
SVT	2.08	1.34	0.02	0.76	1.89	.87	.81–.93
SVT & CIT	3.03	1.44	0.12	0.85	2.46	.95	.91–.99

[a]Standardized differences between the means of the guilty and innocent condition (*d*). Area under the receiver operating characteristic curve (*a*), with its 95% confidence interval.

corresponds to a specificity of 83% (see also MacLaren, 2001). To determine accuracy rates for the combination of the CIT and SVT, we used the Independent Parallel Testing approach (National Research Council, 2003, p. 367). With this approach, deception is inferred if any of the individual tests is positive. Consequently, overall test outcome is negative only when all individual tests are negative.

The cutoff resulted in correct classification of all of the innocent (100%) and 14 (47%) of the guilty participants for the CIT. For the SVT, it yielded correct classification of all (100%) of the innocent and 8 (27%) of the guilty participants. The combination of the CIT and SVT resulted in correct classification of all innocent (100%) and 17 (57%) of the guilty participants.

Defining guilt and innocence using the criteria based method described above has the disadvantage that it relies on a single, arbitrary cutoff point. An alternative approach to describing detection efficiency that does not have this disadvantage is signal detection theory (SDT; National Research Council, 2003). This method defines detection efficiency in terms of the degree of separation between the distributions of the detection measure for the innocent and the guilty conditions. To produce a single detection score for the combination of the two tests, the SVT *z*-score was multiplied by -1 and added to the CIT *z*-score. Subsequently, the distance between the centers of the distribution of the innocent and the distribution of the guilty was computed in terms of standard deviation (*d*), and the area under the (empirical) Receiver Operating Characteristic (ROC) curve (*a*) was computed. These statistics are presented in Table 1 (top panel). Table 1 reveals that *d* values for the CIT and SVT were 1.62 and 0.79, respectively. The *d* value for the combination of CIT and SVT was 1.42. The areas under the ROC curve were .86 for the CIT, .70 for the SVT, and .84 for the combination.

These results indicate that the SVT can be used to detect deception in a typical forensic setting, even though sensitivity was modest. The signal detection parameters revealed no incremental validity of the SVT over the CIT. To conceptually replicate these results, we conducted a second experiment.

In this second experiment, a number of methodological improvements and extensions were made. First, to allow for generalization of the results, a community sample was used, an incentive for beating the test was given, and a feigned amnesia paradigm was applied. To increase statistical power, only guilty participants were included. Furthermore, any possible carryover effect due to the fixed order in Experiment 1 was addressed by using different questions for each test and balancing the order of the tests. Also, the Psychopathic Personality Inventory (PPI; Lilienfeld & Andrews, 1996) was included as a measure of psychopathic traits.

Previous research has shown that psychopathy might be a moderating factor in detecting concealed information. Hyporeactivity is a prominent feature of psychopathy (Lorber, 2004), and recently Verschuere, Crombez, Declercq, and Koster (2005) showed that prisoners who scored high on certain subscales of the PPI exhibited both a decreased overall electrodermal orienting response and decreased differential electrodermal orienting responses to the relevant and irrelevant CIT answers. The PPI was included to examine whether this hyporeactivity phenomenon in high PPI individuals could be replicated and if addition of a SVT could be a potential solution for the reduced detection efficiency that it implies.

EXPERIMENT 2

Method

Participants

Participants were 60 people (18 men) recruited through advertisement in local newspapers. The mean age was 33 years ($SD = 9.1$). All participants read and signed a letter of informed consent before participating. The experiment was approved by the Faculty's ethical committee.

Measurements

The Dutch translation of the Psychopathic Personality Inventory (PPI; Jelicic, Merckelbach, Timmermans, & Candel, 2004; Lilienfeld & Andrews, 1996) was used to assess psychopathic traits. Following Benning, Patrick, Hicks, Blonigen, and Krueger (2003) and Verschuere et al. (2005), we calculated the Fearless Dominance factor and the Impulsive Antisocial factor by summing scores across the appropriate subscales while compensating for the fact that these subscales consist of different numbers of items. Physiological measures were identical to those in Experiment 1.

Procedure

Participants who responded to the newspaper advertisement were contacted for an appointment. During this initial contact, they were asked to supply autobiographical information (e.g., place of birth, mother's maiden name, etc.) that was subsequently used as stimulus material in the experiment.

Upon arrival in the laboratory, participants were asked to fill out the PPI. Subsequently, they were given written instructions.

These instructions explained that in some circumstances, claiming memory problems can have beneficial effects. An example of how feigning memory problems after a traffic accident could increase compensation payments paid by the insurance company was given. Next, participants were instructed to feign complete memory loss of their identity and told that the experiment was designed to test new methods to detect their deceit. They were explicitly told to try to beat the test and were promised a €5 reward if they succeeded.

Initially, participants provided the experimenter with a possible total of 24 autobiographical details. On the basis of these details, 18 questions were constructed, divided into three sets of 6 questions, such that the different sets all contained questions of a similar nature (e.g., each set contained the same number of names of relatives). Subsequently, one set was used for the CIT and two for the SVT. The order of the two tests was counterbalanced. The remainder of the procedure was identical to Experiment 1, with the exception that the critical alternative was randomly presented at any position except for the first.

Response Scoring and Data Analysis

Response scoring and data analysis were similar to Experiment 1. Because only "guilty" participants were included, signal detection parameters were derived differently (see below).

Results and Discussion

SVT and CIT scores within guilty participants were uncorrelated ($r = -.04$, $p = .77$). Using identical criteria as in Experiment 1 resulted in correct classification of 39 (65%) of the participants for the CIT and 38 (63%) of the participants for the SVT. The combination of the CIT and SVT correctly classified 53 (88%) of the guilty participants.

To derive signal detection indices, a number of previous studies that included only guilty participants compared the distribution of the standardized critical items to the distribution of the average standardized control items (e.g., Elaad & Ben-Shakhar, 2006; Gronau, Ben-Shakhar, & Cohen, 2005; Verschuere et al., 2005; see also Ben-Shakhar, 1985). As we will argue below, however, this procedure is suboptimal and overestimates detection efficiency.

The standardization procedure described by Ben-Shakhar (1985) entails subtracting the mean and dividing the outcome by the standard deviation of responses to *all* alternatives from either the response to the critical alternative or from the response to the control alternatives. As a consequence, all information in the data set is used to derive the distribution of the critical item. Applying the same procedure to the control items can thus, by definition, not result in a distribution of control items containing unique information. In fact, each participant's score for the control items, together forming the "innocent" group, is linearly dependent on that participant's score on the critical items. This is because the score on the control items is simply the score on the critical item, divided by $-(N-1)$, where N denotes the total number of unique stimuli. For the demonstration please refer to Appendix A. Furthermore, as a consequence of the averaging over the standardized control items, the standard deviation of the distribution of control scores becomes approximately $\sqrt{(N-1)}$ times smaller than would have been the case if they had been obtained from a single item (e.g., the item that was critical in the "guilty" group). Thus, in case of successful detection, this procedure renders a distribution of the control item with a negative mean and a smaller standard deviation than a hypothetical group of truly innocent participants. The latter would have a mean of 0 and an unbiased standard deviation. As a consequence of this negative mean and smaller standard deviation, signal detection parameters will be unjustly inflated.

Alternatively, we chose to base our signal detection parameters on a comparison with a simulated innocent group consisting of 60 participants (see also Carmel et al., 2003). Such a group was created for both tests by randomly drawing values from their respective distributions, and treating these values in exactly the same manner as the values measured for guilty participants. For the SVT, this entailed drawing 60 values from the binomial distribution with $N = 12$ and $p = .5$. For the CIT it entailed the following steps. First, five values were randomly drawn from a standard normal distribution (mean = 0, standard deviation = 1). Then, one value (as the "response" to the critical item) was standardized relative to the mean and standard deviation of all five responses. This way, a standardized score for one innocent person for one question was derived. This process was repeated six times (to simulate six questions), and these six values were averaged to represent a score for one innocent participant. Based on this procedure, the d values were 1.68, 1.89, and 2.46 for the CIT, SVT, and their combination, respectively, and a values were .88, .87, and .95. These values, with their corresponding 95% confidence interval are presented in Table 1 (bottom panel). Statistical testing of these a values revealed that the combination of the two tests outperformed the CIT alone ($z = 2.24$, $p = .01$; see Hanley & McNeil, 1983).

Fifty-seven participants filled out the PPI completely. Mean total score was 341 ($SD = 39$; range = 256–420). Internal consistency was high (Cronbach's alpha = .89). The Fearless Dominance and Impulsive Antisocial subscales were uncorrelated, $r = .17$, $p = .21$. To investigate the relationship between psychopathic personality traits and overall physiological responding, we computed the correlation between these PPI subscales and the unstandardized mean skin conductance response. Neither the Fearless Dominance factor nor the Impulsive Antisocial factor significantly correlated with overall physiological responding ($r = -.08$, $p = .57$ and $r = -.04$, $p = .75$, respectively). Similarly, the correlations between the two subscales and the detection measures for CIT and SVT did not attain significance ($r = -.24$, $p = .07$ and $r = -.11$, $p = .41$ for CIT and $r = -.07$, $p = .62$ and $r = -.11$, $p = .42$ for SVT with the Fearless Dominance and the Impulsive Antisocial subscales, respectively).

General Discussion

This study aimed to investigate whether SVT can be used to detect deception and whether combining it with a CIT would yield detection efficiency superior to that of the CIT alone. First of all, the results from both experiments show that the SVT can be used to detect deception. Furthermore, we found that combining the two tests yielded superior detection efficiency, but only in the feigned amnesia experiment.

In both experiments, the accuracy rate of the SVT in detecting deception was similar to that found in studies on false claims of amnesia (e.g., Jelicic et al., 2004a, 2004b; Merckelbach et al., 2002). This is not surprising, because the instructions to the participants in all these studies were highly similar. That is, the instruction to feign amnesia for a mock crime is a special instance

of instructing participants to lie (Christianson & Merckelbach, 2004).

The accuracy rates obtained with the CIT were equivalent to those found in earlier studies as well (e.g., Carmel et al., 2003). Importantly, the instructions to the guilty participants in our Experiment 1 contained no specific information addressed by the subsequent CIT items (e.g., the instructions read "steal the money ... " and not "steal the 20 euro ... "). Therefore, our Experiment 1 would qualify as a procedure that Carmel and coworkers (2003) termed a "valid mock crime," and it yielded detection rates similar to those obtained by these authors.

The ability of both the SVT and CIT to differentiate between guilty and innocent participants is also evident from the *d* values. In terms of Cohen (1988), the value of 0.79 for the SVT in Experiment 1 represents a moderate to large effect size, whereas the 1.62 for the CIT in Experiment 1 and 1.68 and 1.89 found in Experiment 2 all represent a large effect size.

The detection efficiency found in Experiment 2 was higher than that in Experiment 1, particularly for the SVT. This may explain why the incremental validity was limited to Experiment 2. The predictive validity of the SVT in Experiment 1 may simply have been too low to establish a significant incremental validity given the number of participants.

There are three factors that may have contributed to the difference in detection efficiency between the two experiments. To begin with, it might be that because of its personal relevance, the autobiographical paradigm yields higher accuracy than the mock crime paradigm. At first sight, this might seem difficult to reconcile with the findings of Ben-Shakhar and Elaad (2003), who found that mock crime studies yield higher accuracy than studies using the personal item paradigm. It should be noted, however, that our personal item paradigm was adapted such that participants were specifically instructed to feign amnesia whereas in many other studies, participants are merely instructed to deny recognition. This way our personal item paradigm more closely resembles a mock crime than a typical personal item paradigm. Furthermore, in their study on the validity of reaction times in the detection of concealed information, Gronau et al. (2005) found a similar pattern. In that study, reaction times differed between relevant and irrelevant items when they denoted personally significant information, but not when they pertained to mock crime details. Second, participants in Experiment 2 were promised a financial incentive for beating the test. This may have increased accuracy through an increase in motivation and is probably ecologically valid, because in typical applied settings, great interests are at stake. Finally, the fact that participants in Experiment 2 were drawn from a community sample may have boosted detection rates. After all, people from a community sample are less likely to understand the rationale of the SVT and they have less knowledge of the phenomenon of random performance. However, both Jelicic et al. (2004a) and Merckelbach et al. (2002) also found sensitivities on the order of 60% in an undergraduate sample.

The continuous scores on the CIT and SVT within the guilty group were independent ($r = -.11$ in Experiment 1 and $r = -.04$ in Experiment 2). Also the dichotomized scores ("innocent" vs. "guilty") showed independence: It appeared that the actual probability of being declared guilty on either test (.57 in Experiment 1, .88 in Experiment 2) was very close to the expected probability given independence of the probability of a guilty SVT outcome and the probability of a guilty CIT outcome (.61 in Experiment 1, .87 in Experiment 2).[2] One implication of this independence is that, to the extent that false negative outcomes of the CIT are caused by too small responses in the underlying psychophysiological mechanisms, adding SVT might be a solution. A second practical implication is that, because the two tests measure different mechanisms, they may not be susceptible to the same countermeasures. Countermeasures refer to anything that an individual might do in an effort to defeat or distort a polygraph test (Honts, Devitt, Winbush, & Kircher, 1996). To the extent that various physiological indices are manifestations of the same underlying mechanism (i.e., orienting response), any countermeasure aimed at interfering with this mechanism (e.g., counting backwards from 200 by 7) is likely to have similar undermining effects on these indices. In contrast, adding independent tests may limit the effects of such countermeasures, although this is an issue that warrants systematic empirical testing.

The psychopathy scores obtained in Experiment 2 did not show any link to overall psychophysiological reactivity or with the detection measure of the CIT. This failure to replicate the findings of Verschuere et al. (2005) may have various reasons. For one thing, Verschuere et al. had a prison sample, whereas the current study relied on a community sample. Although one would expect more extreme psychopathy scores in a prison sample, inspection of the data does not confirm this. The mean total score of the community sample in Experiment 2 ($M = 341$, $SD = 39$) was not dramatically lower than that reported by Verschuere and coworkers for their prison sample ($M = 350$, $SD = 40$). Another possible explanation might be that in our second experiment, contrary to the study by Verschuere et al. (2005), a monetary incentive was promised. It could be argued that psychopaths do not show underarousal under motivational conditions (Arnett, 1997; Verschuere, Crombez, Koster, & De Clercq, 2007). Furthermore, our failure to find an association between psychopathy scores and psychophysiological reactivity is in line with earlier work of Gudjonsson (1982) and Balloun and Holmes (1979), who also found no effect of personality on the detection of concealed information. The exact relation between the PPI and autonomous reactivity remains unclear and merits further research.

Another important point of consideration is the practical applicability of the SVT. When combining this procedure with a CIT, little extra effort is needed. Assuming that in the process of preparing CIT questions, the crime scene has been visited or the criminal records and files have been inspected, little extra effort is needed to create additional SVT items. Some authors (e.g., Podlesney, 1993, 2003) have argued that in real-life cases, it is often difficult to formulate sufficient questions with equally plausible answer items. On the other hand, SVT items can be constructed using only two plausible answer alternatives. They allow, for example, yes/no answer options. This makes it relatively easy to develop sufficient items.

When using these tests as forensic tools, one needs to keep in mind that with the cutoff points chosen, and even with SVT and CIT combined, specificity is higher than sensitivity. This implies that both measures can best be used as challenge tasks in the forensic domain. Thus, failing to pass the SVT or the CIT is a

[2]If the probabilities of being declared guilty on each test are independent of each other, under the parallel testing rule, the expected probability of being declared guilty on either or both is given by: seA+(1 − seA) × seB, where seA represents the sensitivity of test A and seB represents the sensitivity of test B.

strong indication of guilt, but passing both tests is not a strong indication of innocence (see also Denney, 1996).

In the analysis of the data from Experiment 2 we encountered a problem with the method that is often used to derive signal detection parameters. As a solution, we proposed a simulated "innocent" group. Comparing the two methods by cross checking our own data from Experiment 1 supported our conclusion that an analysis based on the guilty participants' standardized critical items and the guilty participants' standardized control items leads to an overestimation of the signal detection parameters. This method yielded an *a* of .95 and a *d* of 2.21, whereas the empirical *a* and *d* were .86 and 1.62, respectively. Comparison of the standardized critical item of the guilty participants to a simulated group of innocent participants yielded an *a* of .85 and a *d* of 1.55, indicating that our simulation procedure yields better estimates of the fully empirical signal detection parameters. It is also noteworthy that the mean and standard deviation of the empirical group of innocents in Experiment 1 and the simulated group of innocents in Experiment 2 were highly similar. The overestimation of signal detection parameters that results from Ben-Shakhar's (1985) method might also explain the high values found in the study by Gronau et al. (2005), and might also explain the finding by Elaad and Ben-Shakhar (2006) of better detection efficiency in their experiment with only guilty participants than in their experiment including guilty and innocent participants. We recommend that future research using Ben Shakhar's (1985) standardization procedure should either incorporate both guilty and innocent subjects or compare the distribution of the guilty participants to that of a simulated group of innocent participants.

Finally, we can make two recommendations for future research. First, data from Experiment 2 showed that addition of a behavioral measure increased detection efficiency. Addition of a respiration measure, however, has also been shown to increase detection efficiency (Ben-Shakhar & Dolev, 1996; Elaad et al., 1992; but see Verschuere et al., 2007). Future studies could make a direct comparison between the incremental validity of a psychophysiological measure like respiration with that of behavioral measures like the SVT. Secondly, recent studies have shown that peripheral details do not serve as good CIT items (Carmel et al., 2003; Jokinen, Santtila, Ravaja, & Puttonen, 2006). Of special interest in this matter is a study by Lieblich, Ben Shakhar, and Kugelmass (1976), who showed that detection of relevant autobiographical information like names of relatives was better than detection of less relevant information like favorite brand of cigarettes. To the extent that the detection efficiency of the CIT is determined by salience of the test items, reserving the most salient items for the CIT while including the less salient items in the SVT may boost detection efficiency even more then found in Experiment 2.

Altogether, our results suggest that it is worthwhile to combine several different types of lie detection and that testing for concealed information need not be confined to indices measuring orienting response.

REFERENCES

Arnett, P. A. (1997). Autonomic responsivity in psychopaths: A critical review and theoretical proposal. *Clinical Psychology Review*, *17*, 903–936.

Balloun, K. D., & Holmes, D. S. (1979). Effects of repeated examinations on the ability to detect guilt with a polygraphic examination: A laboratory experiment with a real crime. *Journal of Applied Psychology*, *64*, 316–322.

Ben-Shakhar, G. (1985). Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology*, *22*, 292–299.

Ben Shakhar, G. (2002). A critical review of the Control Questions Test (CQT). In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 103–126). San Diego: Academic Press.

Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effects of mental countermeasures. *Journal of Applied Psychology*, *81*, 273–281.

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*, 972–977.

Ben Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology*, *88*, 131–151.

Ben Shakhar, G., Gronau, N., & Elaad, E. (1999). Leakage of relevant information to innocent examinees in the GKT: An attempt to reduce false-positive outcomes by introducing target stimuli. *Journal of Applied Psychology*, *84*, 651–660.

Benning, S. D., Patrick, C. J., Hicks, B. M., Blonigen, D. M., & Krueger, R. F. (2003). Factor structure of the psychopathic personality inventory: Validity and implications for clinical assessment. *Psychological Assessment*, *15*, 340–350.

Blackstone, W. (1882). *Commentaries on the laws of England* (3rd ed.). London: Murray.

Brady, J. P., & Lind, D. L. (1961). Experimental analysis of hysterical blindness. *Archives of General Psychiatry*, *4*, 331–339.

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*, 261–269.

Christianson, S., & Merckelbach, H. (2004). Crime-related amnesia as a form of deception. In P. A. Granhag & L. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 195–228). Cambridge, UK: Cambridge University Press.

Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Denney, R. L. (1996). Symptom Validity Testing of remote memory in a criminal forensic setting. *Archives of Clinical Neuropsychology*, *11*, 589–603.

Doob, A. N., & Kirschenbaum, H. M. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Administration*, *1*, 187–293.

Elaad, E. (1990). Detection of guilty knowledge in real-life criminal investigations. *Journal of Applied Psychology*, *75*, 521–529.

Elaad, E. (1998). The challenge of the Concealed Knowledge Polygraph Test. *Expert Evidence*, *6*, 161–187.

Elaad, E., & Ben-Shakhar, G. (2006). Finger pulse waveform length in the detection of concealed information. *International Journal of Psychophysiology*, *61*, 226–234.

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

Fiedler, K., Schmid, J., & Stahl, T. (2002). What is the current truth about polygraph lie detection? *Basic and Applied Social Psychology*, *24*, 313–324.

Frederick, R. I., Carter, M., & Powel, J. (1995). Adapting symptom validity testing to evaluate suspicious complaints of amnesia in medicolegal evaluations. *Bulletin of the American Academy of Psychiatry and the Law*, *23*, 227–233.

Gronau, N., Ben-Shakhar, G., & Cohen, A. (2005). Behavioral and physiological measures in the detection of concealed information. *Journal of Applied Psychology*, *90*, 147–158.

Gudjonsson, G. H. (1982). Some psychological determinants of electrodermal responses to deception. *Personality and Individual Differences*, *3*, 381–391.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.

Haughton, P. M., Lewsley, A., Wilson, M., & Williams, R. G. (1979). A forced-choice procedure to detect feigned or exaggerated hearing loss. *British Journal of Audiology*, *13*, 135–138.
Hira, S., & Furumitsu, I. (2002). Polygraphic examinations in Japan: Application of the Guilty Knowledge Test in forensic investigations. *International Journal of Police Science & Management*, *4*, 16–27.
Hirota, A., Sawada, Y., Tanaka, G., Nagano, Y., Matsuda, I., & Takasawa, N. (2003). A new index for psychophysiological detection of deception: Applicability of normalized pulse volume. *Japanese Journal of Physiological Psychology and Psychophysiology*, *21*, 217–230.
Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology*, *11*, 967–974.
Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the Concealed Knowledge Test. *Psychophysiology*, *33*, 84–92.
Jelicic, M., Merckelbach, H., Timmermans, M., & Candel, I. (2004). De Nederlandstalige versie van de Psychopathic Personality Inventory Psychodiagnostisch gereedschap [The Dutch version of the Psychopathic Personality Inventory: Some psychometric results]. *De Psycholoog*, *39*, 604–608.
Jelicic, M., Merckelbach, H., & Van Bergen, S. (2004a). Symptom validity testing of feigned amnesia for a mock crime. *Archives of Clinical Neuropsychology*, *19*, 525–531.
Jelicic, M., Merckelbach, H., & van Bergen, S. (2004b). Symptom validity testing of feigned crime-related amnesia: A simulation study. *The Journal of Credibility Assessment and Witness Psychology*, *5*, 1–8.
Jokinen, A., Santtila, P., Ravaja, N., & Puttonen, S. (2006). Salience of guilty knowledge test items affects accuracy in realistic mock crimes. *International Journal of Psychophysiology*, *62*, 175–184.
Lieblich, I., Ben Shakhar, G., & Kugelmass, S. (1976). Validity of the guilty knowlegde technique in a prisoners' sample. *Journal of Applied Psychology*, *61*, 89–93.
Lieblich, I., & Ninio, A. (1972). Detection of suppressed involvement with information through a forced number-guessing technique. *Acta Psychologica*, *36*, 381–387.
Lieblich, I., Shaham, E., & Ninio, A. (1976). Effects of time stress and stimulus-response set size on the efficiency of detection of involvement with suppressed information through the use of the forced number-guessing technique. *Acta Psychologica*, *40*, 75–84.
Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment*, *66*, 488–524.
Lorber, M. F. (2004). Psychophysiology of aggression, psychopathy, and conduct problems: A meta-analysis. *Psychological Bulletin*, *130*, 531–552.
Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*, 385–388.
Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*, 258–262.
Lykken, D. T. (1998). *A tremor in the blood*. Reading, MA: Perseus Publishing.
MacLaren, V. V. (2001). A quantitative review of the guilty knowledge test. *Journal of Applied Psychology*, *86*, 674–683.
Merckelbach, H., Hauer, B., & Rassin, E. (2002). Symptom validity testing of a feigned dissociative amnesia: A simulation study. *Psychology, Crime, and Law*, *8*, 311–318.
Meyers, J. E., & Diep, A. (2000). Assessment of malingering in cronic pain patients using neuropsychological tests. *Applied Neuropsychology*, *7*, 133–139.
Münsterberg, H. (1908). *On the witness stand*. New York: The McClure Company.
Nakayama, M. (2002). Practical use of the concealed information test for criminal investigation in Japan. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 49–86). San Diego: Academic Press.
National Research Council. (2003). *The polygraph and lie detection*. Committee to Review the Scientific Evidence on the Polygraph. Division of Behavioral and Social sciences and Education Washington, DC: The National Academic Press.
Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology*, *9*, 501–552.
Podlesney, J. A. (1993). Is the guilty knowledge technique applicable in criminal investigations? *A review of FBI case records. Crime Laboratory Digest*, *20*, 57–61.
Podlesney, J. A. (2003). A paucity of operable case facts restricts applicability of the guilty knowledge technique in FBI criminal polygraph examinations. *Forensic Science Communications*, *5*, Retrieved April, 16, 2007, from http://www.fbi.gov/hq/lab/fsc/backissu/july2003/index.htm.
Rosen, G. M., & Powel, J. E. (2003). Use of a Symptom Validity Test in the forensic assessment of Posttraumatic Stress Disorder. *Anxiety Disorders*, *17*, 361–367.
Siegel, S. (1956). *Nonparametric statistics*. New York: McGraw-Hill.
Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology*, *67*, 391–400.
Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. (2004). Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology*, *41*, 461–466.
Verschuere, B., Crombez, G., Declercq, A., & Koster, E. H. W. (2005). Psychopathic traits and autonomic responding to concealed information in a prison sample. *Psychophysiology*, *42*, 239–245.
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*, 309–318.
Volokh, A. (1997). n Guilty men. *University of Pennsylvania Law Review*, *146*, 173–211.
Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, *77*, 65–72.

(Received October 9, 2006; Accepted April 16, 2007)

APPENDIX A

The purpose of this appendix is to demonstrate the linear dependency of the standardized response to a critical item (probe) and the standardized response to control items (irrelevants), when both are derived from the same data set.

For each participant and question, we measured the responses to $N = 5$ unique stimuli, that is 1 probe (p), and $N - 1$ irrelevants ($i_1, \ldots, i_{N-1}$). Following Ben-Shakhar (1985), these responses were transformed to z-scores for the irrelevants:

$$z_j = (i_j - X)/sdx, j = 1, \ldots, N - 1, \quad (1)$$

and for the probe:

$$z_p = (p - X)/sdx, \quad (2)$$

where X and sdx denote the average and standard deviation across all N stimuli, respectively.

In order to derive an ROC curve and compute the area under this curve (a) and d, Ben-Shakhar used the reponses of "guilty" subjects to the irrelevant answer-stimuli to generate an "innocent" group (representing a group for whom all answer-stimuli would be

irrelevant). For this, the average of the irrelevant z-scores was used:

$$Zi = \sum (z_1, \ldots, z_{N-1})/(N-1). \tag{3}$$

By definition, the average of all z-scores is 0:

$$\sum (z_1, \ldots, z_{N-1}, z_P)/N = 0. \tag{4}$$

This means that

$$\sum (z_1, \ldots, z_{N-1}, z_P) = 0, \tag{5}$$

and also that

$$\sum (z_1, \ldots, z_{N-1}) + z_P = 0. \tag{6}$$

Rewriting (3) as

$$\sum (z_1, \ldots, z_{N-1}) = Zi \times (N-1) \tag{7}$$

and filling in in (6) gives

$$Zi \times (N-1) + z_P = 0, \tag{8}$$

$$Zi = z_P/-(N-1), \tag{9}$$

demonstrating the linear dependency of Zi and zp within one question. This linear dependency remains when the standardized responses are averaged across multiple questions.