Running head: CIT WITH MULTIMODAL MEASUREMENT

A Concealed Information Test with multimodal measurement

Wolfgang Ambach *, Stephanie Bursch *, Rudolf Stark **, Dieter Vaitl */**

* Institute for Frontier Areas of Psychology and Mental Health, Freiburg, Germany
** Bender Institute of Neuroimaging, University of Giessen, Germany

Text pages: 40 (this version)
Figures: 4
Tables: 2

Correspondence: Wolfgang Ambach
Institut für Grenzgebiete der Psychologie und Psychohygiene,
Freiburg, Germany
Wilhelmstr. 3A
D-79098 Freiburg
Germany
Telephone: +49-(0)761-20721-25
Fax: +49-(0)761-20721-99
E-mail: ambach@igpp.de

Abstract

A Concealed Information Test (CIT) investigates differential physiological responses to deed-related ("*probe*") vs. *irrelevant* items. The present study focused on the detection of concealed information using simultaneous recordings of autonomic and brain electrical measures. As a secondary issue, verbal and pictorial presentation were compared with respect to their influence on the recorded measures.

Thirty-one participants underwent a mock-crime scenario with a combined verbal and pictorial presentation of nine items. The subsequent CIT, designed with respect to event-related potential (ERP) measurement, used a 3-3.5 second interstimulus interval. The item presentation modality, i.e. pictures or written words, was varied between subjects; no response was required from the participants. In addition to electroencephalogram (EEG), electrodermal activity (EDA), electrocardiogram (ECG), respiratory activity, and finger plethysmogram were recorded.

A significant *probe*-vs.-*irrelevant* effect was found for each of the measures. Compared to sole ERP measurement, the combination of ERP and EDA yielded incremental information for detecting concealed information. Although, EDA per se did not reach the predictive value known from studies primarily designed for peripheral physiological measurement.

Presentation modality neither influenced the detection accuracy for autonomic measures nor EEG measures; this underpins the equivalence of verbal and pictorial item presentation in a CIT, regardless of the physiological measures recorded.

Future studies should further clarify whether the incremental validity observed in the present study reflects a differential sensitivity of ERP and EDA to different sub-processes in a CIT.

Key Words: Concealed Information Test; Deception; Verbal stimuli; Pictorial stimuli; Orienting reflex; Perception.

Introduction

In the last two decades, the psychophysiological detection of concealed information classically investigated using peripheral physiological measures has been extended by the measurement of brain potentials and functional magnetic resonance imaging. With respect to the different dependent measures recorded, the present work focused on the simultaneous measurement of autonomic and brain electrical responses in a Concealed Information Test (CIT) and questioned for incremental information obtained from multimodal measures. As a secondary issue, verbal and pictorial presentations were compared with respect to their influence on the detection of concealed information.

The Concealed Information Test

The Concealed Information Test (CIT), originally called Guilty Knowledge Test (Lykken, 1959), aims at differentiating 'guilty' subjects deceptively denying their deed-related knowledge from 'innocent' subjects truthfully denying knowledge. In order to detect concealed information, it compares a subject's physiological responses towards a number of presented items, typically combined with yes-or-no questions (for a review, see Elaad, 1998; Ben-Shakhar and Elaad, 2003). Each deed-related, *'probe'* item is combined with several (typically four or five) unknown, *'irrelevant'* items of the same category (e.g. keys, tools), resulting in a set of items for each category. The items are presented setwise with the *probe* item never being presented first; also, the first item of each set is discarded from evaluation. It is assumed that only a person with specific knowledge will distinguish between *probe* and *irrelevant* items. This distinction will be reflected in physiological measures even when the subject is trying to conceal his or her crucial knowledge. Subjects possessing the critical knowledge will tend to show the largest response within a set to the *probe* item. In contrast, for a person without specific knowledge, a random pattern of response measures is expected to occur without a preference for either the *probe* or the *irrelevant* items. If a CIT comprises six or seven sets of questions, the presence or absence of concealed knowledge can be assessed with considerable accuracy, i.e. relatively high correct-classification rates for guilty and innocent subjects (for meta-analyses on the accuracy of the CIT, see MacLaren, 2001 and Ben-Shakhar and Elaad, 2003). Deceptive answering is thereby not a necessary condition for the detection of concealed information, yet it has been shown to enhance the differential physiological response to *probe* compared to *irrelevant* items (Furedy and Ben-Shakhar, 1991).

Physiological measures in the detection of concealed information

Classically, the CIT is based on differential responses in peripheral physiological measures, which mainly reflect functions of the autonomic nervous system. Besides skin conductance, which yields the greatest effect sizes between *probe* and *irrelevant* items, phasic heart rate (pHR), respiration line length (RLL), and finger pulse waveform length (FPWL) are known to be valid indicators of concealed information (Bradley and Janisse, 1981; Bradley and Rettinger, 1992; Elaad and Ben-Shakhar, 2006).

In the last two decades, a new approach based on EEG measurement has emerged in research on the detection of concealed information (for reviews, see Rosenfeld, 2002; Rosenfeld, 2005). The endogenous components of ERPs, such as the P300, are generally considered as correlates of higher-order cognitive processes, e.g. attention, stimulus evaluation, or context updating (Polich, 1986; Duncan-Johnson and Donchin, 1982; Donchin and Coles, 1988). The P300 is typically elicited by the presentation of meaningful, rare stimuli randomly interspersed in a sequence of frequent non-meaningful stimuli (Duncan-Johnson and Donchin, 1977). In a CIT, the P300 is superimposed over the brain waves between 300 and 800 ms after stimulus presentation as a positive potential with its maximum at Pz, which is sensitive to concealed knowledge vs. irrelevant material (Rosenfeld et al., 1987; Farwell and Donchin, 1991). For statistics on an individual level, the method was refined in order to specify confidence intervals for the individual peak measure (Rosenfeld et al, 1991; Allen and Iacono, 1997; Abootalebi et al., 2006). The number of presented stimuli was increased for EEG studies to improve the signal-to-noise ratio; interstimulus intervals (ISIs) were shortened (e.g. 864 stimuli per subject; ISI 1.85 seconds; Farwell and Donchin, 1991); verbal answers were abandoned in order to reduce motor artifacts. *Target* stimuli as an additional class of stimuli requiring a unique response from the subject were used in the majority of these studies; they focused the subject's attention to the display; also, brain electrical responses to these *target* stimuli served as a reference for analytic procedures (Farwell and Donchin, 1991). Besides the broad use of this three-class paradigm, a comparable detection accuracy was found using a two-class paradigm without *target* stimuli (correct detection rates for autobiographical knowledge were 90% and 82%, respectively; Rosenfeld et al., 2006).

The Combination of autonomic and central nervous measures

The measurement of the P300 amplitude, which mirrors cognitive processing rather than merely the orienting response (Rösler et al., 1987), and the measurement of peripheral responses, which mainly reflect autonomic functions and basal reflexes, might complement each other rather effectively. However, combining EEG and peripheral measurement within the same experiment generally entails some difficulties: EEG evaluation on a single-trial basis is not very promising due to a rather low signal-to-noise ratio. Therefore, in most ERP studies, a large number of stimuli are presented with short ISIs in order to obtain an adequate number of valid ERPs per condition. In contrast, EDA responses show a markedly slower characteristic, which is opposed to the use of short ISIs. First, the overlap of EDA responses elicited with short ISIs makes it difficult to quantify each single response (Barry, 1993). Second, response amplitudes become critically smaller with increasing presentation frequency (Barry et al., 1993). Third, EDA responses are known to habituate when a large number of stimuli are presented. Whereas the problem of overlapping responses can be solved by decomposing skin conductance (Lim et al., 1997; Alexander et al., 2005; Ambach et al., 2008b), the two further problems inevitably limit the number of presented stimuli and the presentation frequency.

Typically, pHR measured in CIT studies shows a bi- or multiphasic course and a differential response between item types over a period of 15 seconds after stimulus presentation (Gamer et al., 2006; Ambach et al., 2008a). RLL is usually scored over 15 or 10 seconds (Timm (1982a; 1982b); Gamer et al., 2006); for FPWL, a typical scoring window is 15 seconds (Hirota et al., 2003; Elaad and Ben-Shakhar, 2006). For these measures, shorter scoring intervals e.g. 3-4 seconds after stimulus onset have to be used. But nevertheless, overlapping influences of preceding trials are likely to increase error variance; additionally, omitting relevant parts of the scoring interval and eliciting responses with an untypically high frequency are likely to reduce effect sizes. In sum, the incremental validity of autonomic measures in a CIT with short ISIs, primarily designed for ERP measurement, is unproven so far.

Simultaneous EDA and EEG measurement has been studied in the field of orienting theory (Elton et al., 1983; Lyytinen et al., 1987; Bahramali et al., 1997; Lim et al., 1999); methodological refinements were introduced particularly in experiments with short ISIs (Barry, 1993; Lazzaro et al., 1999; Williams et al., 2002). Only a few recent studies have combined ERP and peripheral physiological measurement in the detection of concealed information (Gamer and Berti, 2009; Matsuda et al., 2009; conference abstract by Meijer et al., 2006). They had different focuses but they all found incremental validity from combined measurement. With the exception of Meijer and colleagues, they used long ISIs, which favor peripheral measurement rather than ERPs. Interestingly, these studies shed light on the different CIT sub-processes reflected by central nervous and peripheral measures. Gamer and Berti (2009) did not find the typical *probe*-vs.-*irrelevant* difference for P300; they discussed their relatively long ISIs as a possible reason and suggested that further studies should investigate a combined measurement with the use of shorter ISIs. To our knowledge, the present study is the first to investigate multimodal measurement in a CIT designed for ERP measurement.

Verbal and pictorial stimuli

Presentation modalities vary between studies of deception detection. While the combination of a scenario-like encoding procedure (e.g. a mock crime) with real objects and a text-based interrogation is quite common in a CIT (e.g. Jokinen et al., 2006), fewer authors used pictures of objects as stimuli (e.g. Lefebvre et al., 2007; Verschuere et al., 2004); most ERP studies in this field used written words (e.g. Rosenfeld et al., 2006). Several studies investigated the effect of verbal vs. pictorial stimuli on the accuracy of the autonomic-based CIT. While some authors found a better detectability of concealed information with verbal than pictorial stimuli using EDA (Ben-Shakhar and Gati, 1987; Gati and Ben-Shakhar, 1990), later studies did not confirm a difference between modalities for EDA (Ben-Shakhar et al., 1996), pHR

(Rill et al., 2003), or reaction times (Seymour and Kerlin, 2007). It may be questioned whether these two stimulus types are differently effective in CIT based on ERP measurement. The recent studies with combined ERP and peripheral measurement used either visual objects in a card test (Gamer and Berti, 2009) or spoken numbers (Matsuda et al., 2009), which hampers a direct comparison. In the present study, we were interested in a possible influence of pictures vs. written words as stimuli in a mock-crime study using a CIT with combined measurement.

Aim of the present study

The present study was designed to investigate the incremental value of peripheral physiological measures in addition to P300 measurement in a CIT. The number of presented items and the ISI were primarily chosen with respect to the elicited EEG responses, including their scoring and analysis. A two-class paradigm, including only *probe* and *irrelevant* items allowed to abandon the demand of verbal or motor answers.

As a secondary question, the aim was to compare the effects of verbal and pictorial stimuli both presented visually, on the multimodal measures recorded in a CIT. A variant of the mock-crime scenario accomplished the parallel encoding of specific knowledge as written words and as viewed real objects.

Our expectations were:

1. The P300 component of event-related potentials as well as electrodermal activity (EDA), phasic heart rate (pHR), respiration line length (RLL), and finger pulse waveform length (FPWL) should differentiate between *probe* and *irrelevant* items in both presentation modalities.

2. The combination of the P300 and the autonomic measures was expected to yield incremental information and enhance CIT accuracy over sole P300 measurement. However, due to the study design directed to ERP measurement, the effect sizes for the autonomic measures were expected to stay below those obtained from their sole measurement. We questioned which of the two opposing effects would prevail.
3. We questioned whether verbal or pictorial stimulus material would lead to a better detection of concealed information considering each of the measures.

Materials and Methods

Subjects

Thirty-one healthy students of various faculties (11 m, 20 f; mean age 25.0 ± 4.6 years; 29 right-handed) voluntarily participated in the study. They were paid 17 Euros, with an additional incentive of 3 Euros. Informed consent was obtained from all subjects.

Mock-crime scenario

In a shielded room, subjects had to handle nine objects according to a task list. For each participant, one particular object had been randomly drawn from each of the nine object categories. The written instruction did not suggest an association with criminal actions. In order to present all objects also as written words, each object was mentioned four times within the text of the instruction list; the pertaining term was emphasized by bold type. Subjects took one object after the other into their hands, looked at it in detail, and deposited it on top of a storage rack. Then they packed all nine objects attentively into a suitcase placed in the same room. Object categories each comprising four objects were household items, key pendants, boxes, office material, cosmetics, wooden toy fruits, drink packages, sweets, and pieces of clothing. Finally, subjects came across a small box containing three Euros. They handed the money together with the filled suitcase to the experimenter who kept it until the end of the experiment. Participants were told that they would receive the money at the end of the experiment provided they performed well in the task they would face in the second part of the experiment. They were not informed about details of the task and no specification of 'performing well' was made.

Concealed Information Test

A variant of the CIT which consisted of nine item categories each comprising four item presentations was used. Either item pictures or written words were presented foveally on a 19’’ monitor at a distance of 90 cm; picture size was 6.0° by 8.0° of visual angle. Onset asynchrony was 3.0–3.5 seconds; each picture was presented for 800 milliseconds and followed by a blank screen.

Each of the *probe* items was combined with three *irrelevant* items which belonged to the same item category (e.g. key pendants, cosmetics); these had some categorical similarity but were all unknown to the subjects. The nine categories were presented in succession. This was repeated six times, which resulted in a total of 216 item presentations per run; the total experiment consisted of two runs with a five-minute break. Sequence of categories and order of items within each category were counterbalanced.

Subjects were assigned randomly to one of two groups; all of them performed the same mock-crime scenario, yet they viewed the items in the CIT in different modalities: Sixteen subjects, assigned to the 'word' group, saw the CIT items as written words, identical to the terms from the mock-crime instruction list. The other fifteen participants, assigned to the 'picture' group, saw the mock-crime items as pictures in the CIT. Subjects from both groups had pairwise identical item sets in the mock crime and identical item series in the CIT. Presentation modalities are illustrated in figure 1.

Figure 1. Item presentation in the Concealed Information Test in the 'word' group (left) and in the 'picture' group (right).

Subjects were instructed to attend to all presented items, pictures, or words and to hide the knowledge obtained from the mock-crime scenario. They were also instructed to take care that their physiological reactions and their brain waves would not reveal them. An active response was not required. No question was combined with the presentations. Subjects were told that inattentiveness would be detected in the measures and would lead to the loss of the three Euro reward.

Procedure

In order to help subjects make a clear distinction between the mock crime and the CIT, the two parts of the experiment were guided by two experimenters. After subjects had given their informed consent, the first experimenter led them to the experimental room; there, the second experimenter informed them about the subsequent mock-crime scenario. After the mock-crime procedure, subjects were led back to the first experimenter who performed the 'psychophysiological investigation'. EEG and EOG electrodes were attached and the leads for the peripheral measures were hooked up. The CIT was then initiated; the two main runs were preceded by a short training run consisting of three sets each comprising four *neutral* items (without relevance). After completing both main runs of the CIT and being disconnected from the leads, subjects performed a memory test. In this test, all four pictures of each category, one item category after the other, were simultaneously presented on the screen and subjects were asked to identify the *probe* item within each of the nine categories.

Physiological measurement

The physiological recordings took place in a dimly lit, electrically and acoustically shielded experimental chamber (*Industrial Acoustics GmbH*, Niederkrüchten, Germany). Subjects sat in an upright position so that they could comfortably see the monitor. Temperature in the cabin was set at 21℃ at the beginning of the first run, with an increase of maximum 2℃ throughout the course of the experiment.

EEG data were recorded from nineteen electrodes according to the standard 10-20 positions (Jasper, 1958), against linked mastoids as reference. Impedances were kept below 5 kΩ; Ag-AgCl electrodes were attached to the skin with the use of Easy Caps and abrasive gel (both by Falk Minow, München, Germany). Vertical and horizontal oculogram was recorded. A Synamps amplifier (NeuroScan, Inc.) was used; EEG was band-pass filtered at 0.15–70 Hz, A/D-converted, sampled at 500 Hz and saved to disc using the software Acquire (Version 4.3.1; Neurosoft, Inc.).

Skin conductance (EDA), respiratory activity, electrocardiogram (ECG), and finger plethysmogram were registered. Physiological measures were A/D-converted and logged by the *Physiological Data System I 410-BCS* manufactured by *J&J engineering (Poulsbo, Washington).* The A/D-converting resolution was 14 bit, allowing skin conductance to be measured with a resolution of 0.01 µS. All data were sampled with 510 Hz. Triggers indicating stimulus onsets were registered with the same sampling frequency.

For skin-conductance recordings, standard Ag/AgCl electrodes (*Hellige;* diameter 0.8 cm), isotonic signa electrode creme (*Parker Laboratories Inc.*) and a constant voltage of 0.5 volts were used. The electrodes were fixed at thenar and hypothenar sites of the nondominant hand.

For registration of thoracic and abdominal respiratory activity, two PS-2 biofeedback respiration sensor belts (*KarmaMatters,* Berkeley, California) with a built-in length-dependent electrical resistance were used. They were fixed at the upper thorax and the abdomen.

ECG was measured with *Hellige* electrodes (diameter 1.3 cm) according to Einthoven II.

Finger pulse signal was transmitted by an infrared system in a cuff around the middle finger of the nondominant hand.

Data processing

EEG data from one subject were lost due to a technical failure. Ocular artifacts were eliminated using a rejection criterion of 80µV in the HEOG and / or VEOG. Artifact-free segments of 1.7 seconds length beginning 100 ms before stimulus onset were the data base for the statistical analyses. Data were filtered with a low cutoff of 0.3 Hz (12 dB / oct.) and a high cutoff of 6 Hz (12 dB / oct.).

EDA data from three subjects were discarded from analysis because of complete non-responding. Skin conductance reactions were assessed via a computerized method based on the decomposition of overlapping reactions as proposed by Lim et al. (1997). This method was chosen because subsequent physiological reactions were expected to overlap considerably due to the short ISIs. Lim (1999) showed in an experiment with short ISIs that conventional trough-to-peak evaluation is inadequate with short ISIs, because the first of two subsequent reactions causes a diminishing bias in the estimation of the second one. The size of this bias is determined by the size of the first reaction and by the time interval between both reactions. Decomposition aims at overcoming this problem of overlapping EDA reactions. The decomposing algorithm employed in the present study has initially been used by Ambach et al. (2008b). In a curve-fitting procedure guided by minimizing error squares, model coefficients are first optimized for each subject. Then, in a second step, all trials are evaluated by decomposing EDA using each subject's individual model coefficients. The use of individually constant templates for the calculation of EDA responses reflects the assumption of individual response characteristics of the electrodermal system. Third, magnitudes of all EDA responses elicited within a time window of 0.5 to 3.0 seconds after item presentation are additively combined. As the decomposed responses are not calculated as trough-to-peak amplitudes but rather as relative activity, the unit of these measures is arbitrary and depends on each subject's individual template. (Transformation into µS is not required because of subsequent within-subject standardization.)

Respiratory data of two subjects had to be discarded from evaluation because of sensor artifacts. After low-pass filtering, the total RLL was automatically computed over a time interval of 3.5 seconds after trial onset; data from the thoracic and the abdominal channel were averaged. The RLL measure integrates information about frequency and depth of respiration. The method was derived from Timm (1982a; 1982b); in the present study, a shorter time window was chosen due to short ISIs[1].

ECG data obtained from two subjects had to be excluded from analysis because of frequent extrasystoles. After notch filtering at 50 Hz, R-wave peaks were automatically detected and visually controlled. The R-R intervals were transformed into heart rate (HR) and real-time scaled (Velden and Wölk, 1987). The HR during the last second before trial onset served as pre-stimulus baseline. The pHR was calculated by subtracting this baseline value from each second-per-second post-stimulus value. For extracting the trial-wise information of the phasic HR, the mean change in HR within 4 seconds after trial onset, compared with the pre-stimulus baseline, was calculated (Bradley and Janisse, 1981) [1].

[1] The shorter time window for RLL, pHR, and FPWL was chosen in order to avoid overlaps - although effect sizes are expected to decline when shortening the scoring window.

Finger pulse waveform data from three subjects were excluded because of insufficient signal transmission of the sensor. The finger pulse waveform length (FPWL) within the first 3.5 seconds after trial onset was calculated and subjected to further analyses. The FPWL comprises information about both HR and pulse amplitude[1].

A within-subject standardization of measured values has been proposed by Lykken and Venables (1971). According to Ben-Shakhar (1985), Gronau et al. (2005), and Gamer et al. (2006), the physiological and behavioral measures are z-transformed for each subject, each run, and each data channel. All *probe* and *irrelevant* trials of one run of the experiment (not including the first trials of each stimulus category) were used to calculate individual means and standard deviations (Ambach et al., 2008a). The z-transformed values were used in subsequent statistical analyses.

Statistical analysis

For the analysis of EEG signals on an individual basis, the bootstrapped amplitude difference (BAD) method was applied on the artefact-free *probe* and *irrelevant* EEG segments (Rosenfeld et al., 2004; Abootalebi et al., 2006). This non-parametrical method calculates the confidence intervals for the *probe* and *irrelevant* P300 for each subject and each EEG channel and allows to assess the statistical significance of the individual *probe*-vs.-*irrelevant* difference. Correct classification of a subject as knowledgeable was based on a significance criterion of $p < 0.1$. While the results from all EEG channels were visualized, the statistical analysis was limited to channel Pz, where P300 is typically largest. P300 amplitude was determined as a peak-to-peak measure (Soskins et al., 2001): in detail, it was defined as the difference between the maximally positive segment average of 100 ms within a window from 400 to 1000 ms post-stimulus and the maximum negative 100-ms segment in the window bounded by P300 latency and 1600-ms post-stimulus (Soskins et al., 2001). In addition to the bootstrap method and in order to integrate the P300 into the multimodal evaluation, P300 amplitudes were determined for each subject's mean ERPs, according to the above scoring method, separately for *probe* and *irrelevant* trials. These P300 amplitude values were subjected to the statistical procedure described below.

All further statistical analyses were performed with *SPSS*, Version 12.0 (*SPSS Inc.*, Chicago). For each physiological and behavioral measure, a 2 x 2 x 2 ANOVA was first calculated with the within-subject factors Probe (*probe* vs. *irrelevant* trials) and Run (first vs. second run) and the between-subjects factor Group ('word' vs. 'picture' group); the factor Group was later dropped from the analyses. Because a mean difference between the two instructional groups was not of interest, ANOVAs for the peripheral measures were calculated on the basis of z-values from a within-run standardization; consequently, no main effects for the factors Group and Run are reported. Significance level for the assessment of main and interaction effects was set to 0.05. Follow-up t-tests for matched samples (one-tailed, significance level 0.05) were carried out when interactions were found. Cohen's d was calculated as estimate of effect size (Cohen, 1988; Rosnow and Rosenthal, 1996).

For an integration of the multimodal measures, the incremental validity of each measure and the validity of an optimized combination of the central nervous with the peripheral physiological measures were additionally evaluated using a binary logistic regression model. This procedure adopted from applied psychophysiology aimed at identifying the contribution of the various measures included as predictors in the model to classify subjects as '*guilty*' or '*innocent*'. Since all subjects had deed-related knowledge, hypothetical distributions of responses to *probe* and *irrelevant* stimuli for an equal number of '*innocent*' participants were calculated (Carmel et al., 2003; Allen and Iacono, 1997; Meijer et al., 2007). For this purpose, the numeric results obtained from *probe* trials were replaced by numeric results drawn randomly (with replacement) from the results obtained from the same subject's *irrelevant* trials. First, a stepwise inclusion of parameters was performed following a Wald statistic with probabilities for inclusion and exclusion set to 0.05 and 0.10, respectively. Second, in order to estimate the predictive value of each measure per se, they were included singularly in an analogous regression analysis. Shrinkage correction of the regression model was performed using the hold-one-out method (Copas and Corbet, 2002; Gamer et al, 2006); each subject was once held out of the calculation, resulting in 32 different logistic regression models for the 'word' group and 30 for the 'picture' group. The calculation of sensitivity and specificity was then based on the data from the held-out subjects. According to the equal numbers of *guilty* and hypothetical *innocent* subjects, subjects were first classified as *guilty* or *innocent* using a $p < 0.5$ criterion. The ROC curve reflects the '*guilty*'-vs.-'*innocent*' differentiation capability of the single measures and their optimal combination for all possible cutoff-points; the area under the ROC curve, varying between 0 and 1 with a chance level of 0.5, serves as an overall index of detection accuracy (Bamber, 1975; Ben-Shakhar and Elaad, 2003; Gronau et al., 2005).

Results

In the final memory test, 99.3% of *probe* items were identified correctly (99.3% in the 'word' and 99.3% in the 'picture' group); categories with false identification were discarded from evaluation.

Table 1 summarizes means and standard errors of means of raw scores for each data channel, collapsed over both runs of the experiment.

	'word' group				'picture' group				over groups			
	probe items		*irrelevant items*		*probe* items		*irrelevant items*		*probe* items		*irrelevant items*	
	Mean	(SEM)	Mean	(SEM)	Mean	(SEM)	Mean	(SEM)	Mean	(SEM)	Mean	(SEM)
P300 [µV]	11.57	0.53	10.22	0.59	11.73	0.61	10.06	0.42	11.65	0.40	10.14	0.37
EDA [nS]	156.32	38.26	146.12	37.57	114.32	30.32	97.50	26.70	135.32	24.10	121.81	23.03
RLL [arb. units]	1021.2	128.04	1047.6	134.2	1236.2	153.0	1259.6	159.3	1128.7	98.7	1153.6	104.2
pHR [1/min]	-0.32	0.149	0.13	0.077	-0.34	0.217	0.15	0.10	-0.33	0.14	0.14	0.06
FPWL [arb. units]	36592	10167	37055	10186	34426	4503	33996	4494	35509	5311	35526	5193

Table 1. Means and standard errors of means (SEM) of raw scores for each data channel. *Probe* vs. *irrelevant* items are contrasted separately for both groups and over groups.

Figure 2 illustrates the differential response to *probe* vs. *irrelevant* items for the 'word' and the 'picture' condition. Estimated effect sizes are depicted for each of the physiological measures.

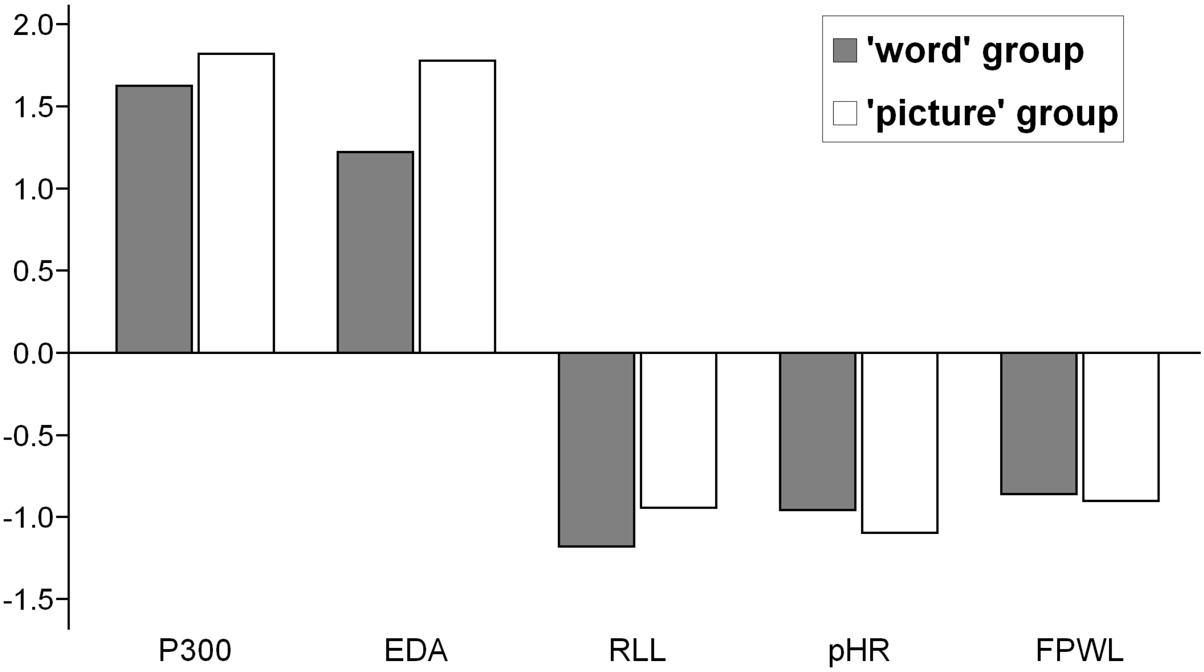

Figure 2. Effect sizes for the differential responses to *probe* vs. *irrelevant* items: for the 'word' and the 'picture' group, Cohen's d is depicted for P300 amplitude, electrodermal activity (EDA), phasic heart rate (pHR), respiration line length (RLL), and finger pulse waveform length (FPWL).

Event-related potentials

Figure 3 shows the EEG grand means at electrode site Pz for *probe* and *irrelevant* items separately for the 'word' and the 'picture' group. Regarding the course of ERPs, it might be suspected that verbal stimuli led to a differentiation between *probe* and *irrelevant* stimuli that occurred earlier within a trial, i.e. after about 250 ms for the verbal vs. 300 ms for the pictorial condition. With respect to trial-wise response scoring however, the conditions did not differ in significance or size of the effect. An evaluation of the sub-components of the P300, or ERP components other than the P300, was not subject of our hypotheses.

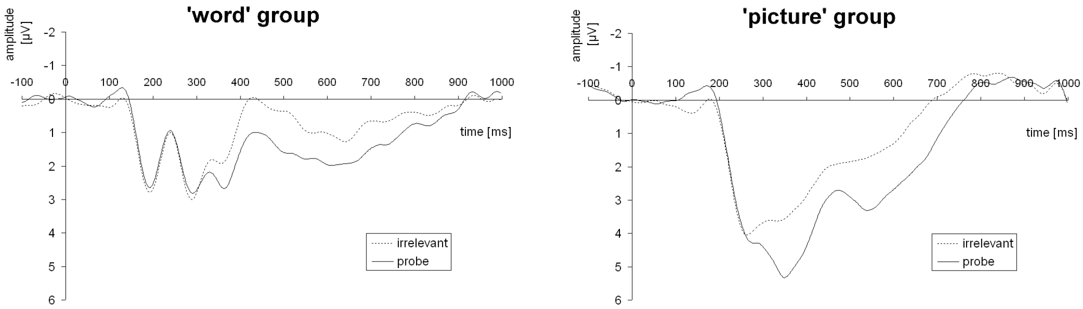

Figure 3. Grand means of event-related potentials for *probe* trials (solid lines) and *irrelevant* trials (dashed lines). Figure 3a: 'word' group; Figure 3b: 'picture' group.

For both conditions, grand means of the event-related potentials at the electrode Pz showed a larger positivity for *probe* than for *irrelevant* trials within a time window from 300 to 800 ms.

All statistics, which will be reported hereafter, were first performed with Group as a between-subjects factor ("word" group vs. "picture group"); a t-test was performed to compare the bootstrap results for P300 from both groups. Since it emerged that for none of the measures a Group main effect or interaction with either Probe or Run was found, this factor was dropped from the analyses and all participants were thereafter analyzed as one group. As the ANOVA was based on within-run standardized data, a possible main effect for Run is outside of scope.

The BAD procedure, which was applied to assess the significance of the peak-to-peak amplitude difference between *probe* and *irrelevant* trials for each subject, revealed a significantly positive amplitude difference ($p < 0.1$) in 22 of 30 subjects. Assuming the mutual independence of the individual bootstrap results, the individual significance levels can be combined towards an overall assessment of significance via z-values (originally from Stouffer et al., 1949). The significance levels p obtained for each subject were transformed into z-values by the inverse Gaussian function. Overall assessment of the significance of differential responding to *probe* vs.

irrelevant stimuli for the entire study was then accomplished by collapsing the z-values over subjects according to $z_{overall} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i$ where $z_{overall}$ reflects the overall significance for the study in terms of a z-standardized value, n is the number of participants, and z_i is the z-value for the i-th subject. The resulting overall z-value was 7.61, which corresponds with a significance level of $p < 0.001$.

ANOVA for P300 amplitude showed a main effect for Probe ($F_{1,29} = 24.83$; $p < 0.001$; $d = 1.69$) with the larger mean amplitude for *probe* than for *irrelevant* trials. No interaction Run by Probe ($F_{1,29} = 2.13$) was found.

Skin conductance

After decomposition and z-standardization of EDA responses, ANOVA showed a main effect for Probe ($F_{1,27} = 16.38$; $p < 0.001$; $d = 1.55$) and a Probe by Run interaction ($F_{1,27} = 16.46$; $p < 0.001$); reactions to *probe* items were greater than to *irrelevant* items in the first but not in the second run ($t_{27} = 4.32$; $p < 0.001$; $d = 1.08$ and $t_{27} = 0.81$; $p > 0.1$, respectively)[2].

[2] In an additional analysis, EDA effect sizes and classification rates were determined separately for each run. Although the second run, per se, showed a poor item-type differentiation, analyzing just the first run did not outperform the analysis of both runs together.

Respiration

ANOVA for RLL data showed a main effect for Probe ($F_{1,28} = 7.93$; $p < 0.01$; $d = 1.06$). No interaction Run by Probe ($F_{1,28} = 2.13$) was found. Respiration was lower in *probe* than in *irrelevant* trials.

Heart rate

Second-per-second values of HR after trial onset for probe and irrelevant items are depicted for both experimental groups in Figure 4. A decelerative HR response was seen in *probe* but not in *irrelevant* trials; the patterns for both groups were visually identical.

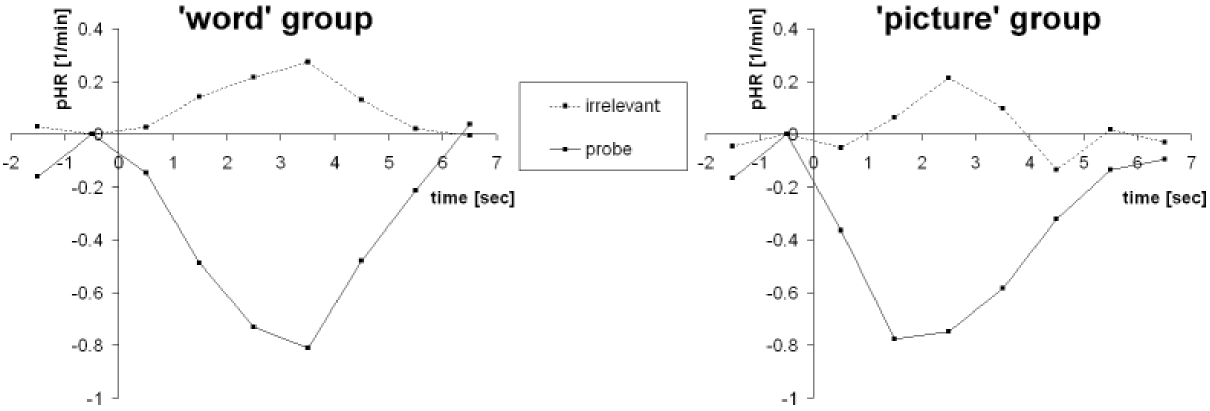

Figure 4. Second-per-second values of heart rate following trial onsets for *probe* trials (solid lines) and *irrelevant* trials (dashed lines). Figure 4a: 'word' group; Figure 4b: 'picture' group.

ANOVA for pHR data showed a main effect for Probe ($F_{1,28} = 7.83$; $p < 0.01$; $d = 1.04$) and a Probe by Run interaction ($F_{1,28} = 5.58$; $p < 0.05$). HR decelerations were enhanced after *probe* items compared with *irrelevant* items; this effect was more pronounced in the first than in the second run ($t_{28} = -4.73$; $p < 0.001$; $d = -1.16$ and $t_{28} = -0.91$; $p > 0.1$, respectively).

Finger pulse

ANOVA for FPWL revealed a main effect for Probe ($F_{1,27} = 4.87$; $p < 0.05$; $d = 0.90$); FPWL values were lower in *probe* than in *irrelevant* trials. No interaction Probe by Run ($F_{1,27} = 0.42$) was found.

Logistic regression model

The binary logistic regression and the following results refer to a discrimination between *guilty* and hypothetical *innocent* subjects using P300, EDA, pHR, RLL, and FPWL as predictors. In the first analysis, only P300 was included. The second analysis used a stepwise inclusion following the Wald statistic; this led to the inclusion of P300 and EDA only; none of the further measures pHR, RLL, and FPWL yielded incremental validity. Optimal weights, standard errors, Wald statistic, and significance levels for both analyses are summarized in table 2.

measure	**β**	**SE**	**Wald**	***p***
only P300				
P300	5.490	1.880	8.53	0.004
constant	-0.338	0.329	1.06	0.304
stepwise				
P300	7.273	2.378	9.35	0.002
EDA	10.938	3.630	9.08	0.003
constant	-1.016	0.438	5.38	0.204

Table 2. Physiological measures included in the binary logistic regression analyses: optimal weights (β), standard errors (SE) of β, Wald statistic and significance level (*p*) are listed for the analyses with inclusion of only P300 amplitude (P300) and with stepwise inclusion (Wald statistic) of P300 and electrodermal activity (EDA).

Correct-classification rates (criterion: $p > 0.5$ for classification as *guilty*, according to the percentage of guilty subjects) with P300 were 70.4% for guilty subjects and 73.1% for the hypothetical innocents (shrinkage-corrected using the hold-one-out method), whereas the combination of P300 and EDA, which were included by the Wald statistic, classified 75% of guilty and 81.5% of innocent subjects correctly. Rates of false-positive (classification of an *innocent* subject as *guilty*) and false-negative outcomes (classification of a *guilty* subject as *innocent*), obtained under variation of the cut-off point for decision can be summarized in ROC curves. Areas under these curves, reflecting overall accuracy of subject classification, were 0.738 (0.613 - 0.863) with only P300, and 0.829 (0.715 - 0.943) when combining P300 and EDA. An additional analysis taking the correlation between paired ROC areas into account was performed in order to assess the significance of the difference between ROC areas (Hanley and McNeil, 1983). The analysis exactly following the procedure described by Hanley and McNeil (1983) resulted in a z value of 1.89, corresponding to a significance level of $p = 0.029$. Thus, the incremental validity of EDA over P300 is reflected in the results of the Wald statistic as well as in the difference between ROC areas.

Discussion

The presented CIT study is regarded as a fundamental step contributing to the methodology used in the detection of concealed information and to the integration of different existing approaches. Recent publications underline the interest in studying the combination of EEG and autonomic measurement in deception detection (Gamer et al., 2009; Matsuda et al., 2009). The study focused on investigating the incremental validity of peripheral physiological measurement over ERPs in a CIT with combined measurement. As a secondary issue, the modality of CIT stimuli was varied; either pictures of objects or written words were presented.

Combination of EEG and autonomic measurement

Brain electrical measurement as well as each of the autonomic measures revealed a significant difference between *probe* and *irrelevant* items and thus turned out to be sensitive to concealed information. This seems particularly noteworthy against the background of the short ISIs employed which are untypical with autonomic measurement.

The modality of stimulus presentation in the CIT, i.e. object pictures vs. written words, did not influence the physiological response differences between *probe* and *irrelevant* stimuli; this holds for each of the measures recorded; data from all participants were further analyzed without respect to presentation modality.

Among the dependent measures, P300 peak-to-peak amplitude yielded the greatest overall effect size and was the first predictor included in the regression analysis. A comparison among the single measures with respect to their validity as a single predictor was not the aim of the study; additionally, such a comparison would not be very meaningful since the study design and parameters cannot be chosen optimally for all measures simultaneously.

The main question of the study was whether the autonomic measures would yield incremental information over ERPs. This question was positively affirmed with the help of two regression analyses. Following P300, EDA was included in the stepwise regression analysis as a significant second predictor. This also led to visibly higher correct-classification rates in comparison to sole P300; the corresponding increase in the area under the ROC was significant. Gamer et al. (2009) argue that the finding of incremental validity indicates that the central and the autonomic nervous system are differentially responsive to concealed information. While we confirmed an incremental validity of EDA over ERPs, it is still an open question whether this finding is due to the reflection of different sub-processes of the CIT or just to a reduction of error variance by additional data.

Small effect sizes

Looking at the measures in detail reveals that the effect sizes of each of the measures as well as the correct-classification rates in this study remained below those of most CIT studies; this holds for a comparison with ERP-based as well as with autonomic-based studies.

Several reasons might account for this:

<u>With respect to the autonomic measurement</u>, the use of short ISIs and the presentation of a great number of items is supposed to have diminished the physiological responses as well as item-type effects. Electrodermal as well as cardiovascular and respiratory responses are known to accept a decrease in their phasic amplitudes as well as in their *probe*-vs.-*irrelevant* effect sizes when they are elicited with a higher stimulus frequency. The relatively small raw scores of the

phasic EDA and pHR responses are in line with this; the EDA raw amplitudes are comparable to those from other studies with short ISIs (e.g. Barry et al., 1993) but far smaller than in most autonomic-based CIT studies (e.g. Ambach et al., 2008b; for an overview, see Ben-Shakhar and Elaad, 2003). On the level of data analysis, the inevitable use of uncommonly short scoring intervals, e.g. four seconds for the phasic cardiovascular and respiratory responses, adds to this problem. In addition, habituation effects might be more pronounced than in studies with a smaller number of repetitions of each single stimulus. The observed decrease of the differential (*probe* vs. *irrelevant*) EDA response from the first to the second run provides an impressive evidence of this.

For ERPs, the large number of trials led to a sufficient number of epochs to be averaged and thereby, meaningful effect sizes and correct-classification rates. Yet a comparison with other ERP-based CIT studies (e.g. Rosenfeld et al., 2006) revealed a decreased mean difference of P300 amplitudes between *probe* and *irrelevant* trials and a smaller effect size. One possible reason for this might be the use of a two-class paradigm without target stimuli. Rosenfeld et al. (2006) found this paradigm equivalent from a statistical perspective, yet it might still be questioned whether the omission of targets might have affected the subjects' attention. Yet, whereas Rosenfeld et al. (2006) gave special instructions (announcing unpredictable interruptions in the experiment asking for the name of the last-shown object) in order to maintain the subjects' attention, no such instructions were given in the present study. Hence, the subjects' attention was not fully guaranteed, which is most likely the reason for the rather small effect sizes observed. Furthermore, the omission of a deceptive answer is known to reduce CIT accuracy in the classic, autonomic-based CIT (Furedy and Ben-Shakhar, 1991) as well as in the ERP-based CIT (Verschuere et al., 2009a). One might suspect that the complete omission of answers and the 'quiet' course of the experiment, which were chosen in order to reduce motor artifacts, might have reduced the subjects' attention, motivation and involvement in the experiment even more and might thereby have diminished not only autonomic but also ERP responses. The omission of a criminal context in the mock-crime instruction might have contributed to this.

While it was not surprising to find smaller effect sizes than studies, which used either brain electrical or autonomic measurement singularly, the question of interest was whether such a reduction of effects in the single measures would be outbalanced by the incremental information obtained from the multimodal measurement. As a common procedure in the research on the detection of concealed information, logistic regression analysis was employed to estimate the incremental information of the peripheral measures over ERPs with respect to subject classification as 'guilty' or 'innocent'. While EDA responses yielded incremental validity over ERPs, absolute and differential response amplitudes were visibly lower than in other studies using unimodal measurement; the same holds for the overall correct-classification rates. In sum, the incremental information from the multimodal measurement did not outbalance the drawbacks arising from the trade-off between the adverse requirements from the single measures.

In a CIT study with combined measurement and ISIs of 7 to 9 seconds, Gamer and Berti (2009) did not find a significant P300 amplitude difference between honest denial of knowledge and its concealment; they argue, P300 is sensitive to task relevance rather than to encoded information. In the present study, P300 was found sensitive to the information encoded during the mock crime; this seems notable since no verbal or motor response was required from the subjects. As outlined, this difference might be due to the longer ISIs used which might have reduced differential responding; yet, it might still be argued that P300 depends on task relevance, since in the present study, the knowledge from the mock-crime was linked to a specific instruction, namely to hide the knowledge silently and not to let anything show.

In a further CIT study with combined measurement, Matsuda et al. (2009) presented acoustic stimuli every 22 seconds. They reported differential responding to concealed information for each of the autonomic measures but not for P300 amplitude, which they discuss as being due to a ceiling effect on the P300 amplitude. In the light of the present results, the frequency (and number) of stimulus presentations in studies with combined measurement determines the relative validity of autonomic measures vs. P300 amplitude. As the different approaches suggest, an optimal experimental trade-off between the requirements from both types of measures, which would entail significant and meaningful results for each of the measures, has not been employed so far; this study appears to mark the short end of possible ISIs.

Verbal vs. pictorial stimuli

While for both stimulus presentation modalities (i.e. verbal and pictorial) significant *probe*-vs.-*irrelevant* effects for each measure were revealed by the analyses of variance, no significant difference in these effects was found between modalities. For P300, the results from the bootstrap method confirmed the equivalence of modalities, as did the analyses of variance for each of the measures. Visual inspection of mean EDA responses suggested a greater (*probe*-vs.-*irrelevant*) effect size for the pictorial condition; this difference was however not significant.

It may further be questioned whether the CIT presentation modalities can be compared adequately using the 'double encoding' procedure employed here. The finding that nearly 100% of subjects in both groups remembered all presented stimuli correctly, suggests a sufficient encoding in both conditions with respect to memorability. Beyond memorability however, an equivalent encoding of both modalities was not proven. Not only the presentation modality as such but also the detailed specifications of the item presentation (or the contact with the item) during the encoding phase might have influenced the physiological responses and the observed effects in the CIT. The fact that modality was varied only in the CIT but not in the mock crime, makes it difficult to generalize the finding that CIT accuracy was independent from stimulus modality. With respect to ERP-based CIT however, the results suggest that the verbal material, which was predominantly used so far, might be replaced with ecologically more valid stimuli without accepting a decrease in CIT accuracy. Further ERP studies might clarify this by independently varying stimulus modalities in the encoding procedure and in the CIT.

Conclusions

The present study found incremental validity from EDA over the ERPs measured simultaneously. Yet, the benefit of incremental information did not outbalance the accepted decrease in effect sizes in each of the singular measures. It is not clear whether the combined measurement, per se, was critical for this decrease or whether specific details of the study caused it. On the one hand, it might well be that not the best trade-off between the partly contrary requirements arising from ERP and peripheral measurements (e.g. number and frequency of stimulus presentations) was chosen for the present experimental design. On the other hand, it is possible that specific details of the mock-crime (e.g. the absence of a criminal context, or the double encoding procedure) led to a diminished involvement of the participants; possibly, the omission of answers in the CIT was also disadvantageous.

Two different consequences can be drawn with respect to further studies: Firstly, stimulus presentation should be modified towards slightly longer ISIs in order to entail greater autonomic responses and to allow for longer scoring intervals. Secondly, a two-part CIT study could be performed with sequential instead of simultaneous recordings, analogously to a study investigating the RT-based and the classic CIT separately (Verschuere et al., 2009b). Each part should be optimized with respect to the measures recorded; stimulus number, stimulus frequency, and answer demands should be determined with respect to the requirements from either EEG or peripheral measurement.

As a concluding suggestion with respect to the increasing number of fMRI studies in the research on deception detection, a combined measurement including brain imaging, and the experimental demands arising from it, might be studied in future deception-detection projects.

References

Abootalebi, V., Moradi, M.H., Khalilzadeh, M.A., 2006. A comparison of methods for ERP assessment in a P300-based GKT. Int. J. Psychophysiol. 62, 309-320.

Alexander, D.M., Tengrove, D., Johnston, P., 2005. Separating individual skin conductance responses in a short interstimulus-interval paradigm. J. Neurosci. Methods 146, 116 - 123.

Allen, J.J.B., Iacono, W.G., 1997. A comparison of methods for the analysis of event-related potentials in deception detection. Psychophysiology 34, 234-240.

Ambach, W., Stark, R., Peper, M., Vaitl, D., 2008. An interfering Go/No-go task does not affect accuracy in a Concealed Information Test. Int. J. Psychophysiol. 68, 6-16.

Ambach, W., Stark, R., Peper, M., Vaitl, D., 2008. Separating deceptive and orienting components in a Concealed Information Test. Int. J. Psychophysiol. 70, 95-104.

Bahramali, H., Gordon, E., Lim, C.L., Li, W., Lagopoulos, J., Leslie, J., Rennie, C., Meares, R.A., 1997. Evoked related potentials associated with and without an orienting reflex. Neuroreport 8, 2665-2669.

Bahramali, H., Lim, L.C., Rennie, C., Meares, R., Gordon, E., 2001. ERPs associated with and without an "orienting reflex" in patients with schizophrenia. Int. J. Neurosci. 108, 163.

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. 12, 387-415.

Barry, R.J., Maltzman, I., 1985. Heart rate deceleration is not an orienting reflex; heart rate acceleration is not a defensive reflex. Pavlov. J. Biol. Sci. 20, 15-28.

Barry, R.J., Feldmann, S., Gordon, E., Cocker, K.I., 1993. Elicitation and habituation of the electrodermal orienting response in a short interstimulus interval paradigm. Int. J. Psychophysiol. 15, 247-253.

Barry, R.J., 1996. Preliminary process theory: towards an integrated account of the psychophysiology of cognitive processes. Acta. Neurobiol. Exp. 56, 469-484.

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

Ben-Shakhar, G., Frost, R., Gati, I., Kresh, Y., 1996. Is an apple a fruit? Semantic relatedness as reflected by psychophysiological responsivity. Psychophysiology 33, 671-679.

Ben-Shakhar, G., Elaad, E., 2003. The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. J. Appl. Psychol. 88, 131-151.

Bradley, M.T., Janisse, M.P., 1981. Accuracy demonstrations, threat, and the detection of deception: cardiovascular, electrodermal, and pupillary measures. Psychophysiology 18, 307-315.

Bradley, M.T., Rettinger, J., 1992. Awareness of crime-relevant information and the Guilty Knowledge Test. J. Appl. Psychol. 77, 55-59.

Brainerd, C.J., Reyna, V.F., 1990. Gist is the grist: fuzzy-trace theory and the new intuitionism. Dev. Rev. 10, 3-47.

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. J. Exp. Psychol. Appl. 9, 261-269.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences, 2nd edn., Lawrence Earlbaum Associates, Hillsdale, NJ.

Copas, J.B., Corbett, P., 2002. Overestimation of the receiver operating characteristic curve for logistic regression. Biometrika 89, 315-331.

Donchin, E., Coles, M.G.H., 1988. Is the P300 component a manifestation of context updating? Behav. Brain. Sci. 11, 355-425.

Duncan-Johnson, C.C., Donchin, E., 1977. On quantifying surprise: the variation of event-related potentials with subjective probability. Psychophysiology 14, 456-467.

Duncan-Johnson, C.C., Donchin, E., 1982. The P300 component of the event-related brain potential as an index of information processing. Biol. Psychol. 14, 1-52.

Elaad, E., Ben Shakhar, G., 1989. Effects of motivation and verbal response type on psychophysiological detection of information. Psychophysiology 26, 442-451.

Elaad, E., 1998. The challenge of the Concealed Knowledge Polygraph Test. Expert Evid. 6, 161-187.

Elaad, E., Ben-Shakhar, G., 2006. Finger pulse waveform length in the detection of concealed information. Int. J. Psychophysiol. 61, 226-234.

Elton, M., Schandry, R., Sparrer, B., 1983. A comparative investigation of ERP components and the SCR in a habituation and dishabituation paradigm. Int. J. Neurosci. 22, 55-62.

Farwell, L.A., Donchin, E., 1991. The truth will out: interrogative polygraphy ("lie detection") with event-related brain potentials. Psychophysiology 28, 531-547.

Furedy, J.J., 1986. Lie detection as psychophysiological differentiation: some fine lines. In: Coles, M.G.H., Donchin, E. & Porges, S.W. (Eds.), Psychophysiology: Systems, Processes, and Applications -A Handbook. New York: Guilford, pp. 683-701.

Furedy, J.J., Ben Shakhar, G., 1991. The Roles of Deception, Intention to Deceive, and Motivation to Avoid Detection in the Psychophysiological Detection of Guilty Knowledge. Psychophysiology 28, 163-171.

Gamer, M., Berti, S., Vossel, G., 2005. Combined event-related brain potential and skin conductance response parameters during processing of concealed information. J. Psychophysiol. 19, 117.

Gamer, M., Berti, S., 2009. Task relevance and recognition of concealed information have different influence on electrodermal activity and event-related potentials. Psychophysiology, in press.

Gamer, M., Rill, H.G., Vossel, G., Gödert, H.W., 2006. Psychophysiological and vocal measures in the detection of guilty knowledge. Int. J. Psychophysiol. 60, 76-87.

Gamer, M., Bauermann, T., Stoeter, P., Vossel, G., 2007. Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. Hum. Brain Mapp. 28, 1287-1301.

Gamer, M., Gödert, H.W., Keth, A., Rill, H.G., Vossel, G., 2008. Electrodermal and phasic heart rate responses in the Guilty Actions Test: comparing guilty examinees to informed and uninformed innocents. Int. J. Psychophysiol. 69, 61-68.

Gati, I., Ben-Shakhar, G., 1990. Novelty and significance in orientation and habituation: a feature-matching approach. J. Exp. Psychol. Gen. 119, 251-263.

Gronau, N., Ben Shakhar, G., Cohen, A., 2005. Behavioral and physiological measures in the detection of concealed information. J. Appl. Psychol. 90, 147-158.

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.

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. Jpn. J. Physiol. Psychol. Psychophysiol. 21, 217-230.

Jasper, J.J., 1958. The 10/20 international electrode system. Electroencephalogr. Clin. Neurophysiol. 10, 371-375.

Jokinen, A., Santtila, P., Ravaja, N., Puttonen, S., 2006. Salience of guilty knowledge test items affects accuracy in realistic mock crimes. Int. J. Psychophysiol. 62, 175-184.

Lazzaro, I., Gordon, E., Li, W., Lim, C.L., Plahn, M., Whitmont, S., Clarke, S., Barry, R.J., Dosen, A., Meares, R., 1999. Simultaneous EEG and EDA measures in adolescent attention deficit hyperactivity disorder. Int. J. Psychophysiol. 34, 123-134.

Lefebvre, C.D., Marchand, Y., Smith, S.M., Connolly, J.F., 2001. Determining eyewitness identification accuracy using event-related brain potentials (ERPs). Psychophysiology 44, 894-904.

Lim, C.L., Rennie, C., Barry, R.J., Bahramali, H., Lazzaro, I., Manor, B., Gordon, E., 1997. Decomposing skin conductance into tonic and phasic components. Int. J. Psychophysiol. 25, 97-109.

Lim, C.L., 1999. Dynamics of SCR, EEG, and ERP activity in an oddball paradigm with short interstimulus intervals. Psychophysiology 36, 543-551.

Lykken, D.T., 1974. Psychology and the lie detector industry. Am. Psychol. 29, 725-739.

Lykken, D.T., 1959. The GSR in the detection of guilt. J. Appl. Psychol. 43, 385-388.

Lykken, D.T., Venables, P.H., 1971. Direct measurement of skin conductance: a proposal for standardization. Psychophysiology 8, 656-672.

Lyytinen, H., Näätänen, R., 1987. Autonomic and ERP responses to deviant stimuli: analysis of covariation. Electroencephalogr. Clin. Neurophysiol. Suppl. 40, 108-117.

MacLaren, V.V., 2001. A quantitative review of the Guilty Knowledge Test. J. Appl. Psychol. 86, 674-683.

Matsuda, I., Nittono, H., Hirota, A., Ogawa, T., Takasawa, N., 2009. Event-related brain potentials during the standard autonomic-based concealed information test. Int. J. Psychophysiol. 74, 58-68.

McCandliss, B.D., Cohen, L., Dehaene, S., 2003. The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn. Sci. 7, 293-299.

Meijer, E., Smulders, F., Johnston, J., Merckelbach, H., 2007. Combining P300 and forced choice in the detection of concealed information. Psychophysiology 44, 814-822.

Meijer, E., Smulders, F., Merckelbach, H., 2006. Combining P300 and SCR in the detection of concealed information. Psychophysiology 43, 66.

Merzagora, A.C., Bunce, S., Izzetoglu, M., Onaral, B., 2006. Wavelet analysis for EEG feature extraction in deception detection. Conf. Proc. IEEE Eng. Med. Biol. Soc., 2434-2437.

Nuñez, J.M., Casey, B.J., Egner, T., Hare, T., Hirsch, J., 2005. Intentional false responding shares neural substrates with response conflict and cognitive control. Neuroimage 25, 267-277.

Polich, J., 1986. Attention, probability, and task demands as determinants of P300 latency from auditory stimuli. Electroencephalogr. Clin. Neurophysiol. 63, 251-259.

Rill, H.G., Gamer, M., Gödert, H.W., Vossel, G., Ashcroft, K.R., Busuttil, A., 2003. Detecting concealed information using the guilty knowledge test with visual stimuli: a study of the effects of linguistic versus photographic presentation. Abstr. Br. Psychophysiol. Soc. 30th. 105.

Rosenfeld, J.P., Angell, A., Johnson, M., Qian, J.H., 1991. An ERP-based, control-question lie detector analog: algorithms for discriminating effects within individuals' average waveforms. Psychophysiology 28, 319-335.

Rosenfeld, J.P., 2002. Event-related potentials in the detection of deception, malingering, and false memories. In: Kleiner, M. (Ed.), Handbook of Polygraphy. Academic Press, New York, pp. 265-286.

Rosenfeld, J.P., Soskins, M., Bosh, G., Ryan, A., 2004. Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology 41, 205-219.

Rosenfeld, J.P., 2005. "Brain Fingerprinting": a critical analysis. Sci. Rev. Ment. Health Pract. 4, 20-37.

Rosenfeld, J.P., Biroschak, J.R., Furedy, J.J., 2006. P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. Int. J. Psychophysiol. 60, 251-259.

Rosnow, R.L., Rosenthal, R., 1996. Computing contrasts, effect sizes, and counternulls on other people's published data: general procedures for research consumers. Psychol. Methods 1, 331-340.

Rösler, F., Hasselmann, D., Sojka, B., 1987. Central and peripheral correlates of orienting and habituation. Electroencephalogr. Clin. Neurophysiol. Suppl. 40, 366-372.

Schillaci, M., Vendemia, J.M.C., Buzan, R.F., 2004. Modeling deception in a visual two-stimulus paradigm as a two-state system. Conf. Proc. Amer. Phys. Soc. abstract R1.209.

Seymour, T.L., Kerlin, J.R., 2008. Successful detection of verbal and visual concealed knowledge using an RT-based paradigm. Appl. Cogn. Psychol. 22, 475-490.

Soskins, M., Rosenfeld, J.P., Niendamm, T., 2001. Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: complex vs. simple paradigms. Int. J. Psychophysiol. 40, 173-180.

Stouffer, S.A., Suchman, L.C., Devinney, S.A., Star, A., Williams, R.M., 1949. The American Soldier: Adjustment during army life: Studies in social psychology in World War II, Vol. 1. Princeton University Press, Princeton, NJ.

Suzuki, R., Nakayama, M., Furedy, J.J., 2004. specific and reactive sensitivities of skin resistance response and respiratory apnea in a japanese Concealed Information Test (CIT) of criminal guilt. Can. J. Behav. Sci. 36, 202-209.

Timm, H.W., 1982a. Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. J. Appl. Psychol. 67, 391-400.

Timm, H.W., 1982b. Analyzing deception from respiration patterns. J. Police Sci. Adm. 10, 47-51.

Travis, F.T., Orme-Johnson, D.W., 1998. Field model of consciousness: EEG coherence changes as indicators of field effects. Int. J. Neurosci. 49, 203-211.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., Frackowiak, R.S., 1996. Functional anatomy of a common semantic system for words and pictures. Nature 383, 254-256.

Velden, M., Wölk, C., 1987. Depicting cardiac activity over real time: a proposal for standardization. J. Psychophysiol. 1, 173-175.

Verschuere, B., Crombez, G., De Clercq, A., Koster, E.H.W., 2004. Autonomic and behavioral responding to concealed information: differentiating orienting and defensive responses. Psychophysiology 41, 461-466.

Verschuere, B., Crombez, G., De Clercq, 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.W., Van Bockstaele, B., De Clercq, A., 2007. Startling secrets: startle eye blink modulation by concealed crime information. Biol. Psychol. 76, 52-60.

Verschuere , B., Rosenfeld, J.P., Winograd, M., Labkovsky, E., Wiersema, R., 2009a. The role of deception in P300 memory detection. Legal Criminol. Psych. 14, 253-262.

Verschuere, B., Crombez, G., Degrootte, T., Rosseel, Y., 2009b. Detecting concealed information with reaction times: validity and comparison with the polygraph. Appl. Cogn. Psychol. 23, 1-11.

Williams, L.M., Bahramali, H., Hemsley, D.R., Harris, A.W., Brown, K., Gordon, E., 2003. Electrodermal responsivity distinguishes ERP activity and symptom profile in schizophrenia. Schizophr. Res. 59, 115-125.

Wölk, C., Velden, M., Zimmermann, U., Krug, S., 1989. The interrelation between phasic blood pressure and heart rate changes in the context of the 'baroreceptor hypothesis'. J. Psychophysiol. 3, 397-402.