Psychophysiological and vocal measures in the
detection of guilty knowledge

Matthias Gamer*, Hans-Georg Rill, Gerhard Vossel*, Heinz Werner Gödert

Department of Psychology, Interdisciplinary Research Group Forensic Psychophysiology, Johannes Gutenberg-University Mainz,
Staudingerweg 9, D-55099 Mainz, Germany

Received 28 December 2004; received in revised form 28 February 2005; accepted 5 May 2005
Available online 6 July 2005

Abstract

The Guilty Knowledge Test (GKT) and its variant, the Guilty Actions Test (GAT), are both psychophysiological questioning techniques
aiming to detect guilty knowledge of suspects or witnesses in criminal and forensic cases. Using a GAT, this study examined the validity of
various physiological and vocal measures for the identification of guilty and innocent participants in a mock crime paradigm. Electrodermal,
respiratory, and cardiovascular measures successfully differentiated between the two groups. A logistic regression model based on these
variables achieved hit rates of above 90%. In contrast to these results, the vocal measures provided by the computerized voice stress analysis
system TrusterPro were shown to be invalid for the detection of guilty knowledge.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Guilty Knowledge Test; Guilty Actions Test; Polygraph; Psychophysiological detection; TrusterPro; Vocal measures; Voice stress

1. Introduction

The field of detection of deception is widely discussed in
the scientific community as well as in applied research. Deception is supposed to play an important role in
numerous interpersonal situations, and until now several
behavioral, verbal, and psychophysiological cues have been
examined as potential indicators of human lies. The most
systematic implementation of these ideas was done in
forensic science, where deception can entail serious
consequences and the inability to detect deceit may harm
society. Although it certainly plays a crucial role in
everyday life, detection of deception has mostly been
attributed to the forensic domain. Therefore, we will
concentrate mainly on forensic applications. Two lines of
research, which overlap with each other in varying degrees,
can be differentiated: the development of questioning

Most researchers in this area are convinced that until now
no specific lie response has been found and that it may
indeed never be found (Lykken, 1998, p. 63 ff.). This means
that one cannot ask “Did you murder Mr. X?” and expect an
unequivocally interpretable behavioral, verbal, or physio-
logical response from the suspect leading to a reliable
diagnosis about his truth status. Special questioning

The two most important questioning techniques are the
Control or Comparison Question Test (CQT), which is
widely applied in North America and Israel, for instance,
and the Guilty Knowledge Test (GKT), which is mainly
used in Japan. Already in the 1970s, a huge controversial
debate about the rationale and the psychometric quality of
the CQT started in the scientific literature and has continued
until now. The main issue of this debate concerns the
inadequate standardization of the so-called control questions
which are no controls in the scientific sense. It is not the purpose of this report to reiterate the discussion on the CQT. Interested readers are referred to more detailed publications on this controversial discussion (e.g., Ben-Shakhar, 2002; Furedy, 1996; Honts et al., 2002; Iacono, 2000; Iacono and Lykken, 2002; Raskin and Honts, 2002).

The GKT is a somewhat indirect test of guilt because it primarily aims to detect crime relevant knowledge instead of directly asking for participation in the crime (Lykken, 1959, 1974). Therefore, it consists of several multiple-choice questions, each containing one relevant alternative (i.e., a feature of the crime under investigation known to be “correct”, also referred to as “key item”) and several irrelevant alternatives. Following the standard procedure, all items are denied by the suspect. If, for example, a robbery of a fuel station is examined, a typical GKT-question could be: “Which car was used for the robbery of the fuel station last night?” If in fact a red BMW was used, proper items for this question could be “(a) a green Ford?”, “(b) a blue Mercedes?”, “(c) a red BMW?”, “(d) a yellow Chrysler?”, “(e) a black Pontia?”. According to the assumptions of the GKT, only the culprit should be able to differentiate relevant and irrelevant items correctly and thus show more pronounced physiological responses to the relevant item. A non-systematic response pattern, on the other hand, would indicate innocence, as all items are homogeneous for persons without knowledge about the crime (Lykken, 1974). Typically, the whole GKT consists of at least four to six (or even more) multiple-choice questions to prevent accidentally stronger responses to relevant items in the group of innocent subjects. In contrast to the CQT, the GKT can be regarded as a scientifically controlled diagnostic test for the presence of information. It may be performed independently of any interrogation or pre-test communication between examiner and examinee.

The GKT has been extensively evaluated and discussed in the research literature and there seems to be a consensus that the rationale and the psychometric quality of this technique are not as problematic as those of the CQT (for an overview see Ben-Shakhar and Elaad, 2003; MacLaren, 2001). By contrast, the limitations of the GKT primarily concern its practical implementation, although this technique is routinely used in police investigations, e.g., in Japan (Nakayama, 2002). One of the main assumptions of the GKT is that guilty suspects are able to differentiate between relevant and irrelevant items (Lykken, 1959, 1974). Thus, it is necessary to identify details from the crime scene that the perpetrator would recognize in a GKT-examination and to conceal these details from the public to prevent innocents from becoming informed about the crime. Bradley and his colleagues introduced a potentially useful approach to solve these problems at least partially (Bradley et al., 1996; Bradley and Rettinger, 1992; Bradley and Warfield, 1984). Instead of asking only for crime-relevant knowledge, they proposed to ask for knowledge and participation, to force perpetrators to lie as they deny the question containing the critical item, e.g., “Which car did you use for the robbery of the fuel station last night?” Innocents on the other hand answer each question truthfully. Bradley and Rettinger (1992) named this “new” GKT version the Guilty Actions Test (GAT). Research on the GAT in direct comparison to the GKT (Bradley et al., 1996) reveals that deception plays a significant role in creating the pattern of differential responding to the relevant items. The GAT thus does protect innocents, who are informed about the relevant details of the crime, more effectively than the GKT. When considering perpetrators or uninformed innocents however, both tests produce largely comparable results. A second major issue besides the research on several questioning techniques concerns the search for indicators of deception. All of the techniques specified above typically work with psychophysiological variables, more precisely with indicators of the activity of the autonomous nervous system. Usually, measures of the electrodermal, the respiratory, and the cardiovascular system are used in the examinations. In the CQT, for instance, the diagnostic decision is usually based on skin conductance responses (SCRs) or skin resistance responses (SRRs), respiratory suppression, and relative arterial blood pressure (e.g., Kircher and Raskin, 1988, 2002a). Starting from the assumption that all of these variables measure the same construct but cover partially different aspects, some research aimed at finding combinations of these measures for an optimal differentiation of guilty and innocent suspects. Particularly with regard to the development of computer methods for the psychophysiological detection of deception, systematic strategies for the combination of different measures have been proposed and evaluated. The most common statistical methods in this area are discriminant analyses (e.g., Kircher and Raskin, 2002a) and logistic regressions (e.g., Olsen et al., 1997), both achieving largely similar results (Kircher and Raskin, 2002a). However, the assumptions underlying the use of logistic regressions are less restrictive than those underlying the use of discriminant analyses. In particular, and in contrast to the discriminant analysis, the logistic regression model does not demand multivariate normal distribution in the predicting variables. Regardless of the statistical methods applied, the research on the CQT indicates that different weights of the various psychophysiological variables are necessary to obtain an optimal decision rule. Typically, higher weights are assigned to SCRs or SRRs in comparison to respiratory and cardiovascular measures, indicating the diagnostic significance of electrodermal responses in the CQT (e.g., Kircher and Raskin, 1988).

For the GKT and the GAT, comparable multimodal psychophysiological studies are rare and until now little is known about how to combine different variables in terms of discriminant or regression analyses (one exception is a promising study by Honts et al., 1996). Ben-Shakhar and Elaad (2003) regarded electrodermal responses as the most important measure in detecting guilty knowledge and
reported impressive effect sizes for the differentiation of guilty and innocent subjects on the basis of this variable alone. On the other hand, indicators of respiratory activation became increasingly important for the GKT and the GAT. For guilty subjects, a significant respiratory suppression to relevant items as compared to irrelevant items could be observed in both, mock crime studies conducted in a laboratory context (e.g., Bradley and Rettinger, 1992) as well as in field studies (Elaad et al., 1992). A decision rule based on the means of electrodermal and respiratory measures achieved higher hit rates than each of the variables alone. In addition, it seems that basing the diagnostic decision on both physiological systems may successfully reduce the effectiveness of countermeasures, used especially by guilty suspects in order to manipulate their physiological responsiveness to appear innocent (Ben-Shakhar and Dolev, 1996). The third important physiological system in detection of deception—the cardiovascular system—has been widely ignored in research on the GKT and the GAT. The relative arterial blood pressure, which is regarded as a very important measure in the CQT, was measured in only a few mock crime GKT-studies (e.g., Elaad and Ben-Shakhar, 1989; Podlesny and Raskin, 1978) with disappointing results. But Elaad and Ben-Shakhar did not report the results in detail and Podlesny and Raskin only examined 20 subjects with the GKT. Overall, the validity of the relative arterial blood pressure in detecting guilty knowledge remains unknown. A more promising and scientifically less controversial measure may be the phasic heart rate (HR). The pioneer work on this variable in conjunction with the GKT stems from Bradley and his colleagues (Bradley and Ainsworth, 1984; Bradley and Janisse, 1981) who found a significant deceleration in the phasic pulse rate following the relevant items for guilty subjects. Moreover, high hit rates could be achieved using a combination of SRRs, respiration cycle length, and pulse rate deceleration with equal weights (Bradley and Ainsworth, 1984). In contrast to these promising results, Podlesny and Raskin (1978) and Balloun and Holmes (1979) could not confirm the validity of heart rate changes in GKT examinations.

Therefore, the first purpose of the present study was to clarify which physiological measures are valid in detecting guilty knowledge, and how a statistical combination of several variables could enhance the overall hit rates. Additionally, a more reliable method of obtaining phasic heart rate changes based on an electrocardiogram (ECG) was used.

All of the above-mentioned traditional questioning techniques relying on psychophysiological measures show various disadvantages: The measurement is relatively complex and expensive, only “yes”, “no” or repetitions of items are allowed as answers, and the examinees are aware of being tested. One technique supposed to solve these problems is called voice stress analysis (VSA). The basic assumption of this application is that truthful and deceptive statements differ in various characteristics of the voice signal. Therefore, a microphone or a recorded verbal statement is supposed to be sufficient to determine the truth status of the subject. Most of the VSA-systems are based on the so-called microtremor analysis which goes back to the physiologist Lippold (Lippold, 1970, 1971; Lippold et al., 1957). He found tiny periodical contractions in many muscles of the human body with a frequency of about 8 to 12 Hz. Besides other factors, a decrease of the blood circulation caused this tremor to diminish or even to disappear. In the 1970s this result was implemented very uncritically in the detection of deception. It was assumed that the microtremor, which was thought to affect the muscles responsible for the voice production, could be extracted out of the voice signal. Changes in the blood circulation due to psychological stress were supposed to influence the power in this small frequency domain and thus serve as a valid index of deception. Consequently, the first VSA-system, which was called Psychological Stress Evaluator (PSE), did only measure the power in the vocal frequency range of about 8 to 12 Hz. Decreases in this variable during verbal statements should indicate deception. Despite quite disappointing results in numerous scientific studies on the validity of various systems (e.g., the Computer Voice Stress Analyzer, CVSA; Cestaro, 1995, 1996; or the PSE; Horvath, 1978, 1979; Nachshon and Feldman, 1980), the microtremor analysis in detection of deception is still popular and until now several manufacturers produce VSA-systems which meanwhile are sold as user-friendly computer programs.

Other attempts to develop VSA-devices that should be valid for the detection of deception have been made by the Israeli companies Trustech and Nemesysco. Their software TrusterPro (also known as Vericator) or the newer version TiPi is not solely based on the analysis of the microtremor in the human voice, but also takes the whole frequency range of verbal statements into account (Trustech Ltd., 1998). Another feature of TrusterPro is its supposed ability to detect the origin of stress in the voice, e.g., by differentiating between emotional and cognitive stress and also by identifying intentional lies. Controlled studies on the TrusterPro software are relatively rare and hard to find. Until now only two abstracts of studies on the Vericator device have been published in peer-reviewed journals. The first research (Sommers et al., 2002) aimed to determine the psychometric quality of the Vericator. The authors conducted three experiments and found neither a hit rate above chance level nor a satisfying reliability of the software in a parallel test design. In the second study (Brown et al., 2002), the Vericator device was tested in a mock crime scenario. Again, the hit rates of the device did not exceed chance level. In contrast to these disappointing results, van Damme (2001) reported very high detection rates in a study available on the Internet which has been claimed to be “the only research conducted so far to test our technology the way it is really supposed to be used” by the manufacturer of TrusterPro (retrieved November 29, 2004, .
from http://www.nemesysco.com). Overall, van Damme reported the results of 311 examinations in the laboratory and the field. Additionally, he did not only vary the interrogation techniques (e.g., structured interview, CQT) but also the mode of analysis in the TrusterPro software. Altogether, he reported hit rates of about 80% for the laboratory situation and above 90% for the field context. Unfortunately, no detailed information is provided concerning the experimental design of the studies, the establishment of the validation criterion in the field studies, the concrete application of the software, and the analyses of the TrusterPro measures and reports. In sum, the validity of the TrusterPro software might at least be regarded as doubtful.

The second purpose of the present study was derived from the unsatisfactory standard of knowledge concerning the diagnostic utility of the VSA in structured interrogations. At this time it is widely unknown, if TrusterPro is able to detect concealed information in GKT or GAT examinations, and also if the results of this VSA device could improve the detection rates based on common physiological measures. No empirical study has been conducted so far to clarify the interrelation between the physiological variables on the one hand and the vocal measures calculated by a VSA device on the other hand.

2. Method

2.1. Participants

A total of 60 male subjects participated voluntarily in the experiment in exchange for reward of at least 5 EUR. They were recruited by means of flyers, placards and internet advertisements. Their mean age was 26.5 years (S.D.=6.3 years) with a range of 19 to 56 years. Most of them were students of different fields (mainly biology and jurisprudence). On the arrival for the experimental session, all participants signed consent forms indicating that participation was voluntary and that they could withdraw from the experiment at any time.

2.2. Instruments

Skin conductance, thoracic and abdominal respiration, and relative arterial blood pressure were registered by the Computerized Polygraph System (CPS, Stoehting Company, cf. Kircher and Raskin, 2002b). Skin conductance was measured by a constant voltage system (0.5 V) using a bipolar recording with two Ag/AgCl electrodes (0.8 cm diameter) filled with 0.05 M NaCl electrolyte. The electrodes were placed at the palmar surface of the medial phalanx of the second and third fingers of the left hand. Respiration was recorded by two Pneumotrace II transducers attached around the chest and the abdomen with Velcro straps. The relative arterial blood pressure was measured with a blood pressure cuff that was wrapped around the upper right arm and inflated to about 60 mm Hg. In addition to the CPS, laboratory equipment was used to register an electrocardiogram (ECG). The measurement was accomplished by two Hellige Ag/AgCl electrodes filled with electrode paste and attached to the manubrium sterni and the left lower rib cage. The reference electrode was placed at the right lower rib cage. ECG-data were registered by the Varioport-device (Vitaport system, Becker Meditec) with a sampling rate of 512 Hz. All verbal statements of the subjects during the examination were recorded in the WAV-Format (8-bit, 11.025 kHz) using an active microphone and the TrusterPro software in the Online Mode (software version 6.30) according to the manufacturer’s recommendations (Trustech Ltd., 1998).

The measurement was conducted in a soundproof chamber with participants seated in a semi-reclining chair. The temperature remained nearly constant (\(M=23.7\ ^\circ\text{C},\ S.D.=1.1\ ^\circ\text{C}\)). All recording and programming equipment was located outside the chamber, but the subjects could be observed via a video system. The activity monitor of the CPS mounted under the rear chair legs was used to detect slight body movements of the subject. A conventional personal computer controlled the auditory stimulus presentation and the timing of the measurement equipment.

2.3. Design

A mock crime procedure was used and participants were randomly allocated to one of the following two conditions (with 30 subjects in each condition): (a) guilty examinees performed a mock theft and acquired knowledge of six critical details during this simulated offence; (b) innocent examinees carried out a specific instruction on the same floor of the building, but remained ignorant of the relevant details of the mock crime scene.

2.4. Procedure

Participants in the mock crime condition were instructed to enter an open access departmental library where they could find a leather jacket hanging nearby the librarian’s desk. They had to take a yellow wallet out of the jacket and “steal” all the money (50 EUR) and an EC bank card of the Sparda-Bank (a German finance company). When opening the wallet, the subjects should notice an eye-catching sticker with the name “Jennie” on the inside (the six critical details, about which questions were asked in the subsequent test, are printed in italics). The wallet itself should be left in the jacket after “stealing” the contents. The subjects had to hide the money and the bank card in their pocket and to return to the examination room.

Participants in the innocent condition were instructed to go to the same floor of the building where the mock theft took place, but they had to find a pin board instead, located a few meters away from the library. At this pin board they had to memorize the content of a specific poster announcing
a psychological congress and finally take this poster with them. After the completion of their instruction, they returned to the examination room, too. The fulfillment of both instructions (guilty and innocent) took about the same amount of time (approximately 5–10 min).

Prior to the test, participants were told that the experiment was designed to check whether they could cope with the polygraph test and convince the examiner that they were innocent. In line with previous research, motivational instructions on self-esteem were given (e.g., Gustafson and Orne, 1963) and a financial reward was offered for a successful performance on the task. Innocent examinees could gain an extra reward of 5 EUR for an innocent test result, whereas an amount of 25 EUR was promised for each guilty participant who would be able to achieve an innocent classification. These manipulations were on the one hand meant to increase the motivation of the examinees and on the other hand to ensure attention to the test procedure.

Following these preparations, an experimenter, who was unaware of the experimental condition the examinee had been assigned to, attached the polygraph devices and conducted the examination. All subjects were interrogated using a GAT with six multiple-choice questions concerning the above-mentioned critical details of the mock crime scene. A GAT rather than a GKT was preferred as questioning technique. It was supposed that the active formulation of the GAT-questions could pose a higher level of threat to guilty participants, which is thought to be essential for most kinds of VSA. The GKT and the GAT are actually equivalent as far as testing uninformed innocent examinees. Each GAT-question consisted of one buffer item following the presentation of the question, four irrelevant items, and one crime-related relevant item. The order of relevant and irrelevant items within each question was randomly determined but remained constant across subjects. The multiple-choice questions were presented as pre-recorded audio samples with an inter-stimulus interval of approximately 22 s. The examinees were requested to respond “No” to every item. The whole GAT was separated into three charts with two multiple-choice questions each. The cardio cuff was deflated in the short breaks between the charts.

After completing the test, all guilty participants were asked to recall the relevant items in form of a multiple-choice test. This memory test was administered to control for the possibility of memory deficits causing guilty subjects to show a lack of differential responding to the relevant items. Finally, all participants were debriefed; they received their preliminary test outcome and were paid according to their experimental condition and their test result.

2.5. Response scoring and analysis

2.5.1. Electrodermal responses

The artifact free amplitude of the largest skin conductance increase that began between 0.5 s and 10.5 s after item onset was calculated automatically by the CPS software (Kircher and Raskin, 2002b).

2.5.2. Respiration

Total respiration line length (RLL) during the interval 0 to 10 s following item onset was calculated by the CPS software for each respiration channel separately (Kircher and Raskin, 2002b). In addition, the mean of the thoracic and the abdominal RLL was computed for further analyses.

2.5.3. Relative arterial blood pressure

The amplitude of the largest baseline increase that began within 12 s after item onset was calculated by the CPS software (Kircher and Raskin, 2002b).

2.5.4. Phasic heart rate

The ECG data were first exported to ASCII-files with a sampling rate of 512 Hz. Afterwards the R-waves were detected with a specially developed computer program and R-R intervals were calculated. After converting these intervals into HR (in beats per minute), a second-by-second sampling according to the formula provided by Velden and Wölk (1987) could be applied. The weighted average of the HR in the last second prior to item onset represented the prestimulus baseline. Poststimulus difference scores (ΔHR) were derived by subtracting the prestimulus baseline value from the HR-score of each poststimulus second. In addition to these detailed trends of the HR, the largest deceleration of the HR within 15 s following item onset was calculated as a more telling single measure for each question (Bradley and Janisse, 1981).

2.5.5. Standardization

In order to eliminate individual differences in responsivity and to permit a meaningful index of the response differences between relevant and irrelevant items for the whole test, standard difference scores for each subject and each measure (SCR amplitude, RLL, baseline increase of the relative arterial blood pressure, and HR deceleration) were calculated according to the following procedure. First of all, the responses to each item were z-standardized based on the mean and the standard deviation of the responses to the irrelevant items. In a second step, difference scores between the response to the relevant item and the mean of the four irrelevant items within each of the six multiple-choice questions were calculated. Afterwards the mean of these measures was computed as an overall index of the differential responsivity in each physiological measure. It was assumed that this aggregated value should be around zero for innocent subjects, as they are supposed to show a non-systematic response pattern to relevant and irrelevant items, whereas a negative value (for a lower RLL or a stronger HR deceleration on relevant items) or a positive value (for higher responses on relevant items in the other physiological variables) should indicate knowledge of
crime-related details. Primarily these standardized response differences were used in all subsequent statistical analyses.

2.5.6. TrusterPro

The verbal answers of each subject following relevant and irrelevant items of the GAT were analyzed by a certified TrusterPro examiner using the TrusterPro Offline-Mode, which is supposed to achieve the highest level of accuracy (Trustech Ltd., 1998). First of all, each verbal answer to the test items was defined as one segment within the TrusterPro software and the remaining part of the WAV-file was defined as noise. Afterwards, the six answers to the relevant items were categorized as relevant segments for the subsequent computations within the software. Three categories of variables were used in the statistical analyses. (1) The software provided an overall diagnosis based on the short classifications of the relevant segments. If at least one of the six relevant segments was classified as D.I. (deception indicated), the subject was thought to be guilty. (2) TrusterPro classified each relevant segment with one of four short diagnoses: D.I. (deception indicated), INC+ (inconclusive plus), INC (inconclusive) or N.D.I. (no deception indicated). The frequency of each of the four diagnoses across the six relevant items within each participant was used for further statistical analyses. (3) TrusterPro computed 21 raw values (e.g., lie stress, global stress, emotional stress, cognitive stress) for each verbal segment to obtain an overall diagnosis. These raw values were standardized using the procedure described above and used for further analyses. A rejection region of \( p < 0.05 \) was used for all statistical tests.

3. Results

3.1. Memory test for guilty participants

The effectiveness of the GKT and the GAT strongly depends on the ability of perpetrators to remember the critical details asked during the test (Carmel et al., 2003). In this study, the recall rate of relevant items was very high for guilty subjects: They remembered 5.27 out of six details on average (S.D. = 0.78).

3.2. Physiological measures

In a first step, the mean \( z \)-standardized response differences between relevant and irrelevant items were compared between the two experimental groups (see Table 1). T-tests revealed statistically significant differences between guilty and innocent subjects for the thoracic RLL: \( t(58)=4.03, p < 0.001, d=1.04 \); for the abdominal RLL: \( t(58)=4.38, p < 0.0001, d=1.13 \); and for the mean of both respiratory channels: \( t(58)=4.88, p < 0.0001, d=1.26 \). Additionally, the SCR data differed significantly between the two experimental groups; \( t(58)=6.71, p < 0.0001, d=1.74 \). By contrast, no significant group difference could be found for the relative arterial blood pressure (“cardio”); \( t(58)=1.09, p=0.28, d=0.28 \). As expected, the typical pattern of differential responses to relevant and irrelevant items was found for most measures in the group of guilty subjects, whereas these differences tended to zero for innocents.

A \( 2 \times 2 \times 16 \) analysis of variance (ANOVA) was performed on the second-by-second \( \Delta HR \) data with the experimental condition (guilty vs. innocent) serving as between-participants factor and the item type (relevant vs. irrelevant) and, respectively, the poststimulus second serving as within-participant factors. Most remarkably, a statistically significant interaction between the factors experimental condition, item type, and poststimulus second could be found; \( F(15, 870)=3.40, p < 0.01, \varepsilon=0.34, f^2=0.08 \). To further investigate this interactive effect, two \( 2 \times 16 \) ANOVAs on the \( \Delta HR \) data with the item type and the poststimulus second serving as within-participant factors were performed separately within each experimental group. For the guilty subjects a significant interaction between the item type and the poststimulus second could be found; \( F(15,435)=6.10, p < 0.001, \varepsilon=0.26, f=0.16 \), additionally, the main effects of item type; \( F(1,29)=41.22, p < 0.001, f=0.43 \), and poststimulus second; \( F(15,435)=17.81, p < 0.001, \varepsilon=0.28, f=0.45 \), were statistically significant. In the group of innocents, the interactive effect between the item type and the poststimulus second was absent; \( F(15,435)=0.93, p=0.57, \varepsilon=0.34 \), and only the main effect of poststimulus second reached statistical significance; \( F(15,435)=14.47, p < 0.001, \varepsilon=0.29, f=0.40 \). These results indicate a differential trend in the \( \Delta HR \) data for relevant and irrelevant items in the group of

<table>
<thead>
<tr>
<th>Measure</th>
<th>Guilty (n = 30)</th>
<th>Innocent (n = 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (S.D.)</td>
<td>M (S.D.)</td>
</tr>
<tr>
<td>Thoracic RLL</td>
<td>0.53 (0.36)</td>
<td>0.08 (0.75)</td>
</tr>
<tr>
<td>Abdominal RLL</td>
<td>0.55 (0.46)</td>
<td>0.09 (0.34)</td>
</tr>
<tr>
<td>Mean RLL</td>
<td>0.54 (0.34)</td>
<td>0.00 (0.49)</td>
</tr>
<tr>
<td>SCR</td>
<td>0.91 (0.60)</td>
<td>0.01 (0.44)</td>
</tr>
<tr>
<td>BP</td>
<td>0.16 (0.44)</td>
<td>0.04 (0.41)</td>
</tr>
</tbody>
</table>

For the respiration line length (RLL) measures, bigger response differences are reflected by smaller values. SCR = skin conductance response, BP = relative arterial blood pressure.
guilty subjects; whereas the ΔHR trends did not differ in the group of innocents (see Fig. 1).

To further evaluate the diagnostic utility of the phasic heart rate, the largest deceleration of the HR within 15 s following question onset was analyzed as a combined value of the HR-responsiveness. Again, the mean z-standardized response differences between relevant and irrelevant items differed significantly between the two experimental groups; $t(58)=2.04$, $p<0.05$, $d=0.53$. Guilty subjects showed differential responses on the two item categories ($M=0.42$, S.D. =0.53), whereas no such difference could be observed for innocents ($M=0.11$, S.D. =0.65). The extraction of dichotomous diagnoses out of several physiological measures can still be regarded as one major problem in psychophysiological detection of deception. One attempt to deal with the problem of arbitrary cutoff points was proposed by Ben-Shakhar and his colleagues (e.g., Ben-Shakhar and Elaad, 2003). They suggested comparing detection efficiencies by means of signal detection theory in order to contrast the whole parameter distribution between the experimental groups. Using this approach, one can directly compare the hit rate and the rate of false positives associated with each possible cutoff. To extract an overall value for the detection efficiency of each physiological measure in this study, separate receiver operating characteristic (ROC) curves were generated on the basis of the mean z-standardized response differences and the area under each curve was computed. This area varies between 0 and 1, whereas an area of 0.5 can be regarded as a random classification. An area of 1 on the other hand indicates that there is no overlap between the distributions for guilty and innocent subjects, and that each participant could be classified correctly based on the selected measure (cf. National Research Council, 2003). In addition to this descriptive value, Bamber (1975) described a method for estimating the variance of the area statistic and for computing confidence intervals for the true area. Using this method, we computed 90% and 95% confidence intervals for each physiological measure. The relevant data are displayed in Table 2.

Confidence intervals indicate that the area under the ROC curves of all physiological variables differed significantly from a chance area of 0.5, except with regard to the

<table>
<thead>
<tr>
<th>Measure</th>
<th>Area</th>
<th>90% CI</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoracic RLL</td>
<td>0.83</td>
<td>(0.74, 0.91)</td>
<td>(0.72, 0.93)</td>
</tr>
<tr>
<td>Abdominal RLL</td>
<td>0.80</td>
<td>(0.71, 0.89)</td>
<td>(0.69, 0.91)</td>
</tr>
<tr>
<td>Mean RLL</td>
<td>0.84</td>
<td>(0.76, 0.92)</td>
<td>(0.74, 0.94)</td>
</tr>
<tr>
<td>SCR</td>
<td>0.89</td>
<td>(0.82, 0.96)</td>
<td>(0.81, 0.97)</td>
</tr>
<tr>
<td>BP</td>
<td>0.45</td>
<td>(0.32, 0.57)</td>
<td>(0.30, 0.60)</td>
</tr>
<tr>
<td>HR</td>
<td>0.67</td>
<td>(0.55, 0.79)</td>
<td>(0.53, 0.81)</td>
</tr>
</tbody>
</table>

For the respiration line length (RLL) and the heart rate deceleration (HR), the signs of each value were reversed before computation in order to score in the same direction as the skin conductance response (SCR) and the relative arterial blood pressure (BP).

$N=60$. 

---

Fig. 1. Phasic heart rate (ΔHR) as a function of poststimulus seconds and item type in both experimental groups. Error bars indicate standard errors of the mean.
relative arterial blood pressure which was again shown to be invalid in conjunction with the GAT. Fig. 2 displays the ROC curves of the diagnostically most relevant measures of each physiological system investigated in this study. The figure illustrates how a displaced cutoff changes rates of hits and false positives.

Although hit rates for guilty and innocent subjects could be extracted out of the ROC curves for each physiological measure separately, it is still unclear how to combine the various measures to obtain individual diagnoses. Using the simplest strategy, one would compute the arithmetic mean of several z-standardized indicators and find a decision on this combined value (e.g., Ben-Shakhar and Dolev, 1996). However, increasing research effort on the CQT suggests that a decision rule based on a statistical combination of several measures with different weights can achieve higher detection rates (e.g., Kircher and Raskin, 1988, 2002a; Olsen et al., 1997). Because of less restricted assumptions as compared to discriminant analysis, a logistic regression model was used in this study to determine the hit rates of a combined value. The experimental group (0 = innocent, 1 = guilty) served as dependent variable and the mean z-standardized response differences of all physiological measures were used as potential predictors. To identify only those variables which account for an additional amount of variance in the criterion, a step-by-step inclusion procedure following the Wald-statistic was applied. Probabilities for inclusion and exclusion were fixed at 0.05 and 0.10 respectively. Three measures were successively included in the logistic regression model: skin conductance, the mean of thoracic and abdominal respiration, and heart rate (Table 3).

The heart rate as an indicator of the activity of the cardiovascular system achieved a slightly lower weight coefficient in comparison to the electrodermal and respiratory measures which were weighted approximately equally. Fig. 3 illustrates the distribution of the probability scores for the two experimental groups in the context of the computed logistic regression model. Setting the cutoff at 0.5 without allowing inconclusive diagnoses resulted in hit rates of 93% for the guilty subjects (sensitivity) and 97% for the innocents (specificity).

Using the same data to fit the logistic regression model and to compute hit rates typically leads to higher validity coefficients than cross-validation on a new sample (Copas and Corbett, 2002). To estimate the shrinkage of the logistic regression model when applied to new data, the hold-one-out method was applied. We computed 60 different logistic regression models based on \( n - 1 = 59 \) cases. The last subject was then classified using the appropriate regression equation and a threshold of 0.5. This procedure yielded a sensitivity of 90% and a specificity of 87% and could be regarded as an indication of the model’s performance when applied to new data.

### 3.3. Vocal measures (TrusterPro)

TrusterPro classifies each analyzed relevant item with one of four diagnoses, ranging from N.D.I. (no deception indicated) to INC (inconclusive) and INC+ (inconclusive...
plus) to, finally, D.I. (deception indicated). Using an overall decision rule, one would classify a subject as guilty if at least one of the six relevant items was classified as D.I. This procedure resulted in rather low hit rates (sensitivity: 30% and specificity: 83%) which were not significantly above chance level; $\chi^2(1, N=60)=0.84$, $p=0.36$.

We further compared the mean frequencies of the four TrusterPro diagnoses across the six relevant items between the two experimental groups. One would assume that guilty subjects should have a greater number of D.I. or potentially INC+ diagnoses, whereas innocents would be expected to have more N.D.I. classifications. However, no statistically significant differences between the groups could be observed in the frequency of the several diagnoses (Table 4).

Table 4
Mean frequencies of TrusterPro diagnoses assigned to relevant items within each experimental condition and t-tests between the groups

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Guilty ($n=30$)</th>
<th>Innocent ($n=30$)</th>
<th>t(58)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>No deception indicated (N.D.I.)</td>
<td>3.33 (1.06)</td>
<td>3.77 (1.17)</td>
<td>1.51</td>
<td>.14</td>
</tr>
<tr>
<td>Inconclusive (INC)</td>
<td>2.10 (1.24)</td>
<td>1.73 (1.08)</td>
<td>1.22</td>
<td>.23</td>
</tr>
<tr>
<td>Inconclusive plus (INC+)</td>
<td>0.23 (0.50)</td>
<td>0.30 (0.47)</td>
<td>0.53</td>
<td>.60</td>
</tr>
<tr>
<td>Deception indicated (D.I.)</td>
<td>0.33 (0.55)</td>
<td>0.20 (0.48)</td>
<td>1.00</td>
<td>.32</td>
</tr>
</tbody>
</table>

The values refer to the frequency of relevant items labelled with the corresponding diagnoses within the participants. As six relevant items were used, the frequencies sum to six.

In addition to the four diagnoses mentioned, TrusterPro measures 21 raw values for each analyzed segment. To search for any significant group differences in these measures, we compared the z-standardized response differences between relevant and irrelevant items for all 21 raw values between the two experimental groups. T-tests revealed no statistically significant difference (see Appendix A).

To find interrelations between physiological measures on the one hand and vocal characteristics on the other hand, correlations between the z-standardized response differences of the 6 physiological variables and the 21 TrusterPro raw values were computed. Only 5 out of 126 correlations fell below an a priori significance level of 0.05.

4. Discussion

The present study aimed to investigate the validity of several psychophysiological and vocal measures in the detection of guilty knowledge. Using a GAT in the laboratory context, electrodermal, respiratory, and cardiovascular measures differentiated between guilty and innocent subjects. Consistent with numerous studies on the GKT and the GAT (for a summary see Ben-Shakhar and Elaad, 2003), guilty participants showed larger skin conductance responses on relevant items than on irrelevant ones, whereas an unsystematic response pattern was observed in the group of innocents. The area under the ROC curve for skin conductance responses was 0.89 which is comparable to the average area statistic of 0.87 reported by Ben-Shakhar and Elaad (2003) for 42 mock crime conditions. Concerning the respiratory measures, significant group differences between guilty and innocent subjects were observed for the RLL measures at the thoracic region as well as at the abdominal region. Guilty participants showed a larger amount of respiratory suppression on relevant items than on irrelevant alternatives, whereas no such response difference was observed in the group of innocents. This result is consistent with other findings (e.g., Ben-Shakhar and Dolev, 1996; Bradley and Rettinger, 1992; Podlesny and Raskin, 1978).

Especially the mean of both respiratory channels seemed to be almost as valid in the detection of guilty knowledge as the electrodermal measure.

Changes in the relative arterial blood pressure were not capable of differentiating between the experimental groups in the present study. However, another variable of the cardiovascular system proved to accomplish this purpose: the phasic heart rate. Guilty participants showed a strong heart rate deceleration following relevant items, whereas no such changes could be observed after irrelevant alternatives. In the group of innocents, the trends in the phasic heart rate did not differ between the two item types and varied only slightly around zero. These results are consistent with GKT-studies taking into account changes in the phasic pulse rate (Bradley and Ainsworth, 1984; Bradley and Janisse, 1981), but they are in conflict with earlier studies by Podlesny and Raskin (1978) and Balloun and Holmes (1979) respectively.

Interestingly, the observed $\Delta$HR trends closely resemble the results of two studies using CQT-interrogations in a mock crime scenario (Podlesny and Raskin, 1978; Raskin and Hare, 1978). In both these studies, an initial HR-acceleration passes into a strong deceleration only in response to relevant questions posed to the group of guilty subjects. Raskin and Hare (1978) interpreted the initial HR-acceleration as “a result of the subject’s preparations to answer and his act of answering” (p. 135), whereas the subsequent HR-deceleration was thought to reflect attentional processes aiming at gathering information from the environment and from one’s own body. This interpretation is plausible in the context of the present study, too. Only guilty subjects are able to identify the critical detail in each GAT-question. It is possible that the more or less automatic redirection of attention following a relevant item aims to control conspicuous physiological responses and thus should help the examinee to successfully pass the test. Although guilty and innocent participants could be differentiated on the basis of the phasic HR in the present study, the effect size of this measure was considerably lower than that of electrodermal and respiratory measures.

A systematical combination of the various physiological variables in the prediction of the truth status yielded a
logistic regression equation with one integrated measure of each physiological system examined. Skin conductance responses and the mean of both respiratory channels were weighted approximately equally, whereas the phasic heart rate seemed to account for a smaller amount of variance in the criterion. However, all of the three mentioned variables significantly enhanced the differentiation between the two experimental groups and on the basis of this regression model hit rates of 93% for guilty subjects and 97% for innocents were computed. To estimate the shrinkage of this model when applied to new data, the hold-one-out method was used. Although sensitivity dropped to 90% and specificity to 87%, the validity of the logistic regression model could still be regarded as satisfactory. The observed hit rates are considerably higher than those reported in the meta analysis on the GKT by MacLaren (2001) which ranged from 76% for guilty examinees to 83% for innocents. Three possible reasons for this discrepancy should be considered. (1) MacLaren used a comparatively simple method to obtain the hit rates of the several studies. This method called Lykken-scoring (Lykken, 1959) only takes into account the relative response differences between relevant and irrelevant items without considering the absolute values of the various physiological measures. The computation of z-standardized response differences used in the present study might reduce the loss of information and increase the validity of the corresponding variable. (2) One inclusion criterion in the meta analysis of MacLaren was the specification of hit rates based on electrodermal responses or a combination of electrodermal and respiratory measures. Other physiological measures were not considered. In the present study, electrodermal, respiratory, and cardiovascular variables were combined to an overall decision in each individual case. If these measures are all predictors of the same construct, but cover at least partially different aspects, higher hit rates might be expected on the basis of this integrative classification rule. The combination of different measures could, on the one hand, compensate a lack of reliability in single values and, on the other hand, increase the explained variance of the criterion. Although cardiovascular measures seemed to be comparatively less important in the present study, their predictive quality might increase in situations where other physiological variables (e.g., the respiratory responses) are manipulated by examinees aiming to appear innocent. This issue is worth examining in the future, as countermeasures raise serious problems for the application of all psychophysiological questioning techniques in the forensic domain (Honts and Amato, 2002). (3) The logistic regression equation of the present study was fitted to the characteristic features of the two groups to be separated from each other. Although the hold-one-out method could be regarded as a first evidence of the model’s performance when applied to new data, it is still necessary to cross-validate the regression equation in an independent sample in order to generalize the hit rates to other settings in the laboratory and, potentially, in the field. It is worth mentioning that the very high hit rates of the logistic regression model in this study might even increase as a function of multiple-choice questions used during the test. If these questions are well-constructed, i.e., all items are equally plausible to innocent examinees and guilty subjects are able to recognize the relevant items with a probability of around 70% or 80%, the use of more questions might increase the validity of the test (Lykken, 1998, p. 288 ff.).

In clear contrast to the promising results of various psychophysiological measures in the detection of guilty knowledge, the voice stress analysis with TrusterPro was shown to be invalid in the present study. While the specificity of the overall diagnoses was satisfactory (83%), only 30% of the guilty subjects were classified correctly. The observed overall hit rate did not exceed chance level. This pattern of results is comparable to research by Sommers et al. (2002) but conflicts with the validity study of van Damme (2001). Deeper analyses, taking into account the frequency of various diagnoses related to the relevant items or computing z-standardized response differences of the TrusterPro raw-values, provided no evidence for the validity of the software in the detection of guilty knowledge, either. TrusterPro is not directly based on the analysis of the microtremor in the human voice as are other VSA-systems (e.g., CVSA or PSE), but it provides raw values that describe the power (SPJ), the distribution uniformity (JQ), and the average range of relatively low frequencies (AVJ). According to their description, these values seem to be related to the same frequency range as are microtremor analyses of other systems. All these raw values failed to differentiate between guilty and innocent participants (see Appendix A). For this reason, our disappointing results might generalize to other VSA-systems that are based on the microtremor analysis.

The most common speculation used to explain several disappointing results with VSA-systems has been frequently expressed by the manufacturers of these devices (e.g., Trustech Ltd., 1998). They claim that their systems are more valid in high-stake situations where there is a substantial amount of threat. There are at least two research results limiting the general validity of this argument. First, it was shown that although the hit rates of TrusterPro were considerably higher in the field situation, satisfactory results with a proportion of around 80% correct decisions could all the same be obtained in the laboratory context (van Damme 2001). Second, a manipulation of the stress level in yet another study did not enhance the sensitivity of TrusterPro which was insufficiently low in all experimental conditions (Sommers et al., 2002). As the results of the present study are comparable to the few research reports on TrusterPro published in recognized journals (Brown et al., 2002; Sommers et al., 2002), a good amount of skepticism
concerning the diagnostic quality of this VSA-device in GKT/GAT-examinations might be appropriate. It remains an interesting challenge, however, to try to explain why the mock crime design of the present study was capable of generating large response differences in the physiological variables, whereas no such pattern could be observed for any of the vocal measures analyzed by the TrusterPro software. Maybe this question could only be answered by explicitly documenting the variables of the VSA-device and relating them to psychophysiological processes supposed to be responsible for the generation of reliable response differences in physiological measures during GKT/GAT-examinations. Although this procedure would, from a scientific as well as from a practical point of view, be quite demanding, it seems an indispensable precondition for any functioning voice-based method for detecting guilty knowledge. Until then, the GKT and the GAT might be effectively applied in police investigations (cf. Nakayama, 2002) and potentially in the court, too (cf. Ben-Shakhar et al., 2002), if individual diagnoses are based on physiological variables which have been proven to be highly valid.

Certainly, the hit rates observed in this study should not be regarded as absolute values that would be found in future examinations inside and outside the laboratory, too. Especially the recall rate of critical details, which is fundamental for an effective differentiation between guilty and innocent subjects, is supposed to be lower in field examinations (Carmel et al., 2003). But the results of this study demonstrate that the systematic combination of different physiological systems and the computation of standardized response differences could be capable of enhancing the detection efficiency of the GAT. This knowledge could be applied to computer methods for the detection of deception that are mainly optimized for CQT-examinations at the moment (Kircher and Raskin, 2002a). Additionally, this study could stimulate independent researchers to cross-validate the resulting regression equation in a different context. In particular, it would be interesting to examine the detection efficiency of this logistic regression equation which was based on an active question formulation originating from the GAT in a comparable study using the passive formulation of the standard GKT. Ideally, further research should directly compare the detection efficiencies of both these question formats within the same experimental setup.

Acknowledgements

We are indebted to Jens Arnold, Marc von der Heiden, and Esther Wetzel for their help and support during data collection and analysis and to Alfred Weber for his helpful comments during the preparation of this article. Additionally, we would like to thank David T. Lykken, John J. Furedy, and an anonymous reviewer for commenting on an earlier version of this manuscript.

Appendix A

Table A1

<table>
<thead>
<tr>
<th>TrusterPro value</th>
<th>Guilty (n=30)</th>
<th>Innocent (n=30)</th>
<th>t(58)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (S.D.)</td>
<td>M (S.D.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lie Stress</td>
<td>−0.12 (0.41)</td>
<td>−0.04 (0.55)</td>
<td>0.65</td>
<td>0.52</td>
</tr>
<tr>
<td>Lie Probability</td>
<td>−0.06 (0.49)</td>
<td>0.02 (0.48)</td>
<td>0.69</td>
<td>0.49</td>
</tr>
<tr>
<td>Global Stress</td>
<td>−0.19 (0.51)</td>
<td>0.02 (0.46)</td>
<td>1.68</td>
<td>0.10</td>
</tr>
<tr>
<td>FRG Stress</td>
<td>−0.05 (0.60)</td>
<td>0.00 (0.45)</td>
<td>0.40</td>
<td>0.69</td>
</tr>
<tr>
<td>Emotional Stress</td>
<td>−0.17 (0.57)</td>
<td>0.01 (0.39)</td>
<td>1.42</td>
<td>0.16</td>
</tr>
<tr>
<td>Cognitive Stress</td>
<td>0.08 (0.38)</td>
<td>−0.01 (0.30)</td>
<td>1.09</td>
<td>0.28</td>
</tr>
<tr>
<td>Thinking Rate</td>
<td>0.02 (0.46)</td>
<td>0.03 (0.39)</td>
<td>0.08</td>
<td>0.93</td>
</tr>
<tr>
<td>Anticipation</td>
<td>0.15 (0.39)</td>
<td>−0.01 (0.42)</td>
<td>1.51</td>
<td>0.14</td>
</tr>
<tr>
<td>SPT</td>
<td>0.14 (0.50)</td>
<td>0.00 (0.37)</td>
<td>1.26</td>
<td>0.21</td>
</tr>
<tr>
<td>SPJ</td>
<td>−0.08 (0.39)</td>
<td>0.03 (0.33)</td>
<td>1.14</td>
<td>0.26</td>
</tr>
<tr>
<td>JQ</td>
<td>−0.19 (0.51)</td>
<td>0.02 (0.46)</td>
<td>1.64</td>
<td>0.11</td>
</tr>
<tr>
<td>AVJ</td>
<td>0.02 (0.46)</td>
<td>0.03 (0.39)</td>
<td>0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>S.O.S.</td>
<td>−0.07 (0.40)</td>
<td>0.05 (0.62)</td>
<td>0.87</td>
<td>0.39</td>
</tr>
<tr>
<td>Fmain</td>
<td>−0.06 (0.51)</td>
<td>−0.03 (0.42)</td>
<td>0.22</td>
<td>0.83</td>
</tr>
<tr>
<td>Fx</td>
<td>−0.05 (0.67)</td>
<td>−0.04 (0.45)</td>
<td>0.04</td>
<td>0.97</td>
</tr>
<tr>
<td>FQ</td>
<td>0.01 (0.39)</td>
<td>−0.03 (0.50)</td>
<td>0.32</td>
<td>0.75</td>
</tr>
<tr>
<td>FFlic</td>
<td>−0.01 (0.53)</td>
<td>0.01 (0.51)</td>
<td>0.16</td>
<td>0.87</td>
</tr>
<tr>
<td>Sub Cog.</td>
<td>−0.13 (0.52)</td>
<td>0.11 (0.48)</td>
<td>1.79</td>
<td>0.08</td>
</tr>
<tr>
<td>Sub Emo.</td>
<td>−0.02 (0.46)</td>
<td>−0.08 (0.57)</td>
<td>0.43</td>
<td>0.67</td>
</tr>
<tr>
<td>F1Q</td>
<td>0.03 (0.46)</td>
<td>0.02 (0.47)</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>MsgX</td>
<td>−0.10 (0.45)</td>
<td>0.08 (0.41)</td>
<td>1.56</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Most TrusterPro raw-values are not described in detail but some information concerning the most important variables is available from Trustech Ltd. (1998).

References


