



Naval Aerospace Medical Research Laboratory



A Comparison of Approaches To Detect Deception

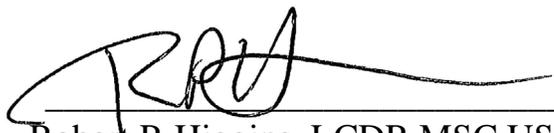
Marcus K. Taylor, Dain S. Horning, Joseph F. Chandler, Jeffrey B. Phillips,
Jasmine Y. Khosravi, Jill E. Bennett, Heather Halbert,
Benedict J. Fern, & Hong Gao



NAMRL Report Number 11-05

Approved for public release; distribution unlimited

Reviewed and Approved
25 Feb 2011



Robert P. Higgins, LCDR MSC USN



The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

This work was funded by work unit number PB802 .

The study protocol was approved by the Naval Aerospace Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects.

I am a military service member of the U.S. Government. This work was prepared as part of my official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

A promising method to detect deception is the Guilty Knowledge Test (GKT) which assesses whether an individual possesses knowledge about a particular crime. Specifically, the GKT involves a series of questions with multiple answers where one answer is relevant to the crime and the others are not. In theory, if a “suspect” is guilty, s/he will recognize the crime-relevant item and display a physiological orienting response that is discernable from responses to irrelevant items. The most widely studied physiological endpoint in conjunction with the GKT is the galvanic skin response (GSR) which reflects activity of the eccrine sweat gland and represents sympathetic modulation of the autonomic nervous system. Little is known regarding the validity of alternate physiological endpoints or if combined endpoints enhance detection accuracy over GSR alone. In this study we compared different physiological approaches to detect deception with the GKT. Secondarily, we explored sociobehavioral moderators of deception, including values, ethnic identity and resilience. Forty-two military men (age 23.9 ± 0.4 years) participated in a mock-crime and then completed a 10-question GKT. Endpoints included GSR, heart rate (HR; reflecting vagal modulation), and finger pulse line length (FPLL; a blood pressure waveform calculation reflecting combined sympathetic-parasympathetic modulation). Separate one-way repeated measures ANOVA with five levels compared mean physiological responses to each question. A common scoring procedure (Lykken, 1959) was applied to classify subjects as guilty or innocent for each physiological endpoint. ROC curves were then constructed to assess the diagnostic value of each endpoint and to determine whether combined measures detected guilt more effectively than GSR alone. Lastly, correlational analyses were used to explore sociobehavioral factors influencing the GKT response. As hypothesized, presentation of

“guilty” items resulted in higher GSR responses ($p < 0.001$) than irrelevant items. Similar nonsignificant patterns emerged for HR and FPLL. GSR performed best with the Lykken approach, correctly identifying guilt and innocence in 47.6% and 90.5% of subjects, respectively. ROC analyses revealed that all three individual endpoints performed better than chance ($p < 0.05$) and that a combined index (GSR + FPLL) enhanced classification power over GSR alone (Area Under Curve: 85.0% versus 79.0%). Although GSR appears to be the most valid individual GKT endpoint of those considered in the present study, a combined index improves classification power over the best individual endpoint. Finally, exploratory analyses suggested that sociobehavioral factors may moderate the human response to guilty knowledge. The current findings confer proof-of-concept for this capability and satisfy a crucial prerequisite for cross-cultural and sociobehavioral comparisons.

INTRODUCTION

Law enforcement and counterterrorism efforts rely upon valid and reliable methods of differentiating guilt from innocence. Operationally, this is often approached through interviews with suspects, their affiliates, or other individuals who may have relevant information. The veracity of this information is a topic of great concern given its ramifications for local, national or international security. The Guilty Knowledge Test (GKT; Lykken, 1959), also termed Concealed Information Test, is a promising method of detecting concealed information. It consists of a series of questions (displayed as pictures or written text) with multiple answers where one answer is relevant to the crime and the other options are not. In theory, if a “suspect” is guilty, s/he will recognize the crime-relevant item and display a physiological orienting response that is discernable from responses to crime-irrelevant items (Sokolov, 1990).

The GKT is typically used in conjunction with tools that measure various physiological responses. The most extensively studied endpoint is the galvanic skin response (GSR) which reflects activity of the eccrine sweat gland and is responsive to psychological stress (MacLaren, 2001; Ben-Shakhar & Elaad, 2003). MacLaren (2001) performed a meta-analysis including 22 GKT studies using this endpoint and found an overall sensitivity of 76%, where 640 of 843 subjects with “guilty knowledge” were correctly identified. In this study, uninformed (i.e., innocent) subjects were correctly identified 83% of the time (336 of 404). A meta-analysis conducted by Ben-Shakhar et al. (2003) reported an overall Cohen’s *d* effect size of 1.55 with the GSR endpoint – a substantial effect when considered within the larger context of the GKT literature.

“Optimal” experimental conditions (i.e., involving motivational instructions, a deceptive verbal response, and at least 5 GKT questions) resulted in the largest effect sizes.

Although validation studies show that the GSR endpoint holds promise as a means to detect guilty knowledge, it is imperative to minimize false-positive (i.e., labeling someone guilty who is in fact innocent) and false-negative identifications (i.e., labeling someone innocent who is in fact guilty) to justify its use in field settings (Kleinmuntz & Szucko, 1984). With this in mind, several researchers have examined with varying success the utility of alternate endpoints. Some of these include cardiovascular indices (e.g., heart rate and blood pressure), respiration, pupil diameter changes, electroencephalographic characteristics (primarily the P300 component of the event-related potential), EKG attributes (e.g., t-wave amplitude (Furedy, Heslegrave, & Scher, 1992), and functional MRI indices (Kozel et al., 2009).

Cardiovascular measures are often used in conjunction with the GKT. Since heart rate typically declines as a byproduct of the orienting response (Sokolov, 1990), GKT studies using this endpoint normally quantify maximal or mean decline in response to a stimulus with a larger decline expected in response to guilty items (Bradley & Janisse, 1981). A handful of studies reflect inconsistent ability of the heart rate response to differentiate guilt from innocence and it is usually outperformed by GSR (Verschuere, Crombez, De, & Koster, 2005; Podlesny & Raskin, 1978). More recently, finger pulse wave forms have also been studied. This endpoint has been operationalized as finger pulse wave length (Elaad & Ben-Shakhar, 2006), also termed finger pulse line length (FPLL; Vandenbosch, Verschuere, Crombez, & De, 2009). FPLL is typically measured with a plethysmograph placed at the top of the finger and is calculated as the signal trace

within a given time frame. As such, FPLL is a composite measure of finger pulse rate and finger pulse amplitude (FPA; systolic minus diastolic blood pressure). Combined with vagally-induced heart rate deceleration, the orienting response is expected to yield a decrease in pulse amplitude; the underlying assumption being that high concentrations of α adrenergic receptors in the finger respond to sympathetic nervous stimulation, leading to peripheral vasoconstriction and increased diastolic pressure. Elaad et al. (2006) examined the utility of FPLL in two experiments. In the first, FPLL outperformed respiration line length (an index of rate and depth of breathing) and was comparable to that of GSR. In the second, FPLL outperformed both GSR and respiration line length. These scientists (Elaad & Ben-Shakhar, 2008) subsequently showed that FPLL performed similarly to GSR and respiration line length, and Vandenbosch et al. (2009) reported that FPLL outperformed independent measures of finger pulse amplitude and heart rate, respectively. Ambach et al. (2008), by contrast, concluded that FPLL performed significantly *worse* than GSR, heart rate, and respiration line length. Although promising, more controlled laboratory studies are needed to systematically evaluate the utility of FPLL – not only as an independent indicator but also as an element of combined proxies to detect guilty knowledge.

In all likelihood, detection accuracy of the GKT may be optimized utilizing combined endpoints. To date, this approach has been taken with varied levels of success. Several studies, for example, have explored the validity of combined GSR and respiration indices (Elaad, Ginton, & Jungman, 1992; Ben-Shakhar & Dolev, 1996; Ben-Shakhar & Elaad, 2002), some finding that combined indices outperform GSR alone. Elaad et al. (1992), for example, correctly classified 75% of subjects using GSR and respiration line

length respectively, but increased classification accuracy to 85% using a combined model. In other work, Gamer et al. (2006) assessed the combined utility of GSR, heart rate, and respiration line length in a GKT, correctly classifying 93% of guilty and 97% of the innocent examinees. Subsequently, this group investigated whether heart rate and respiration indices enhance detection validity over GSR alone. They found that a weighted combination of these measures yielded slightly larger validity coefficients (Gamer, Verschuere, Crombez, & Vossel, 2008). Other studies, however, have failed to demonstrate such improvements (Verschuere, Crombez, Koster, & De, 2007; Bradley & Ainsworth, 1984). Clearly, much remains to be learned regarding which physiological endpoints may comprise an optimal composite index of guilty knowledge.

Altogether, the current literature suggests that the GKT performed in conjunction with GSR holds promise as an instrument to detect guilty knowledge, but much remains to be learned not only of alternate GKT endpoints but also of the possibility that combined measures may enhance classification accuracy. In the present study, we examined the utility of three physiological endpoints in the detection of guilty knowledge with the GKT paradigm and we assessed whether combined indices would improve classification accuracy. It was hypothesized that GSR would perform best as an independent detector of guilty knowledge and that a combined index would enhance classification power over GSR alone.

Finally, although a noteworthy literature suggests that psychopathy tends to moderate deceptive responses (Phinney, 1992; Nunez, Casey, Egner, Hare, & Hirsch, 2005), little is known whether individual differences in healthy populations may play a role. As a secondary purpose (and responding to a dearth of literature on the topic), we

explored potential sociobehavioral moderators of the deceptive response including values, ethnic identity, and resilience.

METHOD

Participants

Forty-two healthy, male active-duty Navy and Marine Corps personnel volunteered to participate in this study. All subjects were awaiting aviation training at Naval Air Station, Pensacola, FL. Exclusion criteria included excessive alcohol consumption (> 3 drinks/day), defective color vision, concurrent ocular pathology, and current diagnosis of heart disease. Inclusion criteria included competence in the English language and permanent residence in United States for at least five years. Subjects were asked to refrain from alcohol consumption and exercise 12 hours prior to data collection, and were asked to provide written confirmation that they have complied with these instructions.

Physiologic Instruments

Galvanic skin response (GSR), electrocardiogram (EKG), and digital blood pressure were recorded concurrently, time marked, and transmitted in real-time to a standard desktop computer via Biopac Systems MP 150 Data Acquisition System (Biopac Systems, Inc. Goleta, CA) at a sampling rate of 62.5 Hz. The NIBP100D-model (Biopac Systems, Inc) non-invasive blood pressure system was used to provide a continuous, beat-to-beat blood pressure signal recorded from the subject's middle and index finger. The system uses a double finger cuff placed on the hand and provides a

continuous blood pressure waveform. High accuracy in comparison to direct (invasive) intra-arterial blood pressure has been demonstrated for this instrument (Fortin et al., 2006; Sackl-Pietsch, 2010).

A sampling rate of 62.5 Hz is considerably lower than usually recommended for EKG analysis; however, data collection at this rate enabled application of a series of custom computer programs (allowing greater control/versatility in performing exploratory analyses), one of which incorporated a simple peak detection algorithm for which the lower rate was found to be optimal. Pre-experiment tests verified that measures obtained with the custom programs (e.g., R-R interval times) did not differ significantly from those obtained using AcqKnowledge with a data collection rate set as high as 1000 Hz.

GSR was measured using Biopac SS3LA Ag/AgCl constant voltage (0.5) electrodes (6.0 mm contact area) attached to the index and middle fingers of the right hand. Attachment sites were scrubbed with a mildly abrasive pad, swabbed with an alcohol prep pad, and then dried with a clean, lint-free gauze pad. A small amount of isotonic gel (Biopac GEL101, 0.5% saline in neutral base) was rubbed into the fingertips, allowing 5 minutes for absorption prior to application of the electrodes. The SS3LA electrode cavities were filled with isotonic gel and the transducers were secured without interrupting normal blood circulation to the finger tips. Signal stability was ensured by applying the GSR electrodes at least 5 minutes prior to initiation of baseline recordings. In addition, a standard deep breath test was performed prior to data collection.

All data were processed with *AcqKnowledge* Software (Biopac Systems, Inc), Microsoft Excel (2003), and MATLAB (Mathworks, Inc., vers. R2009B).

Self Report Instruments

Background Questionnaire

This questionnaire examines basic background, demographic and health information (e.g., age, ethnicity, military occupational specialty) as well as current use of prescription or over-the-counter drugs.

Multigroup Ethnic Identity Measure [MEIM; (Phinney, 1992)].

This scale includes 12 items which measure the extent to which one's ethnic identity is a component in one's self concept. Factor analyses have yielded two subscales, including *ethnic identity search* (a developmental and cognitive component) and *affirmation, belonging, and commitment* (an affective component).

Values Questionnaire (Idler et al., 2003)

This questionnaire is intended to measure ones' value systems, goals, and spirituality. It contains 10 subscales, including *power* (social status, prestige, control or dominance), *achievement* (personal success through demonstration of competence), *hedonism* (pleasure and sensuous gratification), *stimulation* (excitement, novelty and challenge), *self-direction* (independent thought and action choosing), *universalism* (appreciation, tolerance, and welfare for all people), *benevolence* (preservation and enhancement of welfare of people), *tradition* (respect, commitment and acceptance of customs), *conformity* (restraint of actions, impulses and inclinations), and *security* (safety, harmony, and stability of society, relationships and self).

Dispositional Resilience Scale [DRS-15; (Bartone, 1999)]

This 15-item scale includes positively and negatively keyed items and covers three conceptually relevant facets of commitment, control, and challenge. Acceptability internal and test-retest reliability for this scale have been shown.

Procedure

Subjects were asked to perform a mock-crime activity. For this, each subject was instructed to enter a room and open an envelope on a table. Inside of the envelope was an instruction sheet, which directed him to enter a different room occupied by a life-sized male mannequin “victim” dressed in civilian attire. Once in the room, the subject was instructed to assault the victim with a replica pistol until the victim collapsed, and then to search the body and room for specified items. The victim possessed a wallet with a driver’s license, credit card, and currency; as well as a watch and cell phone. Also, on a nearby table there was a folder labeled “confidential” in large red text. The subject was instructed to collect the contents of the wallet, the watch, cell phone and folder, and then to leave the wallet and pistol behind. He was instructed to place the stolen items inside the envelope, close the envelope, and then close the door and exit the room. Each subject was given a total of five minutes to complete the task. The envelope was examined by a member of the research team to ensure that it contained all of the “stolen objects.” Additionally, the mock-crime was recorded by a hidden camera and reviewed by a research team member to confirm that all instructions were followed.

Twenty minutes after completing the mock-crime, subjects were escorted to a climate- and ambient light-controlled, sound-attenuated psychophysiology laboratory for GKT administration. Subjects were first instrumented with equipment and then asked to sit quietly in a comfortable, height-adjustable chair facing a 22-inch computer screen.

Baseline physiological data were then recorded for five minutes. The GKT was administered in a preprogrammed, computerized format with E-Prime software (Psychology Software Tools, Inc., Pittsburgh, PA). A constant background luminance was maintained at 102 ± 2 lux. Subjects were placed 24 inches (61 cm) from the computer monitor and were instructed to minimize movement.

Each subject was informed that he is suspected of committing a crime and that sophisticated equipment is being used to detect whether he is telling the truth or lying. Subjects were instructed, regardless of their innocence or guilt, to give a verbal response of “no” to each item presented (verbal responses are believed to be nonessential to the GKT but have been shown to increase test accuracy (Ben-Shakhar et al., 2003). Thus, subjects told the truth when presented with irrelevant items and lied when presented with the crime-relevant (hereafter referred to as “guilty”) item. Verbal responses were recorded with a high-fidelity microphone and reviewed to ensure compliance with the directions. Ten different questions were presented, each focusing on a different feature of the mock-crime, including: the victim’s stolen driver license, credit card, cell phone, money, office nameplate, shirt, wallet, and watch; as well as the gun used in the crime and the confidential folder stolen from the office. Each question was displayed continuously while five alternate items were displayed for six seconds each. Each item was preceded by a five-second neutral buffer item (a black screen with a white orienting cross). Items included the guilty item, one item selected *a priori* as relevant to another crime that the subject *did not commit* (hereafter referred to as the “innocent” item), and three irrelevant items. The interstimulus interval ranged from 16 to 24 seconds, with a

mean interval of 20 seconds. Order of questions was randomized, as was the order of items within each question.

At the conclusion of the GKT, a computerized multiple-choice recall test was administered. This test consisted of the ten questions given during the GKT, each with five possible answers, including the guilty item, the innocent item, and the three irrelevant items. (If a subject recalled less than 80% of the items, his data were excluded). After the recall test, each subject was asked a series of questions in order to determine the personal relevance of specific GKT items. For example, subjects were asked if they owned a handgun and if they ever lived in any of the states represented by the driver's licenses presented in the GKT. Affirmative answers were then retrospectively compared to the subject's physiological data. If a response exceeded one standard deviation above the mean score for a personally-relevant item, those data were removed from all analyses. Finally, to minimize communication between subjects regarding the study, each subject was asked to sign a statement of non-disclosure indicating that he would not reveal details of the experiment.

Data Processing

Galvanic skin response. Galvanic skin response (Δ GSR) was computed using the maximal change in conductance from 1 to 6 s after stimulus onset (the first second was not analyzed). Baseline (BL) was defined as the average skin conductance over an approximately 0.30 s interval centered on 1 s poststimulus presentation. Δ GSR was defined as the difference between BL and the maximal deflection (either positive or negative) between 1 and 6 seconds. To adjust for overall trends (due to changing baseline eccrine activity) and individual differences in responsivity, the GSR signals were

detrended using MATLAB (linear method with breakpoints set to coincide with the beginning and end of each GKT item presentation) and then transformed to within-subjects standard Z-scores – computed relative to the mean and standard deviation of each subject’s response distribution to all GKT items.

Heart rate response. R-wave peaks in the EKG data were detected using a custom VBA program. R–R intervals were then calculated and converted to instantaneous HR in beats per minute. The last instantaneous HR prior to item presentation served as the prestimulus baseline. The prestimulus baseline value was subtracted from each poststimulus instantaneous HR, giving a series of poststimulus difference scores (Δ HR). The HR endpoint was defined as the average of all Δ HR values within 11 seconds from stimulus onset (Verschuere et al., 2007).

EKG artifact. Due to minimal subject movement, the EKG signal suffered very little noise contamination. Occasionally, however, an R-wave failed to cross the peak threshold that was set in the peak detection algorithm according to individual subject EKG traces. In these cases peak time was determined by graphical inspection. Abnormal heart beats (e.g., ectopic beats) were rare and immediately obvious by inspection of tachograms. At the corresponding time in the raw EKG signal, the abnormal beat (spurious event marker) was removed and an “artificial peak” (a new event marker) was placed halfway between the surrounding normal beats. R-R intervals were then recalculated. Instantaneous HR values were thereby obtained from normal heartbeats.

Finger pulse line length. Our approach was patterned after that of Vandenberg and colleagues with some minor differences. The measurement window for FPLL began at stimulus onset and lasted approximately 11 seconds. Since line length is

disproportionately affected by the measurement's starting point (Vandenbosch et al., 2009; Elaad et al., 2006), we used thirteen 9-second windows, each beginning 0.16 seconds later than the previous one ($13 \times 0.016s \approx 2s$), defining FPLL for a given GKT item as the mean of these 13 length measurements. Each 9-second FPLL was calculated with the following formula:

$$FPLL = \sum_{i=1}^n \sqrt{\Delta x^2 - \Delta y_i^2}$$

where Δx is the constant time difference between two consecutive data points; Δy is the difference in magnitude (of blood pressure readings) between two consecutive data points; and n is the number of data points (after smoothing) in the time period under investigation (i.e., 9 seconds).

Data Reduction and Analyses

Calculation of composite endpoints. As noted earlier, the ΔHR and $\Delta FPLL$ endpoints were expected to indicate guilt by lower rather than higher scores. Therefore, these measures were each subtracted from ΔGSR to generate respective composite scores (See (Elaad et al., 2006)).

Lykken's scoring procedure. A common scoring procedure (Lykken, 1959; Ben-Shakhar, Bar-Hillel, & Kremnitzer, 2002) was used to classify guilt and innocence. According to this procedure, responses of each subject to all of the items within each question are rank-ordered. If the crime-relevant item elicits the strongest response (i.e., largest positive GSR deflection, largest mean HR deceleration, largest decrease in mean FPA or FPLL), a score of 2 is assigned to the question; if it elicits the second strongest response, a score of 1 is assigned. Otherwise a score of 0 is assigned. These scores are

then summed across all ten questions to provide a single detection score for each endpoint. Thus, the detection score ranged from 0 to 20 and a cutoff score of 10 was set. Specifically, a detection score of at least 10 was needed to reach a “guilty” classification for any given item.

ROC analysis. Since accuracy rates calculated using the Lykken method depend on a single arbitrary cut point, we also employed ROC analyses. This approach has been used frequently in GKT studies (Vandenbosch et al., 2009; Gamer et al., 2008; Elaad et al., 2006) and is recommended by the National Research Council (2003) as a relevant method for describing the diagnostic value of polygraph tests. Following signal detection theory, detection efficiency of the ROC curve is described as the degree of separation between the distributions of responses to guilty and irrelevant items. For this purpose, distributions of mean *Z* scores computed for each subject across all guilty items and across all irrelevant items for each physiological endpoint and each composite endpoint were plotted. Based on these distributions, areas under the ROC curves and their corresponding 95% confidence intervals were computed. This area, then, reflects detection efficiency across all possible cut points. It assumes values between 0 and 1, such that an area of 0.5 means that the two distributions (that is, of the mean *Z* scores of the guilty and irrelevant items) are undifferentiated.

Data were analyzed using SPSS software Version 16 (SPSS, Inc., Chicago, IL) Descriptive analyses were conducted (Table 1), after which mean responses for each physiological endpoint were calculated for the five items within each question. Mean responses for the three irrelevant items were collapsed into a single composite variable. Repeated-measures analyses of variance (ANOVA) with Bonferroni-corrected post hoc

paired *t* tests were then used to examine differences between means of the guilty item, the innocent item, and the composite irrelevant item, respectively. Next, Lykken classification scores were computed per the method described above, followed by calculation of percent correct classifications for the guilty and innocent items, respectively. ROC curves were then constructed for each endpoint (GSR, HR, and FPLL); and areas under the ROC curves along with corresponding 95% confidence intervals were computed (Bamber, 1975). Finally, correlational analyses were used to explore potential moderating effects of sociobehavioral factors including values, ethnic identity and resilience. Specifically, Pearson product moment correlations were performed between each total score, subscale, GSR, HR, and FPLL. For these analyses, GSR, HR, and FPLL were calculated as [Item 5 minus (mean of items 1-3)]. Bonferroni corrections were not applied for this exploratory component.

RESULTS

Recall Test. The effectiveness of the GKT may depend on the ability of the individual to remember critical details of a crime (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Gamer, Kosiol, & Vossel, 2010). In this study, recall was high (mean \pm SD 86.1 \pm 16.4%) and was consistent with recall data from previous studies (Gamer, Rill, Vossel, & Godert, 2006).

Subject characteristics are shown in Table 1 and comparisons of mean physiological responses to the guilty, innocent, and irrelevant items for each physiological endpoint are shown in Figure 1. The repeated measures ANOVA on Δ GSR

revealed a significant overall effect ($F = 12.2$, $p < 0.001$, partial $\eta^2 = 0.24$); post-hoc comparisons showed that the mean ΔGSR to the guilty item exceeded that of the innocent item ($p < 0.017$) as well as the composite irrelevant item ($p < 0.017$) (Figure 1A). As hypothesized, no differences were shown between the innocent item and the composite irrelevant item ($p > 0.017$). The repeated measures ANOVA on ΔHR (Figure 1B) displayed a substantial overall trend ($F = 3.1$, $p = 0.06$, partial $\eta^2 = 0.07$). Follow up comparisons confirmed that the mean ΔHR to the guilty item exceeded that of the composite irrelevant item ($p < 0.017$). The observed difference between the guilty item and the innocent item, although noteworthy, did not reach statistical significance ($p > 0.017$). Mean ΔHR of the innocent item, as predicted, did not differ from the combined irrelevant item ($p > 0.017$) (Figure 1B). The repeated measures ANOVA on ΔFPLL revealed a similar overall trend but did not reach statistical significance (Figure 1C).

Rates of correct classification based on the Lykken procedure are presented in Table 2. As expected, GSR performed best with the Lykken procedure, correctly identifying guilt and innocence in 47.6% and 90.5% of subjects, respectively.

The areas under the ROC curves and respective 95% confidence intervals computed for each individual and composite endpoint are displayed in Table 3. ROC curves are plotted in Figure 2. The ROC analyses revealed that GSR was the best independent classifier (Area Under Curve = 0.79, $p < 0.001$) and that two combined endpoints (GSR – FPLL and GSR – HR - FPLL) improved classification accuracy over GSR alone. The GSR – FPLL combined endpoint performed best (Area Under Curve = 0.85, $p < 0.001$). Since seven subjects' data were not useable for the FPLL analysis, the ROC analyses were repeated with a subgroup of subjects for which all endpoint data were

available ($n = 35$), thus permitting a more direct comparison of methods. This adjustment did not result in substantial changes in the AUC values, which suggests robustness of the findings.

Sociobehavioral Moderators. The values subscale *security* (emphasizing safety, harmony and stability of society, relationships, and self) associated with the GSR response such that higher scores related to greater GSR responses to the guilty item (versus mean of the three innocent items) ($r = .39, p < .01$). Similar nonsignificant trends were also observed between ethnic identity and the GSR response (Total MEIM: $r = .24, p = .14$; *affirmation/belonging*: $r = .22, p = .17$; *ethnic identity search* $r = .21, p = .18$). When this analysis was subsequently restricted to Caucasian subjects ($n = 37$), similar effects were observed (Total MEIM: $r = .22, p = .19$; *affirmation/belonging*: $r = .20, p = .17$; *ethnic identity search* $r = .21, p = .22$).

DISCUSSION

The present study was designed to compare the validity of GSR, HR, and FPLL in the detection of deception using the GKT and to determine if combined endpoints improve classification power over the best individual endpoint. As hypothesized, GSR was the best individual performer and combined endpoints improved classification power over GSR alone. These findings confer proof-of-concept for this capability and serve as a crucial prerequisite for performing cross-cultural and sociobehavioral comparisons.

Consistent with our prediction, GSR performed best of the individual endpoints considered in this study, and robustness of this finding was evidenced across three

different analytic approaches. As discussed earlier, GSR is a marker of sympathetic nervous stimulation, is responsive to psychological stress, and is the most extensively studied GKT endpoint. MacLaren's (2001) meta-analysis of 22 GKT studies using the GSR endpoint found that 76% of subjects with guilty knowledge were correctly identified, while innocent subjects were correctly identified 83% of the time. Also, Ben-Shakhar et al. (2003) reported overall effects sizes of 1.55 with the GSR endpoint, with much higher effects achieved under optimal experimental conditions. In the present study, use of the GSR endpoint in conjunction with the Lykken method correctly classified guilt far less consistently than that reported in these meta-analyses (47.6%), but innocence was detected rather well (90.5%). The obvious implication is that, under the conditions of the current study the GSR endpoint has a high likelihood of correctly classifying a person who is innocent of a crime (i.e., true negative) and, by extension, a very low likelihood of incorrectly classifying the innocent individual as guilty (i.e., false positive). However, if an individual is guilty of a crime, there is a distinct possibility that he would be incorrectly classified as innocent (i.e., false negative). As described earlier, the Lykken method is inherently limited in that it is based on a single, arbitrary cutpoint. To address this, ROC analyses were conducted to examine test performance across all possible cutpoints. With an area under the curve of 0.79, the GSR endpoint performed reasonably well, suggesting that for any given false positive rate the true positive rate (i.e., "hits") tended to be fairly high. This is evidenced by a steep slope at the left-hand side of the curve in Figure 2. However, higher areas under the curve (suggesting better classification accuracy across all cutpoints) have been reported for GSR in recent studies

of similar design. Gamer et al. (2008), for example, reported an area under the curve of 0.86 for GSR and Vandenbosch et al. (2009) reported area under the curve of 0.83.

As hypothesized, combined endpoints enhanced classification accuracy over the best single measure (GSR). Specifically, whereas the GSR endpoint produced an area under the curve of 0.79, GSR – FPLL improved the value to 0.85. Interestingly, although GSR – HR – FPLL improved classification accuracy of GSR alone, it did not perform as well as GSR – FPLL (0.83 versus 0.85). Some studies have explored the validity of combined indices (Ben-Shakhar et al., 2002; Ben-Shakhar et al., 1996; Elaad et al., 1992). Elaad et al. (1992) correctly classified 75% of subjects using GSR and respiration line length respectively, but increased classification accuracy to 85% using a combined model. Also, Gamer et al. (2006) assessed the combination of GSR, heart rate, and respiration line length in a GKT, correctly classifying 93% of guilty and 97% of the innocent examinees. These scientists subsequently found that a weighted combination of heart rate and respiration indices enhanced detection validity over GSR alone (Gamer et al., 2008). Other studies, however, have been unsuccessful in demonstrating such improvements (Verschuere et al., 2007; Bradley et al., 1984). Undoubtedly, much remains to be learned regarding which measures comprise an optimal composite index of guilty knowledge. That said, such a composite will likely reflect coactivation of the sympathetic and parasympathetic nervous systems which are believed to underlie both the orienting and fight-or-flight responses. As described earlier, GSR is believed to be sympathetically-modulated, while the HR deceleration response is thought to be primarily parasympathetically-modulated. Interestingly, FPLL changes are sensitive to both sympathetically-modulated decreases in finger pulse amplitude and

parasympathetically-influenced HR deceleration. Other promising endpoints include respiration, pupil diameter changes, electroencephalographic characteristics and selected attributes of the EKG waveform (Furedy et al., 1992). More recently, advances in functional magnetic resonance (fMRI) have been harnessed to detect deception with apparent success (Kozel et al., 2005).

Although most of the current findings were statistically and operationally significant, classification accuracy did not compare favorably in all instances with several recent studies involving mock crime, the GKT and similar physiological endpoints. There is a broad spectrum of possible explanations for this. To begin, variations in equipment, scoring techniques, study design and methodology almost certainly contribute to these discrepancies. For example, many GKT studies utilize a mean of two presentations of each item for any given physiological endpoint. The crime-relevant gun, for instance, would be presented twice within the GKT and the two presentations would then be averaged. In the present study each item was displayed only once which precludes a reliability analysis. On the other hand, this eliminates the risk of habituation effects observed in previous studies – particularly with respect to GSR (Verschuere et al., 2005; Eyal et al., 2006). Also, we did not include an innocent comparison group. Rather, we combined guilty and innocent conditions in a within-subjects design. That is, in addition to displaying items relevant to the crime that the individual *did commit* (i.e., the “guilty item”), another set of ten items was identified a-priori as relevant to a *separate crime that the individual did not commit* (i.e., the “innocent” item). This approach has at least two strengths. First, it permits within-subjects comparison of guilty and innocent conditions which is inherently more powerful. Second, it enhances ecological validity for cases in

which an individual may be guilty of a crime, but *not the crime for which he is being questioned*. This is a plausible scenario when questioning members of organized crime or terror networks. Finally, we calculated standardized GSR scores relative to the mean and standard deviation of all items. By contrast, some authors have standardized GSR scores within each item (Ben-Shakhar, 1985), while others have standardized GSR scores relative to the mean and standard deviation of the irrelevant items (Gamer et al., 2008). We posit that standardizing based on the entire set of items is the most conservative approach since it makes the fewest assumptions regarding expected population differences between the guilty and innocent items. Also, we felt that using the mean and standard deviations within each item would result in too few data points from which to extract a meaningful standardized score. To facilitate comparison and integration of future research, we recommend standardization of data processing methods for key physiological endpoints (e.g., GSR, HR, FPLL, and respiration line length), perhaps via a position statement of best practices reflecting consensus of leading scientists in the field. (See also Gamer et al., 2008 who make similar recommendations). Finally, it is important to note that this subject population is young, healthy, intelligent, and physically fit. Additionally, this group is also generally thought to be more stress-tolerant than the general population – even when matched for age, education, or socioeconomic status. Thus, it is quite possible that this subject pool may differ in physiological reactivity to mild stress and/or responses to guilty knowledge.

Although the extant literature generally suggests that psychopathy influences deceptive responses (Fullam, McKie, & Dolan, 2009; Nunez et al., 2005), little is known whether individual differences in healthy populations may play a role. Interestingly, the

exploratory analyses of sociobehavioral moderators of deception suggested that an individual's values may influence reactivity to the GKT. Specifically, subjects who valued *security* (i.e., safety and stability) demonstrated more discernable galvanic skin reactivity to the guilty items compared to innocent items on the GKT (i.e., were more easily detected). Similar nonsignificant patterns emerged for ethnic identity – a finding that persisted when analyses were confined to a Caucasian subgroup. Since the requirement to detect deception across cultures is implicit to the Global War on Terror, a better understanding of the influence of individual differences in deceptive behavior within healthy populations (e.g., personality, ethnicity, ethnic identity, values and spirituality) is of paramount importance.

The present GKT study most resembles the so-called “realistic” versus “optimal” experimental condition (See Carmel et al., 2003). Under the optimal condition, participants are given motivational instructions, asked to provide verbal deceptive responses, and the GKT includes at least five questions. Under this condition, subjects are also reminded of relevant crime details in advance of the mock-crime to ensure that guilty subjects take notice of all relevant details. Also, guilty subjects are presented with these details just before the GKT is administered (Carmel et al., 2003), typically via instructions specifying all of the relevant items. With this approach, data from subjects who could not recall some of the GKT items are often discarded or reanalyzed adjusting for the missing items (Ben-Shakhar & Gati, 1987). Although the present study included a verbal deception response and contained more than five questions, we were unable to offer a reward to enhance motivation. Some studies show that GKT accuracy increases under conditions of heightened motivation (Elaad & Ben-Shakhar, 1989; Ben-Shakhar et

al., 2003), while other studies have failed to obtain such an effect (Furedy & Ben-Shakhar, 1991). Also, subjects were not reminded of relevant details in advance of the mock-crime or presented with these details prior to GKT administration, and recall performance was not a criterion for exclusion. Mean recall performance, however, was quite good (88.1%) and was consistent with prior studies of similar design. Subjects in this study were instructed to say “no” to every item displayed in the GKT. Thus, subjects actively deceived when presented with the guilty item and told the truth when presented with the innocent items. Active deception may enhance validity of the GKT (Ben-Shakhar et al., 2003), but further research is needed to confirm this. In sum, the current study design is best classified a realistic rather than optimal, which may partially explain the lower classification accuracy than has been achieved in similar studies.

In this study we compared different physiological approaches to detect deception with the GKT. Presentation of guilty items resulted in higher GSR responses than irrelevant items, and this endpoint also performed best with the Lykken approach. ROC analyses showed that combined measures provide more classification power than GSR alone. Although these findings confer proof-of-concept for this capability and satisfy the prerequisite for performing cross-cultural and sociobehavioral comparisons, additional research with consistent methodologies is needed not only to define an optimal composite index to detect deception but also to elucidate individual differences governing this phenomenon. Given the implications for national security, maximal classification accuracy is necessary to warrant use of the GKT in the real world.

Acknowledgments: This work was supported by the Office of Under Secretary of Defense Biosystems Department and the Office of Naval Research Expeditionary Maneuver Warfare and Combating Terrorism Science and Technology Department. We would like to express sincere appreciation to N. Summer Dodson for her analytical expertise.

Table 1**Participant Characteristics**

Characteristic	N (%)	Mean (SE)	Range
Age (years)	42	24.0 (0.4)	22.0-30.0
Body Mass Index (kg/m ²)	42	25.2 (0.4)	21.2-32.2
Years of Military Service	42	2.3 (0.5)	0.0-13.0
Education			
College graduate	42 (100.0%)		
Unreported	0 (00.0%)		
Combat Experience			
Yes	5 (12.2%)		
No	36 (87.8%)		
Ethnicity			
Caucasian	37 (88.1%)		
Hispanic	2 (4.8%)		
African American	1 (2.4%)		
Mixed ethnicity	1 (2.4%)		
Asian	1 (2.4%)		
Handedness			
Left	7 (16.7%)		
Right	35 (83.3%)		

Table 2**Detection rates computed for three physiological endpoints and composite endpoints**

Endpoint	Number (%) correct guilty classifications	Number (%) correct innocent classifications
GSR	20/42 (47.6%)	38/42 (90.5%)
HR	9/42 (21.4%)	39/42 (92.9%)
FPLL	7/35 (20.0%)	34/35 (97.1%)
GSR – HR	12/42 (28.6%)	39/42 (92.9%)
GSR – FPLL	12/35 (34.3%)	35/35 (100%)
GSR – HR – FPLL	14/35 (40.0%)	33/35 (94.3%)

Note: GSR = Galvanic skin response; HR = heart rate; FPLL = finger pulse line length. Number and percent correct guilty classification refers to guilty classification based on Lykken scoring for Item #5 (the “guilty”) item. Number and percent correct innocent classification corresponds to innocent classification based on Lykken scoring for Item #4 (the *a-priori* selected item associated with a crime for which the individual had no knowledge).

Table 3

Area under the ROC curves and related statistics computed for three physiological endpoints and composite endpoints

Endpoint	Number of guilty items	Number of innocent items	Area	Asymptotic sig*	Asymptotic 95% Confidence Interval	
					Lower Bound	Upper Bound
GSR	42	42	0.79	.001	0.69	0.89
HR	42	42	0.65	.020	0.53	0.77
FPLL	35	35	0.72	.002	0.60	0.83
GSR – HR	42	42	0.75	.001	0.65	0.86
GSR – FPLL	35	35	0.85	.001	0.76	0.93
GSR – HR – FPLL	35	35	0.83	.001	0.74	0.93

Note: n=42; GSR = Galvanic skin response; HR = heart rate; FPLL = finger pulse line length. * Null hypothesis: true area = 0.5

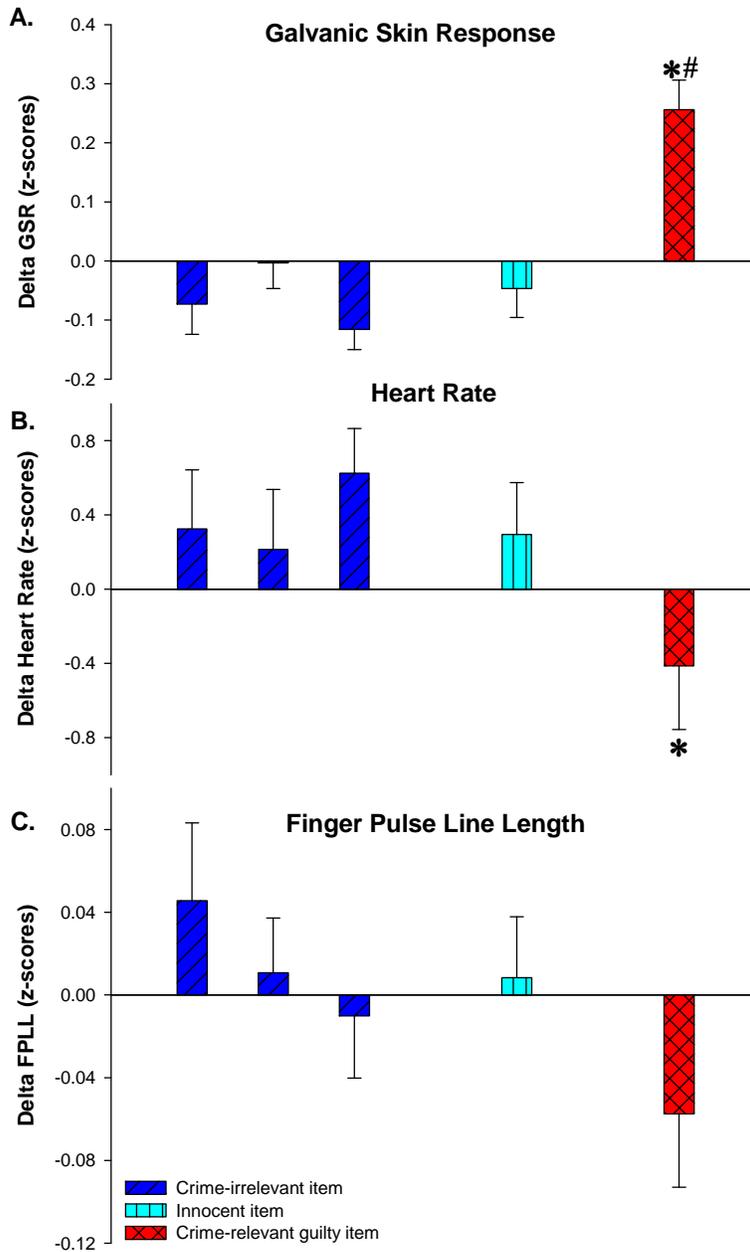


Figure 1. Physiological Responses to the Guilty Knowledge Test. A. Galvanic Skin Response. B. Heart Rate. C. Finger Pulse Line Length. Results are expressed as the mean \pm SEM. Crime-relevant guilty item (red checked bars), innocent item (cyan vertical line bars), and crime-irrelevant item (blue horizontal line bars). * different from composite of three irrelevant items ($p < 0.017$); # different from innocent item ($p < 0.017$).

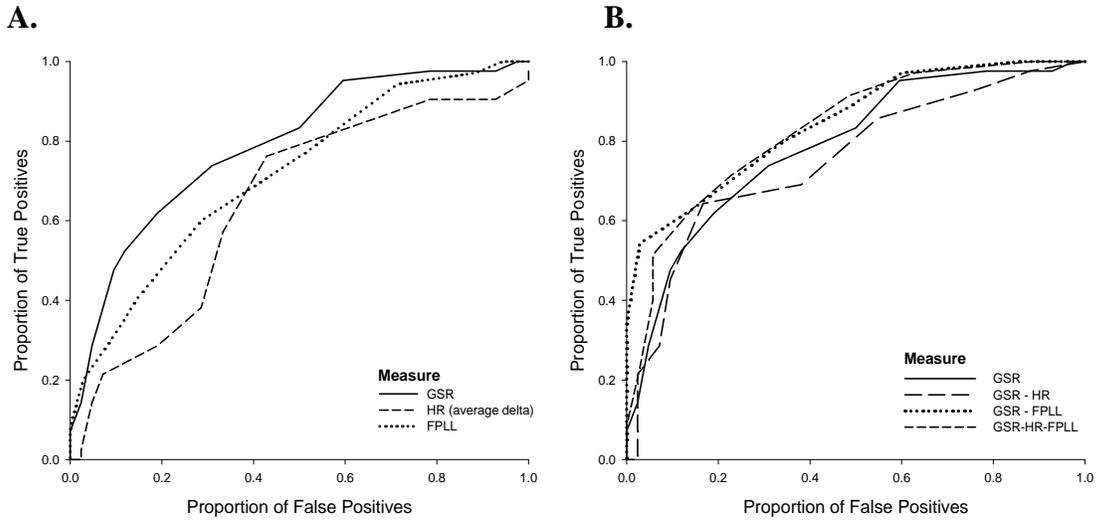


Figure 2. Receiver Operating Characteristic Curves. A. Contrast of distributions of z-standardized response differences in the galvanic skin response (GSR), heart rate (HR), and finger pulse line length (FPLL) between guilty and innocent Conditions. B. Contrast of distributions of z-standardized response differences in GSR and three combined measures.

Reference List

- 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.
- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387-415.
- Bartone, P. T. (1999). Hardiness protects against war-related stress in Army Reserve forces. *Consulting Psychology Journal: Practice and Research*, *51*, 72-82.
- Ben-Shakhar, G. (1985). Standardization within individuals: a simple method to neutralize individual differences in skin conductance. *Psychophysiology*, *22*, 292-299.
- Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: reconsidering the use of the guilty knowledge technique in court. *Law Hum.Behav.*, *26*, 527-541.
- Ben-Shakhar, G. & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: effects of mental countermeasures. *J.Appl.Psychol.*, *81*, 273-281.
- Ben-Shakhar, G. & Eyal, E. (2002). Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: a reexamination. *J.Appl.Psychol.*, *87*, 972-977.

- Ben-Shakhar, G. & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. *J.Appl.Psychol.*, 88, 131-151.
- Ben-Shakhar, G. & Gati, I. (1987). Common and distinctive features of verbal and pictorial stimuli as determinants of psychophysiological responsivity. *J.Exp.Psychol.Gen.*, 116, 91-105.
- Bradley, M. T. & Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. *Psychophysiology*, 21, 63-71.
- Bradley, M. T. & Janisse, M. P. (1981). Accuracy demonstrations, threat, and the detection of deception: cardiovascular, electrodermal, and pupillary measures. *Psychophysiology*, 18, 307-315.
- 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.
- Elaad, E. & Ben-Shakhar, G. (1989). Effects of motivation and verbal-response type on psychophysiological detection of information. *Psychophysiology*, 26, 442-451.
- Elaad, E. & Ben-Shakhar, G. (2006). Finger pulse waveform length in the detection of concealed information. *Int.J.Psychophysiol.*, 61, 226-234.
- Elaad, E. & Ben-Shakhar, G. (2008). Covert respiration measures for the detection of concealed information. *Biol.Psychol.*, 77, 284-291.

- Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *J.Appl.Psychol.*, *77*, 757-767.
- Fortin, J., Marte, W., Grullenberger, R., Hacker, A., Habenbacher, W., Heller, A. et al. (2006). Continuous non-invasive blood pressure monitoring using concentrically interlocking control loops. *Comput.Biol.Med.*, *36*, 941-957.
- Fullam, R. S., McKie, S., & Dolan, M. C. (2009). Psychopathic traits and deception: functional magnetic resonance imaging study. *Br.J.Psychiatry*, *194*, 229-235.
- 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.
- Furedy, J. J., Heslegrave, R. J., & Scher, H. (1992). T-wave amplitude utility revisited: some physiological and psychophysiological considerations. *Biol.Psychol.*, *33*, 241-248.
- Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Actions Test. *Biol.Psychol.*, *83*, 101-107.
- Gamer, M., Rill, H. G., Vossel, G., & Godert, H. W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *Int.J.Psychophysiol.*, *60*, 76-87.
- Gamer, M., Verschuere, B., Crombez, G., & Vossel, G. (2008). Combining physiological measures in the detection of concealed information. *Physiol Behav.*, *95*, 333-340.

- Idler, E. L., Musick, M. A., Ellison, C. G., George, L. K., Krause, N., Ory, M. G. et al. (2003). Measuring multiple dimensions of religion and spirituality for health research: Conceptual background and findings from the 1998 General Social Survey. *Research on Aging, 25*, 327-365.
- Kleinmuntz, B. & Szucko, J. J. (1984). Lie detection in ancient and modern times. A call for contemporary scientific study. *Am.Psychol., 39*, 766-776.
- Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biol.Psychiatry, 58*, 605-613.
- Kozel, F. A., Laken, S. J., Johnson, K. A., Boren, B., Mapes, K. S., Morgan, P. S. et al. (2009). Replication of Functional MRI Detection of Deception. *Open.Forensic Sci.J., 2*, 6-11.
- Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385-388.
- MacLaren, V. V. (2001). A quantitative review of the guilty knowledge test. *J.Appl.Psychol., 86*, 674-683.
- National Research Council (2003). *The polygraph and lie detection*. Washington, DC.: The National Academies Press.

- Nunez, 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.
- Phinney, J. S. (1992). The multigroup ethnic identity measure: A new scale for use with diverse groups. *Journal of Adolescent Research*, *7*, 156-176.
- Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, *15*, 344-359.
- Sackl-Pietsch, E. (2010). Continuous non-invasive arterial pressure shows high accuracy in comparison to invasive intra-arterial blood pressure measurement. Unpublished manuscript.
- Sokolov, E. N. (1990). The orienting response, and future directions of its development. *Pavlov.J.Biol.Sci.*, *25*, 142-150.
- Vandenbosch, K., Verschuere, B., Crombez, G., & De, C. A. (2009). The validity of finger pulse line length for the detection of concealed information. *Int.J.Psychophysiol.*, *71*, 118-123.
- Verschuere, B., Crombez, G., De, C. A., & Koster, E. H. (2005). Psychopathic traits and autonomic responding to concealed information in a prison sample. *Psychophysiology*, *42*, 239-245.
- Verschuere, B., Crombez, G., Koster, E. H., & De, C. A. (2007). Antisociality, underarousal and the validity of the Concealed Information Polygraph Test. *Biol.Psychol.*, *74*, 309-318.

REPORT DOCUMENTATION PAGE

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB Control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD MM YY) 30 12 10	2. REPORT TYPE Technical Report	3. DATES COVERED (from – to) 01 04 09 – 01 05 10
--	---	--

4. TITLE A Comparison of Approaches to Detect Deception	5a. Contract Number: 5b. Grant Number: 5c. Program Element Number: 5d. Project Number: 5e. Task Number: 5f. Work Unit Number: PB802
---	--

6. AUTHORS Marcus K. Taylor, Dain S. Horning, Joseph F. Chandler, Jeffrey B. Phillips, Jasmine Y. Khosravi, Jill E. Bennett, Heather Halbert, Benedict J. Fern, Hong Gao	
--	--

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Aerospace Medical Research Laboratory 280 Fred Bauer Street Building 1811 Pensacola, FL 32508	8. PERFORMING ORGANIZATION REPORT NUMBER NAMRL 11-05
--	--

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES) Office of Naval Research One Liberty Center 875 N. Randolph Street, Suite 1425 Arlington, VA 22203-1995	10. SPONSOR/MONITOR'S ACRONYM(S) ONR 11. SPONSOR/MONITOR'S REPORT NUMBER(s)
---	---

12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT A promising method to detect deception is the Guilty Knowledge Test (GKT) which assesses whether an individual possesses knowledge about a particular crime. Specifically, the GKT involves a series of questions with multiple answers where one answer is relevant to the crime and the others are not. In theory, if a “suspect” is guilty, s/he will recognize the crime-relevant item and display a physiological orienting response that is discernable from responses to irrelevant items. The most widely studied physiological endpoint in conjunction with the GKT is the galvanic skin response (GSR) which reflects activity of the eccrine sweat gland and represents sympathetic modulation of the autonomic nervous system. Little is known regarding the validity of alternate physiological endpoints or if combined endpoints enhance detection accuracy over GSR alone. In this study we compared different physiological approaches to detect deception with the GKT. Secondly, we explored sociobehavioral moderators of deception, including values, ethnic identity and resilience. Forty-two military men (age 23.9 ± 0.4 years) participated in a mock-crime and then completed a 10-question GKT. Endpoints included GSR, heart rate (HR; reflecting vagal modulation), and finger pulse line length (FPLL; a blood pressure waveform calculation reflecting combined sympathetic-parasympathetic modulation). Separate one-way repeated measures ANOVA with five levels compared mean physiological responses to each question. A common scoring procedure (Lykken, 1959) was applied to classify subjects as guilty or innocent for each physiological endpoint. ROC curves were then constructed to assess the diagnostic value of each endpoint and to determine whether combined measures detected guilt more effectively than GSR alone. Lastly, correlational analyses were used to explore sociobehavioral factors influencing the GKT response. As hypothesized, presentation of “guilty” items resulted in higher GSR responses (p < 0.001) than irrelevant items. Similar nonsignificant patterns emerged for HR and FPLL. GSR performed best with the Lykken approach, correctly identifying guilt and innocence in 47.6% and 90.5% of subjects, respectively. ROC analyses revealed that all three individual endpoints performed better than chance (p < 0.05) and that a combined index (GSR + FPLL) enhanced classification power over GSR alone (Area Under Curve: 85.0% versus 79.0%). Although GSR appears to be the most valid individual GKT endpoint of those considered in the present study, a combined index improves classification power over the best individual endpoint. Finally, exploratory analyses suggested that sociobehavioral factors may moderate the human response to guilty knowledge. The current findings confer proof-of-concept for this capability and satisfy a crucial prerequisite for cross-cultural and sociobehavioral comparisons.

15. SUBJECT TERMS Guilty Knowledge Test, deception detection, galvanic skin response, blood pressure, counterterrorism, law enforcement

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UNCL	18. NUMBER OF PAGES 37	18a. NAME OF RESPONSIBLE PERSON Officer in Charge	
a. REPORT UNCL	b. ABSTRACT UNCL	c. THIS PAGE UNCL			18b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (850)452-3486	