Journal of Applied Psychology
2005, Vol. 90, No. 1, 147–158

0021-9010/05/$12.00 DOI: 10.1037/0021-9010.90.1.147

Behavioral and Physiological Measures in the Detection of Concealed Information

Nurit Gronau, Gershon Ben-Shakhar, and Asher Cohen
The Hebrew University of Jerusalem

The authors examined the incremental validity of the reaction time (RT) measure beyond that of skin conductance response (SCR) in the detection of concealed information. Participants performed a Stroop-like task in which they named the color of critical and neutral words. Results show that the SCR highly differentiated between the relevant and neutral words. However, the RT demonstrated a significant differentiation only when the critical words denoted personally significant items (e.g., one's own name) and not when they denoted crime-relevant items related to a simulated crime. In both cases, combining the 2 measures yielded no advantage over the use of SCR alone. Thus, although behavioral measures may differentiate between relevant and neutral information in some cases, their practical use is questionable.

Scientists and forensic experts have attempted for many years to develop methods for the purpose of detecting concealed information. Several polygraph techniques assessing physiological responses have been proposed since the beginning of the 20th century for the detection of concealed information (for reviews, see Ben-Shakhar & Furedy, 1990; Lykken, 1998; Raskin, 1989; Reid & Inbau, 1977; Saxe, Dougherty, & Cross, 1985). In particular, two of these techniques have been a major focus of research, discussion, and debate. The Control Question Test (CQT) is the most widely used method of psychophysiological detection, particularly in North America. Although this method is being used extensively for forensic purposes as well as for security screening, it has been severely criticized in the scientific literature (e.g., Ben-Shakhar, 2002; Ben-Shakhar & Furedy, 1990; Iacono & Lykken, 2002; Lykken, 1974, 1998; National Research Council, 2003; Saxe & Ben-Shakhar, 1999; but see also Podlesny & Raskin, 1977; Raskin, 1986, 1989; Honts, Raskin, & Kircher, 2002). The main criticism of this method concerns the use of improper control questions, which enhance the risk for false-positive errors (i.e., innocent suspects classified as guilty).

The second method of psychophysiological detection, known as the Concealed Information Test (CIT), or the Guilty Knowledge Test (GKT), has drawn considerable attention among researchers but has been extensively applied only in Japan (Fukumoto, 1980; Nakayama, 2002; Yamamura & Miyata, 1990). In contrast to the CQT, there is a general consensus that the CIT relies on sound theoretical principles and proper controls, and, therefore, it satisfies the necessary requirements of an objective test (Ben-Shakhar & Elaad, 2002b; Ben-Shakhar & Furedy, 1990; Lykken, 1974, 1998). The CIT contains a series of multiple-choice questions, each having one relevant alternative (e.g., a feature of the crime under investigation) and several neutral (control) alternatives, chosen so that an innocent suspect would not be able to discriminate them from the relevant alternative (Lykken, 1998). Typically, if the suspect's physiological responses to the relevant alternative are consistently larger than to the neutral alternatives, knowledge about the event (e.g., crime) is inferred. As long as information about the event has not leaked out, the probability that an innocent suspect would show consistently larger responses to the relevant than to the neutral alternatives depends only on the number of questions and the number of alternatives per question. Hence, the probability of false-positive outcomes can be controlled such that maximal protection for the innocent subjects is provided.

The CIT has been extensively researched during the past 3 decades, demonstrating high levels of validity (see reviews in Ben-Shakhar & Elaad, 2003; Ben-Shakhar & Furedy, 1990; MacLaren, 2001). The most commonly used measure in CIT studies is the skin conductance response (see Ben-Shakhar & Elaad, 2003). However, the use of other physiological measures in addition to the skin conductance response (SCR) significantly increases detection accuracy in the CIT (e.g., Cutrow, Parks, Lucas, & Thomas, 1972; Elaad & Ben-Shakhar, 1997; Kugelmass & Lieblich, 1968). For example, several studies reported optimal detection rates of "guilty" (informed) and "innocent" (uninformed) subjects with a combined measure of the SCR and the respiration line length (e.g., Ben-Shakhar & Dolev, 1996; Ben-Shakhar, Gronau, & Elaad, 1999; Elaad & Ben-Shakhar, 1997; Timm, 1982, 1987). Similarly, maximal differentiation between responses to relevant and neutral items was found when the SCR was combined with measures such as eye blink rate and finger-pulse volume (Cutrow et al., 1972).

Whereas nearly all CIT studies have used physiological measures, the present study aimed to investigate whether behavioral indices such as reaction time (RT) can also serve as valid measures

Nurit Gronau, Gershon Ben-Shakhar, and Asher Cohen, Department of Psychology, The Hebrew University of Jerusalem.

This research was supported by the Israel Foundations Trustees. We thank Ori Cohen, Yael Errera, Avia Munchik, and Alona Roded for their help in the data collection.

Correspondence concerning this article should be addressed to Nurit Gronau, who is now at the Nuclear Magnetic Resonance Center at Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, or to Gershon Ben-Shakhar, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem 91905, Israel. E-mail: gronau@nmr.mgh.harvard.edu or mskpugb@mscc.huji.ac.il

in the CIT and whether they contribute to information detection when combined with the SCR. On the one hand, RT measurement is costless and can be easily assessed—and therefore may be highly useful in interrogative contexts. On the other hand, behavioral measures such as RT tend to be more voluntarily controlled than autonomic-based physiological measures and thus may be more susceptible to strategic manipulations.

Previous studies investigating the validity of the RT measure in the detection of concealed information have yielded inconsistent results. For instance, Seymour, Seifert, Shafto, and Mosmann (2000) adopted the methodology used by Farwell and Donchin (1991) and instructed participants to commit a mock crime characterized by a specific set of crime-relevant items. After memorizing (independently of the mock crime) a list of target words, the participants participated in a recognition test, in which they were required to press a certain key whenever they encountered a learned target word and press a different key whenever a nontarget word was presented. Most critically, crime-relevant items were embedded among the nontarget words that were to be rejected. Participants were urged to respond as quickly as possible and received a negative feedback ("Too slow") when responses were over 1,000 ms in length.

Results showed that participants were significantly slower to reject crime-relevant items than other (neutral) nontarget words. Furthermore, when informed about the expected pattern of results (i.e., slower RTs for crime-related words) and explicitly asked to avoid detection of the crime-relevant information (though still required to respond quicker than 1,000 ms), participants continued to show slower latencies to the critical items. Thus, RT was found to be a viable measure in the detection of concealed information, and the ability to strategically manipulate one's behavioral responses appeared to be much more limited than typically believed.

More recently, Verschuere, Crombez, and Koster (2004) found similar results using a modified dot-probe task. However, in a different study, these authors showed no differentiation between crime-relevant and neutral information with the RT measure, in contrast to the high efficiency levels obtained with the heart rate and the SCR measures (Verschuere, Crombez, De Clercq, & Koster, 2004). Similarly, Rosenfeld, Soskins, Bosh, and Ryan (2004) reported conflicting results regarding the validity of the RT measure for differentiating between crime-relevant and neutral items. Furthermore, Vincent and Furedy (1992) showed that while enhanced SCRs were associated with deceptive answers to biographical questions, RTs were unaffected by deception.

Given the inconsistent findings regarding the validity of the RT measure for detecting concealed information, the present study aimed to investigate four main questions:

1. What are the conditions under which RTs can differentiate between critical (e.g., crime-relevant) and neutral information? In particular, can the RT measure serve as an adequate alternative to the SCR measure typically used in the CIT?

2. Can RT add to the detection of concealed information when combined with the SCR, similar to various physiological measures that have shown to improve detection efficiency? In other words, does the RT have incremental validity over the standard physiological measure typically used for detection? This question is of special importance from an applied perspective, because the SCR is widely used in polygraph interrogations and is known to have a high validity (e.g., Ben-Shakhar & Elaad, 2003).

3. Given its voluntary nature, is RT affected by behavioral countermeasures (strategic manipulations)? Extensive research has demonstrated that autonomic-based polygraph tests, including the CIT, were vulnerable to countermeasures designed to create responses to the neutral control questions (e.g., Ben-Shakhar & Dolev, 1996; Honts, Devitt, Winbush, & Kircher, 1996). Seymour et al. (2000), however, found no effect of strategic manipulations on RT-based detection efficiency. In this study, we reexamine this question.

4. Although it is well known that the SCR measure undergoes habituation with stimulus repetitions (Sokolov, 1963), does the RT measure show a similar tendency? Recent studies investigating behavioral responses to significant stimuli (in particular, personally significant stimuli that are used as critical items in the CIT) showed an attenuation of the RTs to these stimuli across repetitions (Gronau, Cohen, & Ben-Shakhar, 2003; Harris & Pashler, 2004). Is it possible, then, that the validity of the RT decreases with repeated presentations of the critical information?

To answer these questions, we administered a Stroop-like task in which participants were instructed to respond as fast as possible to colors of centrally presented words. Two experiments were designed. In the first experiment, participants committed a mock-crime that simulated the typical forensic usage of the CIT in criminal investigations. Following the mock crime, participants performed the color-naming task, which comprised crime-relevant items related to the mock crime and neutral items that were irrelevant to the crime. In the second experiment, biographical information about participants was investigated (e.g., one's name), and, thus, the color-naming task consisted of personally related and nonpersonal control words. The personal-item paradigm simulates a special, less common situation, in which the identity of a suspect is the core issue of investigation (for an interesting example, see Lykken, 1991).

The rationale behind the Stroop-like task used in both experiments was based on extensive research showing that participants are generally unable to ignore the words' content despite its irrelevance to the task requirements (e.g., MacLeod, 1991). In the classical Stroop task (Stroop, 1935), for instance, participants typically respond more slowly to a color name that conflicts with the color print of the word than to a name that matches the color print of the word. In the emotional Stroop paradigm, a variation of the Stroop task that is more relevant for the present study, participants tend to respond more slowly to emotionally laden words (e.g., *DEATH*) than to neutral, nonemotional words (e.g., *HOUSE*; see review in Williams, Mathews, & MacLeod, 1996). In a similar fashion, slower latencies are obtained when participants name the color of personally significant items than when naming nonpersonal words (Gronau et al., 2003), suggesting that personally significant as well as emotional words demand greater processing resources than neutral words. In accordance with these findings, we hypothesized that slower RTs would be obtained for the critical items in the CIT (whether crime relevant or personally related) than for the neutral, irrelevant items (Question 1).

In addition to the RT measurement, SCR was recorded during performance of the color-naming task. To assess the incremental validity of the RT measure (Question 2), we computed a combined measure of RT and SCR and compared this measure with the SCR alone. Two groups of participants were included in each experiment, one receiving ordinary polygraph instructions and the other receiving direct countermeasure instructions guiding them to avoid slowing down their responses to the critical items (Question 3). To examine possible habituation effects of the RT measure (Question 4), we ensured that each item was presented several times in each experiment and that differences between blocks were assessed.

Experiment 1

Experiment 1 examined the validity of the RT measure in the detection of concealed information, using a mock-crime setting. In this experiment, participants were requested to commit a "crime" (i.e., steal an envelope containing several valuables), on which they were later investigated. Participants were randomly assigned to five possible theft scenarios, differing in the critical items to be stolen. After committing the crime, the CIT was administered using the color-naming task, in which participants were asked to rapidly name the color of words denoting the critical, as well as the neutral, control items.

Method

Participants. Fifty undergraduate students (35 women and 15 men), with normal or corrected-to-normal sight, participated in the experiment for either course credit or payment. All participants were native Hebrew speakers. The participants were randomly assigned to one of the two instruction conditions, such that each condition consisted of 25 participants.

Apparatus. All stimuli were presented on a color monitor connected to a Pentium II computer. Participants responded vocally, and their responses were recorded by a small microphone attached to the collar of their shirt and connected to the computer. Skin conductance responses (SCRs) were measured by using a constant voltage system (0.5V ASR Atlas Researches) and two Ag/AgCl electrodes (0.8-cm diameter). An electrode paste (Johnson & Johnson K-Y gel) was placed on the electrodes in order to increase skin conductance. The SCRs were defined as the maximal conductance changes obtained from 1 s to 5 s after stimulus onset. The SCRs were computed using an A/D (NB-MIO-16) converter with a sampling rate of 1,000 per second.

Stimuli and design. After performing the mock crime, participants were informed that an envelope had been stolen from an office in the Psychology department. The participants were told that they would be asked about the details of the theft. During the CIT investigation, participants were presented with four classes of items, each targeting a different feature of the mock crime (the name of the professor from whose office the envelope was stolen, the location of the stolen envelope in the office, the exact sum of money, and the type of jewelry found in the envelope). Before each class of items, a sentence appeared on the screen for 3 s, describing the class of items to be presented (e.g., "We will now measure your responses to the type of jewelry found in the envelope"). Following this sentence, six items appeared, beginning with a buffer word, which was in turn followed by one critical (crime-related) word and four neutral words from the same category. The order of the critical and neutral words was determined randomly. Participants were required to name the color of the words as fast as possible. A white fixation point presented for 500 ms preceded the appearance of each word, in order to enhance participants' focus of attention on the stimuli. The words remained on the screen until a vocal response was made. To encourage the participants to respond rapidly, we ensured that a negative feedback ("Slow Response") appeared if participants' responses exceeded 1,500 ms.

The words appeared in one of four colors (red, yellow, blue, and green), which was randomly assigned to each item in each category. The background luminance was dark. The size of the letters was approximately 0.7 cm in height by 0.6 cm in width, corresponding to a visual angle of approximately 0.5×0.43 from a viewing distance of 80 cm.

The experiment consisted of three blocks, each comprising the four categories of items. Thus, all in all, there were 4 (categories) $\times$ 6 (items) $\times$ 3 (blocks) = 72 experimental stimuli. The order of the four categories in each block was randomly determined. The items in each block were presented with random interstimulus intervals (ISIs) ranging from 16 to 24 s, with a mean of 20 s. The three experimental blocks were preceded by a practice block of five trials, in which participants were investigated about the day of the week in which the theft took place.

Procedure. Participants were randomly assigned to one of five theft scenarios that determined the specific details of the theft to be committed. They were instructed to enter a locked room using a key that was handed to them and search for an envelope located in a certain specified location in that room. They were instructed to "steal" this envelope, which contained a sum of money ranging from 5 to 50 NIS (about \$1–\$10 at the time of the experiment) and an article of jewelry, to hide the money and the jewelry in their pocket, and to enter the examination room.

In the next stage of the experiment, the CIT was administered. Participants were seated at a table facing the screen, and a microphone was attached to the collar of their shirt. Two electrodes were attached with masking tape to the volar side of the index and fourth fingers of their left hand. The experimenter informed them that an envelope had been stolen from an office in the Psychology department and that the polygraph test would determine whether they were involved in the theft through an analysis of their physiological and behavioral responses in a color-naming task. Participants were instructed to respond as fast as possible to the color of the words and were promised a monetary bonus of 10 NIS (about \$2 at the time of the experiment) if they managed to avoid detection and to conceal any information related to the theft. In the countermeasures condition, participants were explicitly told that they should try to avoid slowing their responses when naming the color of the critical items in order to escape detection. In the control condition, no direct countermeasure instructions were given, and participants were simply told to try to avoid detection. All participants were warned of responding too slowly and were told that a negative feedback ("Slow Response") would appear if their responses weren't fast enough. Prior to the color-naming task, the participants' skin conductance baseline was recorded for a period of 2 min. During the task, the experimenter sat beside the participants and keyed their responses to the computer in order to keep track of errors.

In the end of the polygraph test, the mean latencies of the critical and the neutral items for each participant were compared. If the former was larger than the latter, participants were informed that they were found guilty. Otherwise, participants were informed that they were detected as innocent, in which case they were paid the 10 NIS bonus. The experiment lasted approximately 45 min.

Results and Discussion

Trials in which RTs were below 300 ms or exceeded 1,500 ms (less than 0.5%) and trials in which participants made errors in the color-naming task (1.5%) were discarded from the calculations. In addition, trials in which participants made noticeable movements were excluded from the SCR measure (2%). A rejection region of $p < .05$ was used for all statistical tests.

To eliminate individual differences in the RT and the SCR and permit a meaningful comparison between the two measures, we computed within-participants Z scores for each measure, relative to the means and standard deviations computed across all responses of each participant within each block (Ben-Shakhar, 1985). To assess the incremental validity of the RT measure, we computed a combined measure as the mean of the SCR and RT Z scores. This measure was also standardized within participants, within each block.

A three-way analysis of variance (ANOVA) was conducted for each measure (RT, SCR, and the combined measure), with two within-subjects factors of stimulus type (crime relevant vs. neutral) and experimental block and one between-subjects factor of instructions (countermeasures vs. control). In addition, effect-size estimates (Cohen's f values) were computed for each factor of the ANOVA conducted for each measure. According to Cohen (1988), the values of $f = 0.1$, $f = 0.25$, and $f = .40$ correspond to small, medium, and large effects, respectively.

Because there were no significant block effects in the RT and SCR measures, and no interaction effects between the block factor and the other factors (Question 4), we computed the means of the crime-relevant and the neutral words across the three blocks (after exclusion of the buffers). These means are presented in Table 1 for each measure, within each instruction condition and across the instruction conditions. Figure 1 presents the raw RT data in the different conditions (the proportion of errors was small and did not differ between the conditions).

Our main research objective was to determine whether the RT measure can differentiate between the critical and the neutral information and whether this differentiation is as effective as that of the SCR (Question 1). We hypothesized that slower latencies would be obtained for the crime-relevant items than for the neutral items. As can be seen from Table 1 and Figure 1, however, this hypothesis was not confirmed. The three-way ANOVA revealed a statistically significant main effect for the stimulus-type factor only with the SCR and the combined measures, $F(1, 48) = 34.10$, $p < .001$, $f = 0.34$; $F(1, 48) = 23.60$, $p < .001$, $f = 0.28$, respectively, but not with the RT measure, $F(1, 48) = 0.77$, $p < .38$, $f = 0.05$. In addition, the Pearson correlation coefficient between the mean standardized SCR and RT to the critical items was nearly zero ($r = 0.01$). To further assess the validity of each of the three measures, we computed a statistic comparing the entire distributions of the mean Z scores for the critical (crime-relevant) and neutral items. On the basis of these distributions, receiver-operating characteristic curves (ROCs) were generated for each condition and each measure, and the areas under these ROC curves, along with the corresponding 95% confidence intervals, were computed (see Bamber, 1975). The area under the ROC curve (also labeled A_g; see Green, 1964) reflects detection efficiency of the critical items across all possible cutoff points. It assumes values between 0 and 1, such that an area of 0.5 means

Table 1
Mean Z Scores (and 95% CIs) for the Crime-Relevant and Neutral Words and Areas Under the ROC Curves for the Three Measures in Experiment 1

Condition	Crime-relevant words	Neutral words	Area under the ROC curve
RT			
Countermeasures	−0.05 (−0.15, 0.06)	−0.06 (−0.10, −0.02)	0.48 (0.31, 0.65)
Control	0.00 (−0.12, 0.13)	−0.07 (−0.11, −0.03)	0.61 (0.43, 0.78)
Across instruction conditions	−0.02 (−0.10, 0.06)	−0.07 (−0.09, −0.04)	0.55 (0.42, 0.67)
SCR			
Countermeasures	0.32 (0.19, 0.45)	−0.10 (−0.15, −0.04)	0.94 (0.88, 1.00)
Control	0.25 (0.08, 0.41)	−0.07 (−0.12, −0.03)	0.76 (0.61, 0.91)
Across instruction conditions	0.28 (0.18, 0.39)	−0.09 (−0.12, −0.05)	0.86 (0.77, 0.94)
Combined measure			
Countermeasures	0.18 (0.08, 0.28)	−0.10 (−0.15, −0.05)	0.86 (0.75, 0.96)
Control	0.15 (0.01, 0.29)	−0.09 (−0.14, −0.05)	0.75 (0.61, 0.89)
Across instruction conditions	0.17 (0.08, 0.25)	−0.10 (−0.13, −0.07)	0.81 (0.72, 0.89)

Note. $n = 25$ in each group. CI = confidence interval; ROC = receiver-operating characteristic; RT = reaction time; SCR = skin conductance response.

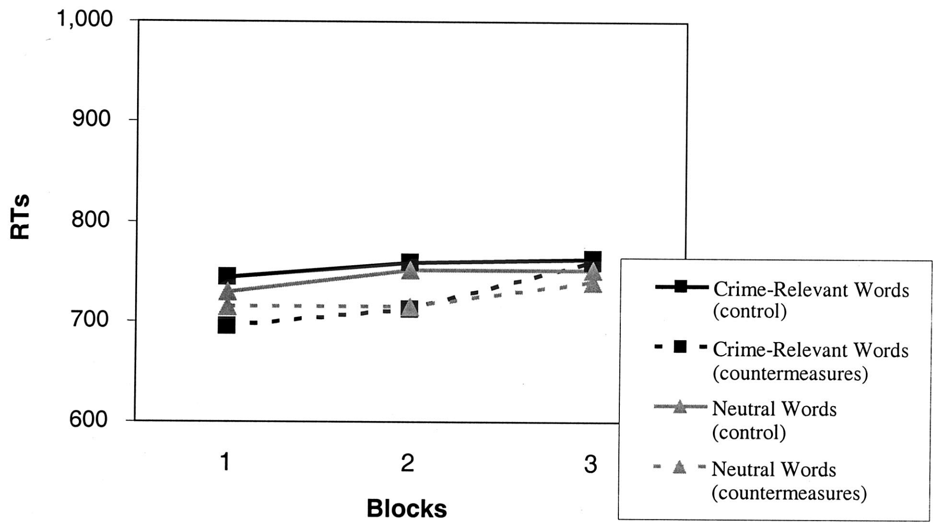

Figure 1. Mean reaction times (RTs; raw scores) for the crime-relevant and neutral words, within each instruction condition and each block of Experiment 1.

that the two distributions (the distributions of the mean *Z* scores of the critical and neutral items) are undifferentiated. An area of 1 indicates that there is no overlap between the two distributions and that they are thus perfectly differentiated. An inspection of Table 1 reveals that in accordance with the ANOVA, the area under the ROC curve showed no differentiation between the crime-relevant and the neutral stimuli with the RT measure (the lower bounds of the confidence intervals in both the countermeasures and the control conditions were lower than 0.5). Detection of the critical information with the SCR measure, in contrast, was significantly greater than chance level. In addition, the combination of the SCR with the RT produced a weaker differentiation between the critical and the neutral information than the use of the SCR as a single measure (e.g., the effect size across instruction conditions decreased from $f = 0.34$ to $f = 0.28$, and the ROC area decreased from 0.86 to 0.81). Thus, in contrast to the findings of Seymour et al. (2000) and Verschuere, Crombez, and Koster (2004), the RT measure was ineffective in detecting the concealed information, and it had no incremental validity above and beyond the SCR (Question 2). In addition, the slight differences in detection efficiency between the countermeasure and the control condition (Question 3) were negligible (the instructions factor produced neither statistically significant main effects nor interactions with stimulus type, for all three measures).

These findings clearly question the use of the RT as a viable measure in information detection. Our research demonstrates that although the SCR measure reliably differentiated between crime-relevant and neutral information, detection efficiency with the RT remained at chance level. Although conflicting with two previous studies (Seymour et al., 2000; Verschuere, Crombez, & Koster, 2004), these results converge with other studies showing no effect of the RT measure in the detection of guilty knowledge (e.g., Rosenfeld et al., 2004; Verschuere, Crombez, De Clercq, & Koster, 2004; Vincent & Furedy, 1992). Thus, it appears that the RT cannot serve as an adequate alternative to the SCR, nor does it contribute to the SCR, in the detection of concealed information.

To further investigate the contribution of behavioral measures to information detection, we ran an additional experiment in which we used the personal-item version of the CIT. In this version, biographical details about participants are investigated, such as their name, profession, date of birth, and so forth (e.g., Ben-Shakhar & Elaad, 2002a, 2003; Ben-Shakhar, Lieblich, & Kugelmass, 1970; Lykken, 1960). To increase the experimental power of the CIT, we chose two highly significant personal details as the items of investigation, the participant's first and last name, each repeated several times in the color-naming task. As previously mentioned, Gronau et al. (2003) have shown, using a similar color-naming task, that personally significant items elicit slower latencies than other nonpersonal items. It is possible, then, that the personal-item paradigm is more sensitive than the mock-crime paradigm in detecting concealed information. Furthermore, the relative efficiency of the RT and SCR measures, and the incremental validity of the RT beyond that of the SCR, was never investigated with the personal-item paradigm. These questions were therefore the main focus of Experiment 2.

Experiment 2

Method

Participants. Forty undergraduate students (23 women and 17 men), with normal or corrected-to-normal sight, participated in the experiment for either course credit or payment. All participants were native Hebrew speakers. The participants were randomly assigned to one of the two instruction conditions, such that each condition consisted of 20 participants.

Stimuli and design. Participants were required to name the color of centrally presented Hebrew words as fast as possible. On each trial, a word appeared on the screen until a vocal response was made, and a negative feedback ("Slow Response") appeared if participants' responses exceeded 1,500 ms. The words appeared in one of four colors: red, yellow, blue, and green.

The words denoted either personally significant or neutral names, forming the personally significant word and neutral-word conditions. The significant words comprised two personal items (the participant's own first name and family name), which were collected prior to the experiment and were incorporated into the computer without notification to the participants. The neutral stimuli denoted nonpersonal items belonging to the same categories (i.e., other first names and family names), which were matched in length to the significant items. Each category included five stimuli consisting of a single significant item (e.g., the participant's first name) and four neutral items (e.g., four other neutral names). The experiment consisted of two blocks. In each block, there were two repetitions of each category. Thus, all in all, there were 2 (categories) × 2 (repetitions) × 5 (items) × 2 (blocks) = 40 experimental stimuli. The stimuli were presented in a random order, with the exception that there was at least one neutral item presented between any two personally significant stimuli. A buffer stimulus denoting a neutral word preceded the 20 stimuli in each block.

The adjustment of a specific color to a specific word was determined randomly. The two experimental blocks were preceded by a practice block of eight trials. These trials consisted of words denoting various nonpersonal object names (e.g., *TABLE*, *PEN*, etc.). All other aspects of the stimulus presentation were identical to those of Experiment 1.

Procedure. Participants were connected to the SCR devices, and a microphone was attached to the collar of their shirt. They were told that they were about to participate in a polygraph test, in which an attempt would be made to detect their personal items (first and family names) through an analysis of their physiological responses and their voice latencies. Participants were strongly urged to conceal their personal items and were promised a monetary bonus of 5 NIS (about $1 at the time of the experiment) if they managed to avoid detection. In the countermeasures condition, participants were also explicitly told that they should try to avoid slowing their responses when naming the color of their own personal items.

Following the practice block, participants were requested to sit at ease for a rest period of 2 min, during which basic skin conductance level was measured. Following this baseline recording period, the experimental stimulus sequences were presented. In the end of the experiment, the mean latencies of the personally significant and the neutral items for each participant were compared. If the former was larger than the latter, participants were informed that their personal items were detected. Otherwise, they were informed that their personal items weren't detected, in which case they were paid the 5 NIS bonus. The experiment lasted approximately 30 min. All other aspects of the experimental procedure were identical to those of Experiment 1.

Results and Discussion

Trials in which RTs were below 300 ms or exceeded 1,500 ms (1%) and trials on which participants made errors in the color-naming task (3%) were discarded from the calculations. In addition, trials in which participants made noticeable movements were excluded from the SCR analysis (less than 2%). All other data analyses followed the same procedures used in Experiment 1.

A three-way ANOVA was conducted for each measure (RT, SCR, and the combined measure), with two within-subjects factors of stimulus type (personally significant vs. neutral) and experimental block (first vs. second) and one between-subjects factor of instructions (countermeasures vs. control). Because there were no statistically significant interactions of the instruction factor with the other two factors, and in order to simplify the data presentation, we display the mean responses of the personally significant and the neutral words in two separate tables. Table 2 presents the standardized means of the significant and neutral words within each instruction condition, across blocks, and Table 3 presents the standardized means within each block, across instruction condi-

Table 2
Mean Z Scores (and 95% CIs) for the Personally Significant and Neutral Words and Areas Under the ROC Curves, Within Each Instruction Condition and Across Instruction Conditions, for the Three Measures in Experiment 2

Condition	Personally significant words	Neutral words	Area under the ROC curve
RT			
Countermeasures	0.27 (0.03, 0.51)	−0.06 (−0.13, 0.00)	0.67 (0.49, 0.84)
Control	0.50 (0.34, 0.66)	−0.11 (−0.16, −0.07)	0.95 (0.87, 1.00)
Across instruction conditions	0.38 (0.24, 0.53)	−0.09 (−0.13, −0.05)	0.81 (0.71, 0.92)
SCR			
Countermeasures	0.76 (0.56, 0.95)	−0.20 (−0.26, −0.15)	0.97 (0.92, 1.00)
Control	0.78 (0.58, 0.98)	−0.17 (−0.22, −0.12)	1.00 (0.99, 1.00)
Across instruction conditions	0.77 (0.63, 0.90)	−0.19 (−0.22, −0.15)	0.98 (0.95, 1.00)
Combined measure			
Countermeasures	0.66 (0.50, 0.83)	−0.17 (−0.22, −0.12)	1.00 (1.00, 1.00)
Control	0.80 (0.63, 0.97)	−0.18 (−0.22, −0.15)	1.00 (1.00, 1.00)
Across instruction conditions	0.73 (0.62, 0.84)	−0.18 (−0.21, −0.15)	1.00 (1.00, 1.00)

Note. $n = 20$ in each group. CI = confidence interval; ROC = receiver-operating characteristic; RT = reaction time; SCR = skin conductance response.

Table 3
Mean Z Scores (and 95% CIs) for the Personally Significant and Neutral Words and Areas Under the ROC Curves, Within Each Block, for the Three Measures in Experiment 2

Block	Personally significant words	Neutral words	Area under the ROC curve
RT			
First	0.50 (0.34, 0.67)	−0.12 (−0.17, −0.08)	0.88 (0.79, 0.97)
Second	0.25 (0.07, 0.43)	−0.06 (−0.10, −0.01)	0.69 (0.56, 0.83)
SCR			
First	0.81 (0.68, 0.94)	−0.22 (−0.25, −0.18)	1.00 (1.00, 1.00)
Second	0.73 (0.53, 0.92)	−0.16 (−0.20, −0.11)	0.90 (0.82, 0.98)
Combined measure			
First	0.82 (0.70, 0.94)	−0.21 (−0.25, −0.18)	1.00 (1.00, 1.00)
Second	0.63 (0.47, 0.79)	−0.14 (−0.19, −0.10)	0.93 (0.86, 0.99)

Note. $N = 40$. CI = confidence interval; ROC = receiver-operating characteristic; RT = reaction time; SCR = skin conductance response.

tions (confidence intervals in parentheses). Figure 2 presents the raw RT data, along with the error proportions, in the different conditions.

The three-way ANOVAs revealed statistically significant main effects for the stimulus-type factor for all three measures, $F(1, 38) = 30.09$, $p < .001$, $f = 0.43$; $F(1, 38) = 144.24$, $p < .001$, $f = 0.95$; and $F(1, 38) = 185.62$, $p < .001$, $f = 1.08$, for the RT, SCR, and the combined measure, respectively. Similarly, the ROC areas were significantly greater than 0.5, across instruction conditions, in all three measures (see Table 2). In addition, a three-way ANOVA conducted on the proportion of errors revealed significantly more errors for the personally significant words than for the neutral words, $F(1, 38) = 9.43$, $p < .01$, $f = 0.24$. All other main and interaction effects were not statistically significant. Thus, in contrast to the results of our previous experiment, and in accordance with the results of Gronau et al. (2003), the behavioral indices (RTs and proportion of errors) successfully differentiated between the critical (personally relevant) and the neutral items in the present experiment. Most important, however, this differentiation was, overall, larger with the SCR measure than with the RT (Question 1). In fact, detection efficiency with the SCR measure reached nearly perfect (or perfect) levels in both instruction conditions (see ROC areas in Table 2). Furthermore, an inspection of Table 3 reveals that even when the SCR produced less than perfect detection efficiency (e.g., in the second block), the RT had only a marginal contribution (the slight increase in the ROC area of the combined measure relative to the SCR was not statistically significant, $Z = 0.85$, $p > .05$). Thus, it appears that the RT measure had no incremental validity relative to the SCR (Question 2). In addition, as in Experiment 1, the two measures were not correlated ($r = 0.08$).

As for the instruction manipulation (Question 3), a smaller differentiation was found between the personally significant and the neutral words in the countermeasure condition than in the control condition, implying that participants in the countermeasure group were able to control their responses to the critical items more efficiently. In fact, the ROC area of the RT measure in the countermeasure condition was close to chance level (see the lower bounds of the confidence interval). The difference between the countermeasure and the control groups, however, was statistically significant only in the ROC analysis ($Z = 2.87$) but not in the ANOVA (the main effect of the instruction factor and its interaction with the stimulus-type factor failed to reach statistical significance, in all three measures). As previously mentioned, Seymour et al. (2000) demonstrated that detection of information based on RTs is resistant to countermeasures, at least when participants are requested to respond fast (under 1,000 ms). These authors argued that conscious strategies require time for their implementation, and therefore participants are unable to differentially alter their responses when they are limited to short responses. The results of the present study suggest that although participants responded more slowly to significant than to neutral items in both the countermeasure and the control conditions, the differentiation between the two types of items was nevertheless smaller in the former condition. Thus, it appears that the RT measure may be more vulnerable to strategic manipulations than previously found. Note, however, that the countermeasures effect occurred predominantly in the second block (see Figure 2), implying that participants are able to control their responses only after some practice and exposure to the significant stimuli.

Finally, to examine possible effects of stimulus repetition on the validity of the RT measure (Question 4), we examined differences in the responses to the personally significant stimuli between the two blocks (see Table 3). The three-way ANOVA revealed a statistically significant main effect for the block factor in the RT measure, $F(1, 38) = 5.59$, $p < .02$, $f = 0.19$, but not in the other two measures, $F(1, 38) = 0.09$, $p < .76$, $f = 0.02$; $F(1, 38) = 2.67$, $p < .11$, $f = 0.13$, for the SCR and combined measure, respectively. More important, the Block × Stimulus Type interaction was statistically significant for the RT and the combined measures,

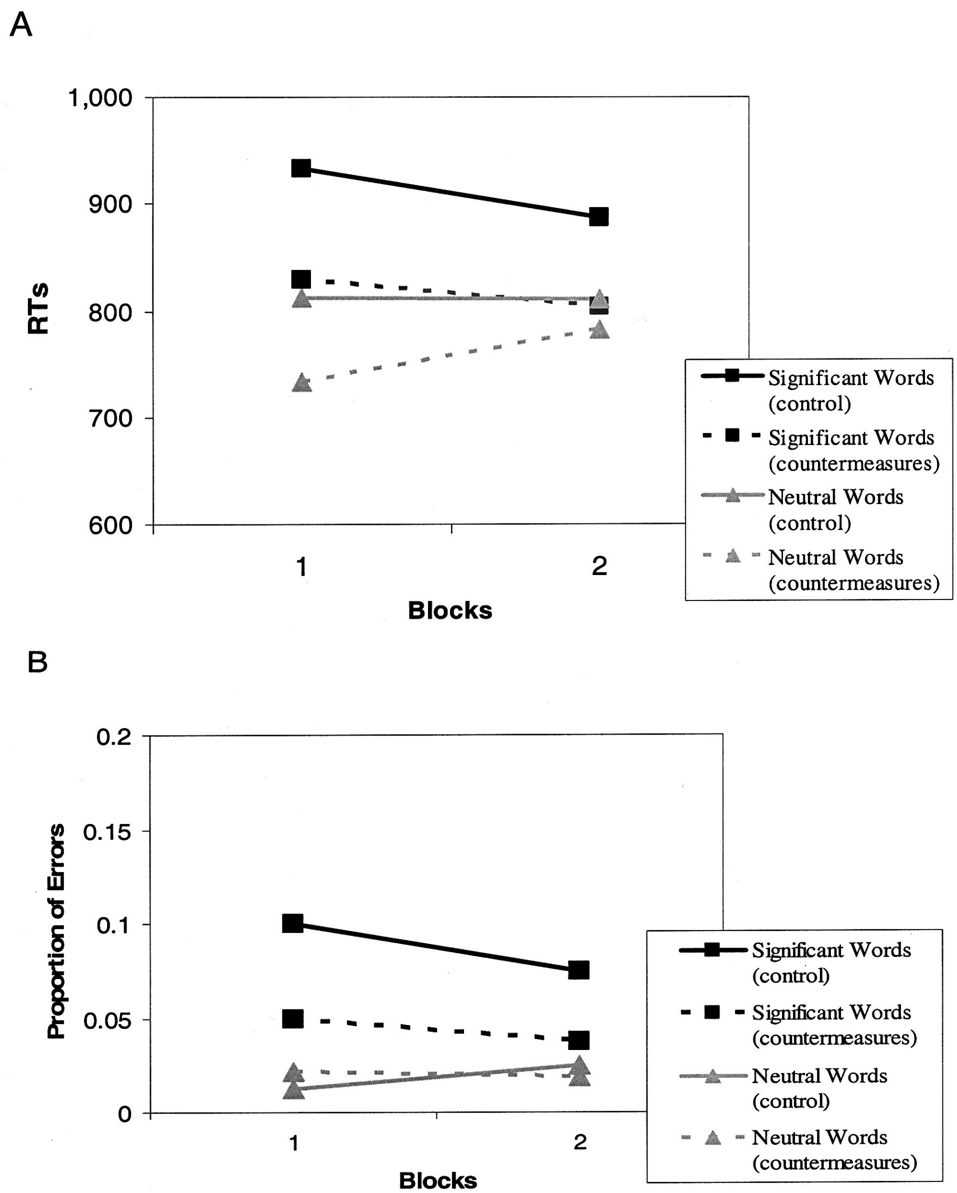

Figure 2. A: Mean reaction times (RTs; raw scores) and B: proportion of errors for the personally significant and neutral words, within each instruction condition and each block of Experiment 2.

$F(1, 38) = 7.29$, $p < .01$, $f = 0.21$; $F(1, 38) = 5.47$, $p < .03$, $f = 0.19$, respectively, indicating that a larger differentiation was found between the personally significant and the neutral items in the first block than in the second block. The SCR showed a similar tendency, but the effect was small and not statistically significant, $F(1, 38) = 1.65$, $p < .21$, $f = 0.10$. In the ROC analysis, there was a significant decline in the ROC area of the second block in all three measures ($Z = 2.49$, $Z = 2.53$, and $Z = 2.22$ for the RT, SCR, and the combined measure, respectively). These results suggest that similarly to the habituation process typically observed with the SCR measure (e.g., Sokolov, 1963), the efficiency of the RT measure may also decrease following repeated presentation of the critical information.

General Discussion

Two experiments investigated whether RT can serve as a valid measure in the detection of concealed information and whether RT contributes to psychophysiological detection of information. Participants performed a Stroop-like task in which they responded to the color of critical (crime relevant or personally significant) and neutral words. Whereas the critical words elicited larger SCRs than the neutral words in both experiments, differential RTs to the critical items were obtained only in Experiment 2. In addition, although the two measures were not correlated, there was no evidence that the RT measure has an incremental validity beyond the SCR.

Practical Implications of the Research

The results of the two experiments are only partially consistent with the results reported by Seymour et al. (2000) and by Verschuere, Crombez, and Koster (2004). In particular, whereas these authors demonstrated that RTs and error rates appear to be viable measures for detecting guilty knowledge, the present study suggests that this result is not sufficiently robust. A similar pattern of inconsistent results with the RT measure, across experiments, was reported recently by Rosenfeld et al. (2004). More important, we showed that the RT measure is much less effective than the SCR, typically used in the CIT. In Experiment 1, RT showed no differentiation between the critical and the neutral words, whereas a considerable differentiation was found in the SCR (an overall effect size of $f = 0.34$, which is considered a relatively large effect size by Cohen, 1988). In Experiment 2, detection efficiency with both measures was higher than that of Experiment 1, but nevertheless, the RT continued to show lower detection levels than the SCR.

These results are in accordance with the results of Vincent and Furedy (1992) and of Verschuere, Crombez, De Clercq, and Koster (2004), indicating that RTs cannot provide an adequate alternative to the SCR measure. Moreover, even in Experiment 2 in which both measures were effective in the detection of the critical items, RTs did not contribute to the SCR measure. Thus, in contrast to the advantage obtained by combining SCR with other physiological measures, such as respiration (Ben-Shakhar & Dolev, 1996; Ben-Shakhar et al., 1999; Cutrow et al., 1972; Elaad & Ben-Shakhar, 1997; Kugelmass & Lieblich, 1968; Timm, 1982, 1987), the present research showed no incremental contribution for the RT above and beyond the SCR.

Our findings demonstrate the importance of examining the question of incremental validity when developing new tests or new measures for detection of information. For example, many research efforts have been devoted to examine detection of guilty knowledge with ERP measures (see Allen & Iacono, 1997; Allen, Iacono, & Danielson, 1992; Boaz, Perry, Raney, Fischler, & Shuman, 1991; Farwell & Donchin, 1991; Rosenfeld, Cantwell, Nasman, Wojdac, Ivanov, & Mazzeiri, 1988; Rosenfeld et al., 2004). Although most of these studies reported impressive levels of detection efficiency, none of them has compared the ERP with traditional measures of detection and examined the incremental validity of the ERP. This is particularly important from an applied perspective because measuring EEG during realistic interrogations is much more complex and expensive than the use of electrodermal and respiration measures. More recent studies that explored additional measures for the detection of guilty knowledge, or deception, such as fMRI (e.g., Lee et al., 2002; Spence, Farrow, Herford, Wilkinson, Zheng, & Woodruff, 2001) or facial-skin surface temperature (Pavlidis, Eberhardt, & Levine, 2002), face a similar problem.

The present research also revealed that behavioral measures may be susceptible, at least to a certain extent, to strategic manipulations. Whereas Seymour et al. (2000) argued that participants are unable to deliberately alter their responses when limited to short latencies, we demonstrated that this may not be the case. Using a similar instruction manipulation to Seymour et al. (i.e., informing participants of the expected pattern of results and instructing them to avoid slowing their responses to the critical items), we found that detection was in fact less efficient with the countermeasures group than with the control group. Although the effect of countermeasures was statistically significant only in the personal-item paradigm (with the ROC analysis), this finding implies that RTs may be prone to countermeasure manipulations even at relatively short latencies. In this respect, RTs do not differ from autonomic measures that are known to be affected by countermeasures (e.g., Ben-Shakhar & Dolev, 1996; Honts et al., 1996) and from ERPs that have been recently shown to be affected by strategic manipulations (Rosenfeld et al., 2004).

Finally, our study shows that RTs to critical stimuli—in particular, personally significant items—may be sensitive to stimulus repetition. This finding converges with previous studies demonstrating that although the behavioral effects of task-relevant stimuli (e.g., color words in a color-naming task) remain relatively stable over time, enhanced latencies to personally related items (e.g., one's own name) and to emotionally charged words attenuate dramatically with practice (Gronau et al., 2003; Harris & Pashler, 2004). Thus, it seems that habituation processes known to influence autonomic measures (e.g., Sokolov, 1963) may also affect RTs to significant stimuli in the CIT.

One unresolved question of the present study concerns the differences in the results of the two experiments. Whereas the mock-crime paradigm failed to produce a behavioral effect in the detection of the crime-relevant items, RTs did differentiate between the personal and the neutral items in the personal-item paradigm. One possible explanation for this discrepancy is that while the critical items in the mock-crime paradigm differed from the neutral items only in their relevance to the experimental (mock-crime) context, the critical items in the personal-item paradigm were also more familiar and subjectively frequent than the neutral words. That is, the enhanced latencies elicited by the personally significant words may have reflected, in fact, familiarity, rather than mere knowledge or task relevance. Some support for this argument comes from a previous study (Klein, 1964) showing that slower latencies were obtained in a color-naming task to frequent, familiar words than to rare words or nonwords. However, more recently, Gronau et al. (2003) demonstrated that mere familiarity or lexical frequency cannot account for the behavioral effects of personally significant items in the Stroop-like task. In particular, highly frequent (nonpersonal) words were shown to elicit shorter, rather than longer, latencies than neutral and personally significant words. Thus, it appears that the slower RTs obtained for the personal items in the present study cannot be accounted for by mere familiarity.

Another possible explanation for the differences in the results of the two experiments is that the personal-item paradigm is more sensitive to the detection of the concealed information than the mock-crime paradigm. Indeed, detection efficiency was higher in the former paradigm with both the RT and the SCR measures (the overall RT effect size increased from $f = 0.05$ in Experiment 1 to $f = 0.43$ in Experiment 2, and the overall effect size of the SCR increased from $f = 0.34$ to $f = 0.95$). On the face of it, this possibility seems inconsistent with the conclusions made on the basis of the meta-analysis recently reported by Ben-Shakhar and Elaad (2003). These authors found that, on average, mock-crime studies produced larger detection efficiency than personal-item studies. However, as indicated by Ben-Shakhar and Elaad (2003), psychophysiological differentiation varied considerably as a function of the type of personal information used. In this study, only the most sensitive personal items were used (i.e., participants' first and family names), and this may explain the high detection efficiency obtained in Experiment 2.

Finally, it should be noted that the color-naming task may be less sensitive to potential behavioral effects, at least when it concerns the mock-crime paradigm, than other behavioral paradigms. For instance, whereas the words used in the present experiment were irrelevant to task requirements, they constituted the core elements in other memory-based paradigms, such as the memory-interference task used by Seymour et al. (2000). It is possible that directly responding to the content of a word (e.g., determining whether it has appeared in a target list), rather than trying to avoid it (e.g., in the color-naming task), yields better detection of the concealed information. In addition, despite the difficulty of ignoring the words' content in the color-naming task, participants may be able to deliberately attenuate word processing (e.g., by blurring their eyes such that reading of the words is difficult but colors can still be distinguished), thus weakening the ecological validity of the Stroop-like task in real-life situations. However, further research is required in order to fully understand the differences between the various behavioral paradigms.

Theoretical Implications of the Research

The present study focused on the validity of behavioral measures for detecting critical information with the CIT paradigm. However, modern conceptions of validity (e.g., Messick, 1995) require a theoretical understanding of the methods, or tests, in addition to a demonstration of criterion-related validity. The present study may shed further light on the psychological principles underlying the detection of concealed information.

The rationale behind the CIT is based on theory and research on the orienting response (OR; e.g., Sokolov, 1963, 1966). The OR is a complex of physiological and behavioral reactions evoked by novelty and significance (Bernstein, 1979; Gati & Ben-Shakhar, 1990; Maltzman, 1979; Sokolov, 1963). Lykken (1974) was the first to note that the OR's sensitivity to significant stimuli endows it with the potential for disclosing concealed information. Most studies investigating the OR focused on its physiological aspects; however, several studies also examined the relation between its physiological (e.g., SCR) and behavioral (e.g., RT) components. For example, Siddle and Packer (1987) demonstrated that the elicitation of an OR by an unexpected target was associated with slower latencies to a probe item appearing immediately after the target, suggesting that the OR reflected a shift of attention toward the unexpected stimulus (see also Siddle, 1991; Siddle & Jordan, 1993; Siddle, Jordan, & Lipp, 1993). Other studies have argued for a more complex relationship between attentional allocation and the OR (Dawson, Filion, & Schell, 1989; Filion, Dawson, & Schell, 1994; Filion, Dawson, Schell, & Hazlett, 1991; Siddle, Lipp, & Dall, 1996).

A complex relation between the OR (reflected by SCR) and RT was more recently demonstrated by Gronau et al. (2003). In their study, capture of visual attention by personally significant and task-relevant stimuli was investigated using both physiological and behavioral measures. Results showed that when words appeared centrally (i.e., in an attended location), relatively large SCRs and RTs were obtained to the personally significant stimuli, implying a positive relation between the two measures. These effects disappeared when the significant words appeared in a peripheral location, outside the focus of attention. Task-relevant stimuli (i.e., color words in a color-naming task), however, affected performance even when positioned peripherally, indicating a capture of visual attention. It is interesting to note that the effect of task-relevant distractors (whether presented centrally or peripherally) on performance was not reflected by an enhanced SCR, suggesting a dissociation between OR and visual attention processes. Gronau et al. (2003) suggested that although the OR is not an index of shifts in visual attention per se, it may be related to high-level attentional mechanisms (i.e., executive functions) known to be distinct from visual attention (e.g., Pashler, 1991; Posner & Boies, 1971).

The present study further illustrates the complex relation between OR and RT. We showed that enhanced ORs (SCRs) are not necessarily associated with an increase in RT (Experiment 1). Furthermore, even when the two indices differentiate between personally significant and neutral information (in Experiment 2), the electrodermal component is more sensitive. Most important, the two measures are not correlated. Thus, although the differences between the RT and SCR could be merely quantitative, our results, as well as those of other studies, imply that RT and SCR measures reflect, at least to some extent, different underlying mechanisms. Further research, however, is clearly required to fully understand the architecture underlying OR and behavioral measures.

Conclusion

This study examined the validity of the RT measure in the CIT and its contribution to psychophysiological detection. Whereas RTs differentiated between personally significant and neutral items above chance level, they failed to differentiate between crime-related and neutral information. Furthermore, RTs consistently showed lower detection levels than the SCR measure, and a combined measure based on both behavioral and physiological indices showed no advantage over the SCR alone. These findings suggest that the contribution of the RT measure for the detection of information may be negligible once the SCR measure is available. From a theoretical perspective, the present study suggests that physiological and behavioral aspects of information detection are at least partially dissociated. Further research, however, is required to fully understand the complexity of the OR phenomenon and its relations to cognitive processes such as attention capture.

References

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

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The development and validation of an event-related-potential memory assessment procedure: A methodology for prediction in the face of individual differences. *Psychophysiology, 29,* 504–522.

Bamber, D. (1975). The area under the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology, 12,* 378–415.

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

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

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

Ben-Shakhar, G., & Elaad, E. (2002a). Effects of questions' repetition and variation on the efficiency of the guilty knowledge test: A reexamination. *Journal of Applied Psychology, 87,* 972–977.

Ben-Shakhar, G., & Elaad, E. (2002b). The Guilty Knowledge Test (GKT) as an application of psychophysiology: Future prospects and obstacles. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 87–102). San Diego, CA: Academic Press.

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

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective.* New York: Springer-Verlag.

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

Ben-Shakhar, G., Lieblich, I., & Kugelmass, S. (1970). Guilty knowledge technique: Application of signal detection measures. *Journal of Applied Psychology, 54,* 409–413.

Bernstein, A. S. (1979). The orienting response as novelty and significance detector: Reply to O'Gorman. *Psychophysiology, 16,* 263–273.

Boaz, E., Perry, N. W., Raney, G., Fischler, I. S., & Shuman, D. (1991). Detection of guilty knowledge with event-related potentials. *Journal of Applied Psychology, 76,* 788–795.

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

Cutrow, R. J., Parks, A., Lucas, N., & Thomas, K. (1972). The objective use of multiple physiological indices in the detection of deception. *Psychophysiology, 9,* 578–588.

Dawson, M. E., Filion, D. L., & Schell, A. M. (1989). Is elicitation of the autonomic orienting response associated with allocation of processing resources? *Psychophysiology, 26,* 560–572.

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the Guilty Knowledge Test. *Psychophysiology, 34,* 587–596.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Journal of Applied Psychology, 76,* 788–795.

Filion, D. L., Dawson, M. E., & Schell, A. M. (1994). Probing the orienting response with startle modification and secondary reaction time. *Psychophysiology, 31,* 68–78.

Filion, D. L., Dawson, M. E., Schell, A. M., & Hazlett, E. A. (1991). The relationship between skin conductance orienting and the allocation of processing resources. *Psychophysiology, 28,* 410–424.

Fukumoto, J. (1980). A case in which the polygraph was the sole evidence for conviction. *Polygraph, 9,* 42–44.

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

Green, D. M. (1964). General predictions relating yes–no and forced-choice results. *Journal of the Acoustical Society of America, 35,* 1042.

Gronau, N., Cohen, A., & Ben-Shakhar, G. (2003). Dissociations of personally significant and task-relevant distractors inside and outside the focus of attention: A combined behavioral and psychophysiological study. *Journal of Experimental Psychology: General, 132,* 512–529.

Harris, C. R., & Pashler, H. E. (2004). Attention and the processing of emotional words and names: Not so special after all. *Psychological Science, 15,* 171–178.

Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the Concealed Knowledge Test. *Psychophysiology, 33,* 84–92.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2, pp. 446–483). St. Paul, MN: West Law.

Iacono, W. G., & Lykken, D. T. (2002). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony* (Vol. 2, pp. 483–538). St. Paul, MN: West Law.

Klein, G. S. (1964). Semantic power measured through the interference of words with color naming. *American Journal of Psychology, 77,* 576–588.

Kugelmass, S., & Lieblich, I. (1968). An analysis of mechanisms underlying psychophysiological detection. *Proceedings of the 15th International Congress of Applied Psychology,* 509–512.

Lee, T. M., Liu, H., Tan, L., Chan., C., Mahankali, S., Feng, C., et al. (2002). Lie detection by functional magnetic resonance imaging. *Human Brain Mapping, 15,* 157–164.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44,* 258–262.

Lykken, D. T. (1974). Psychology and the lie detection industry. *American Psychologist, 29,* 725–739.

Lykken, D. T. (1991). Why (some) Americans believe in the lie-detector while others believe in the guilty knowledge test. *Integrative Physiological and Behavioral Science, 26,* 214–222.

Lykken, D. T. (1998). *A tremor in the blood: Uses and abuses of the lie detector.* New York: Plenum Trade.

MacLaren, V. V. (2001). A quantitative review of the Guilty Knowledge Test. *Journal of Applied Psychology, 86,* 674–683.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203.

Maltzman, I. (1979). Orienting reflexes and significance: A reply to O'Gorman. *Psychophysiology, 16,* 274–282.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Nakayama, M. (2002). Practical use of the concealed information test for criminal investigation in Japan. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 49–86). San Diego, CA: Academic Press.

National Research Council. (2003). *The polygraph and lie detection.* Washington, DC: Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioral and Social Sciences and Education, National Academies Press.

Pashler, H. (1991). Shifting visual attention and selecting motor responses: Distinct attentional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 1023–1040.

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002, January 3). Human behavior: Seeing through the face of deception. *Nature, 415,* 35.

Podlesny, J. A., & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin, 84,* 782–799.

Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review, 78,* 391–408.

Raskin, D. C. (1986). The polygraph in 1986: Scientific, professional, and legal issues surrounding applications and acceptance of polygraph evidence. *Utah Law Review, 60,* 29–74.

Raskin, D. C. (1989). Polygraph techniques for the detection of deception. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 317–371). New York: Springer-Verlag.

Reid, J. E., & Inbau, F. E. (1977). *Truth and deception: The polygraph ("lie detector") technique* (2nd ed.). Baltimore: Williams & Wilkins.

Rosenfeld, J. P., Cantwell, B., Nasman, V. T., Wojdac, V., Ivanov, S., & Mazzeiri, L. (1988). A modified event-related potential-based guilty-knowledge test. *International Journal of Neuroscience, 42,* 157–161.

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

Saxe L., & Ben-Shakhar, G. (1999). Admissibility of polygraph tests: The application of scientific standards post-Daubert. *Psychology, Public Policy and Law, 5,* 203–223.

Saxe, L., Dougherty, D., & Cross, T. P. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist, 40,* 355–366.

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge." *Journal of Applied Psychology, 85,* 30–37.

Siddle, D. A. (1991). Orienting, habituation, and resource allocation: An associative analysis. *Psychophysiology, 28,* 245–259.

Siddle, D. A, & Jordan, J. (1993). Effects of intermodality change on electrodermal orienting and on the allocation of processing resources. *Psychophysiology, 30,* 429–435.

Siddle, D. A, Jordan, J., & Lipp, O. V. (1993). Effects of task-relevance and intermodality change on electrodermal orienting and on processing resource allocation. *Journal of Psychophysiology, 7,* 58–64.

Siddle, D. A, Lipp, O. V., & Dall, J. (1996). Effects of intermodality change and number of training trials on electrodermal orienting and on the allocation of processing resources. *Biological Psychology, 43,* 57–67.

Siddle, D. A., & Packer, J. S. (1987). Stimulus omission and dishabituation of the electrodermal orienting response: The allocation of processing resources. *Psychophysiology, 24,* 181–190.

Sokolov, E. N. (1963). *Perception and the conditioned reflex.* New York: Macmillan.

Sokolov, E. N. (1966). Orienting reflex as information regulator. In A. Leontyev, A. Luria, & A. Smirnov (Eds.), *Psychological research in the U.S.S.R.* (pp. 334–360). Moscow: Progress.

Spence, S. A., Farrow, T. F., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. (2001). Behavioral and functional anatomical correlates of deception in humans. *Neuroreport, 12,* 2849–2853.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Timm, H. W. (1982). Effects of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology, 67,* 391–400.

Timm, H. W. (1987). Effect of biofeedback on the detection of deception. *Journal of Forensic Sciences, 32,* 736–746.

Verschuere, B., Crombez, G., De Clercq, A., & Koster, E. H. W. (2004). Autonomic and behavioral responding to concealed information: Differentiating orienting and defensive responses. *Psychophysiology, 41,* 461–466.

Verschuere, B., Crombez, G., & Koster, E. H. W. (2004). Orienting to guilty knowledge. *Cognition and Emotion, 18,* 265–279.

Vincent, A., & Furedy, J. J. (1992). Electrodermal differentiation of deception: Potentially confounding and influencing factors. *International Journal of Psychophysiology, 13,* 129–136.

Williams, M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120,* 3–24.

Yamamura, T., & Miyata, Y. (1990). Development of the polygraph technique in Japan for detection of deception. *Forensic Science International, 44,* 257–271.

Received September 11, 2002
Revision received December 14, 2003
Accepted December 22, 2003 ■