The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review

Gershon Ben-Shakhar[1] and Eitan Elaad [2]

1. Department of Psychology, The Hebrew University of Jerusalem

2. Department of Behavioral Sciences, The College of Judea and Samaria, Ariel, Israel.

Abstract

The authors performed meta analysis based on 169 conditions, gathered from 80 laboratory studies, to estimate the validity of the Guilty Knowledge Test (GKT) with the electrodermal measure. The overall average effect size was 1.55, but there were considerable variations among studies. In particular, mock-crime studies produced the highest average effect size (2.09). Three additional moderators were identified: Motivational instructions, deceptive ("no") verbal responses, and the use of at least 5 questions were associated with enhanced validity. Finally, a set of 10 studies that best approximated applications of the GKT under optimal conditions produced an average effect size of 3.12. The authors discuss factors that might limit the generalizability of these results and recommend further research of the GKT in realistic setups.

Key Words: Electrodermal activity; Guilty Knowledge Test; Meta analysis; Polygraph; Psychophysiological detection; Skin conductance responses

The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review

Deception is a frequent, perhaps essential, feature of human behavior, which may be expressed in a variety of situations (e.g., Saxe, 1991a). Typically, deceptive behavior is employed to gain advantages (e.g., deceptive self-presentation in an employment interview) and avoid embarrassment (e.g., DePaulo et al., 2002). The frequent use of deception in social contexts, highlights the importance of detecting deception. However, research on perceivers' ability to differentiate between truthful and deceptive messages has indicated that, in most cases, people perform this task at chance levels (see, DePaulo, Stone & Lassiter, 1985 for a review), and even professionals whose tasks involve detection of deceit, perform no better than chance in most cases. This was demonstrated with customs officials (Kraut & Poe, 1980) and federal law enforcement officers (DePaulo & Pfeifer, 1986). Furthermore, Ekman and O'Sullivan (1991) examined a number of professional groups who have special interest in detecting deception, including police investigators, polygraphers, judges, psychiatrists and U.S. Secret Service agents. They found that all groups, except for the Secret Service agents, could not differentiate truth-tellers from deceivers at a better than chance level. More recently, Ekman, O'Sullivan and Frank (1999) reported that two law-enforcement groups and a selected group of clinical psychologists correctly identified people who were lying about their opinions.

The poor ability of people as "lie-detectors" may account for the attempts to use various instruments for this purpose. One such attempt is the use of the "psychophysiological lie-detector", known as the polygraph. The idea of using physiological measures for detecting deception, and for discriminating between individuals involved in an illegal activity and innocent suspects, has been very

appealing to law-enforcement agencies for many years (see, Larson, 1932; Marston, 1917, 1938; Reid, 1947; Reid & Inbau, 1977).

The polygraph is a device, which continuously measures several physiological responses (e.g., changes in respiration, in electrodermal activity and in relative blood pressure) to a series of questions presented to an examinee. Typically, inferences about whether an examinee is deceiving or telling the truth are made on the basis of a comparison of the physiological responses elicited in him or her by different types of questions (i.e., relevant versus control questions). Several techniques of psychophysiological detection have been proposed since the beginning of the 20th century (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 the major focus of research, discussion and debate.

The most widely used method of psychophysiological detection (at least in the US, Canada and Israel), labeled the Control Questions Test (CQT), has been extensively debated in the scientific literature (e.g., Ben-Shakhar & Furedy, 1990; Furedy, & Heslegrave, 1991; Iacono, & Lykken, 1997a, 1999; Kleinmuntz, & Szucko, 1984; Lykken, 1974, 1978, 1979, 1998; Podlesny, & Raskin, 1977, 1978; Raskin, 1982, 1986, 1989; Raskin, Honts, Amato, & Kircher, 1999; Raskin, Honts, & Kircher, 1997; Raskin, & Podlesny, 1979; Saxe, 1991b; Saxe, et al., 1985). The CQT is based on a comparison of physiological responses to relevant and control questions, where the former are crime-related questions of the "Did you do it?" type (e.g., "Did you break into Mr. Jones's apartment last Friday night?"), and the latter focus on non-specific misconducts, of a nature as similar as possible to the issue under investigation (e.g., "Have you ever taken something that did not belong to you?"). Critics of the CQT have argued that this method is not based on solid scientific principles, and

relies on improper control questions, which enhance the risk of false positive errors (i.e., innocent suspects classified as guilty). In addition, it is not standardized and therefore vulnerable to various biases (e.g., Ben-Shakhar 2001; Ben-Shakhar, Bar-Hillel & Lieblich, 1986; Furedy & Heslegrave, 1991; Iacono & Lykken, 1997a, 1999; Lykken, 1974, 1998; Saxe & Ben-Shakhar, 1999).

The second method, known as the Guilty Knowledge Test (GKT), or the Concealed Information Test (CIT), has drawn considerable attention among researchers, but has been extensively applied only in Japan (Fukumoto, 1980; Nakayama, 2001; Yamamura & Miyata, 1990). In contrast to the CQT, there is a general consensus that the GKT does rely on solid scientific principles and on proper control questions (e.g., Ben-Shakhar & Elaad, 2001; Ben-Shakhar & Furedy, 1990; Lykken, 1974, 1998). Iacono and Lykken (1997b) conducted two surveys of scientific opinion about the validity of various methods of psychophysiological detection. They found that 77% of the members of the Society for Psychophysiological Research and 72% of the members of the American Psychological Association who responded believed that the GKT (unlike the CQT) is based on scientifically sound psychological principles or theory. For this reason, the focus of this study is just on the GKT.

The GKT (Lykken, 1959, 1960) utilizes a series of multiple-choice questions, each offering one “relevant” answer (e.g., a feature of the crime under investigation) and several “neutral” (control) answers, chosen so that an innocent suspect would not be able to discriminate them from the relevant one (Lykken, 1998). Typically, if a 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 consistently show larger responses to the relevant than to the

neutral alternatives depends only on the number of questions and the number of alternative answers per question. Thus, the rate of false-positive errors can be controlled such that maximal protection for the innocent is provided.

The rationale behind the GKT is based on theory and extensive research on orienting responses (ORs) and habituation processes in humans (e.g., Siddle, 1991; Sokolov, 1963, 1966). The OR is a complex of physiological and behavioral reactions evoked by any novel stimulus or by any change in stimulation (e.g., Berlyne, 1960; Sokolov, 1963). With repeated presentations of stimuli, ORs undergo habituation, which is a gradual decline in response magnitude (Sokolov, 1963). In addition, stimuli that have a signal value for the subject (e.g., the subject's own name) evoke enhanced ORs (Bernstein, 1979, 1981; Maltzman, 1979; Sokolov, 1963). Lykken (1974) was the first to note that this property of ORs endows them with the potential for disclosing guilty knowledge. He argued that: "… for the guilty subject only, the 'correct' alternative will have a special significance, an added 'signal value' which will tend to produce a stronger orienting reflex than that subject will show to other alternatives." (p. 728).

This conceptualization of the GKT in terms of orienting responses to significant stimuli has made this paradigm interesting not just from an applied perspective, but also as a model for studying basic mechanisms underlying orientation behavior. For example, the dichotomization theory (e.g., Ben-Shakhar, 1977; Lieblich Kugelmass & Ben-Shakhar, 1970) has been proposed to account for the enhanced ORs elicited by relevant stimuli (guilty knowledge stimuli) that are familiar to the subjects and are therefore not novel. More recently, Ben-Shakhar and Gati developed a research program for studying the comparator mechanism hypothesized to account for orientation and habituation processes (e.g., Ben-Shakhar & Gati, 1987; Ben-

Shakhar, Gati & Solomon, 1995; Gati & Ben-Shakhar, 1990). Orientation and habituation processes are important from a broader perspective because they are closely related to attention and information processing (Ohmen, 1979, 1992). The relationships between stimulus novelty and significance on the one hand and ORs on the other, is consistent with the common interpretation that ORs reflect attentional processes (e.g., Dawson, Filion & Schell, 1989; Filion, Dawson, Schell & Hazlett, 1991; Kahneman, 1973; Ohman, 1979; Packer & Siddle, 1989; Siddle & Packer, 1987; Siddle & Spinks, 1992).

Since psychophysiological differentiation in the GKT is mediated through a mechanism of orientation, the enhanced responsivity to relevant items need not be attributed to deception, motivation, or fear of punishment as assumed by early theoretical approaches to psychophysiological detection (see, e.g., Davis, 1961). Indeed, Lykken (1974) said of an individual possessing the guilty knowledge: "Whether he is high or low in reactivity, whether he has confidence in the test or not, whether he is frightened and aroused or calm and indifferent, we can still expect that his response to this significant alternative will be stronger than to the other alternatives as long as he recognizes which alternative is 'correct'" (p. 728). Ben-Shakhar and Furedy (1990) called this a cognitive approach to psychophysiological detection, because it relies on what one knows, rather than on one's emotions, concerns, and conditioned responses. Indeed, research demonstrates that relevant information can be detected even when no motivational instructions are given to the subjects, and even when no verbal response is required (e.g., Ben-Shakhar, 1977; Ben-Shakhar & Lieblich, 1982; Elaad & Ben-Shakhar, 1989).

The GKT has been extensively researched during the past three decades, and many studies focusing on its validity and on various factors that might affect it have

been conducted since the pioneering work published in the 1940s and 50s (e.g., Ellson, Burk, Davis, & Saltzman, 1952; Geldreich, 1941, 1942; Lykken, 1959, 1960; Van Buskirk & Marcuse, 1954). However, so far these attempts to estimate the validity of the GKT have not used modern meta-analytic techniques, and consequently information that extends the single study is available only from narrative reviews or from quantitative summaries of a limited number of studies.

For example, Ben-Shakhar and Furedy (1990) reviewed and summarized 10 GKT laboratory experiments and showed that across these studies, 84% of 248 guilty examinees and 94% of 208 innocent examinees were correctly classified. They acknowledged that the number of studies they had analyzed was too small to allow for a statistical examination of the sources of the between-studies variability, but they noted that the two studies that used the largest number of questions (Bradley & Ainsworth, 1984 with nine; Bradley & Warfield, 1984 with ten) demonstrated the highest rates of correct classifications. Consequently, Ben-Shakhar and Furedy (1990) suggested that the number of GKT questions used, is a natural candidate for accounting for at least some of this variability.

The idea that detection efficiency increases with an increase in the number of questions is consistent with the well known psychometric principle, according to which the reliability of any test is an increasing function of the number of its questions. However, it is unclear from the available research, how many GKT questions are required to provide a sufficient level of validity. Furthermore, in many realistic situations the number of proper GKT questions (i.e., questions related to features of the event under investigation that are very likely to be noticed and remembered by the perpetrator of a crime and at the same time are not too salient to be identified by innocent suspects) that can be used may be limited. Thus, there might

be a trade-off between increasing the number of questions and the quality of some of the added questions. This lead Elaad and Ben-Shakhar (1997) to suggest the use of a GKT with repeated presentations of a small number of questions. Although the results obtained in this study were encouraging (indicating that the validity of the GKT increased similarly with questions' repetition and with questions' variation), Ben-Shakhar & Elaad (2002) conducted a constructive replication of their 1997 study and obtained different results, indicating that a GKT based on multiple questions is superior to the use of many repetitions of a single, or a few questions. One goal of this meta-analysis is to clarify how many GKT questions are needed to obtain a sufficient level of validity and whether an increase in the number of repetitions also affects validity.

More recently, Elaad (1998) reviewed 15 mock crime GKT studies and estimated the accuracy rates among guilty and innocent examinees as 81% and 96%, respectively. He noted that in 11 of these 15 studies no false-positive errors were documented. Ansley (1992) conducted a more comprehensive review of 70 GKT and Peak of Tension (POT) studies. However, Ansley (1992) used the traditional, narrative approach, rather than a quantitative meta-analysis. Vrij (2000) presented several reviews summarizing the accuracy rates of the GKT in laboratory studies. The mean correct detection rates ranged between 78% and 86% for guilty subjects, and between 94% and 99% for innocents. Finally, MacLaren (2001) conducted a quantitative review of 50 treatment groups drawn from 22 GKT studies and estimated the accuracy rates among guilty and innocent examinees as 76% and 83%, respectively.

Clearly, these reviews, which either rely on a small number of studies, or use the narrative approach, are insufficient for providing proper estimates of the GKT

validity. In particular, they are insufficient for examining possible moderator variables that might affect the validity of the GKT. Several factors, in addition to the number of GKT questions, might affect the validity of the GKT. Two notable examples are the factors of "motivation to deceive" and "deceptive answers to the GKT questions". Both of these factors have received considerable attention because of their theoretical importance and practical implications. However, a review of the attempts to assess the impact of these factors on detection efficiency reveals many inconsistent results and conclusions.

While some studies demonstrated that the accuracy of the GKT increases under conditions of heightened motivation (e.g., Elaad & Ben-Shakhar, 1989; Gustafson & Orne, 1963, 1965a; Wakamatsu, 1987), other studies failed to obtain such an effect (Beijk, 1980; Furedy & Ben-Shakhar, 1991; Horvath, 1978, 1979; Lieblich, Naftali, Shmueli & Kugelmass, 1974; Kugelmass & Lieblich, 1966). It is of interest to note that the factor of "motivation to avoid detection" has also been found to play an important role in the communication of deception. DePaulo and Kirkendol (1989) reported that individuals who are highly motivated to avoid detection are relatively less successful in their attempts when observers can watch their nonverbal behavior (the "motivational impairment effect").

Similarly, some studies concluded that deceptive answers to the GKT questions are associated with increased detection accuracy (e.g., Elaad, 1987; Furedy & Ben-Shakhar, 1991; Gustafson & Orne, 1965b; Horneman & O'Gorman, 1985), while other studies showed that this factor has no effect on detection accuracy (e.g., Kugelmass, Lieblich & Bergman, 1967). Like motivation, the role of deception is important from a theoretical perspective. If psychophysiological detection is not affected by deception, as concluded by Kugelmass et al. (1967), then knowledge of

the relevant information is sufficient for psychophysiological differentiation. This would imply that psychophysiological differentiation between significant and neutral stimuli is governed solely by a cognitive mechanism.

The goal of this study is to conduct such a meta-analysis of all studies that used some version of the GKT. We believe such an analysis may provide answers to a number of important theoretical and practical questions. Specifically, we hope to achieve the following goals:

1. Providing an estimate for the validity of the GKT based on a large data set, rather than on a single study. More importantly, we hope to identify a sub-set of studies that best approximate realistic conditions under which the GKT is likely to be applied and to provide a validity estimate, generated from these studies. Although, it is clear from previous reviews that the GKT is basically a valid technique, an accurate validity estimate is needed if this test is to become an important investigation tool. At present, polygraph testimony, which is typically based on the CQT, is generally inadmissible in criminal trials (see, Saxe & Ben-Shakhar, 1999). Recently, Ben-Shakhar, Bar-Hillel and Kremnitzer (2002) recommended that the admissibility of evidence obtained from GKTs would be reconsidered. They argued that if properly administered, this method can meet the admissibility criteria set by Daubert vs. Merrell Dow Pharmaceuticals inc. (1993) and may be of considerable aid to the trier of fact. But it is clear that such a recommendation requires good estimates of the validity of the GKT.
2. From previous reviews of GKT studies, it is clear that different results were obtained in different studies, both with respect to the overall validity estimate and the factors that may affect the accuracy of the technique. By applying meta-analytic methods we will examine whether the different results obtained by

various GKT studies reflect sampling errors, or whether systematic variations in several features of these studies (e.g., number of questions used) may account for these differences.

3. If the observed variability in the outcomes of the different GKT studies exceeds the variability that would be expected by sampling error, a search for possible moderators will be conducted. Several factors that may moderate the validity of the GKT can be identified a priori on the basis of previous research and theory. Our previous discussion suggests that the number of GKT questions used, as well as motivation to deceive and deceptive answers to the questions, may moderate detection accuracy with the GKT. As explicated above, the answers to these questions are less obvious than might seem because conflicting results were obtained in individual studies (in particular with respect to the role of motivation). These conflicts are unlikely to be resolved neither by an additional individual study, nor by a narrative review. Providing conclusive answers to the questions of whether and to what extent motivation to avoid detection and verbal deceptive answers to the GKT questions are associated with an increased accuracy is important from both practical and theoretical perspectives.

4. Finally, the present analysis may encourage and direct future research. In particular, further questions, such as whether the moderators identified in this meta-analysis have similar effects on physiological measures other than the electrodermal measure, can be derived from the present results.

Our meta analysis is limited to experimental studies, because only two GKT field studies were published so far (Elaad, 1990; Elaad, Ginton & Jungman, 1992). Consequently, the generalizability of our results to real-life GKT investigations should be considered. In addition, we include results based only on the electrodermal

measure, because the number of GKT studies reporting accuracy results based on other physiological measures is very restricted. From this respect, our results should be treated as a lower bound for the optimal efficiency of the GKT when based on a combination of several physiological measures, as it has been demonstrated that such a combination significantly increases detection accuracy (e.g., Cutrow, Parks, Lucas & Thomas, 1972; Kugelmass & Lieblich, 1968).

We adopted the approach developed by Hunter, Schmidt and their colleagues (e.g., Hunter & Schmidt, 1990) to meta analysis. Thus, we attempt to distinguish real from sampling error variance of the results across studies. Attempts to examine possible moderator variables are made only when the variance across studies clearly exceeds the variance expected just by sampling error (when the homogeneity assumption is rejected). Other artifacts are hardly relevant in this case. First, since the studies included in the present data set are experimental studies, it can be safely assumed that the criterion (experimental versus control groups) is almost perfectly reliable. Second, range restriction also does not apply to the present situation, since participants in GKT experiments are not selected on the basis of their physiological responses. Moreover, even if such a range restriction exists, it is impossible to estimate it, and therefore no attempt was made to correct for range restriction.

Methods

Studies included in the meta-analysis: Following an extensive literature search and consultations with colleagues who work in this area[1], a total of 80 studies, which included 169 different conditions were identified. A study, or an experimental condition within a study, was included if it satisfied the following criteria: (1) It included at least one set of equivalent items (i.e., one GKT question), some of which were relevant (e.g., crime-related) for participants simulating the "guilty suspects". (2)

The proportion of significant items in a given set did not exceed 0.25. Larger proportions are rare and are highly uncharacteristic of the realistic applications of the GKT. Consequently, some of the experimental conditions reported by Ben-Shakhar (1977), Ben-Shakhar, Lieblich and Kugelmass (1975) and by Elaad (1987) were not included. (3) The GKT was administered under standard conditions that characterize the typical realistic application of this method. Thus, for example, conditions where participants were requested to apply countermeasures (e.g., Ben-Shakhar & Dolev, 1996; Elaad & Ben-Shakhar, 1991; Honts, Devitt, Winbush & Kircher, 1996), or were under the influence of drugs (e.g., Iacono, Boisvenu & Fleming, 1984; Iacono, Cerri, Patrick Fleming, 1992; Waid, Orne, Cook & Orne, 1981) or alcohol (Bradley & Ainsworth, 1984) were not included. Similarly, conditions that involved disclosure of the relevant information to the innocent participants (e.g., Bradley & Rettinger, 1992; Bradley & Warfield, 1984) were also excluded. In all these cases, only the control conditions were used. In addition, we excluded conditions that required participants to make non-standard verbal responses to the GKT questions (e.g., respond "yes" to all questions, see Kugelmass et al., 1967; respond with free associations, see Gustafson & Orne, 1965a; respond by repeating the item, see Balloun & Holmes, 1979; Elaad, 1987, or respond "may be" to all items, see Elaad, 1993). (4) Some measure of accuracy rate or differentiation between electrodermal responses to relevant and neutral items (or differentiation between electrodermal responses, elicited by the relevant items, of "guilty" and "innocent" participants) was reported. Unfortunately, several studies (e.g., Gudjonsson & Haword, 1982; Kunzendorf & Bradbury, 1983; Timm, 1982) did not meet this requirement and could not be used. In addition, van Buskirk and Marcuse (1954), Lahri and Ganguly (1978), Konieczny, Fras, and Widacki, (1984) as well as Krapohl (1994) reported only global accuracy rates,

computed on the basis of all physiological measures. Waid, Wilson and Orne (1981) as well as Waid, Orne and Orne (1981) reported accuracy rates derived from two control questions tests and one GKT together, and therefore the validity of the GKT cannot be assessed from their results. Finally, the early study reported by Ruckmick (1938) was excluded because the number of participants and the number of GKT questions were not specified.

Several characteristics were recorded for each experimental condition included in the meta analysis: (1) Number of GKT questions. (2) Number of repetitions of each question. (3) The proportion of relevant items within each question (base rate), was recorded for studies that relied only on "guilty" participants[2]. (4) Number of participants (when both “guilty” and “innocent” participants were used, both numbers were recorded). (5) Type of verbal answers to the GKT questions (1 when a “no” answer was required to all the items, and consequently a deceptive answer was given to the relevant item; 0 when no verbal responses were made to the GKT items). (6) Level of motivation (1 when motivational instructions were given; 0 when no such instructions were applied). A manipulation of motivation level is achieved either through instructions (see, for example, those employed originally by Gustafson & Orne, 1963, in which participants were informed that only people with high intelligence and self-control can avoid detection), or by promising an incentive for the desirable outcome (e.g., Bradley & Warfield, 1984) or a punishment for an undesirable outcome (e.g., Lykken, 1959).

The various GKT studies were classified into 5 general categories:

1. Studies using a card-test paradigm, in which participants are requested to choose one card from a pile of several cards usually containing numbers, but sometimes words or pictures (for early examples of card-test experiments, see Geldreich, 1941,

1942; Ellson et al., 1952). During the second phase of the experiment, participants are asked a series of questions about the various cards (e.g., "Did you choose card no. x?"), while their electrodermal responses to the various questions are being monitored. Typically, card-test experiments include only "guilty" participants (i.e., participants who actually choose a card and therefore have knowledge of the relevant information).

2. The peak of tension paradigm (POT), which is a special case of the card-test procedure, where the various items of each GKT question are presented in a pre-determined order, known to the participants. Typically, the items are numbers (e.g., card numbers) presented either in a descending or ascending order (e.g., Gustafson & Orne, 1964; Horvath, 1978, 1979). Consequently, participants know exactly when the relevant item will be presented, and tension is assumed to accumulate gradually to its peak at the critical moment.

3. The code-words paradigm, in which participants are required to over-learn a series of "code words". At the second stage of the experiment, participants are presented with a series of GKT questions, which includes the learned code words, with several neutral control words added to each code word. This paradigm might include only "guilty" participants (participants who actually learned the code words, see for example, Horneman & O'Gorman, 1987; Thackray & Orne, 1968), but in some studies (e.g., Waid, Orne, Cook & Orne, 1978), a group of "innocent" participants (who did not study any code words) was also included. The code-word paradigm was extended to other stimuli. For example, Ben-Shakhar, Gati and their colleagues (e.g., Ben-Shakhar & Gati, 1987) used both verbal descriptions of people, and their schematic faces, which participants were instructed to memorize.

4. The personal-items paradigm, in which personal items (such as first name, family name, date of birth) are used as the relevant items embedded within several neutral control items of the same categories (e.g., first names, family names, dates of birth of other participants). Some studies that applied this paradigm used only "guilty" participants (e.g., Ben-Shakhar et al., 1975; Lykken, 1960), while others included an additional group of "innocent" participants whose personal items were not presented (e.g., Ben-Shakhar & Elaad, 2002).

5. The mock-crime paradigm, in which participants simulating the guilty commit a mock crime (e.g., a theft of an envelope containing some money and jewelry). The details of the mock crime (e.g., the exact amount of money stolen) serve later as the relevant items of the GKT questions, and each of these items is embedded within several neutral, control items of the same category. Most mock-crime studies include both "guilty" and "innocent" participants (e.g., Bradley & Warfield, 1984; Lykken, 1959), but some (e.g., Bradley, MacLaren & Carle, 1996; Cutrow et al., 1972) included only "guilty" participants, in which case the expected correct detection rate among the innocents is estimated statistically.

The 169 conditions included in the 80 studies compiled for this meta analysis were classified into these 5 categories and are presented in Tables 1 to 5. Separate tables were constructed for studies that included only "guilty" participants and for studies that included both guilty and innocents.

Insert Tables 1-5 about here

Most studies identified were published in refereed journals, but to avoid the "file drawer problem" (e.g., Rosenthal, 1979), we felt it is important to include relevant studies, even if they were published as research reports or in non-refereed journals. Indeed, we identified four Ph.D. dissertations, two of which were not

published in refereed journals (Diaz, 1985; Furumitsu, 1999), four unpublished research reports (Carlton & Smith, 1991; Ellson et al., 1952; Gaines, 1992; Kubis, 1962) and a conference presentation (Horowitz, Kircher & Raskin, 1986). In addition, seven of the studies included in this meta-analysis (e.g., Suzuki et al., 1979a, 1979b) were published in the official journal of the American Polygraph Association (*Polygraph*). A specific analysis will be conducted to examine whether the unpublished results and those published in non-refereed outlets deviate from the results published in refereed journals

Measures of GKT validity

Most GKT studies report their results in terms of detection-accuracy rates. The correct detection rates (CDR) that were reported for each experimental condition of each study are included in Tables 1 -5. Unfortunately, accuracy rates are not very helpful from a meta-analytic perspective because they depend on the specific decision rule for classifying individuals into the categories of "guilty" versus "innocent", employed in a particular study. In addition, accuracy rates may be biased by variations in base rates that were used in the different studies.

It should be noted that the outcome of a GKT test is a measure of relative responding to the significant items versus the neutral-control items. For example, some researchers (e.g., Ben-Shakhar & Lieblich, 1982) used an average of the within-individuals standardized responses to the relevant items. Others (e.g., Ellson et al., 1952) used the mean rank of the responses to the relevant items. A third procedure, that has become very popular among GKT researchers, is the Lykken scoring procedure (e.g., Lykken, 1959). This procedure assigns a score of 2 to each GKT question, if the largest response was elicited by the relevant item; a score of 1, if the second largest response was elicited by that item; and a score of 0 otherwise. These

scores are then summed across all GKT questions (or repetitions) to yield a detection score. Whatever the chosen detection measure, a cutoff point must be set if one wishes to classify individuals into “guilty” versus “innocent”. Clearly, many cutoff points may be defined, and the obtained detection rate reflects the particular choice of a cutoff point. Furthermore, many studies (in particular studies that relied on the card-test paradigm) use only “guilty” participants. In this case, detection rates depend on the proportion of relevant items among all items (the base-rate). Under a simple decision rule, which classifies an individual as “guilty” if the largest mean response was elicited by the relevant item, the base is the expected rate of false-positive outcomes (individuals classified as "guilty", although they have no guilty knowledge). Thus, it might be misleading to average correct detection rates across studies that relied on different decision rules and used different base rates.

One way to get around this problem is by using signal-detection measures. Many studies conducted by Ben-Shakhar and his colleagues (e.g., Ben-Shakhar, 1977; Ben-Shakhar, & Gati, 1987; Ben-Shakhar, Lieblich & Kugelmass, 1970; Elaad & Ben-Shakhar, 1997) used the signal-detection approach. This approach provides measures of detection efficiency that do not depend on a single, arbitrary, cutoff point. Rather, a statistic is derived, which describes detection efficiency by comparing the entire distributions of the detection score of “guilty” and “innocent” participants (or the distributions of the responses to the critical and the neutral items, where only a ”guilty” sample is used). A Receiver Operating Characteristic (ROC) curve is generated on the basis of these distributions, and the area under the ROC curve is computed.

The area under the ROC curve assumes values between 0 and 1, such that an area of 0.5 means that the two distributions (e.g., the distributions of the detection

scores of guilty and innocent examinees) are undifferentiated, and therefore it is impossible to use the responses for detecting whether an examinee is guilty or not. An area of 1 means that there is no overlap between the two distributions, and therefore a perfect classification of guilty and innocent examinees would be possible. A more detailed description of signal detection theory and its applications can be found in several sources (e.g., Bamber, 1975; Green & Swets, 1966; Swets, Tanner & Birdsall, 1961).

Although most GKT studies did not use the signal-detection approach and did not report the area statistic as a measure of detection efficiency, if certain assumptions about the detection score distributions are made, it is possible to derive the area under the ROC curve from the correct detection rates obtained for the "guilty" and "innocent" samples. In fact, the required assumptions are similar to those made routinely for analyses of variance, namely that the detection score distributions for both groups are Normal with equal variances. The distance (in standard deviation units) between the centers of the two distributions (d) can be directly derived under these assumptions (for a more detailed description of this derivation see Ben-Shakhar, Lieblich & Bar-Hillel, 1982). This approach was applied whenever d was not reported, or could not be computed from the data reported in a given study. The area under the ROC curve (a) can be derived from d by the following formula: $a = \Phi(d/\sqrt{2})$, where Φ is the standard Normal ogive function.

The main focus of this meta-analysis was on the d statistic, which describes the strength of the effect (the degree of separation between the two distributions of the detection score). This is a standard measure for strength of effects in psychological experiments (see, Cohen, 1988) and it is convenient for the present purposes because its sampling error can be estimated (see, Hunter & Schmidt, 1990). For studies that

used both "guilty" and "innocent" samples, the d statistic represents the distance (in standard deviations) between the means of the detection score distributions of the two samples. For studies that used only "guilty" participants, d represents the distance between the means of the electrodermal response distributions of the relevant and the control items. In addition to d, and a, a more conventional measure for the validity of the GKT, namely the Point-Biserial correlation between the detection measure and the criterion of guilt versus innocent (r), was also included. This correlation coefficient has slightly different meaning in studies that used both "guilty" and "innocent" samples, and studies that used only "guilty", or knowledgeable participants. In the former case it reflects the correlation between the detection score and the dichotomous variable of experimental condition ("guilty" versus "innocent"), and in the latter case it reflects the correlation between the response magnitude and the dichotomous variable representing the nature of the item (relevant versus control). The Point-Biserial correlation coefficient was derived from d by the following formula: $r=d\sqrt{[pq/(1+pqd^2)]}$, where p is the proportion of “guilty” participants and $q=1-p$.

In some studies, it was possible to derive d directly from the data reported in the original report (e.g., Ben-Shakhar, et al., 1995). This was possible either when d values were described in the report, or when means and standard deviations of the detection score among each group (“guilty” and “innocents”) were available. In those cases, a and r were derived from the d value, regardless of whether or not correct detection rates were reported. In some cases (e.g., Cutrow, et al., 1972; Tackray & Orne, 1968), the mean rank of the critical item was reported. In these cases, d was computed by subtracting the expected value of the mean rank from the observed mean rank and dividing this difference by the standard deviation, where the mean and

standard deviation were computed from the rectangle distribution of the mean rank (i.e., the ranks' distribution that would be obtained for an innocent individual).

The 3 statistics ("d", "a" and "r") were derived for each study from the data reported by the study's authors. If any of these statistics, as well as the correct detection rates, were reported in the original study, they were highlighted in Tables 1-5. If the statistics d, a, and r were not provided in a given study, they were derived from the correct detection rates, as described above. In some studies only the area statistic was reported (e.g., Ben-Shakhar, 1977; Elaad, Bonwitt, Eisenberg & Meytes, 1982) and d, as well as r were derived from it by the two formulas described above.

Four studies reported perfect detection of both "guilty" and "innocent" (e.g., Bradley & Rettinger, 1992). In these cases, d cannot be computed and therefore we adopted a conservative approach and treated the data as if both error rates were 0.5%. For studies that included only a sample of "guilty" participants, the expected false-positive rate was derived from the decision rule for classifying a response as representing a relevant item. Thus, for example, in a typical card-test experiment (e.g., Ellson et al., 1952), the card number producing the maximal mean response is classified as the critical (i.e., chosen) card. Under this rule, the expected rate of false positive outcomes (i.e., the probability that an innocent participant, who did not choose any card, would give a maximal mean response to the relevant card) is equal to the proportion of relevant cards in the set of cards presented (the base rate). In studies that relied on the Lykken (1959) scoring method, the expected rate of false positive outcomes was derived from the theoretical distribution of the Lykken score and the cutoff point used in the particular study (for a detailed demonstration of this procedure, see Timm, 1989). Thus, for studies that used only a "guilty" sample, the observed rate of correct detection and the expected rate of correct detection among

"innocent" participants served to derive d, a, and r. In these cases, values of 0.5 were assigned to p and q in the computation of r.

Results

First, we computed the weighted averages and standard deviations of the detection-efficiency statistics (d, a, and r) across studies, within each type of study (within each individual Table 1 to 5), as well as across all 169 experimental conditions of all 80 studies (across Tables 1-5). As indicated in the tables, some studies used a within-subjects design, and therefore the same participants were included in more than one experimental condition. In those cases, the weight assigned to each experimental condition was updated according to the number of additional experimental conditions that included the same participants (the "corrected number of observations" in Tables 6-9). For example, Elaad (1993) manipulated the verbal answers to the GKT questions, using a within-subjects design with 24 participants examined in two conditions. Both conditions were included in our computations, but the weight assigned to each was 12, rather than 24. We also computed the 95% confidence interval of d (see Hunter & Schmidt, 1990, pp. 437-438), both within each type of study and across all conditions. Confidence intervals provide useful information about differences between different types of studies. For example, non-overlapping confidence intervals provide sharp confirmation for the effect of a moderator variable. The weighted means of the 3 statistics, along with the 95% confidence interval of d, are presented in Table 6 for each type of study, as well as across all 169 conditions.

Insert Table 6 about here

The results displayed in Table 6 demonstrate a fairly high level of validity, which is reflected by an average effect size of 1.55 (almost twice as large as an effect

size of 0.80, which represents a "large effect size" according to Cohen, 1988). In terms of a correlation coefficient, the overall validity is 0.55. However, there seem to be considerable variations in detection efficiency among the various paradigms. To examine these variations more systematically, and search for possible moderators, we adopted the meta-analytic methods recommended by Hunter and Schmidt (1990) and compared the observed variance of the d statistic to the sampling error variance of this statistic. These comparisons were conducted for each type of study as well as across all studies. The results of these comparisons, which are also displayed in Table 6, reveal that the observed variance across all studies is much larger than the sampling error variance. A similar picture emerges within the five study categories, except for the peak of tension paradigm. In addition to these descriptive comparisons, we conducted statistical tests for homogeneity (see, Hunter & Schmidt, 1990, p. 428) within each study category, as well as across all 169 conditions. The homogeneity assumption was rejected ($p<.05$) on all cases, except for the peak of tension paradigm. Consequently it seems justified to search for possible moderators.

Statistical comparisons between the d values obtained under the 5 paradigms (see, Hunter & Schmidt, 1990, p. 438) reveal that the effect size obtained under the mock-crime paradigm is significantly larger ($p<.05$) than the effects obtained under all other paradigms. The average d obtained under the personal-item paradigm is significantly larger than the d values obtained under the code words and POT paradigms, but not when compared to the card-test. The finding that detection of personal items is easier than detection of code words learned during the experiment is consistent with theories and findings in social cognition (e.g., the "self-reference effect", Greenwald, 1981; Kihlstrom & Cantor, 1984). Finally, no statistically significant differences were observed among the first 3 paradigms, which can be

viewed as variations of the card-test procedure.

Before searching for additional moderator variables, we conducted a preliminary comparison between studies published in refereed journals and all other studies. This comparison was conducted across the 5 paradigms, because the number of non-refereed publications within each study type was too small. Fourteen studies (containing 21 conditions, with a total corrected sample size of 664) which were not published in refereed journals were identified. The weighted averages of d, a, and r computed across all these conditions were 1.46, 0.79 and 0.56, respectively. These values are very similar to the weighted averages of these statistics, computed across all studies (1.55, 0.81 and 0.55). The standard deviation of d computed across these 21 conditions was also similar to the overall standard deviation (1.09 vs. 0.91, respectively). Thus, it was concluded that the studies published in non-refereed outlets do not deviate from the complete data-set, and their inclusion in the meta-analysis is justified.

In an attempt to search for additional moderators, 3 factors that might be promising candidates for moderating detection efficiency were identified on the basis of the GKT literature: (1) Motivation to succeed in the polygraph test, where success typically translates into an "innocence" outcome. (2) A deceptive verbal response to the critical item (typically, a "no" answer) versus an absence of an answer. (3) The number of GKT questions.

To examine these factors, we first divided all 169 experimental conditions into two categories: (a) High-motivation conditions, in which motivation was created either through instructions (e.g., Gustafson & Orne, 1963), or through a monetary incentive for producing the desirable outcome (e.g., Bradley & Warfield, 1984). (b) Low-motivation conditions, where no special motivational instructions were provided

and no incentive was promised. The weighted averages of d and a were computed separately within each type of motivation condition. In addition, the 95% confidence interval for d and its residual variance (S_d^2–S_e^2) were computed within each motivational level. These results are displayed in Table 7. Second, the 169 experimental conditions were divided into two categories according to the type of verbal response made: (a) A deceptive verbal response ("no" response) to the relevant item; (b) Absence of verbal responses to the GKT items. The weighted averages of d and a, as well as the 95% confidence interval of d and its residual variance, were computed across experimental conditions within each type of verbal response condition and are displayed in Table 7.

Insert Table 7 about here

Statistical tests conducted to examine the effects of the motivation and verbal response factors demonstrated that high motivation level was associated with larger d than low motivation level (Z=3.12). Deceptive verbal responses to the GKT items were associated with larger d than that obtained in the "silence" condition, but the difference was not statistically significant (Z=1.45). A similar statistical test conducted to compare the two correlation coefficients (0.58 under the "No" response condition vs. 0.50 under the "Silence" condition) did produce a statistically significant outcome (Z=2.66). However, the residual variance within each type of studies is quite large (see Table 7) and the homogeneity hypothesis was rejected in each of the 4 cases. Thus, there might be additional moderator variables that could account for the differences among studies. One possibility is to classify the experimental conditions into 4 categories created by the various combinations of motivational level and verbal-response type. All the statistics displayed in Table 7 were recomputed for each of these 4 categories and are displayed in Table 8.

Insert Table 8 about here

Statistical tests were conducted to examine the effect of motivation within each verbal response condition, and the effect of deceptive verbal response within each motivation condition. Motivational instructions significantly ($p<.05$, one tailed) increased d within each verbal response condition ($Z=1.72$ in the "no" condition, and $Z=1.92$ in the "silence" condition). Deceptive verbal response significantly increased d only under the low motivation condition ($Z=1.89$). These results indicate that an increased motivation level may be more important than requiring a "no" verbal response, because enhanced motivational level produces a fairly large effect size even when no verbal responses are required.

However, the results displayed in Table 8 reveal that even when the type of verbal response and the level of motivation are jointly considered there is still a positive residual variance. Moreover, the homogeneity assumption was rejected for each of the 4 conditions displayed in Table 8. It is possible of course that the data should have been analyzed for each combination of motivation level and type of verbal response within each type of study, but the small number of studies within each cell renders such an analysis useless.

Another potential moderator is the number of GKT questions or repetitions. Since the number of repetitions is negatively related to the number of different questions used (studies that used just one or two questions, typically repeated each question several times, whereas studies that used 10 or more different questions presented each question just once), we used a multiple regression approach and regressed the d value on both the number of questions and number of repetitions. This analysis, which was conducted on 166 conditions for which the information on the numbers of repetitions was available, revealed that although both factors had positive

regression weights, only the number of questions contributed significantly to the prediction of d. The Pearson correlations between the number of GKT questions used and the three detection efficiency measures, computed across all 169 experimental conditions were, 0.35, 0.27 and 0.26 for d, a and r, respectively. Thus the number of questions accounts for about 12% of the overall variance of d, while the marginal contribution of the number of repetitions is negligible. We approached this issue by another method and divided the 169 experimental conditions into two categories: Studies which used a small number of GKT questions (less than 5) and those that relied on at least five different questions. The weighted averages of a and d, the 95% confidence interval of d and its residual variance, were computed, across experimental conditions, within each of these categories, and are displayed in Table 9.

Insert Table 9 about here

Inspection of Table 9 reveals that indeed the number of GKT questions used does make a difference. Across all experimental conditions, effect size estimates of 2.35 and 1.29 were obtained for GKT studies with large and small numbers of GKT questions, respectively. Furthermore, the 95% confidence intervals computed for these two categories don't overlap, so the difference is statistically significant. This result is not surprising and is consistent with the basic psychometric principles.

Finally, we identified a subset of mock-crime studies, which are most relevant for estimating the validity of the GKT under optimal conditions. This subset includes those mock-crime studies that used motivational instructions, a deceptive verbal response, and relied on at least 5 GKT questions. An analysis of this subset, which includes 10 experimental conditions, reveals an average effect size of 3.12 (which is almost 4 times larger than an effect size of 0.80, considered a "large effect size" by

Cohen, 1988), with a 95% confidence interval of 2.27-3.98. The mean a, and r values computed across these 10 studies were, 0.95 and 0.79, respectively.

Discussion

Two major conclusions can be drawn from the results of this meta-analysis. First, these results indicate that the electrodermal measure can provide an efficient means for detecting relevant information and for differentiating between individuals with guilty knowledge and those who do not have that knowledge. Even a crude estimate of the validity of the GKT, computed across all types of studies, represents a relatively high level of validity (an effect size of 1.55 standard deviations, or a correlation coefficient of 0.55). A reliance on a smaller sub-set of studies, which are more similar to and more relevant for possible applications of the GKT, results in a much higher estimate for the validity of this technique. Second, our analysis demonstrates that real variations exist between studies, and several factors that moderate the validity of the GKT were identified.

One such factor is the research paradigm used to study the validity of the GKT. Relatively low levels of validity (effect size estimates ranging between 1.1 and 1.3 standard deviations) were demonstrated for the various variations of the card-test procedure (including the POT and the "code-words" paradigm). These procedures which are characterized by instructing participants to memorize certain items, that don't have any special meaning to the participants outside the context of the experiment, do not resemble in any way the kind of criminal event that is typically the focus of a realistic GKT investigation.

A somewhat higher level of validity (reflected by an average effect size of 1.58 or a correlation of 0.57) was observed for the personal-item paradigm. In this paradigm biographical information is used instead of neutral items memorized for the

sake of the experiment. Clearly, items such as one's own name are more significant for an individual than arbitrary card numbers. This result is consistent with the cognitive approach to psychophysiological detection (see Ben-Shakhar & Furedy, 1990), which postulates that an orienting response elicited by a stimulus reflects the degree to which this stimulus was attended to, or the depth to which it was processed. Biographical information, such as the person's own name, attracts attention (e.g., Moray, 1959) and elicits enhanced ORs (Ben-Shakhar et al., 1975). The idea that self-relevant information is more deeply processed was also supported by experiments demonstrating superior memory for items encoded with respect to the self (the "self-reference effect", see Greenwald, 1981; Kihlstorm & Cantor, 1984; Kihlstorm et al., 1988).

The highest level of validity (an effect size of about 2.09, or a correlation of about 0.65) was observed for the mock-crime procedures. This d value was significantly larger than each of the 4 d values observed under the other paradigms. This result is encouraging from an applied perspective because mock-crime procedures attempt to simulate real criminal events, and participants are required not just to memorize a few items, but for example steal an envelop containing money and jewelry from an office.

An increased motivation to produce a desirable outcome is associated with increased validity. Across all studies, the average d values obtained for the high and low motivational conditions were 1.84 and 1.36, respectively. Furthermore, the 95% confidence intervals of d computed under these 2 conditions don't show any overlap, which means that the difference is reliable. This outcome is consistent with the early findings of Gustafson and Orne (1963, 1965a) as well as several subsequent studies (e.g., Elaad, 1987), but not with other studies (e.g., Horvath, 1978, 1979; Lieblich et

al., 1974). From a theoretical perspective, this result is consistent with the notion that any manipulation that increases the level of significance of the critical items would contribute to an increased orientation, reflected by an increased differential responding to these items.

Furthermore, it is interesting to note that a similar result was reported in studies that explored behavioral (non-physiological) cues of deception. DePaulo and Kirkendol (1989) documented the "motivational impairment effect", which means that individuals who are highly motivated to avoid detection are relatively less successful in their attempts when observers can watch their nonverbal behavior. DePaulo and Kirkendol (1989) suggested that this impairment may reflect the fact that deliberate attempts by liars to control their expressive behavior result in overcontrolled and inhibited behavior, which facilitates detection. Similarly, enhanced motivation may be reflected by attempts to control physiological reactions, which increases emotional arousal and facilitates detection.

It should be noted that only two levels of motivation were included in this analysis. The question of whether the relationship between motivation level and detection efficiency is monotonic, beyond these two levels is very important from an applied perspective because motivation is likely to be much higher under realistic than under simulated conditions, even when the latter include incentives.

Two studies manipulated motivation and stress and included levels that seem to resemble the realistic situation. Kugelmass & Lieblich (1966) tested police trainees who were instructed to take a laboratory test that was presented as part of the selection procedure to the police force. The standard card test was administered under three different conditions: A - subjects were told that they would undergo a test designed just to examine whether the apparatus was properly functioning. B - they

were told that the test was designed to examine whether their responses could be detected by the machine. C - subjects were told that the test was designed to examine whether they were suitable for service in the Israeli Police Force. They were further told that one trait that characterizes a successful policeman, with good chances of promotion, is the ability to control emotions. The results revealed that the number of chosen cards correctly identified through inspection of changes in skin resistance was similar in the three stress conditions. The authors concluded that "within a considerable range of stress no necessary decrease in the detection efficiency of the GSR channel need be expected" (Kugelmass & Lieblich, 1966, p. 215). A subsequent study reported by Bradley and Janisse (1981) supported these conclusions. They threatened half of their subjects with an electric shock if classified guilty by the polygraph. This manipulation did not affect detection efficiency through electrodermal, cardiovascular, and pupillary measures, either in a GKT procedure or under the CQT. Thus, on the basis of these 2 studies it seems that detection efficiency estimated in laboratory experiments can be generalized to situations characterized by much higher levels of motivation and stress.

A deceptive (i.e., "no") verbal response to the GKT questions emerged as another factor moderating the validity of the GKT. However, the effect of this factor on d was statistically significant only under the low motivation condition. This result shows that although a deceptive verbal response is not necessary for producing differential responsivity to the relevant items, it may contribute to enhancing it. This conclusion is consistent with findings reported by Furedy and his colleagues, who used the "differentiation of deception" paradigm, which attempts to isolate the deception factor and examine whether enhanced physiological responding is associated with deception, when other factors (e.g., stimulus significance, relative

frequency of the relevant items) are held constant (Furedy, Davis & Gurevich, 1988; Furedy, Giglioti & Ben-Shakhar, 1994; Furedy, Posner & Vincent, 1991; Vincent & Furedy, 1992).

It should be noted that although several factors moderating the validity of the GKT were identified in this study, it is difficult to make an accurate assessment of the contribution of each factor. In particular, both level of motivation and the number of GKT questions used are not independent of the research paradigm. Mock-crime studies typically involve enhanced levels of motivation and rely on more GKT questions than the various versions of the card-test paradigm. Thus, the advantage of the mock-crime procedure over other paradigms may be related to its use of more questions.

From a practical perspective, the results of this meta-analysis demonstrate that when properly administered, the GKT may turn out to be one of the most valid applications of psychological principles. An overall effect size of 1.55 is already impressive, but it is quite clear that from a practical point of view, only a subset of the studies included in this analysis are relevant. In particular, the various versions of the card-test paradigm seem too remote from the criminal investigation context where the GKT can be used. The personal-items paradigm, might be relevant for some cases, where the identity of an individual is the core issue (for an interesting example of how this paradigm could have been used in the case of John Demjanyuk see, Lykken, 1991). The mock-crime studies, which are based on simulations of real crimes, and particularly the subset of mock-crime studies that used motivational instructions, a deceptive verbal response, and relied on at least 5 GKT questions seem most relevant for estimating the validity of the GKT under optimal conditions. An analysis of this subset reveals that when properly administered the GKT has an impressive potential.

Very few applications of principles derived from the behavioral sciences reach levels of validity larger than 0.70. The effect size estimated from this sub-set of studies is larger than 3.0 standard deviations, which represents an effect size more than 3 times larger than what Cohen (1988) considered as a large effect. However, even within this highly selected subset of studies there seem to exist real variations among studies unaccounted for by a combination of the moderator variables identified by this meta analysis. Perhaps further research will reveal additional moderators that would allow for yet a further increase in the validity of the GKT.

Although the results of this meta analysis display a rather bright picture for the GKT, there are several limitations that should be considered. Particularly, it should be stressed that the studies included in this meta analysis were experimental studies conducted in laboratory conditions, which do not resemble the realistic conditions of criminal investigation in which the GKT may be applied. The experimental situation differs from the realistic one in several important aspects.

First, as noted by Ben-Shakhar and Furedy (1990), factors affecting perception and memory might be crucial for the efficiency of the GKT. Unfortunately, the GKT studies, including those that relied on the mock-crime paradigm, used very simple tasks in which the experimenters guaranteed that all participants learned all the relevant items. Furthermore, they were typically tested immediately after being exposed to the guilty information, so memory did not play an important role in the experimental situation. In real life, things might be entirely different. The guilty person is faced with a complex scene, and it is much more difficult to assume that all details are indeed noticed, processed, and stored in memory. Criminal suspects are rarely tested immediately after committing the criminal act. Typically, they may be tested days, weeks, and sometimes months after the crime was committed. Elaad

(1997) conducted a mock-crime experiment in which participants were tested several days (between 2 days and a week) after committing the mock crime. He found that several participants did not remember 1 or 2 out of the 4 critical items that were used. Future research should be conducted to examine whether and to what extent these factors affect the accuracy of the GKT.

These differences between the experimental and the realistic setups may account for the relatively large rates of false-negative outcomes observed in the two field GKT studies reported thus far (Elaad, 1990; Elaad et al., 1992). While the rates of false-positive errors obtained in these studies were as low as those reported in laboratory experiments (2% in the former study, which relied only on the electrodermal measure, and 5% in the latter study, which utilized a combination of electrodermal and respiration measures), the rates of false-negative errors were much larger (42% in the former study and 20% in the latter).

We reanalyzed the results of the two field studies, using the same methods applied for the meta-analysis. For each study, the r and d values were computed from the published results, based only on the electrodermal measure. The weighted averages of the r and d values, computed across the 2 studies, were 0.60 and 1.49, respectively. These values are more or less equivalent to the respective values obtained in our meta-analysis on the basis of the personal-items studies, but are lower than those obtained for the mock-crime studies.

This discrepancy may suggest that the mock-crime studies lack ecological validity. It should, however, be noted that the use of the GKT in the criminal cases studied by Elaad (1990) and Elaad et al. (1992) was not optimal. In particular, the mean number of questions used in these field studies was rather small (the mean number of questions used by Elaad, 1990 and by Elaad, et al., 1992 were 2 and 1.8,

respectively). In addition, the two field studies were based on GKTs that were administered immediately after a CQT, and this might attenuate the sensitivity of the physiological measures due to habituation. Thus, it is possible that the relatively high rates of false-negative errors and lower detection efficiency obtained in these field studies resulted from the use of a small number of GKT questions and from the manner in which the test was applied. It is clear that more field studies are required, and in particular field studies that will attempt to apply the GKT under optimal conditions.

Second, the GKT may be severely affected by leakage of the relevant information to innocent suspects. Several studies examined this issue (Ben-Shakhar, Gronau & Elaad, 1999; Bradley & Rettinger, 1992; Bradley & Warfield, 1984; Bradley et al., 1996) and demonstrated that the false-positive rate among innocent participants informed about the relevant items ranges between 25% and 50%. Although there is no particular evidence of leakage having occurred in laboratory studies, it could occur in the field. In realistic situations, the critical items may be leaked to innocent suspects, affecting false-positive outcomes, especially if the informed innocent suspects are unable to explain how they became aware of this information. From this respect, the results of the two field studies conducted by Elaad and his colleagues (Elaad, 1990; Elaad, et al., 1992) are encouraging, because the false-positive outcomes reported in these studies were very small and resembled those obtained in experimental studies. This indicates that leakage of relevant information did not play a role in the criminal investigations examined by Elaad and his colleagues. Again, it is clear that more field studies are needed to resolve the question of whether information leakage is a serious threat to the validity of the GKT.

Third, several studies (e.g., Ben-Shakhar & Dolev, 1996; Elaad & Ben-Shakhar, 1991; Honts et al., 1996; Honts, Raskin, & Kircher, 1987, 1994; Kubis, 1962) demonstrated that both the GKT and the CQT are vulnerable to countermeasures (i.e., deliberate attempts by guilty examinees to distort the test results). These techniques rely either on the use of physical means (such as biting one's tongue), or mental means (calling to mind an exciting or frightening event, or engaging in mental activities that require effort) each time a control question is asked. It should be pointed out that mental countermeasures are most detrimental for all psychophysiological-detection techniques because they cannot be detected even by experienced examiners. Two recent studies demonstrated that mental countermeasures can be used effectively under both the GKT and the CQT (Ben-Shakhar & Dolev, 1996; Honts et al., 1996). Clearly, countermeasures may increase false-negative outcomes (guilty suspects classified as "innocents"), but they have no effect on innocent examinees.

So far we discussed several features of the investigation setup that might increase the error rates relative to the simulated GKT studies that constituted the basis for the present meta analysis. But there are additional factors that may positively affect the outcomes of the GKT in applied setups. While our meta analysis was based only on a single physiological measure (changes in electrodermal activity), the usage of the GKT in real-life criminal interrogations is typically based on several physiological responses. Although several studies demonstrated that electrodermal measures are more effective than any other autonomic measure used for psychophysiological detection in the GKT (e.g., Cutrow et al., 1972; Thackray & Orne, 1968; Waid, Wilson & Orne, 1981), it is clear that a detection measure based on a combination of several physiological indices would yield more accurate results. In

particular, a combination of skin conductance changes with a measure of respiration changes (respiration line length – RLL) produced the highest levels of accuracy rates (e.g., Ben-Shakhar & Dolev, 1996; Ben-Shakhar et al., 1999; Elaad & Ben-Shakhar, 1997; Timm, 1982, 1987). For example, Elaad and Ben-Shakhar (1997) reported an increase in the area under the ROC curve from 0.79 to 0.86 when the RLL was added (with an equal weight) to the electrodermal measure.

An additional measure, that has not been applied yet to realistic GKT investigations but has been demonstrated to be very promising in experiments, is derived from electrophysiological brain activity (i.e., Event-Related Potentials -- ERPs). Several studies demonstrated that the P300 component of the ERP, which represents cognitive activity occurring within 300-500 mili-seconds after stimulus onset, can be used successfully in the GKT (e.g., Allen, Iacono & Danielson, 1992; Farwell & Donchin, 1991; Rosenfeld, Cantwell, Nasman, Wojdac, Ivanov & Mazzeiri, 1988).

Our meta analysis demonstrated that the GKT has an excellent potential as an applied method for detecting information and for differentiating between individuals possessing guilty knowledge and innocents. Furthermore, the GKT has various advantages over alternative psychophysiological detection methods because it is based on sound theoretical foundations and a standardized procedure. This raises a question regarding the limited usage of the GKT in criminal investigations in North America.

The difficulty in identifying a sufficient number of salient features of the event, which can be used to formulate proper GKT questions, may be one reason for the limited application of this method. A proper GKT question refers to a specific feature of the event that is very likely to be noticed by a guilty person. Furthermore, it

is crucial that guilty individuals not only notice the designated feature, but also remember it during the polygraph investigation, which may occur long after the event. Podlesny (1993) estimated that the GKT might have been used in only 13% of FBI cases for which polygraphs have been used. But the fact that the GKT has been used for many years by Japanese law enforcement agencies as the preferred method of psychophysiological detection (Fukumoto, 1980; Nakayama, 2001; Yamamura & Miyata, 1990) demonstrates that it is possible to formulate a sufficient number of proper GKT questions in many criminal cases.

Another possible reason for the infrequent application of the GKT in North America might be the strong belief of many polygraph examiners in the validity of the CQT. Indeed, the CQT is much easier to implement than the GKT. There is no need to identify critical features of the event and the relevant questions are simple and straightforward ("Did you do it?). With the CQT, there is no need to inspect the scene of the crime, and the issue of information leakage becomes moot. But the CQT is typically administered in a non-standardized and contaminated fashion (see Ben-Shakhar, 1991; Ben-Shakhar et al., 1986). We believe that only when law-enforcement authorities will understand the flaws involved in the application of the CQT (see, Ben-Shakhar, 2001; Iacono et al., 1997, 1999), and realize the need for a scientifically-based detection method (rather than a confession-inducing technique), will they make the efforts required for implementing the GKT as a standard detection tool.

To follow the Japanese example and apply the GKT as a standard investigative tool in a large number of criminal investigations it will be necessary to modify police practices, such that critical features of the event are identified and concealed at the outset of the investigation. Furthermore, GKTs should be conducted by investigators

who are familiar with the scene of the crime and are trained to look for salient features that could be utilized as GKT questions. Admittedly, even if all these efforts are made, there will still be various criminal cases for which the GKT is not applicable. But we believe that the possibility of applying the GKT, even for a sub-set of criminal investigations, justifies these efforts. Furthermore, with the accumulation of field studies, the GKT may even become admissible evidence in criminal courts (for a recent discussion of this issue, see Ben-Shakhar et al., 2002).

References

References marked with an asterisk indicate studies included in the meta-analysis.

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.

Ansley, N. (1992). The history and accuracy of guilty knowledge and peak of tension tests. *Polygraph*, *21*, 174-274.

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

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.

*Barland, G.H. (1997). Research in electrodermal biofeedback with stimulus tests. *Polygraph, 26*, 1-37.

*Beijk, J. (1980). Experimental and procedural influences on differential electrodermal activity. *Psychophysiology*, *17*, 274-278.

*Ben-Shakhar, G. (1977). A further study of the dichotomization theory in detection of information. *Psychophysiology, 14,* 408-413.

Ben-Shakhar, G. (1991). Clinical judgment and decision making in CQT Polygraphy: A comparison with other pseudoscientific applications in Psychology. *Integrative Physiological and Behavioral Science*, *26*, 232-240.

*Ben-Shakhar, G. (1994). The roles of stimulus novelty and significance in determining the electrodermal orienting response: Interactive vs. additive approaches.

Psychophysiology, *31*, 402-411.

Ben-Shakhar, G. (2001). A Critical Review of the Control Questions Test (CQT). In: M. Kleiner (Ed.). *Handbook of Polygraph Testing*, Academic Press, 103-126.

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: Reconsidering the use of the GKT in court. Manuscript in preparation

Ben-Shakhar, G., Bar-Hillel, M. & Lieblich, I. (1986). Trial by polygraph: Scientific and juridical issues in lie detection. *Behavioral Science and the Law*, *4*, 459-479.

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

Ben-Shakhar, G., & Elaad, E. (2001). The Guilty Knowledge Test (GKT) as an application of psychophysiology: Future prospects and obstacles. In: M. Kleiner (Ed.). *Handbook of Polygraph Testing*, Academic Press, 87-102.

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

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

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

*Ben-Shakhar, G., & Gati, I.(1987). Common and distinctive features of

verbal and pictorial stimuli as determinants of psychophysiological responsivity. *Journal of Experimental Psychology: General*, *116*, 91-105.

*Ben-Shakhar, G., Gati, I., & Salamon, N. (1995). Generalization of the orienting response to significant stimuli: The roles of common and distinctive stimulus components. *Psychophysiology*, *32*, 36-42.

*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. (1982). The dichotomization theory for differential autonomic responsivity reconsidered. *Psychophysiology, 19,* 277-281.

Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, M. (1982). An evaluation of Polygrapher's judgments: A review from a decision theoretic perspective. *Journal of Applied Psychology*, *67*, 701-713.

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

*Ben-Shakhar, G., Lieblich., I. & Kugelmass, S. (1975). Detection of information and GSR habituation: An attempt to derive detection efficiency from two habituation curves. *Psychophysiology*, *12*, 283-288.

*Ben-Shakhar, G., Lieblich, I., & Kugelmass, S. (1982). Interactive effects of stimulus probability and significance on the skin conductance response. *Psychophysiology*, *19*, 112-114.

Berlyne, D.E. (1960). *Conflict Arousal and Curiosity*. New York: McGraw-Hill

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

*Bradley, M.T., & Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. *Psychophysiology, 21,* 63-71.

*Bradley, M.T., & Janisse (1981). Accuracy demonstration, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures. *Psychophysiology, 18,* 307-315.

*Bradley, M.T., MacLaren, V.V. & Carle, S.B. (1996). Deception and nondeception in guilty knowledge and guilty action polygraph tests. *Journal of Applied Psychology, 81,* 153-160.

*Bradley, M.T., & Rettinger, J. (1992). Awareness of crime-relevant information and the guilty knowledge test. *Journal of Applied Psychology, 77*, 55-59.

*Bradley, M.T., & Warfield, J.F. (1984). Innocence, information, and the guilty knowledge test in the detection of deception. *Psychophysiology, 21,* 683-689.

*Carlton, B.L., & Smith, B.J. (1991). *The effects of aural versus visual presentations of questions during a detection of deception task*. Report No. DoDPI91-R-0002, Department of Defense Polygraph Institute, Ft. McClellan, AL 36205, January, 1991.

Cohen, J.E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence 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,* 578-587.

*Davidson, P.O. (1968). Validity of the guilty knowledge technique: The effect of motivation. *Journal of Applied Psychology, 52,* 62-65.

Davis, R. C. (1961). Physiological responses as a means of evaluating information. In: A. D. Biderman & H. Zimmer (Eds.) *The manipulation of human behavior*. New York, Wiley, pp. 142-168.

Daubert v. Merrell Dow Pharmaceuticals. 113 C. Ct. Supp. 2786 (1993).

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.

*Day, D.A., & Rourke, B.P. (1974). The role of attention in "lie-detection". *Canadian Journal of Behavioral Sciences*, *6*, 270-276.

DePaulo, B. M., & Kirkendol, S. E. (1989). The motivational impairment effect in the communication of deception. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 51-70). Dordrecht, The Netherlands: Kluwer Academic Publishers.

DePaulo, B. M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2002). Cues to deception. Manuscript submitted for review.

DePaulo, B. M., & Pfeifer, R.L. (1986). On the job experience and skill at detecting deception. *Journal of Applied Social Psychology*, *16*, 249_267.

DePaulo, B. M., Stone, J. I., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schlenker (Ed.), *The self and social life* (pp. 323-370). New York: McGraw-Hill.

*Diaz, R. (1985). *The effects of pre-experimental expectancy, opinion and demonstration of accuracy on physiological detection of information*. Unpublished Ph.D. theses, Graduate College of Bowling Green State University.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*, 913-920.

Ekman, P., O'Sullivan, M., & Frank, M.G. (1999). A few can catch a liar?

Psychological Science, *10*, 263-266.

*Elaad, E. (1987). *Psychophysiological detection in the guilty knowledge test*. Unpublished Ph.D. thesis, The Hebrew University of Jerusalem, Israel.

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

*Elaad, E. (1993). The role of guessing and verbal response type in psychophysiological detection of concealed information. *The Journal of Psychology, 127,* 455-464.

*Elaad, E. (1994). The Accuracy of human decisions and objective measurements in psychophysiological detection of knowledge. *The Journal of Psychology, 128,* 267-280.

*Elaad, E. (1997). Polygraph examiner awareness of crime-relevant information and the guilty knowledge test. *Law and Human Behavior, 21*, 107-120.

Elaad, E. (1998). The challenge of the concealed knowledge polygraph test. *Expert Evidence*, *6*, 161-187.

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. (1991). Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. *International Journal of Psychophysiology*, *11*, 99-108.

*Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology, 34,* 587-596.

*Elaad, E., Bonwitt, G., Eisenberg, O., & Meytes, I. (1982). Effects of beta

blocking drugs on the polygraph detection rate: A pilot study. *Polygraph, 11*, 225-233.

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

*Ellson, D.C., Burke, C.G. Davis, R.C. & Saltzman, I.J. (1952). *A report of research on detection of deception.* Contract NG onr-18011, Office of Naval Research.

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

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.

*Forman, R.F., & McCauley, C. (1986). Validity of the positive control polygraph test using the field practice model. *Journal of Applied Psychology, 71*, 691-698.

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

*Furedy, J.J., & Ben-Shakhar, G. (1991). The role 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. (1991). The forensic use of the polygraph: A psychophysiological analysis of current trends and future prospects. In: J.R. Jennings, P.K. Ackles and M.G.H. Coles (eds.), *Advances in Psychophysiology*, *4*, Jessica Kingsley Publishers Ltd.

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology, 25*, 683-688.

Furedy, J. J., Gigliotti, F., & Ben-Shakhar, G. (1994). Electrodermal differentiation of deception: The effect of choice vs. no choice of deceptive items. *International Journal of Psychophysiology*, *18*, 13-22.

Furedy, J.J., Posner, R., & Vincent, A. (1991). Electrodermal differentiation of deception: Memory-difficulty and perceive-accuracy manipulations. *Psychophysiology*, *28*, 163-171.

*Furumitsu, I. (1999). *Laboratory investigations in the psychophysiological detection of deception.* A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy, Graduate Department of Psychology, in the University of Toronto.

*Gaines, K.H. (1992). *Utility and numerical evaluation of the guilty knowledge test*. Report No. DoDPI92-R-0004, Department of Defense Polygraph Institute, Ft. McClellan, AL 36205, August, 1992.

*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.

*Gati, I., Ben-Shakhar, G., & Avni-Liberty, S. (1996). Stimulus novelty and significance in electrodermal orienting responses: The effects of adding versus deleting stimulus components. *Psychophysiology*, *33*, 637-643.

*Geldreich, E.W. (1941). Studies of the use of the galvanic skin response as a deception indicator. *Transactions of the Kansas Academy of Science*, *44*, 346-351.

*Geldreich, E.W. (1942). Further studies of the use of the galvanic skin response as a deception indicator. *Transactions of the Kansas Academy of Science*, *45*, 279-284.

*Giesen, M., & Rollison, M.A. (1980). Guilty knowledge versus innocent associations: Effects of trait anxiety and stimulus context on skin conductance. *Journal of Research in Personality, 14,* 1-11.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and Psychophysics*. New York: John Wiley & Sons.

Greenwald, A.G. (1981). Self and memory. In G.H. Bower (Ed*.). The psychology of learning and motivation*. (Vol. 15). New York: Academic Press.

*Gudjonsson, G.H. (1981). *Some psychological determinants of electrodermal responses to deception*. Unpublished Ph.D. thesis, University of Surrey.

Gudjonsson, G.H., & Haward, R.C., (1982). Detection of deception: Consistency in responding and personality. *Perceptual and Motor Skills*, *54*, 1189-1190.

*Gustafson, L.A., & Orne, M.T. (1963). Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology, 47*, 408-411.

*Gustafson, L.A., & Orne, M.T. (1964). The effect of task and method of stimulus presentation on the detection of deception. *Journal of Applied Psychology, 48,* 383-387.

*Gustafson, L.A., & Orne, M.T. (1965a). Effects of perceived role and role success on the detection of deception. *Journal of Applied Psychology, 49*, 412-417.

*Gustafson, L.A., & Orne, M.T. (1965b). The effects of verbal responses on the laboratory detection of deception. *Psychophysiology, 7*, 10-14.

*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. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology*, *1*, 241-247.

Honts, C.R., Raskin, D.C. & Kircher, J.C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, *79*, 252-259.

*Horneman, C.J., & O'Gorman, J.G. (1985). Detectability in the card test as a function of the subject's verbal response. *Psychophysiology, 22*, 330-333.

*Horneman, C.J., & O'Gorman, J.G. (1987). Individual differences in psychophysiological responsiveness in laboratory tests of deception. *Personality and individual differences*, *8*, 321-330.

*Horowitz, S.W., Kircher, J.C., & Raskin, D. (1986). Does stimulation test accuracy predict accuracy of polygraph tests? *Psychophysiology*, *23*, 442 (Abstract).

*Horvath, F. (1978). An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. *Journal of Applied Psychology, 63*, 338-344.

*Horvath, F. (1979). Effect of different motivational instructions on detection of deception with the psychological stress evaluator and the galvanic skin response. *Journal of Applied Psychology, 64,* 323-330.

Huntet, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications, Inc.

*Iacono, W.G., Boisvenu, G.A., & Fleming J.A. (1984). Effects of Diazepam and Methylphenidate on the electrodermal detection of guilty knowledge. *Journal of Applied Psychology, 69*, 289-299.

*Iacono, W.G., Cerri, A.M., Patrick, C.J. & Fleming J.A. (1992). Use of untianxiety drugs as countermeasures in the detection of guilty knowledge. *Journal of Applied Psychology, 77,* 60-64.

Iacono, W. G., & Lykken, D. T. (1997a). 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*. St. Paul, MN: West Law.

Iacono, W.G., & Lykken, D.T. (1997b). The validity of the Lie Detector: Two surveys of scientific opinion. *Journal of Applied Psychology*, *82*, 426-433.

Iacono, W. G., & Lykken, D. T. (1999). Update: 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*. Volume 1. Pocket Part. St. Paul, MN: West Publishing, pp. 174-184.

*Jones, H.E., & Salter, S. (1989). Polygraph accuracy: An analog study. *Polygraph*, *18*, 69-74.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall, Englewood Cliffs, New Jersey.

Kihlstorm, J.F., & Cantor, N. (1984). Mental representations of the self. In L. Berkowitz (Ed*.). Advances in Experimental Social Psychology* (Vol. 17). New York: Academic Press.

Kihlstorm, J.F., & Cantor, N., Albright, J.S.Chew, B.R., Klein, S. B.,& Niedenthal, P.M. (1988). Information processing and the study of the self. In L. Berkowitz (Ed.*). Advances in Experimental Social Psychology* (Vol. 21). New York: Academic Press.

Kleinmuntz, B., & Szucko, J. J. (1984). Lie detection in ancient and modern times: A call for contemporary scientific study. *American Psychologist*, *39*, 766-776.

Konieczny, J., Fras, M., & Widacki, J. (1984). The specificity of so-called emotional traces and certain features of personality in the polygraph examination. The *Journal of Forensic Medicine and Criminology*, 34, 25-30.

Kraut, R. E., & Poe, D. (1980). Behavioral roots of person perception: The deception judgments of customs inspectors and laymen. *Journal of Personality and Social Psychology*, *39*, 784-798.

Krapohl, D. (1994). The detection of information with items of high or low personal significance using a polygraph: Effects of motivation. *Polygraph*, *23*, 242-250.

*Kubis, J.F. (1962). *Studies in lie detection: Computer feasibility considerations*. Technical Report #62-205, prepared for the Air Force Systems Command. Contract No. AF 30 (602) –2270, project No. 5534, Fordham University.

*Kugelmass, S., & Lieblich, I. (1966). Effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology, 50,* 211-216.

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

*Kugelmass, S., Lieblich, I., Ben-Ishai, A., Opatowski, A., & Kaplan, M. (1968). Experimental evaluation of galvanic skin response and blood pressure change indices during criminal interrogation. *The Journal of Criminal Law, Criminology and Police Science, 59*, 632-635.

*Kugelmass, S., Lieblich, I., & Bergman, Z. (1967). The role of "lying" in psychophysiological detection. *Psychophysiology, 3,* 312-315.

Kunzendorf, R.G., & Bradbury, J.L. (1983). Better liars have better imaginations. *Psychological Reports*, *52*, 634.

Lahri, F.K., & Ganguly, A.K. (1978). An experimental study of the accuracy of polygraph technique in diagnosis of deception with volunteer and criminal subjects. *Polygraph*, *7*, 89-94.

Larson, J.A. (1932). *Lying and its detection: A study of deception and deception tests*. Chicago, Ill: University of Chicago Press.

*Lieblich, I., Ben Shakhar, G. & Kugelmass, S. (1976). Validity of the guilty knowledge technique in a prisoners' sample. *Journal of Applied Psychology*, *61,* 89-93.

*Lieblich, I., Kugelmass, S., & Ben Shakhar, G. (1970). Efficiency of GSR detection of information as a function of stimulus set size. *Psychophysiology, 6,* 601-608.

*Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology, 59,* 113-115.

*Lubow, R.E. & Fein, O. (1996). Pupillary size in response to a visual guilty knowledge test: New technique for the detection of deception. *Journal of Experimental Psychology: Applied*, *2*, 164-177.

*Lykken, D.T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43,* 385-388.

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

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

Lykken, D. T. (1978). The psychopath and the lie detector. *Psychophysiology*, *15*, 137-142.

Lykken, D. T. (1979). The detection of deception. *Psychological Bulletin*, *86*, 47-53.

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.

Maltzman, I. (1979). Orienting reflexes and significance: A reply to O'Gorman. *Psychophysiology*, *16*, 274-281.

Marston, W.M. (1917). Systolic blood pressure changes in deception. *Journal of Experimental Psychology*, *2*, 143-163.

Marston, W.M. (1938). *The lie detector test*. New York: Smith.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*, 81-97.

*Moroney, W.E. & Zenhausern, R.J. (1972). Detection of deception as a function of galvanic skin response recording methodology. *Journal of Psychology, 80*, 255-262.

Nakayama. M. (2001). Practical use of the concealed information test for criminal investigation in Japan. In: M. Kleiner (Ed.). *Handbook of Polygraph Testing*, Academic Press, 49-86.

Ohman, A. (1979). The orienting response, attention, and learning: An information-processing perspective. In: H. D. Kimmel, E. H. van Olst & J. E. Orlebeke (Eds.), *The orienting reflex in humans*. Hillsdale, NJ: Erlbaum, 443-471.

Ohman, A. (1992). Orienting and attention: Preferred preattentive processing of potentially phobic stimuli. In B.A. Campbell, H. Hayne, R. Richardson (Eds.). *Attention and information processing in infants and adults: Perspectives from human and animal research* (pp. 263-295). Hillsdale NJ: Lawrence Erlbaum.

*O'Toole, D., Yuille, J.C., Patrick, C.J. & Iacono, W.G. (1994). Alcohol and the physiological detection of deception: Arousal and memory influences. *Psychophysiology, 31,* 253-263.

Packer, J.S. & Siddle, D.A.T. (1989). Stimulus miscuing, electrodermal activity, and the allocation of processing resources. *Psychophysiology*, *26*, 192-200.

*Pennebaker, J.W., & Chew, C.H. (1985). Behavioral inhibition and electrodermal activity during deception. *Journal of Personality and Social Psychology*, *49*, 1427-1433

Podlesny, J.A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations? A review of FBI case records. *Crime Laboratory Digest*, *20*, 57-61.

Podlesny, J.A. & Raskin, D.C. (1977). Physiological measures and the

detection of deception. *Psychological Bulletin, 84,* 782-799.

*Podlesny, J.A. & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology, 15,* 344-359.

Raskin, D.C. (1982). The scientific basis of polygraph techniques and their uses in the Judicial process. In: A. Trankell (Ed.). *Reconstructing the past: The role of psychologists in the criminal trial*. Stockholm: Norsted & Soners.

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.* New York, Springer.

Raskin, D. C., Honts, C. R., Amato, S. L. & Kircher, J. C. (1999). Update: The scientific status of research on polygraph techniques: The case for the admissibility of the results of polygraph examinations. In: D. L. Faigman, D. H. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony*. Volume 1. Pocket Part. St. Paul, MN: West Publishing, pp. 160-174.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). 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*. St. Paul, MN: West Law.

Raskin, D. C., & Podlesny, J. A. (1979). Truth and deception: A reply to Lykken. *Psychological Bulletin*, *86*, 54-58.

Reid, J.E. (1947). A revised questioning technique in lie-detection tests. *Journal of Criminal Law and Criminology*, *37*, 542-547.

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

*Richardson, D.C., Carlton, B.L. & Dutton, D.W. (1990). Blind analysis of skin conductance response (SCR) recordings from a number test. *Polygraph*, *19*, 9-20.

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.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Ruckmick, C.A. (1938). The truth about the lie detector. *Journal of Applied Psychology*, *22*, 50-58.

Saxe, L. (1991a). Lying: Thoughts of an applied social psychologist. *American Psychologist*, *46*, 409-415.

Saxe, L. (1991b). Science and the CQT polygraph: A theoretical critique. *Integrative Physiological and Behavioral Science*, *26*, 223-231.

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., and Cross, T. P. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist*, *40*, 355-366.

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

Siddle, D.A.T., & Packer, J.S. (1987). Stimulus omission and dishabituation of the electrodermal orienting response: The allocation of processing resources.

Psychophysiology, *24*, 181-190.

Siddle, D.A.T., & Spinks, J.A. (1992). Orienting, habituation, and the allocation of processing resources. In B. Campbell, R.Richardson, & H. Hayne (Eds.), *Attention and information processing in infants and adults: Perspectives from human and animal research*. Hillsdale, NJ: Erlbaum.

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 and A. Smirnov (Eds.). *Psychological research in U.S.S.R*. (p.p. 334-360). Moscow: Progress Publishers.

*Steller, M., Hanert, P., & Eiselt, W. (1987). Extraversion and the detection of information. *Journal of Research in Personality, 21,* 334-342.

*Stern, R.M., Breen, J.P., Watanabe, T., & Bradley, P.S. (1981). Effect of feedback on physiological information on responses to innocent associations and guilty knowledge. *Journal of Applied Psychology*, *66*, 677-681.

*Suzuki, A., Watanabe, S., Ohnishi, K., Matsuno, K., & Arasuna, M. (1979a). The objective analysis of GSR in the detection of deception: An analysis of GSR amplitude in terms of rank scores. *Polygraph*, *8*, 53-63.

*Suzuki, A., Ohnishi, K., Matsuno, K., & Arasuna, M. (1979b). Amplitude rank score analysis of GSR in the detection of deception: Detection rates under various examination conditions. *Polygraph*, *8*, 242-352.

Swets, J. A., Tanner, W.P., Jr., & Birdsall, T.C. (1961). Decision processes in perception. *Psychological Review*, *68*, 301-340.

*Thackray, R.I. & Orne, M.T. (1968). A comparison of physiological indices in detection of deception. *Psychophysiology, 4*, 329-339.

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

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

Timm, H.W. (1989). Methodological considerations affecting the utility of incorporating innocent subjects into the design of guilty knowledge polygraph experiments. *Polygraph, 18*, 143-157.

van Buskirk, D., & Marcuse, F.L. (1954). The nature of errors in experimental lie detection. *Journal of Experimental Psychology*, *47*, 187-190.

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

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. John Wiley & Sons, pp. 169-207.

*Waid, W.M., & Orne, M.T. (1980). Individual differences in electrodermal lability and the detection of information and deception. *Journal of Applied Psychology, 65*, 1-8.

*Waid, W.M., Orne, E.C., Cook, M.R., & Orne, M.T. (1978). Effects of attention, as indexed by subsequent memory, on electrodermal detection of information. *Journal of Applied Psychology, 63,* 728-733.

*Waid, W.M., Orne, E.C., Cook, M.R., & Orne, M.T. (1981). Meprobamate reduces accuracy of physiological detection of deception. *Science*, *212*, 71-73.

Waid, W.M., Orne, E.C., & Orne, M.T. (1981). Selective memory for social information, alertness, and physiological arousal in the detection of deception. *Journal of Applied Psychology, 66,* 224-232.

*Waid, W.M., Orne, M.T., & Wilson, S.K. (1979). Effects of level of socialization on electrodermal detection of deception. *Psychophysiology, 16,* 15-21.

Waid, W.M., Wilson, S.K., & Orne, M.T. (1981). Cross-modal physiological effects of electrodermal lability in the detection of deception. *Journal of Personality and Social Psychology, 40,* 1118-1125.

*Wakamatsu, T. (1987). Effects of motivating the suspect to deceive the polygraph test. *Polygraph, 16,* 129-144.

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

Author Note

1. Gershon Ben-Shakhar, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, 91905, Israel

2. Eitan Elaad – Original affiliation: Division of Identification and Forensic Science, Israel National Police H.Q. New affiliation: Department of Behavioral Sciences, The College of Judea and Samaria, Ariel, Israel.

We thank Neta Bar, Yael Errera, Sossy Fuchs, Shira Garber, Michal Maimaran and Alona Roded for their help in the data analysis, and Norman Ansley, Gisli Gudjonsson and Donald Krapohl for providing us with data and information about GKT studies. We are also grateful to Maya Bar-Hillel, Bella DePaulo, David Lykken and Yaacov Schul for their helpful comments on an earlier draft.

Correspondence concerning this article should be addressed to:

Professor Gershon Ben-Shakhar, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, 91905 Israel

Footnotes

[1]We used several literature search procedures. First, we gathered all relevant studies from our own personal files. Second, we conducted a computer-based search of psychological abstracts (PsychLIT) and Dissertation Abstracts International, using the keywords, "GKT", "GKT validity" and "Guilty Knowledge". Third, we inspected reference lists of previous reviews (Ainsley, 1992; Ben-Shakhar & Furedy, 1990; Elaad, 1998; Saxe et al., 1985; Vrij, 2000). Finally, we consulted with various researchers who had published articles in this area.

[2] For these studies, the base-rates were used to estimate the rate of false-positive outcomes (e.g., the proportion of card numbers classified as "chosen cards", where in reality no card was chosen, in a card-test paradigm).

Table 1: GKT studies using the card-test paradigm

Study	*No. of GKT questions*	*No. of Repetitions*	*N*	*Correct Detection Rate (CDR)*	*Base rate*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Barland (1997)	1	3	10*	**0.900**	.128	1	0	2.42	.956	.770
Beijk (1980) Exp. 1	1	3	102	**0.802**	.100	0	0	2.13	.934	.729
Beijk (1980) Exp. 2	1	3	86	**0.756**	.100	0	1	1.97	.917	.701
Beijk (1980) Exp. 3	1	3	40	**0.875**	.100	0	0	2.43	.956	.772
Ben-Shakhar (1977) I	1	2	40		.125	0	0	0.71	**.691**	.335
Ben-Shakhar (1977) II	2	2	40		.250	0	0	0.35	**.597**	.172
Ben-Shakhar (1977) III	1	2	40		.125	0	0	0.69	**.687**	.326
Ben-Shakhar (1994) Figures	1	2	64		.125	0	1	**1.02**	.761	.454
Ben-Shakhar (1994) Words	1	2	64		.125	0	1	**0.94**	.747	.425
Ben-Shakhar & Lieblich (1982) I	1	1	25	**0.680**	.125	0	0	**2.01**	**.890**	.709
Ben-Shakhar & Lieblich (1982) II	1	1	26	**0.539**	.125	0	0	**1.50**	**.870**	.600
Ben-Shakhar & Lieblich (1982) III	1	1	26	**0.308**	.125	0	0	**0.54**	**.690**	.261
Ben-Shakhar & Lieblich (1982) IV	1	1	26	**0.385**	.125	0	0	**0.76**	**.690**	.355
Ben-Shakhar et al. (1982) I	1	2	30		.125	0	0	**0.60**	.663	.287
Ben-Shakhar et al. (1982) II	1	4	40		.250	0	0	**1.05**	.770	.465
Cutrow et al. (1972)	3	1	63*		.200	1	1	**0.82**	.719	.379
Elaad (1993) I	1	2	24*	**0.667**	.120	1	1	1.61	.873	.627
Elaad (1993) II	1	2	24*	**0.417**	.120	0	1	0.97	.752	.436

Ellson et al. (1952) Exp 1 & 2	1	5	33	**0.788**	.167	0	0	1.77	.894	.663
Ellson et al. (1952) Exp 3 I	1	2	8	**0.500**	.167	1	0	0.97	.752	.436
Ellson et al. (1952) Exp 3 II	1	2	8	**0.125**	.167	0	0	-0.18	.452	-.029
Furedy & Ben-Shakhar (1991) I	1	2	20	**0.550**	.200	0	0	**0.86**	.728	.395
Furedy & Ben-Shakhar (1991) II	1	2	20	**0.550**	.200	0	1	**0.84**	.723	.387
Furedy & Ben-Shakhar (1991) III	1	2	21	**0.857**	.200	1	0	**2.02**	.923	.711
Furedy & Ben-Shakhar (1991) IV	1	2	21	**0.619**	.200	1	1	**1.02**	.764	.454
Geldreich (1941) Exp. A	1	1	50	**0.740**	.200	1	0	1.48	.853	.595
Geldreich (1941) Exp. B	1	1	50	**1.000**	.200	1	0	3.41	.922	863
Geldreich (1942) II	1	1	50	**0.860**	.200	1	1	1.92	.911	.692
Gudjonsson (1981) I	1	2	48*	**0.792**	.200	1	0	1.65	.879	.636
Gudjonsson (1981) II	1	2	48*	**0.854**	.200	1	0	1.89	.910	.687
Gustafson & Orne (1963) I	1	5	18	**0.667**	.200	0	1	1.27	.813	.536
Gustafson & Orne (1963) II	1	5	18	**0.333**	.200	0	0	0.40	.610	.196
Gustafson & Orne (1964)	2	2	48*	**0.688**	.143	1	0	1.56	.864	.615
Gustafson & Orne (1965a)	1	1	32	**0.750**	.200	1	1	1.51	.858	.603
Gustafson & Orne (1965b) I	2	1	25*	**0.760**	.143	1	0	1.78	.896	.665
Gustafson & Orne (1965b) II	2	1	24*	**0.583**	.143	0	0	1.28	.818	.539
Horneman & O'Gorman (1985) I	1	1	78*	**0.442**	.167	1	0	0.82	.719	.379
Horneman & O'Gorman (1985) II	1	1	78*	**0.288**	.167	0	0	0.42	.614	.200
Horneman & O'Gorman	1	2	84	**0.538**	.167	1	0	1.06	.773	.468

(1987)										
Horowitz et al. (1986)	1	?	100	**0.510**	.250	?	?	0.70	.687	.330
Kubis (1962)	1	3	20	**0.800**	.100	1	0	2.12	.933	.727
Kugelmass & Lieblich (1966) Exp. 1 Cond. A	1	2	36*	**0.444**	.167	1	0	0.83	.722	.838
Kugelmass & Lieblich (1966) Exp. 1 Cond. B	1	2	36*	**0.528**	.167	1	0	1.04	.767	.461
Kugelmass & Lieblich (1966) Exp. 1 Cond. C	1	2	36*	**0.472**	.167	1	1	0.90	.739	.410
Kugelmass & Lieblich (1966) Exp. 2	1	2	40	**0.500**	.167	1	0	0.97	.754	.436
Kugelmass et al. (1967)	1	2	27	**0.593**	.167	1	0	1.21	.804	.518
Kugelmass et al. (1968)	1	2	62	**0.565**	.167	1	0	1.13	.788	.492
Lieblich et al. (1970) Exp.1 A	1	2	44*	**0.614**	.250	1	0	0.96	.752	.433
Lieblich et al. (1970) Exp.1 B	1	2	44*	**0.523**	.125	1	0	1.21	.805	.517
Moroney & Zenhausern (1972) Numbers	1	3	76*	**0.641**	.100	1	0	1.64	.877	.634
Moroney & Zenhausern (1972) Words	1	3	76*	**0.776**	.100	1	0	2.04	.925	.714
Pennebaker & Chew (1985) I	1	1	30*	**0.683**	.200	1	0	1.32	.824	.551
Pennebaker & Chew (1985) II	1	1	30*	**0.653**	.200	1	0	1.24	.810	.527
Richardson et al. (1990)	1	1	70*	**0.785**	.167	1	0	1.76	.892	.661
Stern et al. (1981)	1	2	16*	**0.313**	.200	1	0	0.35	.598	.172
Suzuki et al. (1979a)	1	4	30	**0.767**	.167	1	1	1.70	.885	.648
Suzuki et al. (1979b)	1	3	50	**0.780**	.167	1	1	1.75	.891	.659

Notes: * - Participants that were included in more than one experimental condition.
Base rate: The proportion of relevant (chosen) cards.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items.
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 2: GKT studies using the "Peak of Tension" (POT) paradigm with "guilty" participants only

Study	*No. of GKT questions*	*No. of Repetitions*	*N*	*CDR*	*Base rate*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Barland (1997)	1	3	10*	**0.500**	.128	1	0	1.14	.791	.495
Gustafson & Orne (1964)	2	2	48*	**0.566**	.143	1	0	1.24	.810	.527
Gustafson & Orne (1965b) I	2	1	27*	**0.741**	.143	1	0	1.71	.887	.649
Gustafson & Orne (1965b) II	2	1	26*	**0.500**	.143	0	0	1.07	.776	.471
Horvath (1978)	1	2	40	**0.688**	.200	1	0	1.33	.826	.554
Horvath (1979)	1	1	32	**0.445**	.200	1	1	0.70	.688	.331
Moroney & Zenhausern (1972)	1	1	76*	**0.398**	.100	1	0	1.02	.761	.454
Richardson et al. (1990)	1	1	70*	**0.625**	.167	1	0	1.29	.818	.542
Waid et al. (1979)	1	2	G:15* I :15*	**0.467** **0.800**	.167	1	1	0.76	.702	.355

Notes: * - Participants that were included in more than one experimental condition.
Base rate: The proportion of relevant (chosen) cards.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.
G – "Guilty" participants; I – "Innocent participants" (did not choose a card).

Table 3: GKT studies using the code-stimuli paradigm with "guilty" participants only

Study	*No. of GKT questions*	*No. of Repetitions*	*N*	*CDR*	*Base rate*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Ben-Shakhar & Gati (1987) Exp 1	1	2	30	**0.500**	.200	0	0	**0.98**	**.740**	.440
Ben-Shakhar & Gati (1987) Exp 2	1	2	30	**0.530**	.200	0	0	**0.94**	**.740**	.425
Ben-Shakhar & Gati (1987) Exp 3	1	2	30	**0.670**	.200	0	0	**1.49**	**.820**	.597
Ben-Shakhar & Gati (1987) Exp 4	1	2	30	**0.670**	.200	0	0	**1.17**	**.780**	.505
Ben-Shakhar & Gati (1987) Exp 5 I	1	2	15	**0.600**	.200	0	0	**1.00**	**.760**	.447
Ben-Shakhar & Gati (1987) Exp 5 II	1	2	15	**0.670**	.200	0	0	**1.44**	**.840**	.584
Ben-Shakhar & Gati (1987) Exp 5 III	1	2	15	**0.600**	.200	0	0	**0.92**	**.680**	.418
Ben-Shakhar & Gati (1987) Exp 5 IV	1	2	15	**0.730**	.200	0	0	**1.53**	**.830**	.608
Ben-Shakhar et al. (1995) 1	1	2	24		.167	0	0	**0.93**	.742	.421
Ben-Shakhar et al. (1995) 1I	1	2	24		.167	0	0	**1.61**	.871	.627
Ben-Shakhar et al. (1996) Exp 1	1	1	40		.200	0	0	**0.46**	.629	.224
Ben-Shakhar et al. (1996) Exp 2 I	1	1	60		.200	0	0	**0.59**	.662	.283
Ben-Shakhar et al. (1996) Exp 2 II	1	1	60		.200	0	1	**0.98**	.755	.440

Ben-Shakhar et al. (1996) Exp 3	1	1	60		.200	0	1	**0.98**	.755	.440
Ben-Shakhar et al. (1996) Exp 4 Verbal	1	1	40		.200	0	1	**0.65**	.677	.309
Ben-Shakhar et al. (1996) Exp 4 Pictorial I	1	1	40		.200	0	1	**0.91**	.739	.414
Ben-Shakhar et al. (1996) Exp 4 Pictorial II	1	1	40		.200	0	1	**0.95**	.749	.429
Diaz (1985)	10	1	40	**0.550**	.150	1	1	1.16	.794	.503
Elaad (1987) I	4	3	20	**0.750**	.387	1	1	0.66	**.682**	.312
Elaad (1987) II	4	3	20	**0.850**	.387	1	0	0.39	**.610**	.191
Elaad (1987) III	4	3	20	**0.600**	.387	0	1	0.35	**.600**	.172
Elaad (1987) IV	4	3	20	**0.300**	.387	0	0	0.07	**.520**	.034
Gati & Ben-Shakhar (1990) Exp 1 Verbal	1	1	32		.200	0	0	**2.12**	.930	.727
Gati & Ben-Shakhar Exp (1990) 1 Pictorial	1	1	32		.200	0	0	**1.54**	.862	.610
Gati & Ben-Shakhar (1990) Exp 2 Verbal	1	1	30		.200	0	0	**2.21**	.941	.741
Gati & Ben-Shakhar (1990) Exp 2 Pictorial	1	1	30		.200	0	0	**1.47**	.850	.592
Gati et al. (1996) I	1	1	30*		.143	0	0	**0.95**	.749	.429
Gati et al. (1996) II	1	1	30*		.143	0	0	**0.91**	.740	.414
Horneman & O'Gorman (1987) I	3	1	84*	**0.524**	.143	1	0	1.13	.788	.492
Horneman & O'Gorman (1987) II	3	1	84*	**0.274**	.143	0	0	0.47	.629	.229

Thackray & Orne (1968)	3	1	30*	.200	1	1	**1.15**	.709	.498

Notes: * - Participants that were included in more than one experimental condition.
Base rate: The proportion of relevant (code) stimuli.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 3a: GKT studies using the code-words paradigm with "guilty" and "innocent" participants

			Guilty Condition		**Innocent Condition**						
Study	*No. of GKT questions*	*No. of Repetitions*	*N of Guilty*	*CDR Guilty*	*N of Innocents*	*CDR Innocents*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Waid et al. (1978) I	5	2	29	**0.793**	11	**0.727**	1	1	1.42	.841	.535
Waid et al. (1978) II	5	5	18	**0.611**	10	**0.900**	1	1	1.56	.865	.599
Waid et al. (1978) III	5	4	15	**0.733**	15	**0.800**	1	1	1.46	.849	.590
Waid et al. (1979)	6	4	15*	**0.533**	15*	**0.933**	1	1	1.58	.868	.620
Waid & Orne (1980) Ex. 1	6	5	18	**0.667**	10	**0.900**	1	1	1.71	.887	.649
Waid et al. (1981)	6	2	22	**0.773**	11	**1.000**	?	0	3.32	.990	.857

Notes: * - Participants that were included in more than one experimental condition.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 4: GKT studies using the personal-items paradigm with "guilty" Participants only

Study	*No. of GKT questions*	*No. of Repetitions*	*N*	*CDR*	*Base Rate*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Ben-Shakhar et al. (1970) Exp. 1	20	1	27	**0.778**	.200	0	0	1.61	.873	.627
Ben-Shakhar et al. (1975) I	1	5	16	**0.470**	.250	0	0	0.60	.663	.287
Ben-Shakhar et al. (1975) II	1	5	16	**0.340**	.200	0	0	0.43	.618	.210
Ben-Shakhar et al. (1975) III	1	5	16	**0.690**	.167	0	0	1.46	.849	.590
Ben-Shakhar et al. (1975) IV	1	5	16	**0.510**	.143	0	0	1.10	.782	.482
Ben-Shakhar et al. (1975) V	1	5	16	**0.620**	.125	0	0	1.46	.849	.590
Cutrow et al. (1972)	3	1	63*		.200	1	1	**1.03**	.767	.457
Elaad (1987) Exp. 1 A	6	2	9	**1.000**	.077	1	1	1.95	**.916**	.698
Elaad (1987) Exp. 1 B	6	2	9	**0.667**	.077	1	0	1.40	**.839**	.573
Elaad (1987) Exp. 1 C	6	2	9	**0.667**	.077	0	1	1.17	**.797**	.504
Elaad (1987) Exp. 1 D	6	2	9	**0.667**	.077	0	0	0.96	**.752**	.433
Elaad (1994)	6	1	32	**0.750**	.077	1	0	2.09	.929	.722
Elaad et al. (1982)	5	2	10		.200	1	0	1.48	**.854**	.594
Gudjonsson (1981)	1	2	48*	**0.833**	.200	1	0	1.81	.900	.658
Lieblich et al. (1974) I	1	10	30	**0.960**	.200	0	0	2.59	.966	.791
Lieblich et al. (1974) II	1	10	28	**0.950**	.200	0	1	2.49	.961	.779
Lieblich et al. (1976)	20	1	29	**0.621**	.200	0	0	1.15	.792	.498
Lykken (1960)	25	1	20	**1.000**	.200	0	1	3.41	.992	.863
Stern et al. (1981)	1	2	16*	**0.563**	.200	1	0	1.00	.760	.447
Thackray & Orne (1968)	3	1	30*		.200	1	0	**1.42**	.841	.578

Notes: * - Participants that were included in more than one experimental condition.
Base rate: The proportion of personal items among the total number of items in each question.

Verbal response: 1 indicates that a “no” answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 4a: GKT studies using the personal-items paradigm with "guilty" and "innocent" participants

			Guilty Condition		**Innocent Condition**						
Study	*No. of GKT questions*	*No. of Repetitions*	*N of Guilty*	*CDR Guilty*	*N of Innocents*	*CDR Innocents*	*Verbal Response*	*MotIvation*	*d*	*a*	*r*
Ben-Shakhar et al. (1970) Exp. 2	20	1	26		7		0	0	1.14	**.791**	.422
Ben-Shakhar & Elaad (2002) I	12	1	24	**0.500**	12	**0.917**	1	1	0.66	**.680**	.297
Ben-Shakhar & Elaad (2002) II	4	3	24	**0.750**	12	**0.833**	1	1	1.33	**.826**	.531
Ben-Shakhar & Elaad (2002) III	12	1	24	**0.792**	12	**1.000**	1	1	2.99	**.983**	.816

Notes: * - Participants that were included in more than one experimental condition.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 5: GKT studies using the mock-crime paradigm with "guilty" Participants only

Study	*No. of GKT questions*	*No. of Repetitions*	*N*	*CDR*	*Base Rate*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Bradley et al. (1996) GKT I	10	1	10	**0.800**	.159	1	1	1.84	.903	.677
Bradley et al. (1996) GKT II	10	1	10	**0.600**	.159	0	1	1.25	.811	.529
Bradley et al. (1996) GAT I	10	1	10	**0.900**	.159	1	1	2.28	.945	.752
Bradley et al. (1996) GAT II	10	1	10	**0.500**	.159	0	1	1.00	.761	.447
Cutrow et al. (1972)	3	1	63*		.200	1	1	**0.87**	.729	.399
Day & Rourke (1974) I	5	1	20	**0.500**	.200	0	0	0.84	.722	.387
Day & Rourke (1974) II	5	1	20	**0.500**	.200	0	1	0.84	.722	.387
Furumitsu (1999) I	1	6	24	**0.667**	.165	1	0	1.40	.839	.573
Furumitsu (1999) II	2	4	32	**0.688**	.146	1	0	1.54	.862	.610
Timm (1987)	5	1	34	**0.500**	.091	1	?	1.33	.826	.554
Wakamatsu (1987)I	1	1	40	**0.575**	.063	1	1	**.895**	.736	.408
Wakamatsu (1987) II	1	1	20	**0.350**	.063	1	0	**.543**	.648	.262

Notes: * - Participants that were included in more than one experimental condition.
Base rate: The proportion of relevant items.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 5a: GKT studies using the mock-crime paradigm with "guilty" and "innocent" participants

			Guilty Condition		**Innocent Condition**						
Study	*No. of GKT questions*	*No. of Repetitions*	*N of Guilty*	*CDR Guilty*	*N of Innocents*	*CDR Innocents*	*Verbal Response*	*Motivation*	*d*	*a*	*r*
Ben-Shakhar & Dolev (1996)	4	2	32	**0.625**	33	**0.788**	1	1	**1.21**	**.670**	.518
Ben-Shakhar et al. (1999)	3	1	36	**0.667**	36	**0.833**	1	1	**1.14**	**.780**	.465
Bradley & Ainsworth (1984)	9	1	8	**1.000**	4	**1.000**	1	1	5.15	1.00	.925
Bradley Janisse (1981)	4	1	96	**0.594**	96	**0.885**	1	?	1.44	.843	.584
Bradley Warfield (1984)	10	1	8	**1.000**	8	**1.000**	1	1	5.15	1.00	.925
Bradley & Rettinger (1992)	10	1	16	**1.000**	16	**1.000**	1	1	5.15	1.00	.925
Carlton & Smith (1991) Aural	5	1	20	**0.800**	20	**0.850**	1	0	1.88	.908	.685
Carlton & Smith (1991) Visual	5	1	20	**0.450**	20	**0.950**	1	0	1.52	.858	.605
Davidson (1968)	6	1	12	**0.917**	36	**1.000**	0	1	3.96	.997	.864
Elaad (1997)	2	3	25	**0.680**	55	**0.930**	1	1	1.94	.915	.696
Elaad & Ben-Shakhar (1997) Exp. 1 I	1	12	19	**0.632**	5	**0.400**	1	0	0.44	**.620**	.176
Elaad & Ben-Shakhar (1997) Exp. 1 II	1	12	19	**0.789**	5	**1.000**	1	1	1.81	**.900**	.665
Elaad & Ben-Shakhar (1997) Exp. 1 III	4	3	15	**0.600**	9	**0.889**	1	0	1.05	**.770**	.392

Elaad & Ben-Shakhar (1997) Exp. 1 IV	4	3	17	**0.600**	7	**1.000**	1	1	1.24	**.810**	.491
Elaad & Ben-Shakhar (1997) Exp. 2 I	4	3	20	**0.750**	5	**1.000**	1	1	3.09	.985	.777
Elaad & Ben-Shakhar (1997) Exp. 2 II	1	12	20	**0.650**	5	**1.000**	1	1	2.7	.972	.734
Forman & Mc-Cauley (1986)	3	1	20	**0.450**	16	**0.940**	?	1	1.42	.841	.577
Gaines (1992)	5	1	40	**0.250**	40	**1.000**	1	0	1.90	.909	.689
Giesen & Rollison (1980)	6	?	20	**0.950**	20	**1.000**	0	1	4.22	.999	.904
Honts et al. (1996)	5	1	10	**0.800**	10	**0.900**	1	1	2.12	.933	.727
Iacono et al. (1984)	10	1	14	**0.786**	12	**1.000**	1	1	3.36	.991	.859
Iacono et al. (1992)	10	1	15	**0.667**	15	**0.867**	1	1	**1.34**	.828	.556
Jones & Salter (1989)	1	?	3	**1.000**	6	**1.000**	?	1	5.15	1.00	.925
Lubow & Fein (1996)	7	1	19	**0.579**	30	**0.933**	?	0	1.70	.884	.637
Lykken (1959) Murder	6	1	25*	**0.880**	24*	**1.000**	0	1	3.75	.996	.882
Lykken (1959) Theft	6	1	25*	**0.920**	24*	**1.000**	0	1	3.98	.997	.893
O'Toole et al. (1994)	5	1	30	**0.767**	16	**0.938**	1	1	2.27	.945	.750
Podlesny & Raskin (1978)	5	1	10	**0.800**	10	**1.000**	1	1	3.41	.992	.863
Steller Et al. (1987)	6	1	47	**0.851**	40	**1.000**	?	1	3.61	.995	.874
Stern et al. (1981)	6	1	13	**0.923**	13	**0.846**	0	0	2.45	.958	.77

Notes: * - Participants that were included in more than one experimental condition.
Verbal response: 1 indicates that a "no" answer to each GKT item was required; 0 indicates absence of any verbal answers to the items
Motivation: 1 indicates that motivational instructions were given; 0 indicates that no motivational instructions were given.

Table 6: Weighted means and standard deviations of the detection-efficiency statistics, computed across studies for each category and across all 5 categories, 95% confidence intervals (CIs) of d and a comparison of the observed variance of d (S_d^2) and the sampling error variance of d (S_e^2) within each category and across all studies

Category	*N of con dit-ions*	*Correc-ted N of obser-vations*	*Mean r*	*Mean* a	*Mean d*	*95% CI of d*	S_d^2	Se^2	*Residual Variance* of d
1. Card- test	57	1978.5	0.525	0.805	1.347	1.18-1.51	0.413	0.150	0.263
2. POT	9	332.5	0.484	0.782	1.125	0.97-1.28	0.059	0.133	0
3. Code-words	37	1145	0.468	0.743	1.158	0.97-1.35	0.347	0.162	0.185
4. Personal items	24	590	0.569	0.839	1.579	1.28-1.87	0.543	0.233	0.310
5. Mock crimes	42	1534	0.645	0.872	2.088	1.73-2.44	1.385	0.179	1.206
Across all Exp-erimental Con-ditions	**169**	**5198**	**0.553**	**0.815**	**1.548**	**1.41-1.69**	**0.834**	**0.181**	**0.653**

Notes: Corrected N - The corrected number of observations was computed, such that each subject participating in k experimental conditions was assigned a weight of 1/k in each condition

Table 7: Weighted averages of d, 95% CIs of d, S_d^2, S_e^2 and the residual variance, computed across experimental conditions within each level of motivation and each mode of verbal response.

	N of conditions[1]	*Corrected N of observations*	*Mean a*	*Mean d*	95% CI for d	S_d^2	Se^2	*Residual Variance of d*
High Motivation	68	2122	0.83	1.84	1.57-2.11	1.304	0.195	1.109
Low Motivation	98	2726	0.81	1.36	1.23-1.49	0.444	0.191	0.253
Deceptive Verbal Response ("no")	91	2770	0.84	1.59	1.42-1.75	0.638	0.186	0.452
No Verbal Response ("silence")	72	2114	0.79	1.39	1.17-1.61	0.874	0.182	0.692

Notes: Corrected N - The corrected number of observations was computed, such that each subject participating in k experimental conditions was assigned a weight of 1/k in each condition.
CI – Confidence interval
[1] - For 3 experimental conditions the level of motivation was not clearly specified, and for 6 experimental conditions the type of verbal response was not specified. These conditions were not included in the respective computations.

Table 8: Weighted averages of d, 95% CIs of d, S_d^2, S_e^2 and the residual variance computed across experimental conditions within each of the 4 study categories created by crossing motivation level with mode of verbal response.

	N of cond-Itions[1]	*Corrected N of ob-servations*	*Mean a*	*Mean d*	95% CI for d	S_d^2	Se^2	*Residual Variance of d*
High Motivation and a Deceptive Verbal Response	43	1242	0.82	1.77	1.47-2.07	1.020	0.208	0.812
High Motivation with No Verbal Response	22	748	0.82	1.74	1.23-2.25	1.484	0.173	1.311
Low Motivation and a Deceptive Verbal Response	46	1278	0.83	1.46	1.28-1.64	0.399	0.197	0.202
Low Motivation with No Verbal Response	50	1366	0.78	1.21	1.03-1.39	0.443	0.187	0.256

Notes: Corrected N - The corrected number of observations was computed, such that each subject participating in k experimental conditions was assigned a weight of 1/k in each condition.
CI – Confidence interval.
[1] - For 8 experimental conditions either the level of motivation or the type of verbal response was not specified. These conditions were not included in the respective computations.

Table 9: Weighted averages of d, 95% CIs of d, S_d^2, S_e^2 and the residual variance computed across experimental conditions within each category defined by the number of GKT questions used (low – 1-4 vs. high – at least 5)

	N of cond-tions	*Corrected N of obser-vations*	*Mean a*	*Mean d*	95% CI of d	S_d^2	Se^2	*Residual Variance of d*
Small number of GKT questions (1-4)	125	3913	0.79	1.29	1.18-1.40	0.395	0.165	0.230
Large number of GKT questions (at least 5)	44	1243	0.88	2.35	2.00-2.70	1.401	0.258	1.143

Notes: Corrected N - The corrected number of observations was computed, such that each subject participating in k experimental conditions was assigned a weight of 1/k in each condition.
CI – Confidence interval.