Journal of Experimental Psychology: Applied
2003, Vol. 9, No. 4, 261–269

Copyright 2003 by the American Psychological Association, Inc.
1076-898X/03/$12.00 DOI: 10.1037/1076-898X.9.4.261

Estimating the Validity of the Guilty Knowledge Test From Simulated Experiments: The External Validity of Mock Crime Studies

David Carmel, Eran Dayan, Ayelet Naveh, Ori Raveh and Gershon Ben-Shakhar
The Hebrew University of Jerusalem

This experiment was designed to examine the external validity of the standard mock-crime procedure used extensively to evaluate the validity of polygraph tests. The authors manipulated the type of mock-crime procedure (standard vs. a more realistic version) and the time of test (immediate vs. delayed) and examined their effects on the validity of the Guilty Knowledge Test (GKT) and the recall rate of the relevant items. The results indicated that only the type of mock-crime affected the 2 outcome variables. The realistic procedure was associated with a lower recall rate and weaker detection efficiency than the standard procedure. However, these effects were mediated by the type of GKT questions used. Practical implications of these results are discussed.

Scientists and forensic experts have attempted for many years to develop instruments and methods for the purpose of detecting deception. One notable approach, based on measuring psychophysiological responses by a polygraph, has spawned several methods over the past century (see, e.g., Marston, 1917; Raskin, 1989; Reid & Inbau, 1977). The most common of these is the so-called Control Questions Test (CQT), which is widely used in criminal investigations in some countries (primarily the United States, Canada, and Israel) and has been extensively debated in the scientific literature (e.g., Ben-Shakhar, 2002; Furedy & Heslegrave, 1991; Honts, Raskin, & Kircher, 2002; Iacono & Lykken, 2002; Lykken, 1974, 1998; Raskin, 1989).

An alternative 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, 2002; Yamamura & Miyata, 1990). Its lack of popularity in applied settings is probably because of its being much harder to implement than the CQT (Podlesny, 1993). But, in contrast to the CQT, there is a general consensus that the GKT relies on proper control questions (e.g., Ben-Shakhar & Elaad, 2002; Ben-Shakhar & Furedy, 1990; Lykken, 1974, 1998).

The GKT (Lykken, 1959, 1960) utilizes a series of multiple-choice questions, each having one relevant alternative (e.g., a feature of the crime under investigation) and several neutral (control) alternatives, chosen so that an innocent suspect would not be able to discriminate them from the relevant alternative (Lykken, 1998). Typically, if the suspect's physiological responses to the relevant alternative are consistently larger than to the neutral alternatives, knowledge about the event (e.g., crime) is inferred. As long as information about the event has not leaked out and assuming that each alternative appears to be equally plausible to an individual with no guilty knowledge, the probability that an innocent suspect would produce consistently 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, and hence it can be controlled such that maximal protection for the innocent is provided. The assumption that each alternative is equally plausible to an innocent suspect (labeled "transparency" by Honts et al., 2002) can be pretested by administering the GKT items to individuals known to be unaware of the crime details.

Extensive research conducted since the early 1960s has demonstrated that the GKT can be successfully used to detect relevant information and discriminate between knowledgeable (guilty) and unknowledgeable (innocent) individuals (e.g., Ben-Shakhar & Furedy, 1990; Elaad, 1998; Lykken, 1959, 1960, 1998). Recently, Ben-Shakhar and Elaad (2003) conducted a meta-analysis of GKT research and showed that under optimal conditions (i.e., using motivational instructions, deceptive verbal responses to the relevant items, and at least five GKT questions) the GKT can reach an average correlation coefficient as high as .79 between the detection measure and the criterion of guilt versus innocence. It should be noted that these impressive detection efficiency estimates reflect asymmetrical error rates, and whereas the rates of false-positive errors (i.e., innocent suspects classified as "guilty") are indeed very small, the rates of false-negative errors (i.e., guilty suspects classified as "innocents") are typically larger. For example, the studies identified by Ben-Shakhar and Elaad (2003) as representing optimal conditions for the use of the GKT produced 4.4% false positives, but 16.8% false negatives. Similar figures were provided in other reviews of the GKT (e.g., Ben-Shakhar & Furedy, 1990, estimated the false-positive and false-negative rates of the GKT as 6.0% and 16.0%, respectively, and Honts et al., 2002, estimated these rates as 1.0% and 14.0%, respectively). Almost all attempts to examine the validity of the GKT were based on simulations (i.e., mock-crime experiments) in which some participants (the guilty)

David Carmel, Eran Dayan, Ayelet Naveh, Ori Raveh, and Gershon Ben-Shakhar, Department of Psychology, The Hebrew University of Jerusalem.

We thank Ori Cohen, Yael Errera, Shira Garber, Avia Munchik, and Alona Roded for their help in the data collection and analysis. We also thank Eitan Elaad, Alona Roded, and Marilyn Smith for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Gershon Ben-Shakhar, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, 91905 Israel. E-mail: mskpugb@mscc.huji.ac.il

are required to commit a mock crime (e.g., to steal an envelope containing a sum of money and a piece of jewelry from a specified office), whereas others (the innocents) do not commit this act. At the second stage of the experiment, a GKT is administered to all participants in a double blind manner (i.e., the test administrator is unaware of the condition to which the participant is assigned), and an attempt is made to differentiate between these two groups on the basis of their relative physiological responses to the relevant details of the mock crime.

However, Ben-Shakhar and Furedy (1990) questioned the external validity of the mock-crime paradigm. Specifically, they argued that,

> Unfortunately, all GKT studies used a very simple task in which the experimenters guaranteed that all subjects learned all the relevant items (e.g., six code words were overlearned by the subjects). Furthermore, the subjects are typically tested immediately after being exposed to the guilty information, thus memory does not play an important role in the experimental situation. In real life, things might be entirely different. The guilty subject is faced with a complex scene, and it might be much more difficult to assume that all details were indeed noticed, processed, and stored in memory. Criminal suspects are very rarely tested immediately after committing the criminal act. Typically, they may be tested days, weeks, and sometimes months after the crime was committed (see Ben-Shakhar & Furedy, 1990, pp. 55–56).

Although in some GKT studies, memory for the critical items was not verified (e.g., Honts, Devitt, Winbush, & Kircher, 1996; Iacono, Boisvenu, & Fleming, 1984), the typical GKT experiment differs from the realistic setup in which the GKT may be used in several important ways. In particular, factors affecting perception and memory that may be crucial for the efficiency of the GKT in applied settings do not play a sufficient role in mock-crime experiments. These critical differences between the applied and the simulated settings were also noted by Honts et al. (2002, p. 457), who wrote that, “In GKT lab studies, the experimenters usually pre-test potential items for their salience and memorability by guilty subjects.” Clearly, this would be impossible in the realistic setting.

These differences between the experimental and the realistic setups may account for the relatively large rates of false-negative outcomes observed in two field GKT studies reported by Elaad (1990) and by Elaad, Ginton, and Jungman (1992). Although 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). These increased rates of false-negative outcomes, relative to those typically obtained in mock-crime experiments, can be accounted for by perception and memory limitations that characterize the realistic criminal situation (e.g., it cannot be ascertained that culprits paid attention to the critical items used in the GKT, and even if they did, it cannot be ascertained that they remember these items when they take the GKT). Innocent suspects, however, have no knowledge of the critical items in the first place, so memory cannot affect their response pattern to the various GKT items.

Recently, Honts et al. (2002) suggested that these high false-negative rates can be explained in terms of poor memory for the crime details, much like the documented fallibility of eyewitnesses (e.g., Loftus & Ketcham, 1991). However, it can be argued (see Ben-Shakhar & Elaad, 2003) 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 (2.0 in Elaad, 1990, and 1.8 in Elaad et al., 1992) was rather small. In addition, the two field studies were based on GKTs that were administered immediately following a CQT, and this may have attenuated the sensitivity of the physiological measures as a result of 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.

However, even if the use of the GKT in the criminal cases examined by Elaad (1990) and Elaad et al. (1992) was not optimal, memory remains a critical factor that should not be ignored. Moreover, there is no reason to believe that culprits would be less vulnerable to memory fallibility than eyewitnesses. Although no research has been conducted on memory of culprits, there is a vast literature on eyewitness memory (e.g., Cutler & Penrod, 1995; Eisen, Quas, & Goodman, 2002). Recently, Wells and Olson (2003) wrote that mistaken eyewitness identification was the largest single factor contributing to the conviction of innocent people. Unfortunately, this factor was largely ignored in GKT research. Research on eyewitness memory also showed that memory for a given event can be distorted by misleading postevent information (e.g., Loftus, 1975; Loftus, Miller, & Burns, 1978; Loftus & Palmer, 1974; Loftus, Schooler, & Wagenaar, 1985; but see also, McCloskey & Zaragoza, 1985a, 1985b). Amato-Henderson, Honts, and Plaud (1996) demonstrated a similar effect with the GKT.

Some studies have examined the effect of a time lag between the mock crime and the test on the efficiency of the GKT. 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 one or two out of the four critical items that were used. More recently, Hira, Sasaki, Matsuda, Furumitsu, and Furedy (2001, 2002) conducted a mock-crime GKT study and used the P300 event-related potential (ERP) recorded at the Pz scalp site as their detection measure. Nine guilty participants were tested both immediately and 1 month after committing the mock crime, and all of them were correctly identified at both time points. In their second study, Hira et al. (2002) retested five of the nine original participants a 1 year later and once again correctly identified all of them. Thus, these studies indicate that the GKT may be effective even when administered a long time after the crime.

The main goals of this study were to systematically compare the standard mock-crime paradigm, typically used in GKT research, with a more realistic version of this paradigm and to examine the effect of time lag between the mock crime and the test. Specifically, the following two factors that differentiate the mock-crime paradigm from the realistic setup are examined:

1. In mock-crime studies, it is typically guaranteed that the guilty participants take notice of all the relevant details and remember them when they take the GKT. This is achieved through instructions that specify all these details precisely (e.g., “The room you are in is a small office in *a hotel*. A person (the mannequin),

whom you know as *Frank*, is seated in this room", see Bradley & Warfield, 1984, p. 684). In addition, in many of the mock-crime studies the guilty participants are presented with these details just before the GKT is administered (e.g., Ben-Shakhar, Gronau, & Elaad, 1999), and in some studies, data from participants who could not recall the relevant items in a postexperiment recall test were discarded (e.g., Ben-Shakhar & Gati, 1987). In this study, we manipulated this factor by using the typical mock-crime procedure for half our participants (the standard condition), while using a more realistic procedure for the other participants (the realistic condition). In the realistic condition, participants were required to enter an office and steal a CD-ROM, but they were not informed about other relevant details (e.g., a picture of a known public figure on the wall, a beverage on the table). Furthermore, in the realistic condition participants were told that they could stay in the office for a limited time (five minutes), after which the room's occupant, a teaching assistant, would return to his office and catch them. In addition, participants in the realistic condition were not reminded of the relevant details before the GKT.

2. As described earlier, whereas in the standard mock-crime paradigm the GKT is administered immediately after the mock crime, in a realistic setting it is usually administered after a long period. We manipulated this factor by administering the GKT immediately for half the participants (immediate condition) and 1 week later for the other half (delayed condition). Thus, we examined the effect of delayed versus immediate GKT on the efficiency of detection—based on the electrodermal measure—under the standard mock-crime procedure, as well as under the more realistic procedure.

Method

Participants

Eighty-four Hebrew University of Jerusalem undergraduate students (52 women and 32 men) participated in the experiment for payment or course credit. Their mean age was 23.3 years ($SD = 2.46$ years). Participants were recruited through ads placed on notice boards throughout the campus.

Apparatus

Skin conductance was measured by a constant voltage system (0.5 V Atlas Researches, Hod Hasharon, Israel). Two Ag/AgCl electrodes (0.8-cm diameter) were used with a 0.05 M NaCL electrolyte. The experiment was conducted in an air-conditioned laboratory, and an NEC CF-500 computer was used to control the stimulus presentation and to compute skin conductance changes. The stimuli were displayed on the computer monitor.

Design

A 2 × 2 between-participants design was used, with the following two factors: (a) type of mock crime (standard vs. realistic procedure) and (b) time of test (immediate vs. delayed). Twenty-one participants were randomly allocated to each of the four conditions created by this design. The data of 1 participant were lost, so only 20 participants were included in the data analyses of the standard mock crime with a delayed GKT. An additional participant did not complete the recall test, so only 20 participants were included in the data analyses of the recall results of the realistic mock-crime condition with a delayed GKT.

Procedure

All participants were instructed to enter the office of a teaching assistant and steal a CD-ROM with a colored case containing a copy of an examination in an introductory psychology course. In the standard mock-crime procedure, all the relevant details were specified in advance. Specifically, participants in this condition were told the following details: The name of the teaching assistant (Amos Lavie), which was also printed on the office door, the color of the CD-ROM's case (blue), and its exact location in the office (on the shelf). They were also asked to pay close attention to other details, such as the type of soft drink (diet Coke) and the name of the newspaper (Haaretz) placed on the desk, and the picture on the wall (the Israeli President, Moshe Katzav). Furthermore, after completion of the mock crime, participants in the standard mock-crime condition were asked to name all the relevant details. If they had trouble remembering any of them (which rarely occurred), they were reminded. Participants in the realistic mock-crime condition were told that they should steal a CD-ROM, which contained the examination in "Introduction to Psychology," from Amos Lavie's office, but none of the other details were mentioned. Furthermore, they were told that they could stay in that office for no longer than 5 min, after which the teaching assistant would return to his office.

In the next stage of the experiment, the GKT was administered to all participants. Participants in the immediate condition took the test immediately after committing the mock crime, and those in the delayed condition took it 1 week later. An experimenter, who was unaware of the experimental condition to which the examinee was assigned, attached the skin conductance response (SCR) electrodes and conducted the GKT examination. Participants were told that the experiment was designed to test whether they could cope with the polygraph test and convince the examiner that they are innocent of stealing the CD-ROM. They were promised a bonus of 10 New Israeli Shekels (about $2) for successful performance of the task. The GKT questions were presented after an initial rest period of 2 min, during which skin conductance baseline was recorded. All examinees were presented with seven different questions, each targeting a different feature of the mock crime (the color of the CD-ROM's case, the name of the teaching assistant, the subject of the examination, the location of the CD-ROM in the office, the soft drink that was placed on the desk, the newspaper placed on the desk, and the name of the person whose picture hung on the wall). The questions were presented on the computer monitor. Each question was followed by two repetitions of a set of five items (the relevant item and four neutral control items), preceded by a neutral buffer item designed to absorb the initial orienting response. The order of the five items within each repetition of the set was randomized. Each question was presented for 2 s, and each item (answer) was presented for 5 s. The interstimulus interval ranged randomly from 11 to 19 s, with a mean of 15 s. Participants were asked to respond verbally, saying "no" to every item. A short, participant-terminated break was given after presentation of four questions. The questions were presented in a predetermined order, counterbalanced across participants within each condition. At the end of the questioning session, a multiple-choice recall test was administered to examine whether participants recalled the relevant items. The recall test consisted of the seven questions given during the GKT, each with six possible answers (the buffer item and the five items used in the GKT). Finally, all participants were debriefed and compensated.

Response Scoring and Analysis

Responses were transmitted in real time to the computer. The maximal conductance change obtained from the examinee, from 1 s to 5 s after stimulus onset, was computed using an A/D (NB-MIO-16) converter with a sampling rate of 1000 Hz. To eliminate individual differences in responsivity and permit a meaningful summation of the responses of different examinees, each participant's conductance changes were transformed into within-examinee standard scores (Ben-Shakhar, 1985). To minimize habituation effects, within-questions standard scores were used (Ben-Shakhar & Dolev, 1996). Thus, the z scores used in this study were computed relative to the mean and standard deviation of the participant's responses to the 10 items of each question (the responses to the buffer stimuli were not included in the standardization). A rejection region of $p < .05$ was used for all statistical tests, and effect size estimates for all the effects examined in the analyses of variance (ANOVAs) were computed by using Cohen's (1988) f values.

Results

We computed the mean standardized response of each examinee across the two presentations of the relevant item within each question and across questions. These means, which were averaged across participants within each condition, are displayed in Table 1. We conducted a 2 × 2 (Type of Mock Crime × Time of Test) between-participants ANOVA on the data of Table 1. The results of this analysis, which are displayed in Table 2, revealed that the type of mock-crime procedure produced a statistically significant and large effect ($f = 0.34$), reflecting a larger relative mean response in the standard ($z = 0.68$) than in the realistic procedure ($z = 0.35$). Neither the time of test nor the interaction between the two factors produced statistically significant effects. The statistical power for detecting a medium effect size ($f = 0.25$) was 62% and for detecting a large effect size ($f = 0.40$) was 95%. These values apply to all subsequent ANOVAs.

In addition to group data, it is interesting to look at classification accuracy of individual examinees. To achieve this goal, we adopted the Lykken (1959) scoring procedure, which has been used in many GKT studies (e.g., Ben-Shakhar & Elaad, 2002; Bradley & Warfield, 1984). By this procedure, the standardized responses to all alternatives of each question are rank ordered. If the relevant alternative elicits the largest response, a value of 2 is assigned to the question; if it elicits the second largest response, a value of 1 is assigned to the question, otherwise a value of 0 is assigned. These values are then summed up across all questions to produce a single detection score. In the present experiment, there were seven questions, with two repetitions of the set of five items (one relevant and four neutral controls). Thus, we first averaged the SCRs to each item across the two repetitions. Then we computed a Lykken detection score (ranging between 0 and 14) for each participant across the seven questions. A cutoff score of 7 was set on this detection measure, such that a detection score of at least 7 yielded a "guilty" classification. Rates of correct classifications based on this procedure are presented in Table 1 as a function of experimental conditions. The overall correct classification rate obtained under the standard conditions (0.76) was larger that that obtained under the realistic conditions (0.52). This difference is statistically significant ($Z = 2.09$). The difference between the immediate and delayed conditions was very small (0.62 and 0.66 in these two conditions, respectively) and not statistically significant ($Z = 0.26$).

Table 1
Means and Standard Deviations of the SCRs to the Relevant Items, Based on all GKT Questions, Standardized Differences Between the Means of "Guilty" and Hypothetical "Innocent" Distributions (d′), Area Under the ROC Curves (a), and Detection Accuracy Rates as a Function of Experimental Condition

Type of mock-crime procedure	Time of test	Mean z score	Standard deviation	Detection accuracy rates	d'	a
Standard	Immediate	0.67	0.54	71.4%	1.41	0.84
Standard	Delayed	0.68	0.54	80.0%	1.43	0.84
Realistic	Immediate	0.38	0.29	52.4%	0.79	0.71
Realistic	Delayed	0.32	0.39	52.4%	0.67	0.68

Note. SCRs = skin conductance responses; GKT = Guilty Knowledge Test; ROC = receiver operating characteristic.

Table 2
Results of a 2 × 2 Between-Participants ANOVA Conducted on the Mean Standardized Response to the Relevant Items, Computed Across all Seven Questions

Source	Sum of squares	Mean squares	F(1, 79)	f
Type of mock crime	2.016	2.016	9.41*	0.34
Time of test	0.005	0.005	0.02	0.02
Interaction	0.011	0.011	0.05	0.03
Error	16.926	0.214		

Note. ANOVA = analysis of variance.
* $p < .05$.

However, the Lyyken scoring procedure may be nonoptimal because it is based on a transformation of the continuous SCRs to a three-level scale, and consequently, valuable information may by lost. In addition, the accuracy rates presented in Table 1 depend on a single arbitrary cutoff point. Therefore, we adopted an additional approach for describing and comparing detection efficiency from signal detection theory (SDT). As discussed extensively in the recent report of the National Research Council (2003), this approach is particularly relevant for describing the diagnostic value of polygraph tests. Typically, detection efficiency is defined in terms of the relationship between the detection measure (in our case, the mean standardized response across all relevant items) and the actual guilt (or knowledge of the relevant items). In SDT terms, this is measured by the degree of separation between the distributions of the detection score of guilty and innocent participants. In this study, only guilty (knowledgeable) participants were included, but the distribution of the detection score among innocent (unknowledgeable) individuals can be estimated if the assumptions on which the GKT rests are met. Specifically, the expected mean standardized response to the relevant items among innocent individuals is zero, because as long as the relevant items have no special meaning for these individuals, there is no reason to expect that they would show systematically different responses to the relevant items than to the neutral items. Thus, we computed the distance (in standard deviation units) between the centers of the two distributions (d') in each condition by subtracting the expected mean z score to the relevant items that would be obtained for innocent individuals (zero) from the actual mean z score to the relevant items obtained in that condition, divided by the standard deviation of the mean z scores (estimated from the total sample). If, in addition, the standard assumptions underlying all parametric statistical tests are made (i.e., that that the two distributions of the detection score are normal with equal variances), the area under the receiver operating characteristic (ROC) curve can be derived from d′ (see Ben-Shakhar & Elaad, 2003). The d' and the area statistic (a) for each condition are displayed in Table 1. An inspection of Table 1 reveals that this additional data analysis is consistent with both the ANOVA and the detection accuracy analysis. Specifically, although the standard procedure clearly differs from the realistic one, there are no differences between the immediate and delayed conditions. The area statistic is about 0.84 and 0.70 for the standard and realistic procedures, respectively, and the effect size measure (d') is about twice as large in the standard than in the realistic procedure.

To examine whether the effect of standard versus realistic mock crime can be explained by memory of the relevant items, we first analyzed the recall data. We computed the mean and standard deviation of the number of correctly recalled relevant items for each experimental condition. These data, which are displayed in Table 3, suggest that recall under the standard procedure was nearly perfect (the average number of items recalled in this condition were 7.0 and 6.9 for the immediate and delayed conditions, respectively. However, we observed a much poorer recall rate under the realistic procedure (4.3 and 4.1 for the immediate and delayed conditions, respectively). The results of an ANOVA conducted on the recall data are displayed in Table 4. These results reflect a strong ($f = 1.77$) and statistically significant effect for the type of mock-crime procedure. Neither the other main effect (timing of the recall test) nor the interaction produced statistically significant outcomes. Thus, the memory results fit well with the results derived from the electrodermal responses to the relevant items.

To further examine this issue, we reanalyzed the SCR data, using for each participant only the responses for correctly recalled relevant items. The mean z scores based on these items only, for each experimental condition, are displayed in Table 5, along with the a and d' statistics, which we recomputed on the basis of the correctly recalled items. An ANOVA, which we conducted on the results displayed in Table 5, is presented in Table 6. This analysis revealed no statistically significant effects, with small effect size estimates ($f = 0.11$ for crime type, $f = 0.06$ for time of test, and $f = 0.09$ for the interaction between these factors). In addition, we computed detection accuracy rates based on the Lykken scoring procedure on the basis of correctly recalled relevant items, which are presented in Table 5. Because different participants recalled different numbers of relevant items, a cutoff point was set individually for each participant. Specifically, for a participant who correctly recalled k relevant items, the detection score range was $0–2k$. This individual was classified as “guilty” if his or her detection score was at least k. These results also show that the differences between the standard (85.4% correct detection rate) and realistic conditions (88.1% correct detection rate) decreased drastically when only recalled items were considered.

Table 3
Means and Standard Deviations of the Number of Correctly Recalled Relevant Items

Type of mock-crime procedure	Time of test	Mean	Standard deviation
Standard	Immediate	7.0	0.00
Standard	Delayed	6.9	0.30
Realistic	Immediate	4.3	1.04
Realistic	Delayed	4.1	1.04

Table 4
Results of a 2 × 2 Between-Participants ANOVA Conducted on the Mean Number of Correctly Recalled Relevant Items

Source	Sum of squares	Mean squares	$F(1, 78)$	f
Type of mock crime	153.067	153.067	258.05*	1.77
Time of test	0.569	0.569	0.96	0.11
Interaction	0.091	0.091	0.15	0.04
Error	42.267	0.593		

Note. ANOVA = analysis of variance.
* $p < .05$.

An inspection of Table 5 reveals that all detection efficiency measures (the z scores, the areas under the ROC curves, the d' values, and the correct detection rates) seem to reflect poor detection efficiency in the realistic-delayed condition relative to all other three conditions, but the differences were not sufficiently large enough to produce a statistically significant outcome and the effect size estimates were small. Thus, this analysis supports the conjecture that the relatively low detection efficiency observed under the realistic mock-crime procedure was accounted for by the weak memory of the relevant items in this condition.

However, the critical question from an applied perspective is whether relevant items, which are likely to be remembered, could be identified a priori. The seven relevant items used in the present experiment can be roughly divided into two categories: (a) Central Items, directly related to the theft (the color of the CD-ROM and its location in the office, the name of the assistant from whom it was stolen, and the topic of the examination); and (b) Peripheral Items, which happened to be in the office (the soft drink and newspaper found on the desk and the picture on the wall). It is reasonable to assume that items of the first category are highly likely to be remembered, whereas items that just happened to be in the office may be overlooked, especially when the theft was committed under time pressure. Indeed, an analysis of the recall data reveals that the four items of Category a produced an average recall rate of 90.2%, whereas the three items of Category b had a 65.9% recall rate.

To examine detection efficiency on the basis of the four GKT questions of Category a, we computed the mean z score across the

Table 5
Means and Standard Deviations of the Detection Score, Based on Correctly Recalled Relevant Items, Standardized Differences Between the Means of “Guilty” and Hypothetical “Innocent” Detection Score Distributions (d'), and Area Under the ROC Curves (a) as a Function of Experimental Condition

Type of mock-crime procedure	Time of test	Mean z score	Standard deviation	Detection accuracy rates	d'	a
Standard	Immediate	0.67	0.54	71.4%	1.30	0.82
Standard	Delayed	0.71	0.54	80.0%	1.37	0.83
Realistic	Immediate	0.61	0.41	76.2%	1.18	0.80
Realistic	Delayed	0.48	0.54	55.0%	0.93	0.74

Note. ROC = receiver operating characteristic.

Table 6

Results of a 2 × 2 Between-Participants ANOVA Conducted on the Mean Standardized Response to the Relevant Items, Computed Across all Correctly Recalled Questions

Source	Sum of squares	Mean squares	*F*(1, 78)	*f*
Type of mock crime	0.266	0.266	0.97	0.11
Time of test	0.078	0.078	0.29	0.06
Interaction	0.175	0.175	0.64	0.09
Error	21.304	0.273		

Note. ANOVA = analysis of variance.

Table 8

Results of a 2 × 2 Between-Participants ANOVA Conducted on the Mean Standardized Response to the Relevant Items, Computed Across the Four Central GKT Questions

Source	Sum of squares	Mean squares	*F*(1, 79)	*f*
Type of mock crime	0.271	0.271	0.95	0.11
Time of test	0.258	0.258	0.90	0.10
Interaction	0.417	0.417	1.46	0.13
Error	22.235	0.285		

Note. ANOVA = analysis of variance; GKT = Guilty Knowledge Test.

four relevant items of Category a. These mean *z* scores, computed across participants within each experimental condition, are displayed in Table 7, along with the *a* and *d′* statistics, which were recomputed on the basis of these four items. An ANOVA conducted on these results (see Table 8) revealed no statistically significant effects, with small effect size estimates for crime type, timing of test, and their interaction (0.11, 0.10, and 0.13, respectively). In addition, we computed detection accuracy rates based on the Lykken scoring procedure on the basis of these four questions (see Table 7). Inspection of the detection rates reveals that the differences between the standard condition (correct detection rate of 68.3%) and the realistic condition (correct detection rate of 78.6%) completely disappeared when only the four central questions were considered. In fact, we observed a larger rate of correct detections under the realistic than under the standard condition, but the difference is not statistically significant.

The results displayed in Table 7 are similar to those described in Table 5, with detection efficiency estimates only slightly smaller. The two standard mock-crime conditions and the immediate-realistic condition produced practically identical detection efficiencies, whereas the delayed-realistic condition seemed to be associated with weaker detection efficiency. The correct detection rate observed under the realistic-immediate condition is the only exception to this pattern. This can be attributed to the relative instability of the detection rate measure, which is based on transforming the continuous SCRs to a discrete measure, and on setting an arbitrary cutoff on the Lykken scores. However, once again the differences were not sufficiently large enough to produce a statistically significant interaction.

Table 7

Means and Standard Deviations of the Detection Score, Based on the Four Central GKT Questions, Standardized Differences Between the Means of "Guilty" and Hypothetical "Innocent" Detection Score Distributions (d′), and Area Under the ROC Curves (a) as a Function of Experimental Condition

Type of mock-crime procedure	Time of test	Mean *z* score	Standard deviation	Detection accuracy rates	*d′*	*a*
Standard	Immediate	0.62	0.64	66.7%	1.17	0.80
Standard	Delayed	0.62	0.54	70.0%	1.17	0.80
Realistic	Immediate	0.62	0.36	90.5%	1.18	0.80
Realistic	Delayed	0.36	0.51	66.7%	0.68	0.68

Note. GKT = Guilty Knowledge Test; ROC = receiver operating characteristic.

Discussion

The results of this experiment indicate that the standard mock-crime procedure may suffer from weak external validity. We systematically compared this procedure with a more realistic type of mock crime and demonstrated that the latter is associated with inferior recall of relevant items and with less efficient detection than the former. Clearly, the realistic procedure used in this experiment is also artificial, but it better resembles realistic conditions because the mock crime was committed under time pressure, the relevant items were not clearly specified before the mock crime, nor were they mentioned again before the examination. The results also demonstrated that the less efficient detection associated with the realistic mock-crime procedure can be accounted for by the poor recall for some of the items. When only correctly recalled items were taken into account, the effect of type of crime was small and not statistically significant (the effect size of type of crime dropped from 0.34, which according to Cohen, 1988, represents a value somewhere between medium and large effect size to 0.11, which is considered as a small effect). The fact that perpetrators of a crime may not remember some features of the crime scene is not surprising, as it is consistent with the vast literature on eyewitness memory (e.g., Eisen et al., 2002; Wells & Olson, 2003).

However, there are also important differences between a perpetrator and an eyewitness. Whereas an eyewitness is exposed relatively briefly to all the features of the event, a perpetrator is intimately familiar with some features (e.g., the weapon used in an armed robbery; the location of the stolen objects). As indicated by Nakayama (2002, p. 53), "Offenders might not remember objects they saw by chance at the crime site, but will often recall the tools they prepared before breaking into a residence." We tried to make a distinction between the former type of details, labeled *peripheral*, and the latter type, called *central*. Although perpetrators resemble eyewitnesses as far as peripheral features are concerned, they differ from eyewitnesses regarding the central features, which typically draw a great deal of attention from the culprit.

Indeed, our results indicate that there are large variations between the various items in terms of their recall rate. In particular, it seems that central items, directly related to the mock crime, are much more likely to be recalled than peripheral items that were present on the crime scene but have no direct relation to the theft.

Our data reveal a large difference in recall rates between central and peripheral items (90% vs. 66%). Focusing on the GKT questions that are based only on central items eliminated the differences between the two types of mock-crime procedures. This effect was particularly impressive in the realistic-immediate condition, in which the use of just four GKT questions was associated with greater detection efficiency than the use of all seven questions (the area under the ROC curves increased from 0.71 to 0.80, the average *z* score increased from 0.38 to 0.62, and the correct detection rate increased from 52.4% to 90.5% when the three peripheral questions were eliminated). It should be noted that this contrasts with the general view that detection efficiency is an increasing function of the number of GKT questions used (see Ben-Shakhar & Elaad, 2003). The impact of using only central items was less salient in the realistic-delayed condition, in which detection efficiency was largely unaffected by eliminating the peripheral questions. Although the interaction between the two factors manipulated in this study did not produce a statistically significant outcome (the effect size associated with this interaction was 0.13), the data displayed in Tables 5 and 7 suggest that the realistic-delayed condition was associated with smaller detection efficiency compared with all other three experimental conditions. A hypothetical explanation for this result could be based on the notion that the more realistic situation was characterized by a relatively shallow processing of the relevant information (e.g., Craik & Tulving, 1975), which was not reflected in the immediate testing, but had some effect on the relative responses to the relevant items when the test was delayed. Further research will be necessary in order to clarify this issue. Nevertheless, these results clearly suggest that proper use of the GKT should be based only on features of the crime scene directly related to the execution of the crime. A similar conclusion was reached by Nakayama (2002), based on his experience in the use of the GKT for criminal investigations by the Japanese police.

Our conclusion that GKT studies suffer from low external validity is based, in part, on our concern that when this method is applied, investigators may be tempted to include some peripheral features of the crime, rather than focus strictly on central items. This may occur because it is very difficult to identify proper GKT items in real polygraph investigations, and because central items are more likely than peripheral items to be leaked out, either through the media or during the course of interrogation. Such a tendency may attenuate the differential responses to these critical items because of disruption of memory, and this attenuation may not be detected by the standard mock-crime experiment. Thus, it is recommended that a policy of relying exclusively on central features will be adopted both in practice and in research. As we demonstrated, this would minimize the effects of attention and memory and would also minimize the differences between the artificial nature of the mock-crime paradigm and the realistic criminal situation. Nevertheless, it should also be recommended that future GKT studies would use less artificial and more realistic versions of the mock crime. In particular, a time pressure element should be introduced into the execution of the mock crime, and no attempt should be made to remind the guilty participants about the nature of the relevant details.

The other factor manipulated in this experiment (time of test) did not produce any statistically significant effects, either on the recall rates or on the detection efficiency measures. Although this result is based on a failure to reject the null hypothesis, it should be emphasized that the effect size obtained was very small (0.02), and although the statistical power for detecting a medium size effect was not high (0.62), the power for detecting a large effect was very high (0.95). Thus, our analysis might have missed a small effect of time delay, but not a large one. This finding, which is rather surprising, is nonetheless consistent with the recent reports by Hira et al. (2001, 2002), who demonstrated a perfect ERP-based detection rate of guilty participants tested 1 month and even 1 year after committing the mock crime. However, Elaad (1997), who also used a more realistic mock-crime procedure than the standard mock-crime paradigm, reported that some relevant items were not recalled when the GKT was administered a few days after the event.

This study focused on the external validity of the mock-crime paradigm, which has been used extensively in the past 3 decades to evaluate the validity of the GKT (e.g., Ben-Shakhar & Dolev, 1996; Ben-Shakhar et al., 1999; Bradley, MacLaren, & Carle, 1996; Bradley & Rettinger, 1992; Davidson, 1968; Lykken, 1959). We showed that the standard mock-crime procedure, applied in most of these studies, may have weak external validity because it does not tap several factors, which operate in the realistic situation and may reduce memory of some relevant items. In particular, in the standard mock-crime paradigm participants are often reminded about the nature of the relevant items to be used in the GKT. Consequently, the validity estimates derived from mock-crime studies may be inflated. This may account for the relatively low validity obtained in the two field studies reported by Elaad and his colleagues (Elaad, 1990; Elaad et al., 1992).

However, the results of the present study also demonstrated that loss of memory for relevant items depends, to a large extent, on the type of items used. Central features of the mock crime that are directly related to the execution of the crime (e.g., the stolen objects, their location, the mode of operation) are much less susceptible to memory decay and may be successfully used days and perhaps even weeks after the event. Future research should further examine the relationship between types of potential GKT questions and their likelihood to be recalled during the test and their contribution to the detection of guilty knowledge.

Another factor that may be a threat to the validity of the GKT in some realistic investigations, but not in mock-crime experiments, is related to criminals who operate frequently (e.g., criminals who burglarize houses or steal cars on a regular basis). This type of criminal may have additional memory problems because of interference among the details of the various events in which they were involved. This factor, which can also affect the external validity of the mock-crime procedure, was not examined in this study and requires a separate investigation.

Finally, two major practical conclusions can be drawn from the present study. First, when constructing a GKT, examiners should formulate only questions that are directly related to the execution of the crime and avoid the use of other features of the crime scene, such as objects that happened to be present there. Second, although in general, increasing the number of GKT questions is extremely desirable, our results demonstrated that there are restrictions to this general rule. Using only the four questions directly related to the theft was associated with better or equal detection efficiency compared with the use of all seven questions. These conclusions imply that the task of constructing a proper GKT may be more difficult than previously believed because it is not always easy to identify a sufficient number of items that are directly related to the execution of the crime. Perhaps detection efficiency could be increased by using a small number of GKT questions (i.e., three or four) based only on the most salient features of the event under investigation and repeating each question a few times, as suggested by Elaad and Ben-Shakhar (1997). Although a subsequent study (see Ben-Shakhar & Elaad, 2002) demonstrated that the use of multiple GKT questions is associated with higher detection efficiency than the use of many repetitions of a few questions, detection efficiency does increase with repetitions, and if only a few proper GKT questions can be identified, repeating them would enhance detection.

References

Amato-Henderson, S., Honts, C. R., & Plaud, J. J. (1996). Effects of misinformation on the concealed knowledge test [Abstract]. *Psychophysiology, 33,* S18.

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

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

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. (2002). Effects of questions' repetition and variation on the efficiency of the Guilty Knowledge Test: A reexamination. *Journal of Applied Psychology, 87,* 972–977.

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

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

Ben-Shakhar, G., & 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., 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.

Bradley, M. T., MacLaren, V. V., & Carle, S. B. (1996). Deception and nondeception in guilty knowledge and guilty actions 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.

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294.

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law.* New York: Cambridge University Press.

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

Eisen, M. L., Quas, J. A., & Goodman, G. S. (2002). *Memory and suggestibility in the forensic interview.* Mahwah, N J : Erlbaum.

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

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. (1997). Effects of items' repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology, 34,* 587–596.

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

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

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, & M. G. H. Coles (Eds.), *Advances in Psychophysiology, 4,* Jessica Kingsley.

Hira, S., Sasaki, M., Matsuda, T., Furumitsu, I., & Furedy, J. J. (2001). Pz-recorded P300 is highly accurate and sensitive to a memorial manipulation in an objective laboratory guilty knowledge test. *Psychophysiology, 38,* S50.

Hira, S., Sasaki, M., Matsuda, T., Furumitsu, I., & Furedy, J. J. (2002). A year after the commission of a mock crime, the P300 amplitudes, but not reaction time, are sensitive guilty knowledge test indicators. *Psychophysiology, 39,* S42.

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

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

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

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7,* 560–572.

Loftus, E. F., & Ketcham, K. (1991). *Witness for the defense.* New York: St. Martin's Press.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 9–31.

Loftus, E. F., & Palmer, J. E. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585–589.

Loftus, E. F., Schooler, J. W., & Wagenaar, W. A. (1985). The fate of memory: Comment on McCloskey and Zaragoza. *Journal of Experimental Psychology: General, 114,* 375–380.

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. (1998). *A tremor in the blood: Uses and abuses of the lie detector.* New York: Plenum Press.

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology, 2,* 117–163.

McCloskey, M., & Zaragoza, M. (1985a). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypothesis. *Journal of Experimental Psychology: General, 114,* 3–18.

McCloskey, M., & Zaragoza, M. (1985b). Postevent information and memory: Reply to Loftus, Schooler, and Wagenaar. *Journal of Experimental Psychology: General, 114,* 381–387.

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

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

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.

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

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

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277–295.

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

Received April 21, 2003
Revision received September 9, 2003
Accepted September 9, 2003 ■