Journal of Experimental Psychology: Applied
2012, Vol. 18, No. 3, 292–300

1076-898X/12/$12.00 DOI: 10.1037/a0028798

Algorithms for Detecting Concealed Knowledge Among Groups When the Critical Information Is Unavailable

Assaf Breska and Gershon Ben-Shakhar
Hebrew University of Jerusalem

Nurit Gronau
Open University

We examined whether the Concealed Information Test (CIT) may be used when the critical details are unavailable to investigators (the Searching CIT [SCIT]). This use may have important applications in criminal investigations (e.g., finding the location of a murder weapon) and in security-related threats (e.g., detecting individuals and groups suspected in planning a terror attack). Two classes of algorithms designed to detect the critical items and classify individuals in the SCIT were examined. The 1st class was based on averaging responses across subjects to identify critical items and on averaging responses across the identified critical items to identify knowledgeable subjects. The 2nd class used clustering methods based on the correlations between the response profiles of all subject pairs. We applied a principal component analysis to decompose the correlation matrix into its principal components and defined the detection score as the coefficient of each subject on the component that explained the largest portion of the variance. Reanalysis of 3 data sets from previous CIT studies demonstrated that in most cases the efficiency of differentiation between knowledgeable and unknowledgeable subjects in the SCIT (indexed by the area under the receiver operating characteristic curve) approached that of the standard CIT for both algorithms. We also examined the robustness of our results to variations in the number of knowledgeable and unknowledgeable subjects in the sample. This analysis demonstrated that the performance of our algorithms is relatively robust to changes in the number of individuals examined in each group, provided that at least 2 (but desirably 5 or more) knowledgeable examinees are included.

Keywords: Concealed Information Test (CIT), psychophysiological detection, group-level analysis, searching CIT

Psychophysiological methods for the detection of deception have been developed and investigated since the beginning of the 20th century (see, e.g., Ben-Shakhar & Furedy, 1990; Marston, 1917; Raskin, 1989; Reid & Inbau, 1977). The present study focused on one of two prominent methods for psychophysiological detection, known as the Guilty Knowledge Test or the Concealed Information Test (CIT). This method, which is designed to detect concealed knowledge rather than deception, is based on sound theoretical principles and proper controls and, therefore, it can protect innocent suspects from false accusations (see Ben-Shakhar, Bar-Hillel, & Kremnitzer, 2002; Ben-Shakhar & Elaad, 2002b; Lykken, 1974, 1998).

The CIT (Lykken, 1959, 1960) uses 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). The relevant items are significant only for knowledgeable (guilty) individuals (in the CIT paradigm, "guilty" participants are knowledgeable and "innocent" participants are unknowledgeable; we use these terms interchangeably) and there is ample evidence, mostly from psychophysiological research on orienting responses, indicating that significant stimuli elicit enhanced orienting responses (e.g., Gati & Ben-Shakhar, 1990; Siddle, 1991; Sokolov, 1963). Thus, if the suspect's physiological responses to the relevant alternative are consistently larger than to the neutral alternatives, knowledge about the event (e.g., crime) is inferred. As long as information about the event has not leaked out, the probability that an innocent suspect would 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; hence, it can be controlled such that maximal protection for the innocent is provided.

Extensive research conducted since the early 1960s has demonstrated that the CIT can be successfully used for detecting relevant information and discriminating between knowledgeable and unknowledgeable individuals (see Ben-Shakhar & Elaad, 2003, for a meta-analysis of CIT studies). During the past decade, the research interest in the CIT seems to be growing and various studies examining both the mechanisms underlying this method and applied questions related to its

This article was published Online First June 18, 2012.

Assaf Breska and Gershon Ben-Shakhar, Department of Psychology, Hebrew University of Jerusalem, Israel; Nurit Gronau, Department of Psychology, Open University, Raanana, Israel.

This research was funded by a grant from the Israel Science Foundation to Gershon Ben-Shakhar. We are grateful to Ewout Meijer, John Kircher, and Peter Rosenfeld for their constructive comments. We thank these individuals for their general contribution and not for any specific comment.

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

possible use as an aid in criminal investigations have been published (for a recent review of this research, see Verschuere, Ben-Shakhar, & Meijer, 2011).

Typically, most research and applications of the CIT have been focused on identifying individuals who committed a crime with the attempt to discriminate them from innocent suspects. This use of the CIT rests on the assumption that salient features of the crime are known to the investigators and, thus, they can be used to formulate the CIT questions (e.g., the type of weapon used, the stolen items). However, there are cases for which the precise details are not available to the investigators. For example, the Japanese police, who use the CIT extensively, apply this method in some cases to retrieve information that is unavailable to the investigators (e.g., finding the location of a murder weapon). This application of the CIT, "the Searching CIT" (SCIT), is described in detail by Osugi (2011). However, research examining the validity of the SCIT as used by the Japanese police is unavailable.

The SCIT may be applied in various other situations, and these applications may have significant implications for security and law enforcement agencies. For example, imagine a terrorist group planting a bomb in a certain location unknown to the investigators. Can this location be detected when suspects are identified and tested using the SCIT to prevent an upcoming explosion? Furthermore, can this method be applied to identify individuals and groups who are suspected in planning a future terror attack?

Recently, these questions have started receiving some research attention. Specifically, Meijer, Smulders, and Merckelbach (2010) used the SCIT with the electrodermal measure to test 12 participants who were informed about the details of a planned terror attack, but these details were not known to the investigator (although it was assumed that the terror-related details were among the different alternatives included in the test). Relying on group averages, these researchers were able to identify the correct alternative in each of the three SCIT questions used. However, in the study, all the examinees were exposed to the critical items, whereas in most real-life cases, it is uncertain whether a given suspect is actually aware of all the critical items. For example, in the terror attack example, some suspects may be only partially aware of the critical information or they may be innocent altogether (not belonging to the terror organization). Therefore, it is important to test the SCIT validity under conditions in which suspects' status (i.e., knowledgeable or unknowledgeable) is unknown to the investigator.

Bradley and Barefoot (2010) used the standard CIT in their study in which the critical information was known to the investigators, but they took advantage of groups sharing critical information. They used both skin conductance responses (SCRs) and heart rates to detect concealed information regarding two events (tea making and bomb making) and obtained much higher detection rates when the responses to the critical items were averaged across groups sharing the same information than when detection was made at the individual level. Specifically, for one event, detection rates of knowledgeable individuals was 47%, whereas 75% of the groups were correctly classified; for the second event, the detection rates for individuals and groups were 44% and 80%, respectively. A more recent study by Meixner and Rosenfeld (2011) was the first to use the SCIT and included a sample of innocents. This study used the P300 component of the event-related brain potentials and compared the largest average P300 amplitude of each participant with the second largest response. Detection was made at the individual participant level, and 10 of the 12 knowledgeable participants were correctly detected, with no false positives. This yielded an area under the receiver operating characteristic (ROC) curve of 0.979. In addition, 58% (21 of 36) critical items were correctly detected.

The application of the SCIT is a relatively new, standard procedure for analyzing data and detecting both critical items and knowledgeable individuals. Thus, the goal of the present study was to develop and test various algorithms for the detection of concealed information (i.e., unknown to the investigators) when suspects are either aware (guilty) or unaware (innocents) of the critical items. Specifically, we examined whether the SCIT can be used to detect critical information unknown to the investigators and, at the same time, differentiate between knowledgeable and unknowledgeable individuals. The proposed algorithms were based on the assumption that individuals who share the same concealed information would demonstrate a similar pattern of differential responses across items. One class of algorithms was based on averaging responses across subjects to identify critical items and then averaging responses across critical items to identify knowledgeable subjects. A second class of algorithms used clustering methods based on correlations between the response profiles of different subjects.

To test the proposed algorithms, we analyzed three data sets from previous CIT studies conducted in our laboratory. Originally, these studies relied on standard methods typically used in CIT research, namely, the original data analysis was designed to distinguish between knowledgeable and unknowledgeable individuals when all the critical items were known to the investigators. In the present study, we reanalyzed these data sets using the SCIT procedure (assuming that the critical items were included in the set of alternatives presented to the examinees, but were unknown to the investigator) in an attempt to identify the unknown critical items and, at the same time, differentiate between knowledgeable and unknowledgeable examinees. The detection efficiency of the various algorithms was compared with the results obtained from the same data sets when all critical items were known to the investigators. In addition, because the success in identifying both critical items and knowledgeable persons may largely depend on the numbers of knowledgeable and unknowledgeable examinees, we applied simulations to examine classification accuracy as a function of different group sizes.

Method

Description of the Data Sets

Data Set 1 was taken from a study designed to examine the effects of questions' repetitions on the outcomes of the CIT (Ben-Shakhar & Elaad, 2002a). Two conditions were extracted from this study, the condition in which each of 12 different questions was presented just once to knowledgeable participants ($n = 24$) and a control condition in which the same 12 questions were presented to unknowledgeable participants ($n = 12$). This experiment relied on the autobiographical version of the CIT in which the questions are related to several autobiographical details (e.g., first name, place of birth) and the critical items are the true biographical details of the examinee. Both SCRs and respiration line length (RLL) were used

in this study. The RLL was defined as the total respiration line length during the 15-s interval following stimulus onset. Typically, significant stimuli induce a suppression of respiration, which is reflected by a shorter RLL. Each measure was standardized within each participant and within each of three blocks containing 24 items (four questions with six alternative items per question). Within-block standard scores (i.e., standard scores based on the mean and standard deviation computed across the 24 items within each block) were used because it has been demonstrated that they are more resistant to habituation effects than standard scores computed across all items of all questions (e.g., Ben-Shakhar & Elaad, 2002a; Elaad & Ben-Shakhar, 1997). Thus, the detection measures were the mean standardized SCR and the mean standardized RLL to the critical items. For the unknowledgeable participants, one item within each question was randomly chosen to serve as the critical item.

Data Set 2 was taken from a study designed to examine the effects of information leakage and the use of crime-irrelevant "target items" on the outcomes of the CIT (Ben-Shakhar, Gronau, & Elaad, 1999). This experiment was based on a mock crime, and the conditions extracted from it were the "guilty" (knowledgeable) and the "innocents" (unknowledgeable); each condition included 36 participants. Three CIT questions were used in this study, and each question included one critical and four neutral control items, each repeated twice. As in the Ben-Shakhar and Elaad (2002a) study, both SCR and RLL were used as dependent measures and responses were standardized within participants and within each question (i.e., the *Z* scores were based on 10 items).

Data Set 3 was retrieved from a recent CIT study (Nahari & Ben-Shakhar, 2011) designed to examine the effects of memory and information leakage on the outcomes of the CIT. Because the original study included several experimental groups and to simplify our analysis, we selected only two groups of participants from this study: "guilty" participants who committed a mock crime ($n = 20$) and "innocent" participants who did not commit the mock crime and were unaware of the critical items ($n = 20$). The mock crime involved a theft from a university office of an envelope containing a sum of money and a jewel. The CIT was constructed from 10 questions pertaining to various features of the crime, including items that were central to the crime (e.g., the type of stolen jewel) as well as peripheral items that were related to the crime only indirectly (the name of a newspaper located on the office desk). SCRs were used in this study and they were standardized within participants and within each of two blocks containing 30 items (five questions with six items per question). Thus, the detection measure used in this study was the mean of the within-block standardized SCR elicited by the critical items.

Description of the Algorithms

We used two classes of algorithms to identify the critical information and to differentiate guilty from innocent individuals. Both classes relied on the assumption that as all guilty participants shared the same critical information, their response to the critical item within each question would be systematically increased relative to their responses to all other items. This increased response stems from an enhanced orienting response to the critical items, whereas variations between neutral items reflect random noise. However, for innocent participants, none of the items were expected to systematically elicit enhanced responses and, thus, responses to all items within each question necessarily would reflect random variations.[1] In Data Sets 1 and 2 in which both SCR and RLL were used, the algorithms were applied separately to each measure, as well as a combined measure defined as the average SCR and RLL *Z* scores (after multiplying the RLL *Z* scores by -1 because critical items are associated with reduced RLLs).

The first class of algorithms consisted of two stages. In the first stage, standardized responses to each item within each question were averaged across all participants (knowledgeable as well as unknowledgeable). These averages were used to identify the critical item within each question. Specifically, the item producing the maximal mean response within each question was labeled as the critical item for that question. A similar procedure was used by Meixner and Rosenfeld (2011), but they applied it at the individual level. To benefit from group data, in the SCIT, it is necessary to average responses across both knowledgeable and unknowledgeable participants because the investigator is unaware of who is guilty and who is innocent. However, we expected that the true critical item would elicit the largest average response because it would certainly elicit a strong response among the knowledgeable suspects for whom it had a special meaning and the unknowledgeable suspects would just add random noise (standardized responses around zero). On the other hand, the average standardized response to a noncritical item should be close to zero, as it reflects random noise for both knowledgeable and unknowledgeable suspects.

In the second stage, two algorithms were employed to compute a detection score for each participant. In the simple algorithm, the *Z* scores of the critical items, identified in Stage 1, were averaged within each participant across all questions to create a detection score for each participant. A second algorithm was based on a weighted average of these *Z* scores, where the weight of each question was the absolute *Z* score difference between the maximal response elicited by the critical item identified in Stage 1 and the next largest response elicited by another item within this question. In Data Sets 1 and 2, this was done separately for each measure as well as for the combined measure (thus each participant had three different detection scores for each algorithm).

The second class of algorithms applied to each data set was based on the intercorrelations between the response profiles of all subject pairs (i.e., for each pair of subjects, a correlation coefficient was computed between their respective responses to all stimuli). The underlying rationale of this algorithm was that two subjects sharing the same concealed information would be expected to show positive correlations between their response profiles across items as they would have increased responses to the same critical items. However, the response profile of an unknowledgeable subject reflects just random variations and, thus, it should not correlate with the response profile of any other subject. Thus,

[1] Although it can be argued that the critical items may be more salient than the neutral ones because of an unsuccessful attempt to equate all items, this was not the case in the present analyses because in all three studies different critical items were chosen for different participants. For example, in mock crime studies, each of the five types of jewels served as the critical (i.e., stolen) item for 20% of the participants and as a neutral item for the rest of the participants. Thus, CIT items were counterbalanced across participants, such that none of them was expected to systematically elicit enhanced responses among the unknowledgeable participants.

the challenge was to identify the cluster of subjects who had positive correlations because they shared this common pattern and to differentiate them from those who had near-zero correlations with other subjects.

To achieve this, we performed a linear decomposition of the original data matrix using principal component analysis (PCA). In this method, the pattern of values of each variable (column of the data matrix) across data points (rows of the data matrix) is assumed to be a linear combination of latent patterns (termed *principal components*), such that each observed variable is a weighted sum of these components. Each component is weighted by a scalar, termed *coefficient*, which reflects the extent to which the latent component affects the observed variable (in the similar factor analysis model, the equivalent of principal components are termed *factors* and the equivalent of coefficients are termed *loadings*). PCA is a mathematical method to decompose the data matrix into its principal components and estimate the coefficients by which each component should be multiplied to reproduce each of the original variables. It is important to note that the components extracted by PCA are those that explain most of the common variance in the data and are arranged in descending order. As such, the first component explains most of the common variance to all variables.

Here, we treated the subjects as variables (columns of the data matrix) and the responses of each subject across items as data points (rows of the data matrix). According to the PCA model, the positive correlations between knowledgeable subjects result from the fact that they all have large coefficients on the same component. In addition, assuming that the difference between knowledgeable and unknowledgeable individuals is the only systematic source of variance in the data, it is expected to be the largest source of variance. Thus, the component that explains the largest portion of the variance (the first component extracted by PCA) should reflect the pattern shared by all knowledgeable subjects, and the coefficient of each subject on this component reflects the extent to which the responses of that subject are explained by this pattern. Thus, the coefficient of each subject on the first component was used as the subject's detection score. In addition, the values of the first component were standardized within each question, and the item with the largest absolute standard score in each question was labeled as the item conveying the critical information in that question. Again, in Data Sets 1 and 2, this was done separately for each measure as well as for the combined measure.

Analyses of the Algorithms' Performance

To evaluate the efficiency of the proposed algorithms, we used the detection scores to discriminate between the knowledgeable and unknowledgeable participants in each data set. For this purpose, we constructed a ROC curve for each detection measure in each data set, and we computed the area under each ROC curve for the different algorithms (for a description of ROC construction in CIT studies, see Ben-Shakhar et al., 1999; National Research Council, 2003). In addition, the obtained ROC areas in each data set were compared with the respective area obtained in the original analysis under the assumption that all critical items were known (the true critical items). This area represents the maximal area that can be obtained from the simple averaging algorithm (see detailed explanation below).

In addition to testing the algorithms on full data sets, we also examined how their performance was affected by the number of guilty and innocent subjects in the different data sets. For the two classes of algorithms to operate, it was required that the response pattern that characterized the guilty subjects would be prominent compared with the random variation in the data. As the response profiles of the innocents reflect random variation, it is reasonable to expect that the larger the ratio of guilty to innocent subjects (G/I ratio), the higher the probability of identifying the critical items and discriminating the two groups. This expectation is based on the notion that increasing the proportion of knowledgeable to unknowledgeable subjects would yield an increase in the signal-to-noise ratio in the data. Thus, it was crucial to test the sensitivity of the algorithms to the G/I ratio. To achieve this, we randomly sampled from each data set subgroups of k guilty and l innocent subjects, such that k and l varied from two to the maximum number of guilty and innocent subjects in that data set. The algorithms and the ROC analysis were applied to each of these subgroups. To obtain stable estimates of the ROC areas, we used 1,000 iterations of this resampling procedure for each k and l combination, within each data set, and calculated the mean area under the ROC curve across these 1,000 iterations. In Data Sets 1 and 2, this analysis was conducted on the combined measure (mean of standardized SCR and RLL), which provided the best results in the initial analysis.

Results and Discussion

The relative efficiency of the algorithms was assessed by the number of critical items correctly detected in each data set, and by comparing the obtained areas under the ROC curves with the original areas derived in each study when all critical items were known. Figure 1 displays the outcome of the three proposed algorithms in detecting concealed information and in discriminating between knowledgeable and unknowledgeable individuals for Data Sets 1, 2, and 3 (Figure 1I, 1II, and 1III, respectively). Table 1 displays the number of CIT questions in which the true critical item of the total number of questions in each data set was correctly detected when using each of the proposed algorithms on each available measure.

Inspection of the Figure 1 reveals that despite obvious differences between the three data sets, most algorithms produced ROC areas that were fairly close to the original one. Indeed, even when examining data sets in which discrimination performance was initially moderate (Data Set 3 and the RLL results of Data Sets 1 and 2), the areas produced by the algorithms were highly similar to the original ones. This implies that the relative efficiency of the algorithms (i.e., the differences in ROC areas obtained in the original analysis, in which the investigator had knowledge of the critical items, and those obtained in the current analysis, in which such knowledge was unavailable) does not necessarily depend on the data being highly discriminative in the first place. To directly compare the performance of the proposed algorithms with a situation in which the critical items are known, we used a statistical test that performs a paired comparison between areas under ROC curves obtained by different classifiers that operate on the same data set (Hanley & McNeil, 1983). This test takes into account the correlation between the detection scores of the two classifiers and computes a Z score for the difference between areas

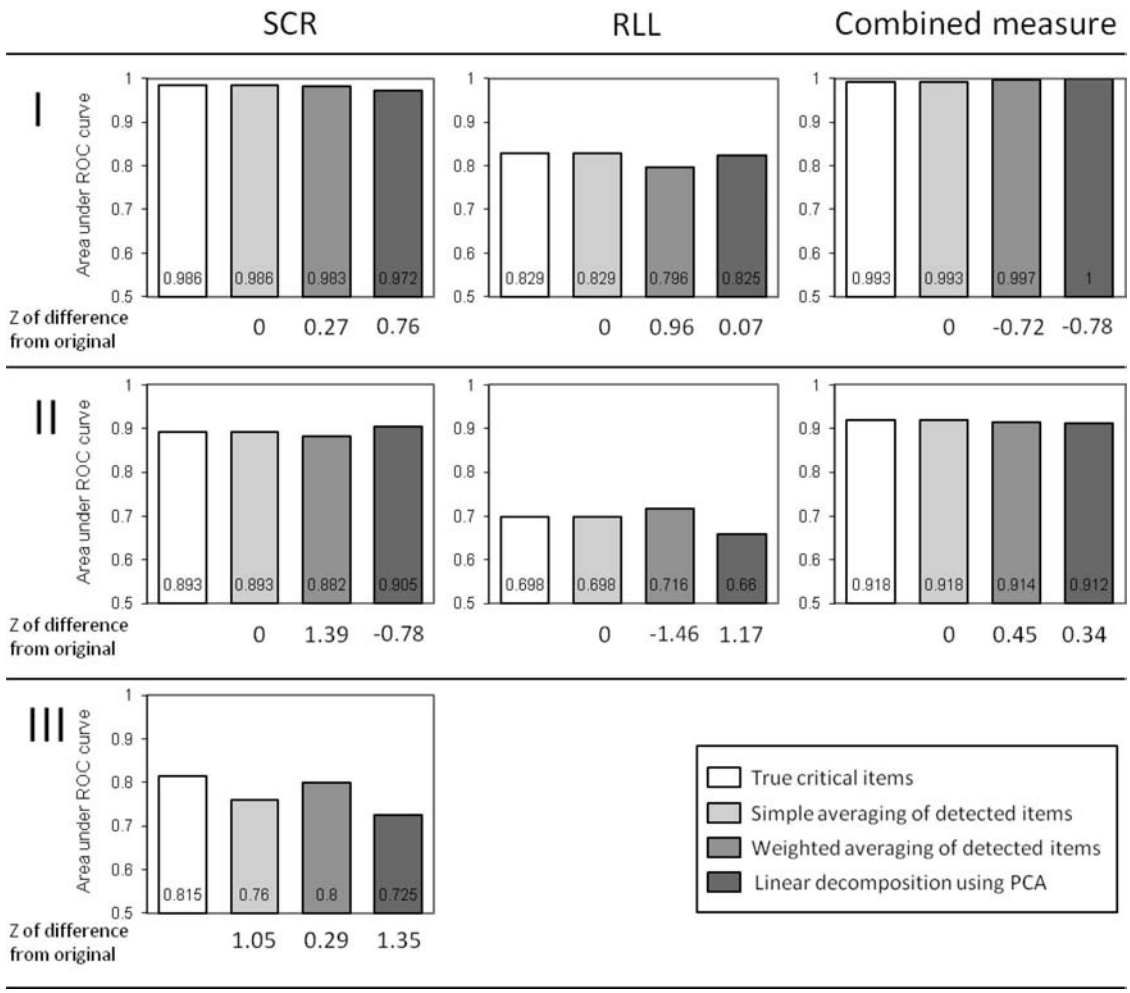

Figure 1. Areas under ROC curves for each physiological measure in each data set when using the true critical items (the original area) and with the three proposed algorithms. Top row: Data Set 1; middle row: Data Set 2; bottom row: Data Set 3. Left column: SCR; middle column: RLL; right column: combined measure (Data Set 3 included only SCR). The numbers below each graph are *Z* scores derived from a statistical test that examined the significance of the difference between the area obtained when using each algorithm and the area obtained when knowledge about true critical items was available (the original area). ROC = receiver operating characteristics; RLL = respiration line length; SCR = skin conductance response.

under the null hypothesis (i.e., the true difference is zero). When comparing the area obtained by each algorithm with the area obtained when the true critical items were known, for each measure in each data set, we found that none of the comparisons yielded a significant difference (Z scores < 1.65, $ps > .05$). The Z scores obtained for these comparisons are displayed in Figure 1.

Note that in Data Sets 1 and 2 the detection efficiency of all three algorithms was very high and in some cases reached perfect levels of discrimination between knowledgeable and unknowledgeable individuals. This is due to the impressive detection efficiency obtained in the original studies, but still it is encouraging to observe that our algorithms are capable of detecting concealed information even under the assumption that the critical information is not known. As all items were correctly detected in these data sets, the simple averaging algorithm always produced ROC areas that were identical to the original values reported in these studies. This is evident because in the original analyses the detection measures were defined as simple averages of the standardized responses across all critical items. This raises the possibility that the other two algorithms may produce even larger ROC areas than those reported in the original studies. Indeed, these two algorithms produced very high area values that, in some cases, exceeded the original ROC area, but the differences were small and not statistically significant.

In contrast, detection efficiency in Data Set 3 was much weaker than in the other two studies, and only 70% of the critical items were correctly identified by the simple averaging algorithm. This is probably due to the fact that Nahari and Ben-Shakhar (2011) used both central (i.e., items directly related to the mock crime, such as the stolen object) and peripheral (i.e., items not directly related to the mock crime such as a picture on the wall) CIT items (see also Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003;

Table 1
Algorithms' Success Rates in Detecting Critical Items

Data set	Measure	Averaging across subjects	Linear decomposition using PCA
1 (12 questions)	SCR	12/12	12/12
	RLL	11/12	10/12
	Combined	12/12	12/12
2 (3 questions)	SCR	3/3	3/3
	RLL	3/3	3/3
	Combined	3/3	3/3
3 (10 questions)	SCR	7/10	10/10

Note. Fractions indicate the number of questions, of the total number of questions in each data set, in which the true critical item was correctly detected with each algorithm. PCA = principal component analysis; SCR = skin conductance response; RLL = respiration line length.

Gamer, Kosiol, & Vossel, 2010). Yet, as mentioned earlier, even when using peripheral items, the ROC area produced by the weighted average algorithm was very close to the original ROC area. The other two algorithms yielded lower detection performance efficiencies, and it is interesting to note that although all critical items were correctly identified by the PCA-based algorithm (see Table 1), the respective ROC area reached only 89% of the original ROC area. This raises the possibility of using a combination of the two types of algorithms, such that at the first stage the critical item in each question will be identified by the PCA-based algorithm and then, at the second stage, the detection score will be defined as an average of the *Z* scores computed across all identified items. Clearly, this proposal requires an additional examination with independent data sets.

In addition to conducting a direct comparison among the different algorithms, Data Sets 1 and 2 allowed for a comparison between the SCR and RLL measures and for an examination of their incremental validity. This comparison revealed that although the SCR better discriminated knowledgeable from unknowledgeable individuals than the RLL, the use of a combined SCR-RLL measure further increased detection efficiency relative to the SCR alone. This advantage of the combined measure was too small to yield statistically significant outcomes (mostly because detection efficiency of the SCR alone approached perfect levels in these two data sets), but it was consistent across algorithms and data sets. Furthermore, this conclusion is consistent with findings of many previous studies (see a recent review by Gamer, 2011). Most relevant for the current study, the superiority of the combined measure remained when applying the CIT to the current situation in which the correct critical items were unknown (i.e., the SCIT). It should be additionally noted that in the current application of the algorithms as well as in the two original studies (Ben-Shakhar & Elaad, 2002a; Ben-Shakhar et al., 1999), the combined measure relied on equal rather than optimal weights. Equal rather than optimal weights were chosen to avoid capitalization on chance and also because equal weights have been demonstrated to perform very well, particularly when sample sizes are not very large (see Dawes, 1979). Detection efficiency of the combined measure with all three algorithms in Data Sets 1 and 2 was particularly impressive. In Data Set 1, all ROC areas for the combined measure exceeded 0.99, and both the weighted averaging and the PCA-based algorithms functioned slightly better than the simple averaging algorithm and consequently produced larger ROC areas than the one reported in the original study. In Data Set 2, all ROC areas were somewhat smaller than those in Data Set 1, probably because only three CIT questions were used in this study (Ben-Shakhar et al., 1999) compared with the 12 questions used by Ben-Shakhar and Elaad (2002a), but even here all areas exceeded the 0.91 value.

However, despite the impressive performance of the three proposed algorithms, it is clear that applying them to a situation in which the precise critical information is unavailable to the investigators may critically depend on the number of knowledgeable and unknowledgeable individuals tested. The present application of the algorithms relied on the number of participants initially allocated to different groups in the three data sets (ranging from 12 to 36 participants per group in each study). However, realistic situations, in which the proposed algorithms could be applied within the SCIT, may be characterized by very different numbers of knowledgeable and unknowledgeable suspects. For example, an attempt to identify a terrorist group may be characterized by a small number of individuals sharing the critical knowledge. Thus, to examine the robustness of the proposed algorithms to variations in the number of both knowledgeable and unknowledgeable suspects, we conducted an additional analysis in which we simulated various combinations of knowledgeable and unknowledgeable examinees and used the resampling method described earlier to estimate detection efficiency of the three algorithms for each combination.

The results of this analysis are presented in Figure 2. The figure includes nine displays (one for each combination of data set and algorithm). Each display includes several graphs showing the mean area under the ROC curve (based on the combined measure for Data Sets 1 and 2 and on the SCR measure for Data Set 3), averaged across 1,000 iterations, as function of the number of unknowledgeable examinees in the sample. Each graph represents a different number of knowledgeable examinees (2, 5, 8, 12, and for Data Set 3 only, also 17). Note that although Data Sets 1 and 2 include larger samples, we do not present results for the entire range of knowledgeable examinees because the results were fully saturated from about 12 subjects. The results for Data Sets 1, 2, and 3 are displayed in Figures 2I, 2II, and 2III, respectively, and the results for the different algorithms are displayed in the columns: The left, middle, and right columns display the results for the simple averaging, weighted averaging, and PCA-based algorithms, respectively.

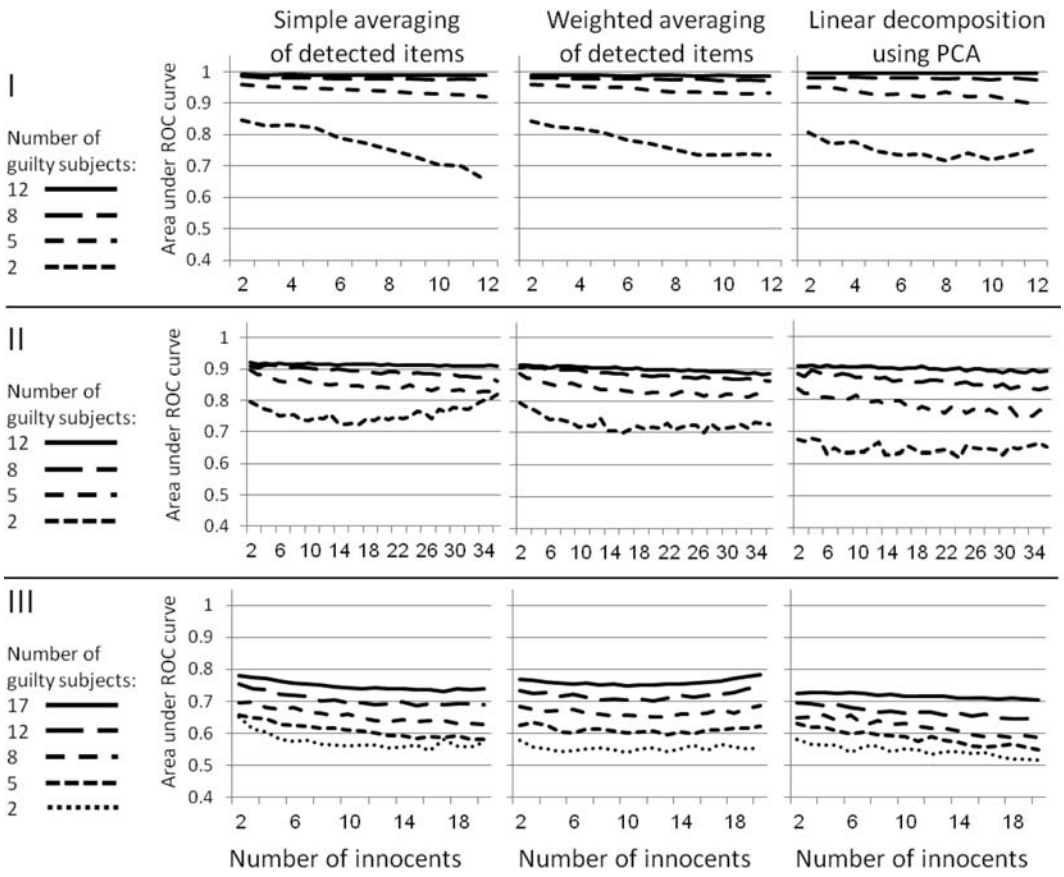

Figure 2. Areas under the ROC curves obtained for each data set with the three proposed algorithms when varying the numbers of guilty to innocent subjects. The combined measure was used for Data Sets 1 and 2 and SCR was used for Data Set 3. Top row: Data Set 1; middle row: Data Set 2; bottom row: Data Set 3. Left column: simple averaging algorithm; middle column: weighted averaging algorithm; right column: PCA-based algorithm. Within each graph, the *x*-axis represents the number of innocent subjects and separate lines represent the number of guilty subjects. ROC = receiver operating characteristics; SCR = skin conductance response; PCA = principal component analysis.

An inspection of these figures reveals two main findings. First, as expected, there was an improvement in performance with the increase in the number of knowledgeable individuals, as reflected by the increasing intercepts of the lines as the number of knowledgeable examinees increases. We did not conduct direct statistical comparisons among the intercepts, but the chance probability that three independent data sets would produce the same rank order of intercepts across the four numbers of guilty subjects was well below the .05 level. This finding supports the idea that with more knowledgeable examinees, the response pattern characterizing these examinees becomes more prominent, resulting in an improved separation between the two groups. It is interesting that, for a given number of knowledgeable individuals, performance was almost unaffected by an increase in the number of unknowledgeable (i.e., "innocents") examinees, as reflected by a slope close to zero for most lines. The only exception to this trend was Data Set 1, in which performance efficiency tended to decline with the addition of unknowledgeable examinees when only two knowledgeable individuals were included. However, this effect was eliminated when at least five knowledgeable examinees were included. This important finding suggests that even though the proportion of knowledgeable to unknowledgeable examinees becomes smaller with more unknowledgeable individuals, the classification efficiency is hardly affected by this decline, as it primarily depends on the number of knowledgeable examinees.

The superiority of Data Sets 1 and 2 is also reflected in this analysis, as detection efficiency in these two data sets approached the maximal possible value when at least five knowledgeable examinees were included in the sample. However, it should be noted that in these two data sets, the proposed algorithms performed above chance level even when only two knowledgeable individuals were examined. Taken together, the efficiency of information detection when the critical items are unknown appears to be high even when relatively small groups of "guilty" examinees share the relevant knowledge.

Our study constitutes an initial attempt to examine the efficiency of the proposed algorithms for detecting concealed knowledge unknown to the investigators. Several questions, however, remain open and should be addressed in future studies. First, the present results were predominantly based on standard CIT studies in which

subjects either performed a mock crime or were tested on their autobiographical details. It is not clear that similar results would have been obtained when testing suspects intending to perform a criminal act (but not actually performing it, as in the terror group example). In other words, we raise the question of whether detecting past actions is equivalent to detecting future intentions. This question was recently addressed by Meijer, Verschuere, and Merckelbach (2010), who conducted a systematic comparison between committing a mock crime and planning a mock crime. These authors demonstrated that the CIT with the SCR measure was similarly effective in both conditions, suggesting that our demonstration can be generalized from detecting involvement in a mock crime to detecting malintentions. This suggestion is also supported by the recent findings reported by Meixner and Rosenfeld (2011) showing impressive detection efficiency of the SCIT with participants who planned a mock terrorist attack.

Another important issue is the appropriateness of comparing the data sets used here with realistic cases. In particular, it may seem that Data Set 1, which was based on autobiographical items, does not resemble the type of items likely to be used in antiterror scenarios. However, names of people known to be important figures in terror (or organized crime) organizations may well serve as critical items for detecting whether suspects belong to such organizations. The other data sets used mock crime scenarios in which the critical items were provided to the "guilty" participants. As noted by Carmel et al. (2003) and Gamer et al. (2010), this may limit the generalizability of mock crime studies to real crimes in which the critical information is acquired incidentally and consequently may not be remembered when suspects are interrogated. However, in terror investigation scenarios, it is reasonable to assume that terror agents well rehearse the critical items.

Another concern is the fact that in realistic situations groups planning terror activities may not share all the information; thus, some members may be aware of some details and others are aware of a different set of details related to the plan. This, of course, introduces an additional difficulty to all the proposed algorithms, and future studies will have to be designed to test the robustness of these algorithms to partial information. It should also be noted that the application of the PCA-based algorithm requires that at least two examinees be aware of each critical item and this, of course, may limit the application of this algorithm.

A successful application of the SCIT as well as the standard CIT depends on a proper choice of items. In particular, there is always a concern that a particular noncritical item will be more arousing than all other items. In the standard CIT in which the critical item is known, this may lead to an increase in false-negative errors as knowledgeable suspects may show large responses to the arousing noncritical item. In the SCIT, on the other hand, this may lead to an increase in false-positive errors because when unknowledgeable suspects show large responses to the arousing noncritical item, this item may be mistakenly identified as the critical item and innocent suspects may be incriminated. In the present study, this concern did not play a role because known innocents were included in all data sets and they were effectively discriminated from the knowledgeable participants both with the standard CIT (the original studies) and when the SCIT was applied. However, in practice, investigators must pay special attention to this issue, and one possible solution is to pretest the CIT questions with a group of known innocents, as suggested by Lykken (1998).

Finally, it should be stressed that the application of the CIT to a situation in which the critical items are unknown to the investigators (i.e., SCIT) requires that forced-choice questions could be formulated such that the critical item is necessarily included among the various alternative items within each question. Clearly, this is a severe limitation, but there are important situations (e.g., detecting suspected terrorists) in which some knowledge is available to the investigative authorities (e.g., through intelligence). In addition, the SCIT can be applied in specific cases in which the critical item is a location (e.g., the location of a bomb, a body, or weapons). In these cases, the relevant space (e.g., a map of a city or a region) can be divided into smaller and smaller subspaces, such that the final identified area will be sufficiently small to be effectively searched. Clearly, this application of the SCIT also will have to be examined in future research.

Summary and Conclusions

The present study provides an initial demonstration for the efficiency of the CIT in situations in which the precise critical items are unknown to the investigators (i.e., SCIT). We proposed three algorithms and tested them on data from previous CIT experiments. The results indicated that, in most cases, the proposed algorithms were nearly as effective as in the original CITs, in which all critical items were known to the investigators. Furthermore, we demonstrated that the performance of these algorithms is relatively robust to changes in the number of knowledgeable and unknowledgeable individuals examined, provided that at least two (but desirably five or more) knowledgeable examinees are tested. Although this demonstration has obvious applied significance, we have raised several factors that may limit the application of the SCIT when the critical items are unavailable to the investigative authorities and have suggested several directions of future research to examine these factors.

References

Ben-Shakhar, G., Bar-Hillel, M., & Kremnitzer, M. (2002). Trial by polygraph: Reconsidering the use of the GKT in court. *Law and Human Behavior, 26,* 527–541. doi:10.1023/A:1020204005730

Ben-Shakhar, G., & Elaad, E. (2002a). Effects of questions' repetition and variation on the efficiency of the Guilty Knowledge Test: A reexamination. *Journal of Applied Psychology, 87,* 972–977. doi:10.1037/0021-9010.87.5.972

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

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. doi:10.1037/0021-9010.88.1.131

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception: A psychophysiological and international perspective.* New York, NY: Springer-Verlag. doi:10.1007/978-1-4612-3282-7

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. doi:10.1037/0021-9010.84.5.651

Bradley, M. T., & Barefoot, C. A. (2010). Eliciting information from

groups: Social information and the Concealed Information Test. *Canadian Journal of Behavioural Science, 42,* 109–115. doi:10.1037/a0018146

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the Guilty Knowledge Test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied, 9,* 261–269. doi:10.1037/1076-898X.9.4.261

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. doi:10.1037/0003-066X.34.7.571

Elaad, E., & Ben-Shakhar, G. (1997). Effects of items' repetitions and variations on the efficiency of the Guilty Knowledge Test. *Psychophysiology, 34,* 587–596. doi:10.1111/j.1469-8986.1997.tb01745.x

Gamer, M. (2011). Detecting concealed information using autonomic measures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 27–45). Cambridge, England: Cambridge University Press.

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Action Test. *Biological Psychology, 83,* 101–107. doi:10.1016/j.biopsycho.2009.11.005

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. doi:10.1037/0096-3445.119.3.251

Hanley, J. A., & McNeil, B. J. (1983). A method for comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148,* 839–843.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43,* 385–388. doi:10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44,* 258–262. doi:10.1037/h0044413

Lykken, D. T. (1974). Psychology and the lie detector industry. *American Psychologist, 29,* 725–739. doi:10.1037/h0037441

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

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology, 2,* 117–163. doi:10.1037/h0073583

Meijer, E., Smulders, F., & Merckelbach, H. (2010). Extracting concealed information from groups. *Journal of Forensic Sciences, 55,* 1607–1609. doi:10.1111/j.1556-4029.2010.01474.x

Meijer, E. H., Verschuere, B., & Merckelbach, H. (2010). Detecting criminal intent with the Concealed Information Test. *The Open Criminology Journal, 3,* 44–47. doi:10.2174/1874917801003020044

Meixner, J., & Rosenfeld, P. J. (2011). A mock terrorism application of the P300-based complex trial protocol for detection of concealed information. *Psychophysiology, 48,* 149–154. doi:10.1111/j.1469-8986.2010.01050.x

Nahari, G., & Ben-Shakhar, G. (2011). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology, 48,* 733–744. doi:10.1111/j.1469-8986.2010.01148.x

National Research Council. (2003). *The polygraph and lie detection.* Washington, DC: National Academies Press.

Osugi, A. (2011). Daily application of the CIT: Japan. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test* (pp. 253–275). Cambridge, England: Cambridge University Press.

Raskin, D. C. (1989). Polygraph techniques for the detection of deception. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 247–296). New York, NY: Springer-Verlag.

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

Siddle, D. A. T. (1991). Orienting, habituation, and resource allocation: An associative analysis. *Psychophysiology, 28,* 245–259. doi:10.1111/j.1469-8986.1991.tb02190.x

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

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). *Memory detection: Theory and application of the Concealed Information Test.* Cambridge, England: Cambridge University Press.

Received August 17, 2011
Revision received March 28, 2012
Accepted April 9, 2012 ■