Detecting deception by physiological and behavioral measurements:
Effects of emotional arousal and realistic mock-crime

By: Yulia Levin

Supervised by: Prof. Gershon Ben-Shakhar, The Hebrew University of Jerusalem
Dr. Nurit Gronau, The Open University of Israel

December 2012
Abstract

This study examined the effects of emotional arousal on memory in detecting concealed information using the electrodermal measure, respiratory measure and Symptom Validity Test. Participants were randomly assigned to two groups: guilty, who committed murder in virtual reality apartment; informed-innocents, who were in the same virtual reality apartment and were exposed to the same details as a guilty group, but their task was with emotionally neutral context (to pass an envelope to character that enters the apartment). Participants were tested immediately or one week later. The results showed that guilty participants remembered more items, especially central items. The tests’ efficiency declined in the delayed condition in both groups. This decline reduced the possibility to accuse suspects who were exposed to crime details but did not commit it. In addition, the results showed that the detection efficiency of combination of physiological measures better than detection efficiency of behavioral measure. Combination of physiological and behavior measures improve the detection efficiency only in the guilty group in delayed testing. These findings suggest that, while time delay may somewhat reduce the ability to detect the guilty, it also diminishes the danger of accusing informed-innocents.
Detecting deception by physiological and behavioral measurements.

Scientists and forensic experts have attempted for many years to develop instruments and methods for detecting deception. Over the years several methods of psychophysiological detection of deception such as the Control Question Technique (CQT) and the Guilty Knowledge Test (GKT) also known as Concealed Information Test (CIT) were developed. Both methods are based on the principle of mirroring psychological processes through a person's physiological responses and compare the suspect's physiological responses to relevant and control questions that are irrelevant for the investigated event. The difference between the methods is expressed in the structure of the relevant and irrelevant questions. Although the CQT method is common in North America and Israel in the applied field, it is subject to criticism about its validity. One of the main arguments against it is the inherent dissimilarities between the control questions and the crime relevant questions that constitute a built-in bias against innocent suspects (Lykken, 1974; Furedy, 1989; Saxe, 1991; Ben-Shakhar, 1991, 2002).

The GKT (or CIT) method has been studied and shown to be of high validity for detecting hidden knowledge. This test has a series of questions with multiple answers: one correct answer to the investigated crime and a number of neutral responses (not related to the investigation). A naïve subject, lacking knowledge about the details of the crime, can't distinguish between the answers (Lykken, 1959, 1960, 1998). If a subject's physiological responses to the relevant answer are consistently larger than to the neutral answers, it can be assumed that he is knowledgeable about crime-related details. The probability that an innocent subject will respond to relevant stimuli
can be controlled by a number of test questions and the number of possible answers each question provides (Ben-Shakhar & Elaad, 2003).

One of the arguments directed against the GKT is the problem that not only the culprit is exposed to details of the crime. A possible leakage of information to innocent suspects may lead to enhanced responses to these details and eventually to a misclassification of the informed innocent suspects as guilty (e.g., Bradley, Barefoot, & Arsenault, 2011). Some studies compared the responses of subjects who were aware of the details of the crime but did not actually perform it (hereinafter – informed innocents) to responses of the subjects who committed the crimes. These studies found that informed innocents showed weaker responses to the relevant items than the guilty ones, but compared to the innocents, had stronger reactions. This makes it difficult to distinguish between guilty subjects and informed innocent subjects (Ben-Shakhar, Gronau, & Elaad, 1999; Bradley, MacLaren & Carle, 1996).

Bradley & Warfield (1984) suggested a change in the GKT and called it Guilty Action Test (GAT). In this test subjects are asked about their actions. For example, in the GAT test subject is asked "Did you steal ₪200?", 300?..., while in the GKT test, the subject is asked, "How much money was stolen? – 300?, 200?...". The main research paradigm that is used to investigate the efficiency of the process is mock crime. In this paradigm some of the subjects are asked to perform a theft (stealing a CD, money, jewelry, etc.) and some are asked only to enter a different room. Then the subjects are asked to answer questions about the details of the event and the room while they are connected to a device that measures their physiological indices. (Eg, see Bradley et al., 1996; Ben-Shakhar et al., 1999).
Gamer, Kosiol & Vossel (2010) used the GAT test to compare “guilty” subjects to “informed innocents”. Subjects were tested immediately after committing the crime or two weeks later. They found that immediately after the crime it was easy to distinguish between the innocent individuals and the guilty but it was difficult to distinguish between guilty subjects and informed subjects. Two weeks later it was easier to distinguish between guilty subjects and informed subjects but it was more difficult to distinguish between innocent subjects and informed subjects. This can be explained by the differences of the memory of the “guilty” and “informed” subjects. The memory of the “guilty” subjects to the relevant items is not damaged even two weeks later, while the memory of the informed subjects is weakened. This explains why it was harder to distinguish between innocent subjects and informed subjects in the delayed examination.

Ben-Shakhar et.al (1999) suggested that factors, such as personal involvement and motivation to avoid detection, can increase the efficiency test of the GKT. Ben-Shakhar & Elaad (2003) performed a meta-analysis of mock crime studies that used the GKT based on skin conductivity. They suggested that increased motivation for concealment leads to attempts to control physiological responses, which in turn, raises the significance of the crime-related items and facilitates the identification.

Another physiological measure used for lie identification is respiration. This measurement combines breathing depth and rate together in a response called "respiratory line length" (RLL). This index has been shown to be relatively unaffected by habituation (e.g., Ben-Shakhar & Elaad, 2002). Some of the studies in this field found that a combination of respiratory index
together with the skin conductance measure improves the efficiency of identifying guilty subjects in mock crime studies (Ben-Shakhar & Elaad, 2002; Ben-Shakhar et al., 1999; Gamer, Verschuere, Crombez & Vossel, 2008).

In addition to physiological measures there have been several attempts to use behavioral measures, such as response time (e.g. Seymour, Seifert, Shafto & Mosmann, 2000; Gronau, Ben-Shakhar & Cohen, 2005) for the detection of concealed knowledge. A behavioral method that can be applied to identify knowledge is the Symptom Validity Testing (SVT), which has been used for the detection of malingering (e.g., Bianchini, Mathias & Greve, 2001). This is a direct test in which subjects are asked several questions about the details of an event they claim not to remember or not to have performed. They are presented with two answer options, one associated with the event details and the other not related to the event. The subject is supposed to choose one answer. Subjects who suffer from amnesia or who do not have the information will randomly select answers and thus will choose the right and the wrong answer with equal probabilities. If the level of subject performance significantly exceeds the chance level, the subject can be classified as a malingerer or impostor. Studies that used the SVT found that 40-59% of imposters exceed the chance level (Jelicic, Merckelbach & Bergen, 2004a, 2004b; Merckelbach, Hauer, & Rassin, 2002).

One of the criticisms of this test is that its effectiveness decreases when it comes to skilled subjects. Currently, the principle of the SVT is publicized and there is a risk that potential examinees will be aware of the principle by which the test runs and apply countermeasures. Thus, when the subject performs the test at chance level the result can indicate that the
subject is innocent, but also that he learned the test principles and used countermeasures (Verschuere, Meijer & Crombez, 2008).

The SVT is based on different psychological mechanisms than those that underlie the physiological measures typically used in the GKT and thus combining the two types of measures may enhance detection efficiency. Indeed, Meijer, Smulders, Johnston & Merckelbach (2007) demonstrated that a combination of both tests increased the percentage of guilty detection in the feigned amnesia paradigm. Nahari & Ben-Shakhar (2011) found that the GKT combined with behavioral tests such as the SVT increases the ability to distinguish between guilty and innocents in a mock crime paradigm.

The mock crime paradigm takes place in unnatural conditions. Typically, participants receive detailed instructions on the items they need to steal and sometimes these items are rehearsed prior to the tests. In realistic situations it is uncertain which crime details will be noticed by the perpetrator and whether they will be remembered when suspects are tested. Thus, it is necessary to test the GKT paradigm under more realistic conditions. Van Oorsouw & Merckelbach (2006), used a more realistic mock crime and required participants to enter a bar, hit a man (a doll of a person), tie him up and finally steal money from the money box. Then in the SVT, the subjects had to fake amnesia when they were asked about the bar, the victim and the crime. There were many items and details that the subjects could pay attention to. Only a small fraction of subjects were identified by the SVT as faking amnesia.

An additional factor that differentiates mock crimes from realistic ones is emotional arousal which may affect perpetrators’ memory and physiological
responses. Laney, Campbell, Heuer & Reisberg (2004) found that subject's memory for events that evoke emotion were better than for neutral events. Nahari & Ben-Shakhar (2011) and Gamer et al. (2010) found that an informed subject's memory of the mock crime was weakened after a week or two when compared to guilty subjects. It is possible that this occurs due to different emotional arousal in informed and guilty subjects.

A number of studies examining the effects of negative emotions on memory have revealed that individuals remember elements that are centrally tied to the emotional item, but they tend to forget elements more peripheral to the emotional aspect of the event (e.g., Loftus, 1979). Kensinger, Garoff-Eaton & Schacter (2007) found that participants better remembered the details of negative objects than of the neutral ones, but they were worse at remembering the details of the backgrounds presented with a negative item compared to those presented with a neutral item. These results emphasize that the effects of emotion on memory for detail can be critically affected by the type of detail being assessed (Kensinger, 2007). Studies that used mock-crime paradigms found that there is a different pattern of memory for central and peripheral items. This pattern is helpful in distinguishing "guilty" and "informed" subjects. (Carmel, Dayan, Naveh, Raveh & Ben-Shakhar, 2003; Gamer et al, 2010; Nahari & Ben-Shakhar, 2011). For example, Peth, Vossel & Gamer (2011) found that central crime details were remembered better than peripheral ones in the immediate and in the delayed conditions. In addition, they found enhanced emotional arousal during the mock crime reduced memory performance for peripheral details.
The aim of the current study was to examine the effects of emotional arousal on memory for details in a mock-crime paradigm. Memory of crime details was investigated under two conditions - 1) immediately after the crime, or 2) a week after the crime. In order to simulate "natural" conditions as much as possible, the 'crime' was conducted in a virtual reality environment - computer-simulated environment that can simulate physical presence in real world, as well as in imaginary worlds. This virtual reality environment is primarily a visual experience, presented through special stereoscopic displays and including sound through headphones.

The crime that participants performed was a murder. It is a morally forbidden crime and one of the most serious crimes. Therefore, it was expected that the murder would induce high emotional arousal. Emotional arousal was measured during the virtual reality environment by physiological measurements such as skin conductance and heart rate, and also by a self-report questionnaire.

The hypothesis was that subjects in the guilty group would be more emotionally involved than the informed group and therefore would show larger differential responses to the relevant items in both the immediate and the delayed tests. In addition, subjects in the guilty group would remember more items, which would be reflected in the later examination. It follows that the difference in memory between the guilty and the informed groups would be maximal in the delayed condition. In addition, this study would compare the detection efficiency of the GKT test based on physiological measurements of skin conductance and respiration and the SVT test. The hypothesis was that the detection efficiency using the physiological measures would be higher.
than that obtained with the SVT but a combination of all measures would lead to the maximal detection efficiency.

**Methods**

**Participants**

Eighty nine Hebrew University of Jerusalem undergraduate students (55 females and 34 males) participated in the experiment for course credit or payment (they received 45 NIS). All participants signed a consent form indicating that participation was voluntary and that they could withdraw from the experiment at any time without penalty and health form indicating that they can participate in the experiment (see appendix A). Twenty eight participants did not have physiological data from the first stage because of technical problems. Nine participants were eliminated from SCR analysis due to unusually high skin resistance levels or excessive movements during the experiment, and four participants were eliminated from RLL analysis due to unusual respiration levels or excessive movements during the experiment.

**Tools**

**First stage**

Virtual reality system - WorldViz company system that includes a head-mounted display through which the subjects experienced the scenario. The head-mounted display has a motion sensor that allows the subject's movement to be monitored while simultaneously adjusting itself to what he sees.
Skin conductance measurements – This measurement was performed by a BIOPAC MP150. Two electrodes, Ag / AgCl with 0.05 M NaCl electrolyte, were placed, on the left index and the left ring finger.

Heart rate – This measurement was performed by a BIOPAC MP150. Two electrodes, AG / AgCl with 0.05 M NaCl electrolyte, were placed on the left wrist and on the left side of neck.

Emotional Questionnaire - This questionnaire was composed of 10 questions. Its purpose was to measure subject's emotional state while being in the virtual reality environment (see appendix B).

**Second stage**

Skin conductance - This measurement was performed by constant electrical current (0.5 V Atlas Researches, Hod Hasharon, Israel) during GKT. Two electrodes Ag / AgCl with 0.05 M NaCl electrolyte, one on the left index finger and the other on the left ring finger.

Measurement of respiration - was done by a resistance band with two electrodes which was attached to the subject's chest during GKT.

The experiment took place in air-conditioned room. The computer was used to display the stimuli and for calculating the conductivity of the skin and the respiratory line length.

**Experiment design**

A 2 x 2 between subjects design was used. The first factor was the group: subjects who had committed a crime - "guilty" subjects and subjects who have not committed a crime but were exposed to the crime scene in the virtual reality environment, but in an innocence context (they didn’t see the crime...
itself)- a group of "informed" subjects. The second factor was the time span of
the CIT and memory test for crime details: immediately after the first part of
the experiment - immediate condition, or a week after the first part – delayed
condition. The subjects were randomly assigned to the four experimental
conditions.

Procedure
The experiment was conducted in two stages:

Stage1:

The first stage of the experiment was in a virtual reality environment.
Participants arrived individually to the laboratory at a predetermined time.
They filled out a health questionnaire that ensured their ability to participate in
the experiment. They were then connected to electrodes that measured skin
conductance and heart rate. During the experiment, subjects had a head-
mounted display; through it they received the instructions about their task and
saw the "crime environment ". This "crime environment" was an apartment; all
the subjects saw the same apartment, but with different details. The virtual
reality environment had five "sets" of details. All sets had the same Item
categories, but they varied by the specific details of the item. Different
subjects could see similar yet different items, such as different color cushions
on the sofa, wall clocks in various forms and different musical instruments.
The subjects were allocated randomly to the different sets.

Subjects that were assigned to the "guilty" group were given
instructions about being a CSI secret agent and that the person who will enter
the apartment is a double secret agent. Their task was to kill the man. The
person that entered the room had to ask the subject "where's the money?" before the subject shot him. The subject had a two minute interval between the end of instructions and the person's entrance into the apartment. In this two minute period the subject could freely walk around the apartment.

The subjects that were assigned to the "informed" group received instructions that they were the company's purchasing managers and that they should transfer money for a big purchase to the person that enters the apartment. The subjects had to pass an envelope containing money to the person entering the apartment after he asks: "Where's the money?". The subjects had a two minutes interval between the end of instructions and the person's entry into the apartment. In this time the subjects could walk freely around the apartment. This apartment was the same apartment with the same items, as the apartment that was seen by the "guilty" participants.

After completing this section, the subjects filled out a questionnaire about their emotional state while being in the virtual reality environment. This part took about half an hour in all conditions.

Stage 2

GKT and SVT were administered to all participants. Participants in the immediate condition took the tests immediately after first stage, while those in the delayed condition took them one week later. An experimenter blind to the subject group and to the participants' relevant items explained to the subjects that a murder had been committed in the virtual reality environment and that they are suspects in committing this murder. He/she explained that the experiment was designed to test whether they could cope with lie detection tests and convince the examiner that they
are innocents. It was emphasized that beating these tests is a difficult assignment that only a few people can succeed in, and were promised a bonus of 10 NIS for a successful performance of the task. Subsequently, the subject was connected to electrodes that measured skin conduction and a belt that measured respiration. The GKT questions were presented after an initial rest period of 2 minutes, during which skin conductance baseline was recorded. The GKT questions appeared on the computer screen and at the same time played through computer speakers (see appendix C). This test had 14 questions with 6 possible answers for each question, such that only one of them was "relevant" (i.e., was seen by the participant in the virtual reality environment during first stage). The first item in each question was always a buffer aimed to absorb the subject's orientation response to the first stimulus. This item was not included in the data analyses. Questions and answers appeared in random order, except two questions. Question 11 was always about the voice of the person that appeared in the virtual reality and the question about the weapon was always last. Subjects were asked to respond verbally, saying "No" to all items. A short break was given after the presentation of seven questions.

After a GKT subjects were disconnected from the electrodes and respiration belt and the SVT was administered. This test had 21 questions (see appendix D). Each question had two possible answers, one correct and one incorrect. The subject was asked to select one of the answers with the mouse. Before the beginning of this test, the experimenter instructed the subjects to prove their innocence and told them that if they don't know the answer, they must guess.
At the end of the questioning session, the experimenter thanked the participants. He/She asked them to wait until the computer program processed the data of the tests and reached a decision as to whether they were found "guilty" or "innocent". This calculation was made in order to know whether the subject deserves the promised bonus. Then, the experimenter told the subject that the experiment was over and asked him to undergo a memory test in which he has to answer only the truth about details that he remembers. The instruction was that if he does not remember or know the answer, he can select "Do not know" option. This test was like the SVT test with 21 questions and two possible answers for each question, such that one of them was correct. The purpose of this test was to check the subject's memory. Finally, all participants were debriefed and compensated.

The second stage of the experiment lasted about an hour.

**Manipulation check: Heart rate and SCR during the virtual reality phase.**

The purpose of the manipulation was to induce a high emotional arousal in the experimental groups. To check whether the manipulation indeed had a significant emotional impact, a control group of "innocents" (unknowledgeable) participants was added to the first stage. Twenty five participants arrived individually at the laboratory and filled out a health questionnaire that assured they were eligible to participate in the experiment. They were then connected to electrodes that measured their skin conductance and heart rate. During this stage of the experiment, subjects had a head-mounted display through which they received instructions about their task. They were told that they were purchasing managers of an art gallery who traveled to a foreign country, in which they were to buy a painting from
the local art gallery. The participants were asked to point to the selected painting when the manager of the gallery entered the room. They had a two minutes interval between the end of the instructions and the entrance of the manager into the gallery. In this time they were able to walk freely around the virtual environment of the gallery and to observe all the paintings. After completing this section, the participants filled out a questionnaire about their emotional state while being in the virtual reality environment. This part lasted about half an hour.

The duration during which participants explored in the virtual reality environment was divided for all groups into 3 phases: (1) baseline; (2) exploring the apartment or the gallery; and (3) the time duration from the entrance of the character to the end of the experiment (including the entrance itself). Mean heart rate and SCR scores were computed for each group in each phase. There were no differences between the groups in the baseline and the exploring phases in both physiological measures.

The following two orthogonal planned contrasts were conducted on the Heart rate and SCR measures for the last phase (i.e., from the entrance of the character tills the end of the experiment): (1) Combined "guilty" and "informed innocents" (knowledgeable participants) were compared with the "uninformed innocents"; (2) "Guilty" group was compared with "informed innocents". The first contrast conducted on the Heart rate measure was statistically significant \[t(79)=1.86\ p<0.05\ d=0.45\], indicating that knowledgeable subjects showed an increased heart rate than unknowledgeable subjects. Mean Heart rate scores within each condition are presented in Table 1. No statistically significant Heart rate differences were
found between the "guilty" and the "informed innocents" groups [t(79)=0.43, p>.05]. These findings do not support the hypothesis that participants in the "guilty" group will show larger responses in the Heart rate measure. These findings show that knowledgeable participants were more emotionally involved than the unknowledgeable participants but there were no differences in the emotionally involvement between the "guilty" group and the "informed innocents" group. Mean SCR scores within each condition are presented in Table 1. None of the planned contrasts for SCR measure was statistically significant [t(79)=0.53 for the first contrast, t(79)=0.65 for the second contrast, p<.05].

A factor analysis conducted on the post-experimental questionnaire revealed two factors: stress (questions 1,3,5,6,8) and identification with role played by the subject (questions 2 and 4). Question 9 and 10 were excluded from the analysis because they served as control questions, checking the visibility and the participants' feeling during the VR. Question 7 was excluded from this analysis because it did not match any of the factors. The same two contrasts described above were conducted on each of these two factors. A statistically significant difference between knowledgeable and unknowledgeable subjects was found for the stress factor [t(118)=2.99, p<0.01, d=0.57], indicating that knowledgeable participants showed higher levels of stress than the unknowledgeable participants. In addition, the "guilty" group displayed a significantly higher level of stress than the "informed innocents" group [t(118)=1.91 p<0.05, d=0.41]. Knowledgeable participants showed significantly smaller identification scores than the unknowledgeable subjects [t(118)=-2.84, p<0.01, d=0.54] and the "guilty" group did not differ
significantly from the "informed innocents" group on this factor \[t(125)=0.06, p>.05\].

In conclusion, the behavioral findings suggest that our manipulation has achieved the desired results (i.e., higher stress levels among the guilty participants than the informed-innocents participants). No differences in stress levels between the two knowledgeable groups were observed with the physiological measures (GSR, Heart rate).

**Table 1.** Means and SDs of Heart rate, SCR, stress and identification factor scores within each experimental group.

<table>
<thead>
<tr>
<th>Identification</th>
<th>Heart rate Mean</th>
<th>SCR Mean</th>
<th>Stress Mean</th>
<th>Identification Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guilty</td>
<td>103.15</td>
<td>7.45</td>
<td>4.33</td>
<td>4.03</td>
</tr>
<tr>
<td>Informed</td>
<td>100.74</td>
<td>7.94</td>
<td>3.87</td>
<td>4.01</td>
</tr>
<tr>
<td>Innocents</td>
<td>92.32</td>
<td>8.06</td>
<td>3.45</td>
<td>4.75</td>
</tr>
</tbody>
</table>

Data analyses of the second stage of the experiment

**SCR**

Responses were transmitted in real time to the computer. SCR was defined as the maximal increase in conductance obtained from the examinee, from 1s to 5s after stimulus onset and computed using an A/D (NB-MIO-16) converter with a sampling rate of 50 Hz. To eliminate individual differences in responsively and permit meaningful comparisons of the responses of different examinees, each participant’s SCR was transformed into within-examinee standard scores (eg. Ben-Shakhar, 1985; Nahari & Ben-shakhar, 2011). To minimize habituation effects, within-block standard scores were used (see Ben-Shakhar & Elaad, 2002; Elaad & Ben-Shakhar, 1997). The 84 items were divided into 2 blocks, each consisting of 42 items. Thus, the Z scores used in
this study were computed relative to the mean and standard deviation of the participant’s responses to the 42 items of each block. Finally, three detection scores were computed for each participant using the average standardized response computed across the relevant items (one across all 13 items, excluding the question about the weapon, one across the 3 central items and one across the 10 peripheral items). The Z score for the question about the weapon was used as an additional detection score.

RLL

Respiration line length (RLL) was scored during the interval 0-12 s following item onset (for a detailed description of RLL computation, see Elaad & Ben-Shakhar, 1997). The RLLs of each participant were transformed into within-examinee standard scores computed within each block of trials, just like the SCRs, but as shorter RLLs are expected when critical items are presented (respiration pause), the RLL Z scores were multiplied by -1. Finally, four detection scores were computed for each participant using the same procedure described above.

Combined measure

Combined detection scores were computed for each participant as the sum of the SCR and RLL detection scores.

SVT

An unknowledgeable individual (uninformed innocent) is expected to guess the answers on the SVT and thus give about 50% correct answers (chance level). It is hypothesized that a person who is aware of the critical items will be unable to ignore this information when answering the SVT and consequently deviate from chance level performance. It is reasonable to
assume that individuals attempting to conceal critical items will display below chance level performance on the SVT (e.g., Verschuere, Meijer & Crombez, 2008).

A detection measure based on the SVT was defined as the absolute deviation of the percent of correct answers from chance level (50%). Specifically, this measure was defined as $|P - 50\%|$, where $P$ is the percent of the participant’s correct answers (Nahari & Ben-Shakhar, 2011). Detection scores were computed for 20 items. The question about a weapon was excluded from this analysis because participants in the "informedinnocents" group were not informed about it. Scores for central items were not computed separately because of small number of items in this category.

**Memory test**

Accuracy scores in the memory test for all items without the weapon question, for central items, for peripheral items were computed as proportion of correct answers.

**Data analysis and statistics**

A mixed 2X2 X2 analysis of variance (ANOVA), with item-type (central, peripheral) serving as a within-subjects factor and group ("guilty" vs. "informed innocents") and time of GKT (immediate vs. delayed) serving as the 2 between-subjects factors. The dependent measures were detection scores constructed for SCR, RLL and combined scores. In addition, a 2X2 ANOVA, with group ("guilty" vs. "informed innocents") and time of GKT (immediate vs. delayed) serving as the 2 between-subjects factors was conducted on the detection score computed across all items (without the question about the weapon) for each measure (SCR,RLL combined measure and SVT). This
Anova was conducted to examine the effects of the between-subject factors across all items because the detection scores for the central and peripheral items were based on very different numbers of items. Finally, another Group X Time 2X2 ANOVA was conducted on the weapon question detection scores (for SCR, RLL and the combined measure scores). Participants of the "informed innocents" condition were asked about the weapon during the GKT but they did not see the weapon during the mock crime and were not informed about it. Thus, they served as an innocent control group in the data analysis of the weapon question.

Results

SCR

The mixed Anova on the SCR detection scores yielded a statistically significant effect only for the item-type factor \[F(1,76)=10.6, \ p<.01, \ \eta_p^2=0.12\], indicating that central items elicited larger response (M=0.36, SD=0.71) than peripheral items (M=0.03, SD=0.3). Neither the interactions of item-type with the other factors nor the between-subject factors produced statistically significant effects. The 2 by 2 ANOVA conducted on the SCR detection score computed across all items without the weapon question yielded a statistically significant effect for the time factor \[F(1,76)=6.05, \ p<.05, \ \eta_p^2=0.074\], indicating that in the immediate condition subjects had higher responses (M=0.17, SD=0.22) than in the delayed condition (M=0.05, SD=0.20). In addition this analysis yielded statistically significant main effect for group \[F(1,76)=5.74, \ p<.05, \ \eta_p^2=0.07\], indicating that participants in the "informed innocents" group (M=0.16, SD=0.23) had higher responses than participants
in the "guilty" group (M=0.05, SD=0.2). This difference between the "informed innocents" group and the "guilty" group is opposite to the hypothesis that expected that responses of "guilty" participants would be higher than response of "informed innocents". This surprising result will be discussed later in the discussion section. There was no statistically significant effect for the Group by Time interaction.

The 2X2 Anova conducted on the weapon question detection score yielded a statistically significant effect for the group factor \( [F(1,74)=13.09, p<.001, \eta^2_p=0.15] \), indicating that the responses in the "guilty" group were higher (M=1.00, SD=1.5) than in the "informed innocent" group (M=-0.02, SD=0.96). There were no significant effects for the time factor or for the Group by Time interaction. The mean SCR scores computed across participants within each condition are presented in Table 2.

**Table 2. Means and SDs of the SCR detection scores within each experimental group.**

<table>
<thead>
<tr>
<th></th>
<th>SCR</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>central</td>
<td>Peripheral</td>
<td>weapon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
</tr>
<tr>
<td>guilty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.11</td>
<td>0.20</td>
<td>0.31</td>
<td>0.64</td>
<td>0.04</td>
<td>0.31</td>
<td>0.99</td>
<td>1.66</td>
</tr>
<tr>
<td>delayed</td>
<td>0</td>
<td>0.18</td>
<td>0.4</td>
<td>0.73</td>
<td>-0.13</td>
<td>0.25</td>
<td>1.02</td>
<td>1.34</td>
</tr>
<tr>
<td>informed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.22</td>
<td>0.23</td>
<td>0.49</td>
<td>0.86</td>
<td>0.15</td>
<td>0.28</td>
<td>0.19</td>
<td>1.22</td>
</tr>
<tr>
<td>delayed</td>
<td>0.10</td>
<td>0.21</td>
<td>0.22</td>
<td>0.58</td>
<td>0.07</td>
<td>0.32</td>
<td>-0.25</td>
<td>0.55</td>
</tr>
</tbody>
</table>

**RLL**

The mixed Anova on the RLL detection scores did not yield statistically significant effects. Anova that was conducted on RLL detection scores for all items without the question about weapon nor Anova that was conducted on
the score of the reaction to the weapon question did not yield any significant results. RLL scores computed across participants within each condition are presented in Table 3.

**Table 3. Means and SD scores of RLL measurement within each experimental group.**

<table>
<thead>
<tr>
<th></th>
<th>RLL</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>central</td>
<td>peripheral</td>
<td>weapon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
</tr>
<tr>
<td>guilty</td>
<td>immediate</td>
<td>0.14</td>
<td>0.26</td>
<td>0.07</td>
<td>0.41</td>
<td>0.16</td>
<td>0.33</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>delayed</td>
<td>0.12</td>
<td>0.26</td>
<td>-0.09</td>
<td>0.65</td>
<td>0.2</td>
<td>0.3</td>
<td>0.41</td>
</tr>
<tr>
<td>informed</td>
<td>immediate</td>
<td>0.11</td>
<td>0.17</td>
<td>-0.01</td>
<td>0.54</td>
<td>0.15</td>
<td>0.25</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>delayed</td>
<td>-0.01</td>
<td>0.24</td>
<td>-0.06</td>
<td>0.55</td>
<td>0</td>
<td>0.28</td>
<td>0.46</td>
</tr>
</tbody>
</table>

**Combined measure**

The mixed Anova on the combined measure detection scores yielded a statistically significant effect for the item-type factor \[F(1,73)=3.09, p<.05\] (one tailed), \[\eta^2_p=0.041\], indicating that subjects responses for central items (M=0.33, SD=0.86) were higher than their response for peripheral items (M=0.12, SD=0.36). There were no statistically significant effects for the Group X Item type interaction, of Time X Item type interaction, or of the triple interaction. In addition, this analysis yielded a statistically significant main effect of the time factor \[F(1,73)=4.37, p<.05, \eta^2_p=0.056\], indicating that in the immediate condition subjects had higher reactions (M=0.32, SD=0.06) than in the delayed condition (M=0.12, SD=0.06). There were no effects for the group factor or for the interaction. 2 by 2 ANOVA conducted on the combined detection scores for all items without the question about the weapon yielded statistically significant main effect for time \[F(1,73)=7.71, p<.01, \eta^2_p=0.09\],
indicating that participants in the immediate condition (M=0.26, SD=0.27) had higher reactions than participants in the delayed condition (M=0.08, SD=0.29). There were no statistically significant effects for group factor or for Time X Group interaction.

The Anova conducted on the weapon question detection scores yielded a statistically significant effect for the group factor \([F(1,71)=7.13, p<.01, \eta^2_p=0.09]\), indicating that the responses in the "guilty" group were higher (M=1.48, SD=1.9) than in the "informed innocent" group (M=0.5, SD=1.17). There were no effects for the time factor or for the interaction. Combined scores computed across participants within each condition are presented in Table 4.

**Table 4. Means and SD scores of combined measurement within each experimental group.**

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>all</th>
<th>SD</th>
<th>central</th>
<th>all</th>
<th>SD</th>
<th>peripheral</th>
<th>all</th>
<th>SD</th>
<th>weapon</th>
<th>all</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
</tr>
<tr>
<td>guilty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.19</td>
<td>0.28</td>
<td>0.36</td>
<td>0.85</td>
<td>0.13</td>
<td>0.38</td>
<td>1.48</td>
<td>2.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delayed</td>
<td>0.08</td>
<td>0.33</td>
<td>0.23</td>
<td>1.05</td>
<td>0.04</td>
<td>0.34</td>
<td>1.47</td>
<td>1.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>informed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.32</td>
<td>0.25</td>
<td>0.53</td>
<td>0.72</td>
<td>0.26</td>
<td>0.32</td>
<td>0.73</td>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delayed</td>
<td>0.08</td>
<td>0.24</td>
<td>0.19</td>
<td>0.8</td>
<td>0.04</td>
<td>0.36</td>
<td>0.24</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SVT**

A 2X2 Anova conducted on the SVT detection scores computed across all items with group and time of testing as between-subject factors yielded a statistically significant main effect for time \([F(1,85)=4.85, p<.05, \eta^2_p=0.05]\), indicating that SVT detection scores for "all" items were higher in the immediate condition (M=0.15, SD=0.11) than in the delayed condition.
(M=0.10, SD=0.08). The detection scores on SVT are higher as the participants' proportion of correct answers departs from chance level.

**Table 5 Means and SD scores of SVT detection score within each experimental group.**

<table>
<thead>
<tr>
<th></th>
<th>SVT Mean</th>
<th>SVT SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>guilty</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>delayed</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>informed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>delayed</td>
<td>0.12</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**Memory test**

A mixed 2X2 X2 Anova, conducted on the recognition accuracy rates, with central and peripheral items serving as a within-subjects factor and group and time as the 2 between-subjects factors yielded the following outcomes: A statistically significant effect for the item-type factor \[F(1,85)=24.92, p<.001, \eta_p^2=0.22\], indicating that memory accuracy for the central items (M=0.63, SD=0.37) was better than for the peripheral items (M=0.44, SD=0.15). Additionally, a statistically significant Group X Item-type interaction was found \[F(1,85)=10.12, p<.005, \eta_p^2=0.10\], indicating that participants memory for central items in the "guilty" group was better (M=0.76, SD=0.28) than in the "informed innocents" group (M=0.5, SD=0.4) and that participants' memory for peripheral items was similar in the "guilty" group (M=0.44, SD=0.16) and in the "informed innocents" group (M=0.43, SD=0.14). The interaction of item-type with time of testing as well as the triple interaction produced very small and non-significant effects. In addition, this analysis revealed a statistically significant effect for the group factor \[F(1,85)=9.63, p<.005, \eta_p^2=0.10\],
indicating that memory of participants in the "guilty" group was better (M=0.60, SD=0.16) than memory of participants in the "informed innocents" group (M=0.47, SD=0.23). There were no statistically significant effects for the time factor or for the Group X Time interaction.

The 2X2 ANOVA, with group and time serving as the 2 between-subjects factors conducted on the memory accuracy rates across all items yielded statistically significant results for time factor \[F(1,85)=4.44, p<.05, \eta^2_p=0.05\], indicating that participants in the immediate condition (M=0.5, SD=0.16) remembered more items than subjects in the delayed condition (M=0.43, SD=0.12). There were no statistically significant effects for the group factor or for the interaction.

**Table 6.** Means and SD scores of memory accuracy rates within each experimental group.

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>central</th>
<th>peripheral</th>
<th>weapon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
</tr>
<tr>
<td>guilty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.52</td>
<td>0.16</td>
<td>0.74</td>
<td>0.25</td>
</tr>
<tr>
<td>delayed</td>
<td>0.45</td>
<td>0.12</td>
<td>0.78</td>
<td>0.31</td>
</tr>
<tr>
<td>informed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.47</td>
<td>0.17</td>
<td>0.51</td>
<td>0.42</td>
</tr>
<tr>
<td>delayed</td>
<td>0.41</td>
<td>0.11</td>
<td>0.47</td>
<td>0.4</td>
</tr>
</tbody>
</table>

**ROC**

Detection efficiency is defined in terms of the relationship between the detection measure and the actual guilt (or knowledge of the relevant items). The ROC curve reflects the degree of distinction between the distributions of the detection score of "guilty" and "innocent" participants. This approach was adapted from Signal Detection Theory – SDT (e.g., Green & Swets, 1966;
Swets, Tanner, & Birdsall, 1961) because it is particularly useful for analyzing psychophysiological as well as behavioral detection data, and it has been applied extensively in this area (e.g., Ben-Shakhar & Elaad, 2003; Nahari & Ben-Shakhar, 2011)

In the present experiment, there were two groups of knowledgeable participants, and the ROC for each of these groups was constructed by comparing the detection score distribution of the knowledgeable participants (either "guilty" or "informed innocents") with a simulated distribution of the detection score among "uninformed innocents" (see, e.g., Meijer et al., 2007). This procedure was applied for all detection measures (SCR, RLL, combined measure and SVT).

For the SCR, RLL and combined measure it entailed the following steps. First, 70 values were randomly drawn from a standard normal distribution (mean=0, standard deviation =1). The entire sample was divided into two blocks. Then each value was standardized relative to mean and standard deviation of its block. All the 70 values were divided to 14 questions, (5 responses for each question). Then, one value (response to the critical item) was randomly picked from all five responses. This way, a standardized score for one innocent person for one question was derived. This process was repeated 13 times (to simulate 13 questions), 13 values were averaged to represent a detection score computed across all items (without the question about the weapon) for one innocent participant. The detection score for the central items was constructed by averaging "critical" values of 3 questions. The detection score for the peripheral items was constructed by averaging "critical" values of 10 remain questions.
The following steps were made for the SVT: First, for each participant values were randomly drawn from the binomial distribution with \( N=20 \) and \( p=0.5 \). Then proportion of correct answers was computed for all 20 items. SVT detection scores were computed as it was described earlier. These detection scores indicate the deviation of subject scores from the chance level. In order to verify that ROCs are not affected by single random draw of simulated innocents, the process of cresting simulated innocents and calculating ROC was repeated 10,000 times. In each step, unknowledgeable distribution was constructed as it described earlier and calculated the respective ROC curve and it's AUC within each experimental condition. Repeating this procedure 10,000 times provided a sampling distribution of the AUC. The mean of this distribution is presented in tables 7 and 8. Next, was calculated minimal AUC value (the lower bound 95% confidence limit), such that the probability that the true AUC value exceeded it is 0.95 (i.e., the value \( C \) for which \( P[\text{population AUC} > C] = 0.95 \)). This minimal value extracted from the AUC sampling distribution and used to test whether the AUC was statistically significant. In addition, ROC curves based on combination of the SCR and SVT and combination of SCR, RLL and SVT were constructed in the same way. In order to combine the SVT scores with physiological measures within-subject Z-scores were constructed for each participant by Z-test for proportion, for the knowledgeable participants and for simulated innocents. The measures were combined by using simple averages of the standardized detection measures.

Table 7 displays the areas under the ROC curves of the SCR and RLL measures as a function of item-types and experimental conditions. Table 8 displays the areas under the ROC curves of the combined physiological
measure, the SVT scores, the combination of SCR and SVT and combination of SCR, RLL and SVT as a function of all items and experimental conditions.

**Table 7.** Areas under ROC curves and corresponding 95% confidence intervals of the SCR and RLL as a function of item-types and experimental conditions.

<table>
<thead>
<tr>
<th></th>
<th>SCR</th>
<th></th>
<th>RLL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>central</td>
<td>peripheral</td>
<td>all</td>
</tr>
<tr>
<td>guilty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.62* (0.51-0.74)</td>
<td>0.62* (0.51-0.72)</td>
<td>0.53 (0.42-0.64)</td>
<td>0.64* (0.54-0.75)</td>
</tr>
<tr>
<td>delayed</td>
<td>0.51 (0.38-0.63)</td>
<td>0.65* (0.56-0.74)</td>
<td>0.64* (0.53-0.75)</td>
<td>0.63* (0.54-0.72)</td>
</tr>
<tr>
<td>informed</td>
<td>0.74* (0.65-0.83)</td>
<td>0.65* (0.57-0.73)</td>
<td>0.64* (0.55-0.74)</td>
<td>0.64* (0.52-0.76)</td>
</tr>
<tr>
<td>immediate</td>
<td>0.61 (0.49-0.67)</td>
<td>0.59 (0.48-0.69)</td>
<td>0.52 (0.40-0.64)</td>
<td>0.5 (0.39-0.61)</td>
</tr>
<tr>
<td>delayed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* ROC area statistically significant from chance level, p<.05.

An inspection of table 7 reveals that in the "informed innocents" group detection efficiency decrease in the delayed condition (on SCR measure from 0.74 to 0.61 for "all" items [t=1.43, p>.05], from 0.65 to 0.59 for central items [t=0.78, p>.05], from 0.64 to 0.52 for peripheral items [t=1.37, p>.05]; on RLL measure from 0.64 to 0.5 for "all" items [t=1.52, p>.05], from 0.65 to 0.51 for peripheral items [t=1.54, p>.05). Although the decrease was not statistically significant, in the delayed condition the area under ROC did not differ from the chance level in the SCR and in the RLL measures. This pattern is absent in the "guilty" group, where a decrease was noticed only for "all" items analysis on the SCR measure (from 0.62 in immediate condition to 0.51 in delayed condition). In addition, an inspection of table 7 reveals that there is ability to differentiate between the guilty subjects and innocents by central items with SCR measurement but this ability is absent with the RLL measurement.
Table 8. Areas under ROC curves and corresponding 95% confidence intervals of the combined measure, SVT, combination of SCR and SVT and combination of SCR, RLL and SVT as a function of experimental conditions.

<table>
<thead>
<tr>
<th></th>
<th>combined SVT</th>
<th>SVT+SCR</th>
<th>SVT+SCR+RLL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
<td>central</td>
<td>peripheral</td>
</tr>
<tr>
<td>guilty</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.69*</td>
<td>0.62*</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>(0.58-0.79)</td>
<td>(0.52-0.70)</td>
<td>(0.49-0.69)</td>
</tr>
<tr>
<td>delayed</td>
<td>0.61*</td>
<td>0.53</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>(0.51-0.70)</td>
<td>(0.44-0.62)</td>
<td>(0.46-0.66)</td>
</tr>
<tr>
<td>informed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate</td>
<td>0.82*</td>
<td>0.72*</td>
<td>0.73*</td>
</tr>
<tr>
<td></td>
<td>(0.74-0.88)</td>
<td>(0.64-0.80)</td>
<td>(0.65-0.81)</td>
</tr>
<tr>
<td>delayed</td>
<td>0.59</td>
<td>0.56</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>(0.49-0.69)</td>
<td>(0.48-0.65)</td>
<td>(0.40-0.62)</td>
</tr>
</tbody>
</table>

* - ROC area statistically significant from chance level, p<.05.

The inspection of table 8 reveals that the ability to differentiate between the "informed innocents" and "uninformed innocents" on combined measure decrease in the delayed condition as it was on SCR and RLL measures. For the "guilty" group there was decrease in the ROC area for all categories in the delayed examination. For central and peripheral items the ability to distinguish the "guilty" participants from "innocents" disappeared in this condition whereas the ability remained for all items.

Results of ROC for SVT reveal that informed participants can be differentiated from "uninformed" in the immediate and in the delayed testing while guilty participants can be differentiate from "uninformed innocents" only in the immediate condition.

The ability to differentiate "guilty" from "uninformed innocents" seems to be better by combined physiological measure than by SVT, but this difference was not statistically significant. The better ability to differentiate
"informed innocents" from "uninformed innocents" in the immediate condition by combined measure compared to SVT was statistically significant, p<0.05. The differentiation between the "informed innocents" and "uninformed innocents" in the delayed condition seems to be better by SVT compared to combined physiological measure but this difference was not statistically significant.

The combination of SCR and SVT measures seems to reveal better detection efficiency compared with combined physiological measure or compared with SVT only for "guilty" group in delayed condition, these differences were not statistically significant. In contrast to the hypothesis, the combination of all three measures did not reveal maximal detection efficiency.

To summarize the results of ROC, the participants in the "informed innocents" group cannot be differentiated by physiological measures from the "uninformed innocents" in the delayed condition. SVT allows differentiation between the "informed innocents" and "uninformed innocents" in the immediate and delayed conditions. It allows differentiation between the "guilty" participants and "uninformed innocents" only in the immediate condition. The combined physiological measure seems to be more effective in differentiation of "guilty" participants than "uninformed innocents" compared to the behavior measure. Opposite to the hypothesis combination of all physiological and behavior measures did not reveal better detection efficiency.

**Discussion**

This study used "realistic" mock-crime paradigm to examine the differences between the "guilty" and "informed innocent" participants. The
realistic conditions were produced by virtual reality environment. "Guilty" participants got scenario of being CSI secret agents and their instructions were to kill the character that will enter the room. The "informed innocents" got scenario of being company's purchasing managers and their instructions were to pass an envelope containing money to character that will enter the room. Participants of both groups were in the same apartment and saw the same character during the virtual reality environment. These participants were either tested immediately afterwards or with a delay of one week.

Consistent with the hypothesis, the results of this experiment indicated that participants in the "guilty" group remembered more items than participants in the "informed innocents" group. The results also revealed a memory decrease in the delayed condition. Nevertheless, memory of central crime details remained stable in the group of guilty participants, while it declined in the "informed innocents" group. These results are consistent with the findings of Gamer et al. (2010), although the two studies used different memory encoding techniques for the "informed innocents" group. In this study the "informed innocents" were exposed to the crime items by the same way as "guilty" participants, while in Gamer et al (2010) "informed innocents" only read about the crime in a newspaper. Gamer et. al (2010) suggested that the differences in memory performance may be caused by the different encoding methods. The results of this study indicate that other factor may account for these differences.

Although this study succeeded to show the differences in memory between the two experimental groups, it succeeded only partly in inducing emotional arousal. The physiological measurements did not show the
expected differences between the groups. The difference was reflected only by stress factor of the participant's self report questionnaire. Peth et al (2011) measured the physiological response during the mock crime and around 60 seconds after. They found a stronger increase in the heart rate in the arousal induction group especially in the phase after committing the mock crime. In this experiment, no physiological measures were taken after committing the crime (the measurement ended at the moment the participant shot and the character fell down). It is possible that measuring the physiological reaction after committing the crime would reveal the expected effect. Another support for this suggestion is the fact that the emotional questionnaire, which was filled several minutes after committing the crime, revealed statistically significant results which indicated that the participants of the "guilty" group were under more stress during the VR stage. The Identification factor of this questionnaire showed that participants in the "guilty" group scored lower than participants in the "innocent" group. Another possible explanation to the fact that "guilty" scenario did not elicit the expected emotional arousal is that the scenario was not realistic enough.

Even though the manipulation did not achieve the desired result, the procedure used in this experiment resembles realistic conditions in the "guilty" group and even more in the "informed innocents" group because the relevant items were not clearly specified before the mock crime and the "informed innocents" were present in the crime scene, rather than being exposed to the crime details by reading about the crime or watching it on a video (Gamer, 2010; Gamer et al, 2010; Nahari & Ben-Shakhar, 2011; Peth et al, 2011). In this study participants in the "informed innocents" group were exactly in the
same apartment as the "guilty" participants and they obtained their knowledge about the items in the same way as "guilty" participants. These realistic features allowed us to examine whether the CIT can be used to distinguish between the two groups.

Participant's scores on SCR revealed a difference between the two groups for all items, excluding the question about the weapon. Surprisingly, not in line with former CIT studies (Gamer 2010; Nahari & Ben-Shakhar, 2011), the "informed innocents" group showed significantly larger differential responses to the critical items than the "guilty" group. Inspection of Table 2 revealed that the relative responses of the "informed" participants were larger than those of the "guilty" participants, when computed across all items and across the peripheral items, in both the immediate and the delayed conditions. In addition, the reactions of the "informed innocent" participants were also larger than the responses of the "guilty" participants, when computed across the central items, in the immediate condition (the groups did not differ significantly in their responses to peripheral and central items). It might be that the differences between these results and those reported in previous studies are due to the differences in memory encoding procedure of the "informed innocents" group. However, these results opposite to the results of the memory test. Participants in the "informed innocents" group remembered fewer items than participants in the "guilty" group but their reactions to the relevant items on SCR measure were higher than reactions of participants in the "guilty" group. It might be that participants in the "informed innocents" group were less stressed and paid more attention to the items in the apartment, these items had no emotional context for them and they did not
intend to remember them. As a result, they had implicit memory for these items that was expressed by physiological index and not by behavior direct test. Other factors that might affect these results are the number of peripheral items compared to number of central items and the difference of presentation of them. The peripheral items were presented to participants during the free time in the apartment (around 2 minutes) while the central items were presented for short time only when the character entered the apartment (besides the name of the character that was mentioned in the instructions). Taken together, it can be that participants in the "guilty" group were focused on the character entrance and paid less attention for the peripheral items in the apartment, the large number of them brought to bias toward peripheral items in the SCR measure. However, the two groups did not differ in the memory accuracy for the peripheral items. Nevertheless, further research is needed in which the exposure of "guilty" and "informed innocents" groups for details of crime will be the same, the number of peripheral and central items will be equal and their presentation way and time will be similar.

Respiration is another physiological measure used in mock-crime studies. In contrast to the studies reported by Gamer et.al (2010) and by Peth et.al (2011) no significant differences between respiratory responses at the immediate and the delayed measurement occasions were found in this study. Moreover, unlike the findings reported by Peth et. al (2011), no differences between central and peripheral items were found. As it was mentioned earlier, it can be that the larger number of peripheral compared to central items caused to such difference between this study and previous studies (Gamer
et.al, 2010; Peth et. al, 2011). In addition, it might be that the RLL measure in this study was not sensitive enough.

All detection measures used in this experiment, except RLL, reflect, as expected, an effect of time, indicating a decrease in the delayed condition. These results are consistent with the temporal decline in various physiological and behavior measures reported by Nahari & Ben-Shakhar (2011) and by Gamer et al. (2010).

The differentiation between central and peripheral items is another important aspect that was examined in the present study. Consistent with former studies (Carmel et. al, 2003; Nahari & Ben-Shakhar, 2011), central items produced more efficient SCR and combined measure detection efficiency and participants remembered them better than the peripheral items. In addition, this study revealed that participants in the "guilty" group remembered better central items than participants in the "informed innocents" group, while no differences were found between the groups in peripheral items.

It is also interesting to examine the differentiation between the "guilty" and the "innocent" participants based on their responses to the question about the weapon used in the crime. Kramer, Buckhout & Eugenio (1990) found that weapon is a salient object that demands a certain amount of attention from observers and that eyewitnesses seem to remember well the weapon itself but not the weapon carrier. In this study both groups were asked about the weapon that was used, but only the "guilty" group was informed about the weapon; hence the "informed innocents" group can serve as an uninformed control group for this question. Analysis of this question revealed
a differentiation between the informed and uninformed participants with the SCR and the combined measures but not with the RLL. In addition, 95% of informed participants remembered the weapon used for murder in both the immediate and in the delayed conditions. It seems that this item has been memorable and so can be used as an efficient index for distinguishing between guilty and innocent suspects. However, suspects cannot be classified according to this single question, it is necessary to combine it with other central questions. Yet, further research is required to examine whether different types of weapon are similarly effective.

This study examined also the efficiency of detecting deception by SVT. The results revealed that in the immediate condition this test can be used to detect deception among "guilty" and "informed innocents" participants but in the delayed condition the efficiency decreased. The ability to differentiate the "guilty" group from "uninformed innocents" by "all" items was not larger than chance level in the delayed condition. These results are in line with previous study reported by Nahari & Ben-Shakhar (2011).

The ability to differentiate between "informed innocent" participants and "uninformed innocent" participants (simulated "uninformed innocents") by ROC was examined in this study, too. The results revealed that the ability to differentiate between "informed innocents" and "uninformed innocents" was present in the immediate condition for all measures while in the delayed condition the ability decreased for all measures and was not significantly larger than chance level for almost all of them. This decrease in the ability to differentiate between "informed innocents and "uninformed innocents" is in line with previous studies (Gamer et. al, 2010; Nahari & Ben-Shakhar, 2011).
It is interesting to note that the participants in "informed innocents" group were exposed to the crime details in this study differently than in the previous studies. In addition, the combination of behavior and physiological measurements, in this study and in study reported by Nahari & Ben-Shakhar (2011), reduced the possibility for false positive results among "informed innocents" participants in the delayed condition. This study examined not only the possibility of information leakage but also the possibility that suspects had a prior knowledge about the place where the crime happened or about the details of the crime. The practical implication of these results is that a great caution must be exercised when there is a suspicion that suspects may have prior knowledge of the crime scene. It was found that when the test was delayed, there was a substantially lower risk for false positive results when a combination of physiological and behavior measurements were used. Further research is needed to examine the memory of informed suspects when they are exposed to the information not only once but several times. Additional research can be conducted to examine physiological responses of eyewitnesses to crime details.

The ability to distinguish "guilty" participants from innocents is a core issue of CIT research. Previous studies found that the discrimination of these two groups remains good even when the test was delayed for two weeks (Gamer et.al 2010, Peth et. al, 2011). In addition, it was found that a combination of behavior and physiological measures increase the ability to discriminate the two groups (Nahari & Ben-Shakhar, 2011). Also, some previous studies found a reduced detection efficiency of "guilty" participants in the delayed testing (Carmel et. al 2003, Nahari & Ben-Shakhar, 2011). This
study found that immediately after committing the mock-crime, the ability to
distinguish between the "guilty" and "innocents" is present for "all" items in all
measures that were used. In the delayed testing the results were not
homogeneous and detection efficiency in some measures was reduced such
that it was impossible to distinguish between the groups. On the other hand,
detection efficiency increased with other measures. In this study combination
of the SVT with SCR measure revealed better ability to differentiate "guilty"
participants from "uninformed innocents" only in delayed condition. This result
is in line with results reported by Meijer et. al (2007) in the experiment used
mock-crime paradigm. Nahari & Ben-Shakhar (2011) found that combination
of physiological and behavior measures increased the detection efficiency in
all conditions. Moreover, the combination of all measures used in this study
did not increase the detection efficiency, as it was expected. Further research
is needed in order to understand the factors affecting detection efficiency of
guilty participants.

It is important to note that in this study all ROC areas were lower than
those that were reported in previous studies (Carmel et. al, 2003; Gamer et.
al, 2010; Nahari & Ben-Shakhar, 2011; Peth et. al, 2011). It might be that the
participants do not perceive the virtual reality environment as something
realistic. It is possible that they feel like playing in computer game. It might be
that realistic mock-crime procedures where participants actually searching for
the items that they need to steal and they actually hold these items in their
hands feel more realistic than murdering somebody in the virtual reality
environment. Further research is needed to examine whereas there are
differences in participants reactions in virtual reality mock-crime and realistic mock-crime.

Conclusions

This study examined the effect of emotional arousal on the ability to differentiate between "guilty" and "informed innocents" participants in virtual reality mock crime. The results revealed that it was difficult to discriminate between "guilty" participants and "informed innocents" in the immediate condition. In the delayed testing, there was a decrease of the physiological reactions and the memory accuracy in both groups. The ability to differentiate the "guilty" participants from unknowledgeable remained almost stable by several measures, whereas the ability to differentiate the "informed innocents" participants from unknowledgeable disappeared in the delayed condition. That means that in the delayed testing the possibility of false positive results among "informed innocents" decreases. In addition, the study found effect for central items in both groups, particularly that "guilty" group remembered them better than "informed innocents" group. It is important to note that in this study emotional arousal was observed only by behavioral measures. Moreover, it might be that virtual reality scenarios were not realistic enough. Further research is needed to examine the effect of emotional arousal in mock-crime paradigm with more realistic scenarios.


Appendix A: Consent and health forms

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Do you agree to have your medical condition treated? (in Hebrew)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Do you agree to participate in medical research?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Do you agree to undergo medical procedures?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Do you agree to receive medical treatment?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Do you agree to have your personal information shared with medical professionals?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Do you agree to participate in psychological research?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If yes, please indicate your consent:

- [ ] Yes
- [ ] No

If no, please indicate the reason:

- [ ] I do not want to participate
- [ ] I do not understand the consent form
- [ ] I do not have time to participate
- [ ] I do not agree to the terms and conditions

Signed:

__________________________

Date: ________________
## Appendix B: Emotional questionnaire

<table>
<thead>
<tr>
<th>Question</th>
<th>Scale 1</th>
<th>Scale 2</th>
<th>Scale 3</th>
<th>Scale 4</th>
<th>Scale 5</th>
<th>Scale 6</th>
<th>Scale 7</th>
<th>Scale 8</th>
<th>Scale 9</th>
<th>Scale 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.quares מידה לשנת בפומת בולך שנזרך במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.quares מידה ההודדות לשנת הזרות שלפ במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.quares התרות בפומת היאצאת במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.quares התרות בפומת י כתובת במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.quares מידה לשנת למולצת תקלה בפומת האעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.quares התרות בפומת הערוך לשנת פומת הערוכת המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.quares מידה לשנת בפומת מثقة במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.quares התרות בפומת יписыва תבשיר במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.quares מידה לשנת התרות בפומת מثقة במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.quares התרות בפומת הערוך או לשנת פומת הערוכת בפומת מثقة במעעיאות המודמח?</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C: GKT questions. Items of marked questions were classified as central.

1. Ми модифіковане бачато в.Merge?
4. Які модифікатори між в.Floor в.Floor.
5. Які модифікатори між в.Floor в.Floor.
6. Які модифікатори між в.Floor в.Floor.
7. Які модифікатори між в.Floor в.Floor.
8. Які модифікатори між в.Floor в.Floor.
9. Які модифікатори між в.Floor в.Floor.
10. Які модифікатори між в.Floor в.Floor.
11. Які модифікатори між в.Floor в.Floor.
12. Які модифікатори між в.Floor в.Floor.
13. Які модифікатори між в.Floor в.Floor.
14. Які модифікатори між в.Floor в.Floor.
Appendix D: SVT questions.

1. מ Mahmoud מהו הבחינת החוזה? justify
2. מה הביא להכרתך? justify
3. איך ספר היה מענה על התרשים הבסיס? justify
4. איך שמעון היה תליי בסולר? justify
5. איך הסיא עוד בפניך חירב? justify
6. איך כל גנייה היה בירור? justify
7. איך תלמידי היה על השרשרת בסולר? justify
8. איך תמנה היחידה התליאו ביד? justify
9. איך דגש של פעולות היה מענה על המודף בסולר? justify
10. איך ארסלי היה על השרשרת במדבב? justify
11. איך הכרות היה מענה על הטרפה בסולר? justify
12. מה שיש להכרתך? justify
13. מה זה זכויות הילולות ביד? justify
14. האם היה מדרגות ביד? justify
15. האם היה כל בירור? justify
16. האם היה מאורר תקווה ביד? justify
17. האם הוכתב דופן בולת לפנים שלגנום? justify
18. האם היה עץ בירור? justify
19. האם הוכתב בוחיק ביד מנווג? justify
20. האם היה מחשב ניד לע השרשרת בסולר? justify
21. איך כל شيء לורה?