한국심리학회지 : 법
The Korean Journal of Forensic Psychology
2019, Vol. 10, No. 2, 89-109

Can virtual mock crime replace actual mock crime? An event-related potential study*

Inuk Song Hyemin Kim Kyoung Eun Lee Eunhee Chang Hyun Taek Kim†

Department of Psychology, Korea University

For detecting deception research, the Concealed Information Test (CIT) is the most widely used method in conjunction with electroencephalography (EEG). Moreover, mock crime scenarios were commonly adopted for providing materials for lying. Mock crime scenarios have relatively higher ecological validity than other paradigms like autobiographical information or card test. However, current mock crime scenarios also have some limitations because of ethical issues, resource issues, and experimental controllability. Virtual reality (VR) is a potential alternative to overcome the disadvantages. Nonetheless, few studies used VR for mock crime, and there is no research on the comparison between 'actual' mock crime and 'virtual' mock crime. In the present study, we developed a high-fidelity virtual environment and used it for the virtual mock crime. Participants were randomly assigned both for the Crime status (innocent or guilty) and the Environment mode (actual or VR). After the scenarios, participants were tested by P300-based CIT with EEG recording. To verify the effects of virtual mock crime on subsequent EEG data during CIT, we focused on the P300 event-related component (ERP) and individual classification using the bootstrapping method in the study. As we hypothesized, the main effect of environment mode was not significantly different, and the interaction between stimuli type (target, probe, irrelevant) and environment mode was also not significant when we exclude one outlier. Furthermore, the accuracy of individual classification was equivalent between the actual and the VR. These results were also supported by ROC analysis and equivalence test. All statistical results suggest that there is no significant difference between actual mock crime and virtual mock crime. In conclusion, the study suggests that the virtual mock crime is a potential alternative method for mock crime scenarios.

Key words : Concealed Information Test, Deception, Virtual Reality, Mock Crime, P300

* 이 논문은 제1저자 송인욱의 석사학위 논문 일부를 발췌 수정한 것임.
This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5A2A01009211).

† 교신저자 : 김현택, 고려대학교 심리학과, 서울특별시 성북구 안암동 5가 고려대학교 구법학관 205호
Tel : 02-3290-2065, E-mail : neurolab@korea.ac.kr

For a long time, people tried to find a way to determine the authenticity of a statement. In these days, detecting deception is still a topic of much interest from people, especially in fields of clinical psychology, forensic psychology, and neurolaw. This attractiveness has led researchers to develop many lie detection methods (For more details, see Rosenfeld, 2018). For example, the most widely studied way is known as the Concealed Information Test (CIT) and also known as the Guilty Knowledge Test (GKT). As a variant of the oddball paradigm (Donchin & Coles, 1988), the test usually consists of three classes of stimuli. Probe is crime-related information (e.g., knife) or significantly important information for suspects (e.g., birthday date). Irrelevant stimuli refer to the stimuli related to non-crime-related or less important information. Target stimuli are recruited to maintain a suspect's attention and served as another reference similar to irrelevant stimuli. CIT assumes that probe stimuli will elicit a different pattern when it compared to irrelevant stimuli if suspects possessed concealed information. For innocent people, they can not distinguish probe from irrelevant. However, for guilty people, probe stimuli are much meaningful, while it causes different neural underpinnings.

In these underpinnings, many researchers have focused on an event-related potential (ERP) component, which is called P300. The P300 is a positive potential of the brain that occurs between 300 and 800ms after stimulus presented. This component is acquired within the parietal area (Usually at Pz electrode), and it is used as the most typical neural underpinning to detect deception. The P300 is known to reflect diverse cognitive activities and revealed a relative huge amplitude compared to other ERP components. Specifically, the P300 component is related to unpredictable or unusual stimuli (Soltani & Knight, 2000), frequency of a specific stimulus given stimuli pool (Vogel, Luck, & Shapiro, 1998), and memory process (Wilding, 1999). Because these cognitive processes are associated with deception, the P300-based CIT has been widely studied, its accuracy has been shown 80% to 95% that is relatively high. Due to these characteristics, many P300-based CIT research is conducted in many laboratories.

Moreover, researchers have begun to broaden our perspectives by looking for other ERP components like N400 (Ganis & Schendan, 2013) and late positive component (LPC; Leng, Wang, Cao, & Li, 2017) as well as other neural features (Gao, Yang, Huang, Lin, Ge, Zheng, & Rao, 2016; Wang, Chang, & Zhang, 2016). In addition, researchers have revised task paradigms to resist countermeasures which is one of most prominent obstacles in practical use (Complex Trial Protocol; Rosenfeld, Labkovsky, Winograd, Lui, Vandenboom, & Chedid, 2008), and recruited various methods like electrodermal activity (EDA; Ben-Shakhar & Elaad, 2003), electrocardial activity (ECG; Gamer, Verschuere, Crombez, & Vossel, 2008), and functional magnetic resonance imaging (fMRI; Gamer, Bauermann, Stoeter, & Vossel, 2007; Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003).

However, compared to the efforts to develop methodologies for detecting deception, there was not much attention to the content of lying. In other words, many researchers have used various material to ask participants to conceal. For example, autobiographical information like participants' birthday date was used to the information that should be hidden (Lee, Liu, Tan, Chan, Mahankali, Feng, & Gao, 2002; Sartori, Agosta, Zogmaister, Ferrara, & Castiello, 2008). In another, sometimes card test was adopted (Kugelmass & Lieblich, 1966; Zaitsu, 2016). However, such content of lying could have a lower degree of ecological validity. There may be differences from the content that will encounter in the practical field. Furthermore, according to Rosenfeld's research, self-referring stimuli made a higher accuracy rate than mock crime scenario (Rosenfeld, Biroschak, & Furedy, 2006). It means that using autobiographical information for lying research could have over-valuated accuracy results. Therefore, well-structured mock crime scenario could be the possible best way when considering ecological validity. For example, Ben-Shakhar & Elaad conducted a meta-analysis on CIT using skin conductance response, they concluded that mock crime studies had highest effect size overall although there was variation by details of the mock crime (Ben-Shakhar & Elaad, 2003). Also, a mock crime scenario can contain some details which can improve emotional arousal. Peth and colleagues showed that sudden appearance of another person to the scene of theft mock crime increased emotional arousal, it leads to sustained CIT accuracy when the test was delayed (Peth, Vossel, & Gamer, 2012).

Despite these advantages, however, it is not without the limitations of the mock crime currently used. First, because of ethical issues, types of mock crime is limited. Therefore, many mock crime scenarios adopted in laboratory studies have used 'theft' mock crime scenarios. Second, there are problems with available spaces and expenses. Third, it is difficult to control external variables during a mock crime.

The new method of mock crime with virtual environment may be one of the possible solutions to overcome these shortcomings. 'Virtual mock crime' can provide various type of scenario that was impossible because of ethical issues. For instance, scenarios that assault others have not been available, but it seems that the scenario will be possible because virtual reality can provide participants a similar but less severe emotional experiences. Using virtual reality is also unconstrained by the problems of place and cost, and can effectively control external variables. However, to our knowledge, only two studies used virtual reality for the mock crime. Mertens & Allen used Quake 3 game engine to develop a high-quality virtual environment (VE).

Guilty subjects were asked to enter an office where usually off-limited, log-on computer, and retrieve some items in the VE. Innocent subjects were asked to enter the same VE, but they navigated it only (Mertens & Allen, 2008). However, the research was focused on the effects of countermeasures rather than focused on virtual mock crime. Another virtual mock crime was adopted by Hahm and colleagues. They used the virtual library where guilty participants were asked to conceal the roll of bill to the specific item (e.g., bag). Innocent subject did not experience the VE (Hahm, Ji, Jeong, Oh, Kim, Sim, & Lee, 2009). Therefore, the effects of virtual mock crime on subsequent results of the test are still unknown. Because the VE has a lower presence than the real environment and the physical stimuli given to subjects are different, it is necessary to verify the effect of these factors on the test results.

Consequently, we focused on the subsequent effects of virtual mock crime in the study. For this purpose, we used theft mock crime scenario and P300 metric, which are the most widely used in detecting deception field. We hypothesized that virtual mock crime would produce similar results of actual mock crime in specificity (Innocent participants were classified as innocent correctly) because it is the same in the virtual reality that innocent subjects do not see the crime related objects. Also, we assumed that the sensitivity (Guilty participants were classified as guilty correctly) of the EEG results would be similar or lower because VE would provide a lower presence level and this may reduce attention. However, we also thought that an equal level of sensitivity would possible because the VE we created had a sufficient level of fidelity. To confirm our hypothesis and find out the usefulness of virtual mock crime, we verified the EEG results in various aspects including P300 amplitude, individual classification using a bootstrapping method, and receiver operating characteristic (ROC) curve analysis. Furthermore, the equivalence test was used to confirm whether there is no statistical differences of EEG results between virtual mock crime and actual mock crime.

Method

Virtual Environment for Mock Crime

To compare actual mock crime with virtual mock crime, a virtual environment (VE) was developed as a mirror of the real environment. Using Matterport pro 3d camera (Sunnyvale, Calif., USA), the place of actual mock crime was captured and the data were imported into Unity 3D (Unity 5.3.0f4) game engine. After refining 3d mesh data, the first-person controller was inserted. Therefore, participants can navigate the VE in the first-person perspective of their avatar. Background sound was recorded in the real environment and inserted in the VE. To

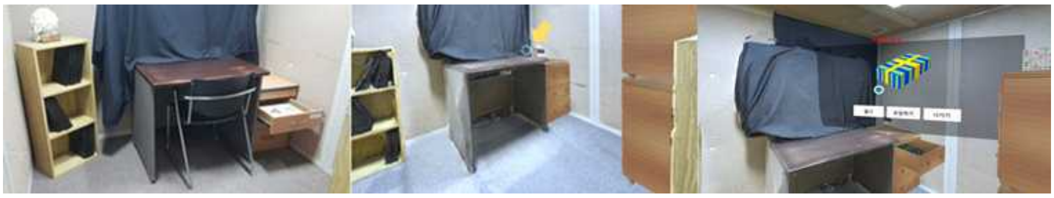

Fig 1. Scenes of the mock crimes. The real environment (left) and the virtual environment (middle, right). Blue circle appear when participants gaze interactive objects (yellow arrow in middle). A panel window was opened when participants interact with the objects (right).

interact with objects in VE, eye gaze pointer function was used. It means that a blue circle appeared when participants look at interactive objects in VE (Fig 1). In addition, a panel window was opened when participants interacted with the objects. In panels, participants could decide their behavior by pressing a keyboard button. There was 10 minutes time limit of mock crime, the remaining time was displayed in red on the screen. The created VE was presented using Head-mount display (HMD). A comparing clip is available at the following address: youtu.be/XydTHxqUmDY.

Participants

Sixty-two right-handed participants (mean age = 24.5, S.D. = 2.14, 31 female) were recruited by posting a notice on online board of Korea University. All participants had normal vision and no psychiatric, neurological disease history. These criteria for excluding were written on the notice and confirmed by a short interview before the experiment. In addition, they were not under the influence of medication, illness (e.g., common cold). Participants informed that they would receive 40,000 KRW if they would be acquitted by following lie detection test. Otherwise, they would receive 20,000 KRW and stay 30 minutes after the test as a penalty. In fact, All amounts paid after the end of the study. All participants provided a written informed consent before performing the experimental procedures. In the analysis, two participants were excluded because one participant failed to mock crime, another participant had an excessive artifact in the EEG data. Finally, 60 participants were included in the analysis. This study was approved by the Ethics Committee of Korea University (KU-IRB-15-155-A-1-(E-P-1)).

Experimental Group

In this study, 2 by 2 factorial design was used with ‘Crime status (guilty or innocent)’ and ‘Environment mode (actual or VR)’. Each participant received a ‘guilty’ mission or 'innocent' mission. In the innocent mission, participants were required to wrap a small gift box (containing ring or watch with counterbalancing) that prepared to give to a professor. Specifically, they had to search the gift box and packing materials (ribbon, big gift box, shredding paper for decoration), to put the small gift box to the large one with shredding paper, and to wrap the big gift box with ribbon decoration. In the guilty mission, several procedural components were added. They had to pretend to wrap gift box, but steal the contained gift secretly. For this purpose, their mission included turning off mock CCTV switch that located at the around entrance of the professor's office.

'Environment mode' means the participants conducted the mission in a real environment (actual) or a virtual reality environment (VR). Actual group performed the mission at the office of professor. VR group conducted the mission in another experiment room using HMD. Another experiment room was a soundproof room with black curtains to prevent extraneous noise. Participants were allocated to one of four groups randomly with considering gender ratio (each n=15).

Procedure

Participants were informed of the purpose and procedure of the experiment. At this time, an experimenter emphasized that participants would be judged by a result of EEG lie detection test, and they would get a penalty if they are sentenced as guilty. After writing consent forms, they were also required to fill the self-reported questionnaires.

The questionnaires consisted of measures that could affect CIT results: Handedness Scale, Behavioral Inhibition System/Behavioral Activation System (BIS/BAS; Carver & White, 1994), Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), Self-monitoring (Snyder & Gangestad, 1986), Machiavellianism IV (Mach IV; Christie, Geis, & Berger, 1970), Risk-taking (RTQ; Knowles, Cutter, Walsh, & Casey, 1973), Beck Depression Inventory (BDI; Beck, 1985), Beck Anxiety Inventory (BAI; Beck, 1967), and Psycho-pathic Personality Inventory-Revised (PPI-R; Lilienfeld & Widows, 2005). BIS/BAS scores are known to be associated with EEG asymmetry. The emotion-related measures (PANAS, BDI, BAI) were used for excluding the influence of emotions at the time of participation in the experiment. Machiavellianism and risk-taking tendency are related to the proficiency of lying. Also, psycho-pathic trait is known to have a less physiological response when lying.

Besides, participants allocated to VR conditions filled Simulator sickness questionnaire (SSQ; Kennedy, Lane, Berbaum, & Lilienthal, 1993) additionally before and after VR missions and the difference score was calculated to measure potential sickness related to virtual

Table 1. Self-reported questionnaire

	Guilty- actual	Guilty- VR	Innocent- actual	Innocent- VR
Age	22.4 (2.2)	23.9 (2.0)	23.7 (2.1)	23.4 (3.9)
Gender(M/F)	9 / 6	8 / 7	7 / 8	7 / 8
Handedness Scale	28.7 (5.1)	27.4 (7.8)	33.4 (13.3)	29.4 (9.38)
BIS	20 (2.3)	19 (2.5)	19 (1.6)	20 (2.1)
BAS	11 (2.0)	12 (2.8)	12 (1.7)	11 (2.5)
PANAS-positive	24 (5.5)	24 (4.8)	24 (7.2)	25 (7.7)
PANAS-negative	15 (4.0)	15 (6.0)	14 (2.9)	16 (5.0)
Self-Monitoring	8.2 (2.5)	8.8 (2.3)	7.8 (2.1)	7.7 (1.0)
Mach-IV	58 (4.1)	59 (3.1)	60 (5.6)	57 (4.3)
Risk-Taking	59 (7.0)	64 (5.0)	63 (5.3)	62 (7.3)
BAI	7 (5.2)	11 (9.7)	6.7 (4.2)	7.6 (5.1)
BDI	8.2 (4.2)	7.9 (8.3)	5.7 (4.6)	6.5 (5.3)
PPI-R	58 (11.0)	56 (6.9)	55 (8.8)	56 (8.9)
SSQ		-3.3 (3.4)		-2.8 (4.4)

* each indicates mean and (S.D)

reality experience. There are no significant differences between groups in the self-reported questionnaires (all $p>.05$, Table 1).

Participants chose between two envelopes. It was mentioned that one envelope contains guilty mission, another has innocent mission. However, the envelopes have the same mission papers actually according to the already allocated group. The experimenter explained their mission and reaffirmed whether the participants understood the content well. After the explanation, they asked to write down the mission procedure briefly in the blank space in below of mission paper.

In the case of virtual reality groups, participants had a practice session. This is a process to accommodate VE, such as walking corridors and open a drawer in other virtual lecture room. By this practice, participants became accustomed to interaction by eye gaze pointer function and to movement by keyboard input in VE.

Each participant performed the mission according to the group assigned. After the mission, participants had a waiting time of 20 minutes, and they moved to EEG recording room. In the EEG room, they had a brief interview with another experimenter and conducted a CIT with EEG recording.

Concealed Information Test

In the study, CIT was presented by Korean language. The stimuli were presented as a series of words with inter-stimulus-interval (ISI). The content of a sentence was 'Did you steal XXX?' in Korean grammar order. In other words, it was presented as such 'Dangsineun (you)' / 'XXX (item)' / 'Humchyeosseubnikka (steal)?' (Fig 2). It was designed to prevent eye movement and brain activities related to behavioral responses. Subject word and object word were presented 1000ms each, verb word as 500ms. ISI was randomly selected between 1050, 1150, 1250, and 1350ms. When a verb word appeared, participants had to respond button within 1000ms and trials with later response were excluded from further analysis. Response buttons were'e (yes)' and 'i (no)' of a keyboard, and counterbalanced from the break time in the middle. Participants had to press 'no' button except target item was presented in object word. If the target item was presented, participants were requested to press 'yes' button regardless of the claim that they did not steal things.

In object word position, stimuli were ring, watch, wallet, necklace, perfume, fountain pen, and belt. All items had been selected with considering the number of characters and price values through the pre-survey. In these items, the probe stimulus was an item stolen by guilty participants in the mission (ring or watch). For innocent participants, one of two items was randomly selected as a probe for further analysis. Other items were served as irrelevant items except for a selected one target item. A target item was selected from irrelevant items randomly. For instance, in the case of a guilty participant who had stolen a ring, an example of the composition of all the stimuli was as follows: probe (ring), target (wallet), irrelevant (watch, necklace, perfume, fountain pen, and

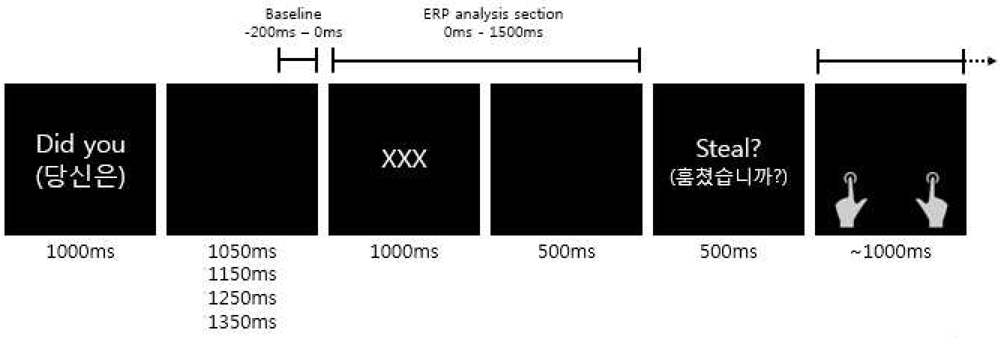

Fig 2. Overview of Concealed Information Test.

belt). Probe and irrelevant stimuli required to press 'no' button, but target item required to 'yes' button following experimenter's direction. This item was also counterbalanced between participants.

The CIT procedure had 6 blocks in the study, each block was consists of 77 trials. Therefore, total number of trials was 462 ((1 target + 1 probe + 5 irrelevant) x 11 repetitions x 6 blocks). When every block ended, 30 seconds of recess time was provided. And the middle of the test, or the end of the third block, the participants could rest 5 minutes.

All stimuli were presented in light gray letters on a black background, and the visual angle of stimuli was 2.1°. We used a 17 inch CRT monitor to minimize monitor refresh delay.

EEG Acquisition and Analysis

EEG signals were acquired by SynAmps RT NeuroScan 64-channel EEG system (NeuroScan Compumedics, Charlotte, NC, USA). Reference electrodes were located in ear lobes and averaged ear-lobe values served as the reference (linked-earlobe). A ground electrode was in the size between FPz electrode and Fz electrode. Electrooculogram (EOG) channels were located at upper/down-side of the left eye, and at left/right-side of both eyes. We used monopolar EOG acquisition to improve the quality of subsequent processing of independent component analysis (ICA; Bigdely-Shamlo, Mullen, Kothe, Su, & Robbins, 2015). All electrode impedances were lower than 5kΩ. The sampling rate was 1000Hz.

Along with EEG acquisition, electrocardiogram (ECG) were recorded simultaneously with the same sampling rate. This was conducted by BIOPAC system (model MP150, Santa Barbara, CA, USA), ECG electrodes were attached to the lower part of the collarbone and the belly side. the ECG data were used to exclude the effect of cardiac signals on EEG data (Park, Correia, Ducorps, & Tallon-Baudry, 2014).

Acquired EEG and ECG data were down-sampled to the 250Hz sampling rate. The EEG data were band-pass filtered at 0.1 – 30Hz (Luck, 2014; Mertens & Allen, 2008), and ECG data filtered at 1 – 40Hz (Park et al., 2014). We used FIR filter in EEGLAB (Delorme, & Makeig, 2004). We conducted ICA decomposition, and ICA components which have significantly positive correlation with EOG, ECG signals were removed. We used SASICA plugin to select ICA components (Chaumon, Bishop, & Busch, 2015). After artifact correction, continuous data were epoched to –200 ms to 1500 ms range to the object word stimuli (e.g., ring, watch). Epochs that have ±75 μV amplitude and the max slope exceed 45 μV/slope were excluded from further analysis.

In the remaining trials, P300 amplitude was calculated and bootstrapping classification was conducted using signals at the Pz electrode. We used the peak-to-peak method which is recommended in concealed information test researches to calculate P300 amplitude (Soskin, Rosenfeld, & Niendam, 2001). This method regards the amplitude as the difference between the highest peak amplitude and lowest amplitude which is located after the highest peak in time series. the ERP signal was moving-averaged with 100 ms time window that step toward 4 ms. After that, the highest amplitude was searched within 300 ms to 800 ms after stimulus onset. And the lowest amplitude was searched to 1500 ms. Calculated P300 amplitudes were analyzed by repeated measure ANOVA with ‘Crime status (guilty or innocent)’ and ‘Environment mode (actual or VR)’ as between-subjects factors and ‘Stimuli type (target, probe, guilty)’ as a within-subjects factor. A Greenhouse-Geisser correction was used where the assumption of sphericity has been violated (Mauchly’W = .689, df = 2, p < .01).

Individual Classification

In detecting deception research, classifying participants correctly to guilty or innocent for specific criteria is as important as the statistical results. This is because the research field aims to be used in the practical field (Ben-Shakhar, 2012). For individual classification (sentence as guilty or innocent), we used bootstrapped amplitude difference (BAD; Farwell & Donchin, 1991) based on peak-to-peak amplitude. This method is not only widely used but also showed the overall highest accuracy in our previous study (Song, Kim, Lee, Chang, & Kim, 2018). For each stimulus type (target, probe, irrelevant), all trials were selected randomly with replacement. The number of selection followed the actual number of trials for that subject respectively. The selected trials were averaged, and this regarded as one re-sampled trial. During 100 iteration, P300 amplitudes of probe and irrelevant were calculated using the peak-to-peak method. In each iteration, bootstrapping index was 1 if probe amplitude > irrelevant amplitude. Otherwise, bootstrapping index was 0. Finally, subjects were sentenced as guilty if the index is over 90, or as innocent is the index is less than 90 (Mertens & Allen, 2008; Rosenfeld, Soskins, Bosh, & Ryan, 2004). For more reliable results, we repeated BAD analysis 10 times and the average values were used to the final bootstrapping index (For more details, see Song et al., 2018).

ROC curve Analysis

Receiver operating characteristics (ROC) analysis was also conducted to investigate the trade-off between sensitivity and specificity for every possible cut-off criterion. In other words, this analysis shows overall classification efficiency. Each of Innocent-actual and Innocent-VR groups was served as specificity criteria group, and the other groups was compared to the specificity. ROC analysis was based on the bootstrapping index. We used parametric ROC. In addition, the area under curves (AUC) were compared with 'Environment mode' using MedCalc (Schoonjans, F.R.A.N.K., Zalata, Depuydt, & Comhaire, 1995). When Innocent groups served as a specificity group, the difference of AUCs between 'actual' and 'VR' was calculated using Hanley & McNeil method (McNeil & Hanley, 1984).

Equivalence Test

To make a more reliable conclusion, we tested the equivalence test on P300 amplitudes of each stimulus and on bootstrapping index. This is because insignificant statistical results could come from the lack of statistical power rather than there is no effect. Therefore, we conducted the equivalence test using 'two one-sided tests' between the actual groups (n=30) and the VR groups (n=30). Following the recommendation of 'smallest effect size of interest' based on given sample size (Lakens, Scheel, & Isager, 2018) and alpha level (α=.05), we assumed the smallest effect size as 0.8 Cohen's d. The analysis was conducted using a TOSTER package implemented in R. As the following convention, the higher p-value is presented in the result section.

Result

P300 Amplitude

ERP waveforms of each group are presented in Fig 3. As we hypothesized, all groups showed relatively higher P300 waveform of the target stimulus and a relatively small increase in the irrelevant stimulus. Guilty groups revealed

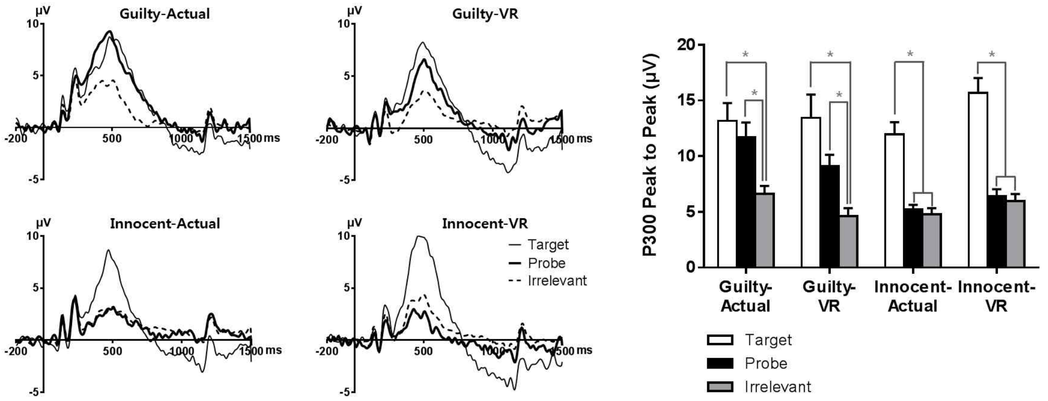

Fig 3. Grand average ERPs (left) and P300 amplitudes (right) of four groups. The bars of P300 amplitudes are SEM.

Table 2. Mean and S.D of the peak-to-peak P300 amplitude

	Target	Probe	Irrelevant
Guilty-actual	13.2 (6.2)	11.7 (5.1)	6.6 (2.7)
Guilty-VR	13.4 (8.1)	9.1 (4.0)	4.6 (2.7)
Innocent-actual	11.9 (4.3)	5.2 (1.6)	4.8 (2.1)
Innocent-VR	15.7 (5.2)	6.4 (2.3)	6.0 (2.6)

Table 3. Result of repeated measure ANOVA and Post hoc

Variable	F	*df*	*p*	η^2	Observed Power
Stimuli type	118.796	(1.52,85.41)	.000***	.680	1.000
Crime status	2.379	(1,56)	.129	.041	.329
Environment mode	.097	(1,56)	.757	.002	.061
Stimuli type × Crime status	147.053	(1.52,85.41)	.000***	.190	.988
Stimuli type × Environment mode	3.877	(1.52,85.41)	.035*	.065	.606
Stimuli type × Crime status × Environment mode	.045	(1.52,85.41)	.918	.001	.056
Variable	t	*df*	*p*		
Target (guilty vs innocent)	-.318	58	.752		
Probe (guilty vs innocent)	4.895	58	.000***		
Irrelevant (guilty vs innocent)	.402	58	.689		
Target (actual vs VR)	-1.268	58	.210		
Probe (actual vs VR)	.636	58	.528		
Irrelevant (actual vs VR)	.619	58	.538		
Target (Innocent-actual vs Innocent-VR)	-2.137	28	.042*		

* p<.05, ** p<.01, *** p<.001

increased P300 waveform of probe stimulus, and innocent groups have the irrelevant-level waveform of the probe (Table 2). A repeated measure ANOVA showed a significant main effect of Stimuli type (*F*1.52, 85.41) = 118.796, p < .01, η^2 = .680). And two interactions were significant: Stimuli type × Crime status (*F*1.52, 85.41) = 13.154, p < .01, η^2 = .190), Stimuli type × Environment mode (*F*1.52, 85.41) = 3.877, p < .05, η^2 = .065). The main effect of Crime status and of Environment mode were not significant (Table

3). A three-way interaction of Stimuli type × Crime status × Environment mode was also not significant. As we expected, P300 amplitudes have different patterns between guilty groups and innocent groups. Moreover, the differences in amplitude by Stimuli type was significant. However, a significant interaction of Stimuli type × Environment mode suggests that there could be a possible different amplitude pattern between the actual and the VR. We checked amplitudes again and found there was an outlier in Innocent-VR group whose P300 amplitude of target over 27 μV which is a comparatively extreme value. After exclude one outlier, the interaction was not significant (F(1.53, 84.56) = 3.06, p = .07, η^2 = .049).

To scrutinize the P300 amplitude difference, post hoc analysis was conducted (Table 3). Probe amplitudes were significantly different between guilty groups and innocent groups. Other stimuli were not different significantly. For guilty groups, there were significant differences between target-irrelevant, probe-irrelevant. And for innocent groups, target-probe, target-irrelevant were significantly different (all $p < .05$). In addition, the significant interaction of Stimuli type × Environment mode was mainly caused by a significant difference in the target amplitudes between Innocent-actual and Innocent-VR. However, when outlier was excluded, the difference was not significant (t(27) = -1.722, p = .089).

Individual Classification

Individual classification by BAD method is presented in Table 4. In Guilty-actual and Guilty-VR, every 2 participants were incorrectly classified as 'innocent'. The accuracy was 86.7%. In Innocent-actual and Innocent-VR, every 3 participants were classified as 'guilty'. The accuracy was 80%.

ROC curve Analysis

ROC curve analysis was also conducted to investigate statistical classification efficiency (Fig 4). ROC analysis was based on the bootstrapping index, which had been calculated at individual classification analysis. When the specificity was Innocent-actual, the AUC was following (AUC, Confidence Interval): Guilty-actual (0.7847, 0.6303-0.9243), Guilty-VR (0.8047, 0.7149-0.8993), Innocent-VR (0.4920, 0.3441-0.7084). In addition, if the specificity was Innocent-VR AUC was following: Guilty-actual (0.8027, 0.6925-0.9298), Guilty-VR (0.8231, 0.6754-0.9166). Overall, Guilty groups were effectively distinguished from innocent groups even compared to Innocent-VR. Comparison between AUCs revealed that there was no significant difference between actual and VR (z-statistics = 0.293, p = 0.76).

Table 4. Individual classification with BAD

BAD results (Bootstrapping index)				
Num	Guilty-actual	Guilty-VR	Innocent-actual	Innocent-VR
1	G (100)	G (99)	I (34)	I (28)
2	I (58)	G (99)	I (59)	I (88)
3	G (100)	G (100)	G (99)	I (23)
4	G (90)	G (100)	I (86)	G (96)
5	G (100)	G (99)	I (57)	I (65)
6	G (99)	G (90)	G (96)	I (88)
7	G (100)	G (100)	I (86)	I (81)
8	G (94)	G (100)	I (79)	I (74)
9	G (98)	G (99)	I (85)	I (82)
10	G (100)	I (63)	I (47)	I (83)
11	G (93)	I (79)	I (79)	I (37)
12	G (99)	G (100)	I (57)	G (92)
13	I (73)	G (99)	G (97)	G (93)
14	G (100)	G (100)	I (5)	I (67)
15	G (99)	G (100)	I (88)	I (41)
Total				
G	13	13	3	3
I	2	2	12	12
Accuracy				
	86.7%	86.7%	80%	80%

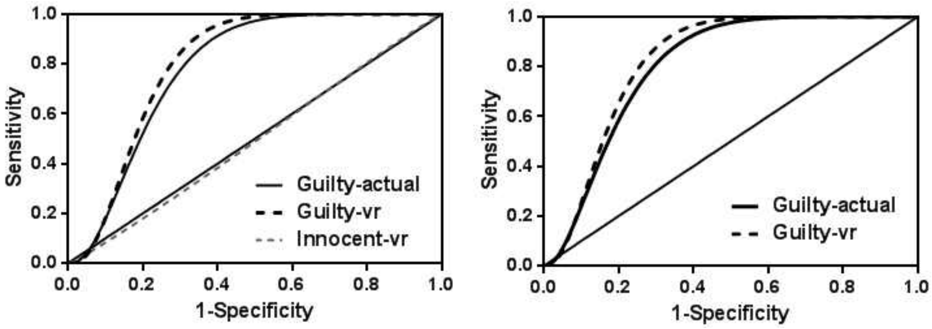

Fig 4. The results of ROC curve. Specificity group was selected innocent group each. When Innocent-actual served as specificity group (left), and Innocent-VR served as (right)

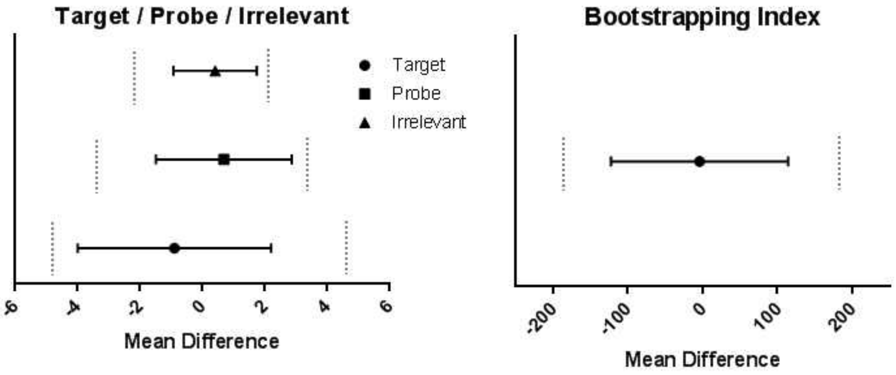

Fig 5. The results of equivalence test. horizontal lines indicate 95% confidence interval, vertical lines indicate the equivalence bounds

Equivalence Test

Equivalence tests were revealed that P300 amplitudes of each stimuli were significantly equivalent between actual and VR groups: target (t(58) = 2.529, p = 0.007), probe (t(58) = -2.452, p = 0.008), irrelevant (t(58) = -2.468, p = 0.008). In addition, bootstrapping index was also significantly equivalent (t(58) = 3.03, p = 0.001). In Fig 5, 95% confidence intervals were presented as horizontal lines. vertical lines indicate that equivalence bounds. All CIs were located within equivalence bound range, which means the mean of each groups (actual, VR) is statistically equivalent.

Discussion

In this study, a high-fidelity VE was developed using Matterport 3d camera and Unity 3D engine. As a comparison with previous research using virtual reality for mock crime, the VE in the current study was more realistic. Furthermore, to our knowledge, this is the first study to compare the effects of virtual mock crime compared to the real environment. At the debriefing session, at the end of CIT, the experimenter asked VR participants to rating fidelity of the VE by 10 Likert scale. Most participants responded 9 or 10 points. The average score was 9.1 (1.0 S.D).

To validate the potentiality of virtual mock crime, we tested the P300 metrics in several ways including parametric statistics, individual classification, ROC curve, and equivalence test. These results suggest that there was no statistical difference between actual mock crime and virtual mock crime when we focused on P300 amplitude. In repeated measure ANOVA, when excluding one outlier, only the main effect of Stimuli type and the interaction between

Stimuli type × Crime status were significant. This is consistent with our hypothesis that there would be no significant differences in P300 amplitude when we use virtual mock crime in sensitivity and specificity. In individual classification, guilty groups correctly classified at 86.7%, and innocent groups classified at 80%. There was no difference of ratios between actual and VR. These results also confirmed by ROC curve analysis and equivalence test. In the ROC curve, AUC was significantly above the chance level when we compared guilty groups and innocent groups even when Innocent-VR was used as the specificity group. There was no statistically significant difference between AUCs of the actual and the VR. All P300 amplitudes and bootstrapping index were significantly equivalent.

Considering that P300 amplitude is the most widely used metric in CIT research, we believe that virtual mock crime is likely to be used as an alternative method for the actual mock crime. Well-established virtual mock crime will enable mock crime scenarios that we have not been able to use before. Furthermore, virtual mock crime will be helpful in limited space and expense. The virtual also can help to control extraneous variables.

However, there are limitations in this study, and we would like to discuss it. First, the comparison between virtual mock crime and actual mock crime should be conducted in more various ways. For example, we only adopted the peak-to-peak method to calculate P300 amplitude and BAD method for individual classification because these methods are the most widely used in CIT research field. However, as our previous research (Song et al., 2018), there are other methods to calculate P300 amplitude and individual classification (e.g., Base-to-peak P300 amplitude, Bootstrapped correlation method). For a more rigorous comparison, several methods should be included in further research. In addition, we only focused on P300 ERP component, but it would become better to include other ERP components to analysis (e.g., N400, LPC), other metrics (e.g., P300 onset latency, power spectra of frequency bands), and other analysis methods (e.g., source localization, connectivity analysis). For example, the N400 component at the frontal area is known as reflecting familiarity-based recognition (Rugg & Curran, 2007), and conscious control processing (Proverbio, Vanutelli, & Adorni, 2013). Compared with the actual environment, it may be possible that VE provides worse circumstances for encoding peripheral objects and affects subsequent N400 amplitude in detecting deception test using peripheral objects to classify guilty participants. LPC has been known as associated with emotional aspects like feeling guilty (Leng et al., 2017). If the VE provided a more emotionally neutral environment, a lower level of LPC might have been introduced. Therefore, these components need to be proven for the usefulness of virtual mock crime. Further studies will require more diverse analysis.

In addition, the subjects performed mock crimes and received detecting deception test (CIT) on the same day. Although we put 20 minutes waiting time, this is the relatively small interval compared to the practical field. It is possible that the performance of the CIT may vary over time depending on the adopted Environment mode. Because the senses such as the olfactory and tactile are not provided in the VE, the encoding process is slightly different, and it can affect processes of memory encoding and subsequent memory retrieval. In fact, multisensory learning is known as a more effective way for learning (Shams & Seitz, 2008). Therefore, the efficiency of virtual mock crime should be reviewed when a lie detection test is conducted with a relatively long time interval.

Nevertheless, this study is significant in that it is the first study to examine the effects of virtual mock crime on subsequent CIT results. In order to thrive in the field of lie detection research, it is necessary to give attention to not only develop detection methods but also provide proper scenarios with high ecological validity. Virtual reality is a promising tool for providing mock crime scenarios.

Reference

송인욱, 김혜민, 이경은, 장은희, & 김현택 (2018). P300-CIT 부트스트랩 분석 비교. 한국심리학회지: 법, 9(2), 75-99.

Beck, A. T. (1967). Depression: Clinical, experimental, and theoretical aspects (Vol.32): University of Pennsylvania Press

Beck, A. T. (1985). Anxiety disorders and phobias: A cognitive perspective. New York: Basic Books.

Ben-Shakhar, G. (2012). Current research and potential applications of the concealed information test: an overview. *Frontiers in psychology, 3*, 342.

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *Journal of Applied Psychology, 88*(1), 131.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. Frontiers in neuroinformatics, 9, 16.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of personality and social psychology, 67*(2), 319.

Chaumon, M., Bishop, D. V., & Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods, 250*, 47-63.

Christie, R., Geis, F. L., & Berger, D. (1970). Studies in machiavellianism: Academic Press NewYork.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods, 134*(1), 9-21.

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating?. *Behavioral and brain sciences, 11*(3), 357-374.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("Lie detection") with Event Related brain potentials. *Psychophysiology, 28*(5), 531-547.

Gamer, M., Bauermann, T., Stoeter, P., & Vossel, G. (2007). Covariations among fMRI, skin conductance, and behavioral data during processing of concealed information. *Human brain mapping, 28*(12), 1287-1301.

Gamer, M., Verschuere, B., Crombez, G., & Vossel, G. (2008). Combining physiological measures in the detection of concealed information. *Physiology & behavior, 95*(3), 333-340.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: an fMRI investigation. *Cerebral cortex, 13*(8), 830-836.

Ganis, G., & Schendan, H. E. (2013). Concealed semantic and episodic autobiographical memory electrified. *Frontiers in human neuroscience, 6*, 354.

Gao, J. F., Yang, Y., Huang, W. T., Lin, P., Ge, S., Zheng, H. M., ... & Rao, N. N. (2016). Exploring time-and frequency-dependent functional connectivity and brain networks during deception with single-trial event-related potentials. *Scientific reports, 6*, 37065.

Hahm, J., Ji, H. K., Jeong, J. Y., Oh, D. H., Kim, S. H., Sim, K. B., & Lee, J. H. (2009). Detection of concealed information: combining a virtual mock crime with a P300-based Guilty Knowledge Test. *Cyberpsychology & behavior, 12*(3), 269-275.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology, 3*(3), 203-220.

Knowles, E. S., Cutter, H. S., Walsh, D. H., & Casey, N. A. (1973). Risk-taking as a personality trait. *Social Behavior and Personality: an international journal, 1*(2), 123-136.

Kugelmass, S., & Lieblich, I. (1966). Effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology, 50*(3), 211.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269.

Lee, T. M., Liu, H. L., Tan, L. H., Chan, C. C., Mahankali, S., Feng, C. M., ... & Gao, J. H. (2002). Lie detection by functional magnetic resonance imaging. *Human brain mapping, 15*(3), 157-164.

Leng, B., Wang, X., Cao, B., & Li, F. (2017). Frontal negativity: An electrophysiological index of interpersonal guilt. *Social neuroscience, 12*(6), 649-660.

Lilienfeld, S. O., & Widows, M. R. (2005).

Psychological assessment inventoryrevised (PPI-R). Lutz, FL: Psychological Assessment Resources.

Luck, S. J. (2014). An introduction to the event-related potential technique. MIT press.

McNeil, B. J., & Hanley, J. A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical decision making, 4*(2), 137-150.

Mertens, R., & Allen, J. J. (2008). The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology, 45*(2), 286-298.

Park, H. D., Correia, S., Ducorps, A., & Tallon-Baudry, C. (2014). Spontaneous fluctuations in neural responses to heartbeats predict visual detection. *Nature neuroscience, 17*(4), 612.

Peth, J., Vossel, G., & Gamer, M. (2012). Emotional arousal modulates the encoding of crime related details and corresponding physiological responses in the Concealed Information Test. *Psychophysiology, 49*(3), 381-390.

Proverbio, A. M., Vanutelli, M. E., & Adorni, R. (2013). Can you catch a liar? How negative emotions affect brain responses when lying or telling the truth. *PloS one, 8*(3), e59383.

Rosenfeld, J. P. (Ed.). (2018). Detecting concealed information and deception: Recent developments. Academic Press.

Rosenfeld, J. P., Biroschak, J. R., & Furedy, J. J. (2006). P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. *International Journal of Psychophysiology, 60*(3), 251-259.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure resistant, accurate, P300 based method for detection of concealed information. *Psychophysiology, 45*(6), 906-919.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300 based tests of detection of concealed information. *Psychophysiology, 41*(2), 205-219.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in cognitive sciences, 11*(6), 251-257.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological science, 19*(8), 772-780.

Schoonjans, F. R. A. N. K., Zalata, A., Depuydt, C. E., & Comhaire, F. H. (1995). MedCalc: a new computer program for medical statistics. *Computer methods and programs in biomedicine, 48*(3), 257-262.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in cognitive sciences, 12*(11), 411-417.

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of personality and social psychology, 51*(1), 125.

Soltani, M., & Knight, R. T. (2000). Neural origins of the P300. *Critical Reviews^{TM} in Neurobiology, 14*(3-4).

Soskins, M., Rosenfeld, J. P., & Niendam, T. (2001). Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: complex vs. simple paradigms. *International Journal of Psychophysiology, 40*(2), 173-180.

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance, 24*(6), 1656.

Wang, H., Chang, W., & Zhang, C. (2016). Functional brain network and multichannel analysis for the P300-based brain computer interface system of lying detection. *Expert Systems with Applications, 53*, 117-128.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063.

Wilding, E. L. (1999). Separating retrieval strategies from retrieval success: An event-related potential study of source memory. *Neuropsychologia, 37*(4), 441-454.

Zaitsu, W. (2016). External validity of Concealed Information Test experiment: comparison of respiration, skin conductance, and heart rate between experimental and field card tests. *Psychophysiology, 53*(7), 1100-1107.

1 차원고접수 : 2019. 06. 04.
심사통과접수 : 2019. 06. 26.
최종원고접수 : 2019. 07. 11.

한국심리학회지 : 법
The Korean Journal of Forensic Psychology
2019, Vol. 10, No. 2, 89-109

가상현실 모의범죄는 실제 모의범죄를 대체할 수 있는가? 사건관련전위 연구

송인욱　김혜민　이경은　장은희　김현택

고려대학교 심리학과

거짓말 탐지 연구 분야에서, 숨김정보검사(Concealed Information Test; CIT)는 뇌전도(EEG) 측정과 결합하여 가장 널리 사용되는 방법이다. 또한 모의범죄 시나리오는 참가자에게 거짓말을 할 내용을 제공하기 위해 널리 사용된다. 모의범죄 시나리오는 자전적 정보 등을 사용하는 것보다 더 높은 생태학적 타당도를 지닌다는 장점이 있으나, 윤리적인 문제, 현실적 자원의 문제, 실험 통제 가능성 측면에서 몇 가지 한계점을 가지고 있다. 가상현실(Virtual reality; VR)은 이러한 단점들을 극복할 가능성이 있는 대체방법이다. 그럼에도 불구하고, 모의범죄에 가상현실을 적용한 연구는 극히 드물며, '실제' 모의범죄와 '가상' 모의범죄를 직접적으로 비교한 연구는 아직까지 이루어지지 않았다. 본 연구에서는, 실제 모의범죄가 수행된 공간을 측정하여 만든 높은 충실도의 가상현실을 제작하고 이를 모의범죄에 사용하였다. 참가자는 '유죄' 또는 '무죄' 시나리오에 참가하였으며, 각 시나리오는 '실제' 또는 '가상현실'로 제공되었다. 모의범죄 시나리오를 마친 후, 참가자는 뇌전도 측정과 함께 숨김정보검사를 받았다. 가상현실 모의범죄를 사용하는 것이 이후의 숨김정보검사 및 뇌전도에 미칠 수 있는 영향을 알아보기 위하여, 본 연구는 P300 사건관련전위 및 부트스트래핑 방법을 사용한 개인 판별분석에 초점을 맞추었다. 본 연구진이 세운 가설대로, 모의범죄를 제시하는 방법의 차이에 따른 주효과는 유의미하지 않았으며, 극단적인 P300 진폭 크기를 보인 한 명을 제외하였을 경우 실험 자극(목표, 탐침, 무관련)과 모의범죄 제시 방법의 교호 효과도 유의하지 않았다. 뿐만 아니라, 실제 모의범죄와 가상현실 모의범죄 집단들에 대한 개인 판별분석은 동일한 수준의 정확도를 보여주었다. ROC 분석 및 동등성 검증 분석 또한 위 방법들의 결과를 지지하였다. 본 연구에서, 모든 결과는 실제 모의범죄와 가상현실 모의범죄를 사용하는 것이 차이가 없는 것으로 나타났다. 이러한 결과는 가상현실 모의범죄가 실제 모의범죄의 대안으로 사용될 수 있음을 시사한다.

주요어 : Concealed Information Test, Deception, Virtual Reality, Mock Crime, P300