Legal and Criminological Psychology (2009), 14, 279–292

The British Psychological Society

www.bpsjournals.co.uk

Effectiveness of pupil diameter in a probable-lie comparison question test for deception

Andrea K Webb[1], Charles R. Honts[2]*, John C. Kircher[1], Paul Bernhardt[3] and Anne E. Cook[1]

[1]University of Utah, Salt Lake City, Utah, USA
[2]Boise State University, Boise, Idaho, USA
[3]Frostburg State University, Frostburg, Maryland, USA

Purpose. There were three objectives of this study: (1) To assess the possibility of using pupil diameter as an index of deception in the context of a comparison question polygraph test. (2) To determine if pupil diameter would make a significant contribution to an optimal multivariate classification equation in combination with the traditional predictor variables used in field polygraph practice. (3) We explored the possibility of replacing one or more of the traditional predictor variables with pupil diameter.

Methods. We used a laboratory mock crime experiment with 24 participants, half of whom stole $20 (US) from a secretary's purse. Participants were tested with a comparison question test modelled after standard field practice. Physiological measures were taken with laboratory quality instrumentation. Features were extracted from the physiological measures. Those features were subjected to a number of different statistical analyses.

Results. Innocent participants showed larger increases in pupil diameter in response to probable-lie questions than to relevant questions. Guilty participants did not show differential responding to the question types. The additional of pupil diameter to a multivariate classification model approached, but did not reach significance. Subsequent analyses suggest that pupil diameter might be used to replace the traditional relative blood pressure measure.

Conclusions. Pupil diameter was found to be a significant predictor variable for deception. Pupil diameter may be a possible replacement for the traditional relative blood pressure measure. Additional research to explore that possibility would seem to be warranted.

Despite decades of research, there is a long-standing and heated debate on the validity of the comparison question test (CQT) for psychophysiological deception detection (PDD; Honts, Raskin, & Kircher, 2005; Iacono & Lykken, 2005; National Research

*Correspondence should be addressed to Professor Charles R. Honts, Psychology Department, Boise State University, 1910 University Drive MS-1715, Boise ID 83725-1715, USA (e-mail: chonts@boisestate.edu).

DOI:10.1348/135532508X398602

Council, 2003). Proponents argue that decision accuracies approaching 90% can be achieved with the CQT (Honts *et al.*, 2005). Critics claim that decision accuracy is about 84% for deceptive individuals and is no better than chance (50%) for truthful subjects (Iacono & Lykken, 2005). In addition to the debate over accuracy, the rationale underlying the CQT has been challenged and argued in the scientific literature (Ben-Shakhar & Furedy, 1990; Honts *et al.*, 2005; Iacono & Lykken, 2005). Despite this controversy, the CQT is used extensively throughout the world for criminal investigations (Raskin & Honts, 2002), and the consequences of decision errors in these settings can be serious. The present study was designed to test if pupil diameter is diagnostic of deception in a CQT and if it can be used to improve on the accuracy achieved by traditional measures of physiological arousal.

The CQT includes several types of questions, only two of which are used to assess credibility, relevant questions and probable-lie comparison questions. Relevant questions directly and unambiguously address the matter under investigation (e.g. Did you take any of the missing money?), whereas probable-lie comparison questions pertain to the matter under investigation only in a general way and cover a long period of time (e.g. Before the age of 30, did you ever take something that did not belong to you?). Probable-lie comparison questions are intentionally vague and difficult to answer truthfully with an unqualified 'No.' The examiner maneuvers the subject into a quick 'No' response through the demand characteristics of the pre-test interview by telling the subject that these questions are used to assess character and determine if the person is the type of person who might have committed the crime. Innocent subjects answer relevant questions truthfully, but they are assumed to be deceptive in their answers to the comparison questions. The rationale of the CQT predicts that innocent subjects will be more concerned about the comparison questions and will respond more strongly to them than to the relevant questions. In contrast, guilty subjects answer the relevant questions deceptively, and because relevant questions are more salient, guilty subjects are expected to react more strongly to the relevant questions than to the comparison questions. In the field, the question sequence is presented at least three times, providing three or more sets of recordings of physiological activity.

Physiological measures traditionally used in the CQT include thoracic and abdominal respiration, skin conductance, relative blood pressure, and vasomotor activity. These measures have utility for detecting deception individually and in combination (Kircher, Kristjansson, Gardner, & Webb, 2005; Raskin & Honts, 2002). It is currently hypothesized that these physiological measures reflect activation in both affect and information processing (Handler & Honts, 2008a, 2008b). All sources agree that there are classification errors with the CQT and that accuracy might be improved by a new dependent measure that might capture discriminative variance not encompassed by the traditional measures.

Vrij (2008) notes that lying is likely to be more cognitively demanding than truth telling. Vrij also notes that aspects of the deceptive context modulate the cognitive load experienced by the liar. Vrij goes on to describe six factors that may affect the cognitive load on the liar. There is a long history of research that demonstrates that task-evoked changes in pupil diameter are reliable and valid indicators of cognitive load. Increases in pupil diameter are associated with task difficulty in recall and transformation of digit strings (Kahneman & Beatty, 1966), mental multiplication (Ahern & Beatty, 1979; Hess & Polt, 1964), sentence processing (Just & Carpenter, 1993; Schluroff, 1982), letter processing (Beatty & Wagoner, 1978), and lexical translation (Hyona, Tommola, & Alaja, 1995). If deception is more cognitively demanding than being truthful as suggested by Vrij (2008), then increases in pupil diameter may provide an independent diagnostic measure in the CQT that may be based primarily on the cognitive component. It seems clear that there is substantial difference in cognitive load for relevant and comparison questions for innocent examinees. For the innocent, the relevant questions although affectively loaded represent a simple cognitive task, the individual is truthful and this should require relatively little effort to process. The comparison questions with their ambiguous nature, long time period and assumed deceptive response should result in a considerable amount of cognitive load as memory is scanned and the response considered. Predictions for the guilty are much less clear. The guilty respond to both question types with deception and both may present problems with considerable cognitive load, although the rationale of the CQT predicts that the comparison questions will have less affective power for the guilty.

Although previous studies on the detection of deception have measured pupil diameter, the findings are sparse and, to some extent, mixed. Heilveil (1976) asked participants questions about themselves and subsequently had them rate their responses as completely deceptive, partially deceptive, or completely true. The pupil was most dilated in the intervals participants reported being deceptive. Dionisio, Granholm, Hillix, and Perrine (2001) measured pupil diameter while participants made truthful and deceptive responses regarding episodic and semantic information. Deception was associated with the greatest increase in pupil size, but there was no difference in pupil size for the two types of information. Bradley and Janisse (1979) and Janisse and Bradley (1980) measured pupil diameter while participants were administered a concealed information test and found that pupil diameter was diagnostic of deception. Of particular relevance to the present study, Bradley and Janisse (1981) conducted a mock-crime experiment in which guilty participants were instructed to steal a dollar and conceal it in their pocket. Innocent participants did not steal the dollar. Guilty and innocent participants then were given both concealed information tests and CQTs. Pupil diameter was measured during the first 4 s following question onset. For the concealed information test, pupil diameter reliably discriminated between guilty and innocent participants. Classification accuracy was 80% for innocent participants but only 33% for guilty participants. For the CQT, pupil diameter did not reliably discriminate between the groups.

The first objective of the present study was to reevaluate the possibility that changes in pupil diameter are diagnostic of deception in CQTs. Although the prediction was tested previously (Bradley & Janisse, 1981), we used a stronger manipulation of guilt, we introduced stronger incentives to pass the test (Kircher, Horowitz, & Raskin, 1988), we used newer technology for measuring pupil diameter, the CQT was not preceded by a concealed information test, and we measured pupil diameter during a longer time window (8 s following question onset). Participants might process information for some time after the question is asked, and a longer time window was used to capture that information.

The second objective was to test if a measure of pupil dilation would make a significant contribution to an optimally weighted combination of traditional physiological measures. Based on prior research (Kircher & Raskin, 2002), we expected that the combination of electrodermal, cardiovascular, and respiration measures would accurately predict group status (guilt). A goal in the present study was to determine if changes in the pupil provided new information about-group membership, beyond that already available in the traditional measures.

The third objective was to test if any of the traditional physiological measures could be replaced by a measure of pupil dilation without sacrificing predictive validity. Whereas traditional measures require site preparation, proper placement of multiple transducers, and in some cases may be uncomfortable (Podlesny & Kircher, 1999), pupil size may be measured safely, remotely, and unobtrusively.

Method

Participants

Newspaper advertisements were used to recruit participants from the general community. The advertisement stated that \$30 would be earned for 2 h of participation, and there was an opportunity to earn a \$50 bonus. Participants were eligible for participation if they were between the ages of 18 and 65, were fluent in English, were not taking any prescription medications, did not have significant medical problems, were male, and had not previously taken a polygraph test. Participants were 24 males between the ages of 18 and 53 ($M = 32.04$, $SD = 9.42$).

Procedure

The experimental procedures were approved by the Institutional Review Board at the University of Utah. In response to the newspaper advertisement, potential participants called a secretary who described the experiment and payment and ensured eligibility. Participants were given a date and time to report to a room in a building on campus. When a participant arrived for his appointment, an envelope with his name on it was taped to the door. The instructions in the envelope told the participant to enter the room, close the door, read and sign an informed consent form, fill out a questionnaire, and play a cassette recorder to receive further instructions over earphones.

Guilty participants were instructed to commit a mock theft of \$20 from a wallet in a purse in a secretary's office and to prepare an alibi in case they were caught in the secretary's office. They went to the secretary's office on a different floor of the building and asked the secretary (a confederate) for directions to the office of Dr Mitchell. The secretary told the participant that there was no Dr Mitchell in the department. The participant thanked the secretary and left. The participant waited in the hall for the secretary to leave her office. When she left, the participant entered the office, searched the desk for the purse, and took a \$20 bill from the wallet in the purse. Guilty participants concealed the money on their person and reported to a room to await the polygraph examiner.

Innocent participants were told that other participants took money from a secretary's purse but that they were innocent and would commit no crime. After listening to this description of the crime, the innocent participant left the area for 15 min and reported to the room to await the polygraph examiner.

All participants were told that they would be given a polygraph test by an expert polygraph examiner who did not know if they stole the \$20 from the secretary's purse. In fact, the examiner was unaware of the participant's guilt or innocence. The examiner was aware of the proportion of guilty and innocent participants in the study. In the field, most polygraph examinees are highly motivated to appear truthful on the polygraph test. In the present study, all participants were told that they would receive a \$50 bonus if they could convince the examiner of their innocence.

When the polygraph examiner arrived, he obtained biographical information from the participant and then attached the sensors. The examiner was a male doctoral level experimental psychologist who was trained to conduct CQT polygraph examinations within our laboratory. Every effort was made to model testing procedures in common application in the field. Although the polygraph test rarely immediately follows the commission of the crime in a field setting, it was beyond the scope of the present study to implement a delay between the mock crime and the polygraph examination. Following standard field practice, a preliminary numbers test was administered, and then all of the CQT test questions were reviewed with the participant. For experiment design purposes, the CQT question sequence was presented four times, resulting in four series of physiological data rather than the traditional three series usually collected in the field (see below). The question sequence is presented in Table 1. Following the examination, the probability of truthfulness was computed using algorithms described in Kircher and Raskin (2002). The participant was paid on the basis of the computer decision and debriefed.

Table 1. Question sequence

1. (Buffer) Do you understand that I will ask only the questions we have discussed?
2. (Sacrifice relevant) Do you intend to answer truthfully all of the questions about the theft of the $20?
3. (Neutral) Is today __?
4. (Probable-lie) Between the ages of __ and __, did you ever lie to get out of trouble?
5. (Relevant) Did you take that $20?
6. (Neutral) Do you live in the United States?
7. (Probable-lie) Before the age of __, did you ever take something that didn't belong to you?
8. (Relevant) Do you have that $20 with you now?
9. (Neutral) Is your first name __?
10. (Probable-lie) During the first __ years of your life, did you ever do anything that was dishonest or illegal?
11. (Relevant) Did you take that $20 from the purse?

Apparatus

The Computerized Polygraph System (CPS) Lab version (CPS-LAB; Scientific Assessment Technologies, Salt Lake City, UT) was used to configure the data collection hardware, specify storage rates for the data, build protocols to collect the data, and collect, edit, and score the data.

Pupil diameter was obtained with the Eye Dynamics Department of Defense Polygraph Institute Eye Data System (Eye Dynamics Inc, Torrance, CA). An IR/Video ENG Goggle used a miniature video-camera to magnify an image of the right eye on a video monitor. The goggles blocked all ambient light from entering the eyes. A red LED was constantly illuminated inside the participant's visual field for two charts and was not illuminated for the remaining two charts. For half of the innocent and half of the guilty participants, the LED was illuminated during the first and third repetition of the question sequence (chart) and was not illuminated during the second and fourth charts. For the remaining participants, the LED was illuminated during the second and fourth charts and was not illuminated during the first and third charts. The LED was designed to constrict the pupil slightly and avoid the possibility that the pupil would be completely dilated in a completely darkened visual field (J. A. Stern, personal communication, April, 1997). The Eye Dynamics system stored pupil diameter at 60 Hz for 10 s that began at the onset of each test question. Respiration, skin conductance, and relative blood pressure were recorded using standard field transducers and data collection parameters.

Response curves

Software was developed that averaged successive 60 Hz samples from the Eye Dynamics system to reduce the sampling frequency to 10 Hz. A pupil diameter response curve was computed for each test question. The pupil diameter at question onset was subtracted from each post-stimulus value for an interval that began at question onset and ended 8 s later. Similarly, the 1,000 Hz samples for respiration and SC also were reduced to 10 Hz for a period that began at question onset and ended 20 s later. For the cardiograph, CPS-LAB identified the time and level of each systolic and diastolic point in the cardiograph record and computed a weighted average for each of 20 post-stimulus seconds. Second-by-second systolic and diastolic response curves were averaged to obtain a mean cardiograph response curve (Kircher & Raskin, 1988).

Feature extraction

CPS-LAB was programmed to extract the following features:

Amplitude was extracted from the pupil, SC, and cardiograph response curves. CPS-LAB identified low and high points on the response curve and then computed the difference between each low point and every succeeding high point. Peak amplitude was the greatest observed difference.

Area under the response curve was extracted from the pupil response curve. Area under the curve was measured from the lowest point following response onset until it returned to the level at response onset or until the eighth second following question onset, whichever occurred first.

Excursion was obtained from the thoracic and abdominal respiration signals. Excursion was the sum of absolute linear differences between successive pairs of 100 ms time samples from question onset for 10 s.

Differential reactivity

For each feature, a measurement was obtained for each comparison and each relevant question on each of four charts of recorded physiological activity. Each participant provided 24 measurements for each channel of physiological data (three comparison and the three relevant questions on each of the four charts). The 24 measurements of a feature for a participant were converted to z scores. Thoracic and abdominal respiration excursion scores are highly correlated (Kircher & Raskin, 1988, 2002). To reduce the number of variables, reduce multicolinearity, and increase reliability, the z scores for thoracic and abdominal measurements were averaged.

The mean of the 12 z scores for relevant questions was subtracted from the mean of the 12 z scores for comparison questions. The difference provided a mean index of differential reactivity to comparison and relevant questions for each feature for each participant. The index of differential reactivity is analogous to the numerical score obtained by an examiner in a field polygraph setting. The sign of the index indicates which question type produced the larger response. For all features except respiration excursion, a large measured response was indicative of physiological arousal. For respiration excursion, arousal was indicated by a relatively small measured response (respiratory suppression). To achieve a common direction for predicted effects, the sign of the mean difference between responses to comparison and relevant questions was reversed for respiration excursion. Thus, for all measures, a positive difference was expected for innocent participants (comparison > relevant), and a negative difference was expected for guilty participants (comparison < relevant).

Results

Pupil responses to comparison and relevant questions are presented in Figures 1 and 2 for guilty and innocent participants, respectively. Responses to neutral questions were not included in the statistical analyses but are presented in the figures for completeness. On average, the pupil response to comparison questions peaked at about 5 s following question onset, whereas the response to relevant questions peaked between 2 s and 3 s following question onset. The mean length of the comparison questions ($M = 16.33$ words, $SD = 1.53$) was over twice the mean length of relevant questions ($M = 7.00$ words, $SD = 1.73$), and the times at which the pupil response peaked corresponded closely with the amount of time it took to ask the respective comparison ($M = 4.33$ s, $SD = .58$) or relevant questions ($M = 2.67$ s, $SD = .58$).

Repeated measures analysis of variance (RMANOVA) was used to test for effects of guilt, question type, and illumination on pupil responses to comparison and relevant questions. The factors were guilt (guilty and innocent); question type (comparison and relevant); illumination (LED illuminated and LED not illuminated); and time (10 samples per second for 8 s). Initially, repetition was included as a factor in the design, because the questions were presented twice in the illumination condition and twice in darkness. However, RMANOVA revealed no main effect of repetition on pupil diameter and no

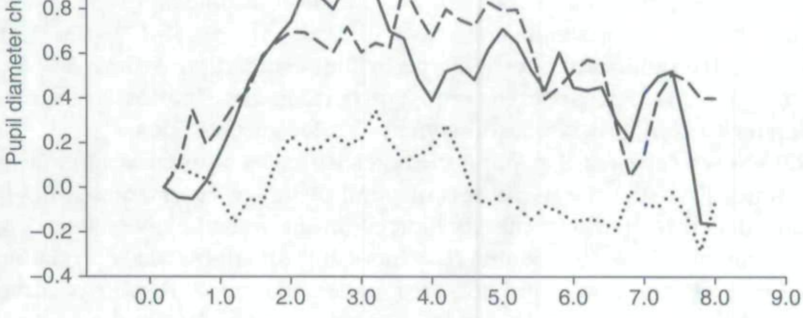

Figure 1. Pupil responses to test questions for the guilty group ($N = 12$).

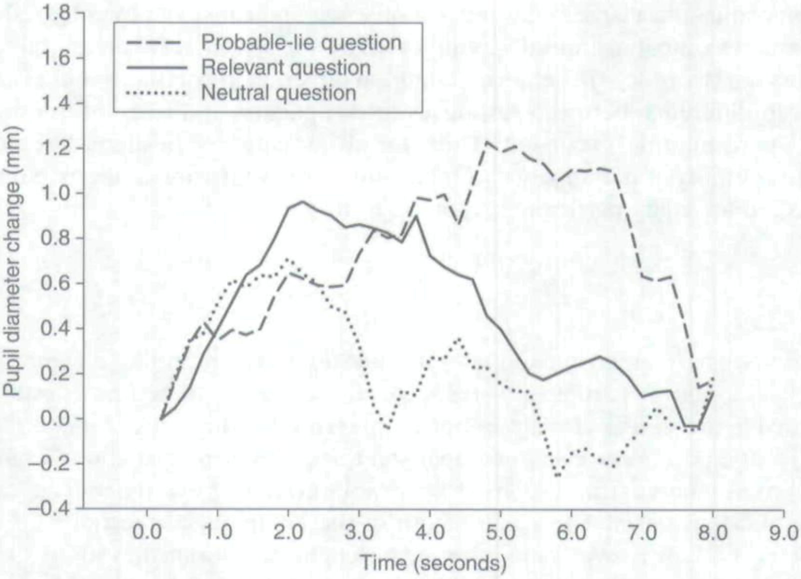

Figure 2. Pupil responses to test questions for the innocent group ($N = 12$).

meaningful interaction with any other factor. To simplify the analysis and presentation of results, the data were pooled (averaged) over repetitions, and repetition was dropped as a factor. The Huynh–Feldt correction was applied to reduce the degrees of freedom for tests involving time.

The guilt by question type by time interaction was significant, $F(3.88, 85.44) = 2.55$, $p < .05$, $\eta^2_{\text{partial}} = .10$, as was the main effect of time, $F(5.86, 128.90) = 6.05$, $p < .05$, $\eta^2_{\text{partial}} = .22$, and the question type by time interaction, $F(3.88, 85.44) = 3.30$, $p < .05$, $\eta^2_{\text{partial}} = .13$. These effects are illustrated graphically in Figures 1 and 2. Guilty and innocent participants responded differently to comparison and relevant questions, and pupil diameter changed over time. The four-way interaction between illumination, question type, time, and guilt was marginally significant, $F(5.66, 124.45) = 2.18$, $p = .05$, $\eta^2_{\text{partial}} = .09$.

Tests of simple effects were conducted to assess the effects of question type and time for each group separately. The only significant effect for guilty participants was a main effect of time, $F(6.19, 68.05) = 3.03$, $p < .05$, $\eta^2_{\text{partial}} = .22$. Pupil diameter changed over time for guilty participants, but not as a function of question type. For innocent participants, there was a significant effect of time, $F(4.38, 48.17) = 3.48$, $p < .05$, $\eta^2_{\text{partial}} = .24$, and a significant question type by time interaction, $F(5.31, 58.36) = 5.62$, $p < .05$, $\eta^2_{\text{partial}} = .34$. As expected, innocent participants showed larger changes in pupil diameter to comparison questions than to relevant questions.

The RMANOVA revealed that pupil diameter varied as a function of guilt, question type, and time. To assess the usefulness of pupil diameter for discriminating between truthful and deceptive participants, its indices of differential activity were correlated with group membership (0 = guilty, 1 = innocent). Point-biserial correlations were obtained for peak amplitude and for area under the pupil response curve (pupil diameter area). Table 2 shows the point biserial correlations for measures of pupil dilation in the first column. The first column also shows the point biserial correlations for SC, cardiograph, and respiration.

Table 2. Correlations among the physiological measures and guilt

	Guilt	Pupil diameter amplitude	Pupil diameter area	SC amplitude	Cardiograph amplitude
Guilt	–				
Pupil diameter amplitude	.42*	–			
Pupil diameter area	.61**	.69**	–		
Skin conductance amplitude	.59**	.29	.51*	–	
Cardiograph amplitude	.18	.17	.13	.22	–
Respiration excursion	.65**	.40	.47*	.65	.21

$*p < .05$; $**p < .01$.

Pupil diameter amplitude (peak diameter) is often seen in the pupillometry literature. In the present investigation, it did not appear that differences in peak diameter were as great as differences in area under the response curve. These impressions were confirmed by the point biserial correlations for pupil diameter amplitude and area in Table 2. Since pupil diameter amplitude and area were highly correlated ($r = .69$) and the area under the response curve measure was more highly correlated with the criterion ($r = .61$), only pupil diameter area under the response curve was retained for further analyses.

A hierarchical regression analysis was performed to test if pupil diameter could be used in combination with the SC, cardiograph, and respiration measures to improve discrimination between the guilty and innocent groups. The criterion was a dichotomous variable that distinguished between guilty (coded 0) and innocent participants (coded 1). The adjusted R^2 for the combination of SC amplitude, cardiograph amplitude, and respiration excursion was .39. When pupil diameter area was added to the regression model, the adjusted R^2 increased to .46. The 7% increase in R^2 with the addition of pupil diameter approached significance, $F(1, 19) = 3.77, p = .07$.

To explore the possibility that pupil diameter could replace a traditional polygraph measure, pupil diameter area was added to each pair of traditional measures. The adjusted R^2 for these models are presented in Table 3.

There was little difference between the two models that contained pupil diameter, respiration, and either SC amplitude or cardiograph amplitude. There was no significant difference between the adjusted R^2 values for the three models. The model that accounted for the greatest proportion of variance contained pupil diameter area, SC amplitude, and respiration excursion. These preliminary findings suggest that if pupil diameter were to replace one of the traditional measures, it would probably be relative blood pressure.

Discussion

The goals of the present study were to determine if pupil diameter is diagnostic of deception and if it could be used in a comparison question test to improve prediction of guilt status. We also evaluated the possibility that pupil diameter could replace a traditional physiological measure that requires direct application of sensors to the participant. Pupil diameter was as highly correlated with deception ($r = .61$) as skin conductance ($r = .59$), and skin conductance invariably is the best traditional indicator

Table 3. Regression models and adjusted R^2 values

Model	Adjusted R^2
Pupil diameter area SC amplitude Respiration excursion	.49
Pupil diameter area Cardiograph amplitude Respiration excursion	.47
Pupil diameter area SC amplitude Cardiograph amplitude	.40

of deception in laboratory and field research on polygraph techniques (Kircher & Raskin, 2002). Adding pupil diameter to a regression equation that contained SC, cardiograph, and respiration measures increased the proportion of variance explained appreciably but not significantly.

The present findings confirm and extend the results of prior research on pupil diameter, cognitive effort, and the detection of deception. For innocent participants, pupil diameter was greater for comparison questions than for relevant questions. This result is consistent with the underlying rationale of the CQT. Innocent examinees were deceptive only to comparison questions. If deception requires more cognitive effort than being truthful and if the pupil reflects changes in cognitive effort, then increases in pupil size should be greater for comparison questions than for relevant questions. On the other hand, guilty participants did not show differential pupil responses to comparison and relevant questions. In contrast to innocent participants, guilty participants gave deceptive responses to both question types, and it may be that the pupil responses simply reflect this. The traditional physiological changes monitored by the polygraph may show greater differentiation between comparison and relevant questions for guilty people because relevant questions include a greater affective component than do comparison questions, which are relatively benign. These speculations deserve additional research.

Although effort was made to simulate a field setting, this was a laboratory mock crime experiment. Unlike in a field situation, there are no consequences for failing the polygraph other than not receiving the monetary bonus. This may be another reason why guilty subjects did not show differential pupil diameter responses to the questions types. Although the question type by time interaction was not significant for guilty subjects, the difference between pupil responses to comparison and relevant questions was highly diagnostic of group membership. Pupil size was diagnostic primarily because the innocent subjects showed substantially larger responses to comparison questions than to the relevant questions.

Our results were not consistent with those of Bradley and Janisse (1981). In their study, pupil diameter did not discriminate between guilty and innocent participants with the CQT. Although sampling variability may account for the discrepant results, the studies used different procedures to establish guilty and innocent treatment conditions, and methodological differences may affect the results obtained from laboratory mock crime experiments. It has been found that more realistic mock crimes and stronger incentives to pass the test are predictive of higher polygraph accuracy in laboratory studies (Kircher *et al.*, 1988). Bradley and Janisse instructed their guilty participants to steal one dollar and then open the door of the room where the money had been located and wait for the examiner. In contrast, our guilty participants were instructed to construct an alibi, go to another floor of the building, wait for a secretary to leave her office unattended, and steal $20. The level of involvement in the present experiment may have been greater than in the Bradley and Janisse study. Moreover, Bradley and Janisse recruited college students for their participants, whereas in the present study, participants were recruited from the general community for pay and were motivated to pass the test by the promise of a substantial monetary bonus. Half of the participants in the Bradley and Janisse (1981) study were motivated to pass the test by the threat of an electric shock. They were told they would receive a painful electric shock if deemed guilty, although no one actually received such a shock. Motivation to avoid a painful electric shock and motivation to obtain a monetary bonus may be different. Most of the participants in the present study were unfamiliar with the university setting and had no planned contact with anyone except the victim before they arrived at the laboratory for their polygraph examination. Community samples are more representative of the target population in terms of age, education, and life experience than are college students, and community samples tend to show larger effects (Kircher *et al.*, 1988).

It should be noted that although the present results are very suggestive that pupil diameter in this context indexes cognitive load, changes in pupil diameter also are sometimes associated with emotional arousal (Stern, Ray, & Quigley, 2001), and, as noted above, emotional arousal plays a major role in some theoretical discussions of polygraph techniques (Ben-Shakhar & Furedy, 1990; Handler & Honts, 2008a, 2008b; Kircher, 1981; Podlesny & Raskin, 1977; Raskin, 1979). The present data do not unambiguously indicate if the observed pupil responses reflected affective or cognitive processes.

Our results also suggest that it might be possible to replace the relative blood pressure with pupil diameter without sacrificing accuracy. A replacement for the relative blood pressure would be useful because the cuff becomes uncomfortable for some subjects if it is inflated for more than a few minutes. Moreover, the use of an inflated cuff limits the number of questions that may be asked before it is deflated (Podlesny & Kircher, 1999). However, before a traditional measure with considerable prior empirical support is replaced with a new one, the results should be replicated in other laboratories and in field settings.

One limitation of the present study concerns the device used to measure pupil diameter. The goggles could have been distracting or uncomfortable for some participants. The experimental design did not permit a test of the effect of the goggles on other physiological measures. Remote eye-tracking instruments have been used to measure pupil diameter unobtrusively (e.g. Bernhardt, Dabbs, & Riad, 1996), and with improved technology might be used in place of measures that require contact sensors, such as the cardiograph or SC. Additionally, remote eye-tracking devices can track eye-movements as well as pupil size as a participant reads text or views images on a computer screen. Several new techniques that use oculomotor measures of eye position and pupil size to detect deception have been reported (Marchak, 2006; Webb *et al.*, 2006).

Because the sample size was small, efforts were made to minimize potential sources of variance in results by testing only males. Further research is needed to determine if similar effects are obtained from females. Additionally, the sample size was small, and the power to detect an improvement in classification accuracy with the addition of pupil diameter was limited. A large sample of participants should be used to reassess the possibility that pupil diameter adds to a combination of optimally weighted traditional measures.

Three other issues deserve mention. As mentioned previously, skin conductance typically is the best traditional indicator of deception. In the present study, the correlation between respiration and guilt was higher than the correlations between the other measures and guilt. This rarely is seen in the laboratory or the field and is likely due to sampling variability. As noted in the Methods, the polygraph examiner in this study was a male doctoral level experimental psychologist. Although this examiner was highly trained to administer examinations in our laboratory, he was not a field trained polygraph examiner and this could be raised a criticism of this study. We would note that recent research has failed to find significant CQT accuracy differences between an experienced field examiner and students examiners (both male and female) who had similar training to the examiner in this study (Honts *et al.*, 2009). Lastly, it also was noted that the mean length of relevant and comparison questions was different and peak pupil responses closely corresponded with the amount of time it took to ask the question. Question length was a confound in the present study, and future work could attempt to equate question length, although doing so might move the test further from typical field situations.

It also should be noted that pupil diameter may be sensitive to attempts to employ countermeasures during a polygraph examination. Attempts to use countermeasures should require cognitive effort, as evidenced by increases in pupil diameter, because participants must monitor the question sequence and employ the countermeasure at the appropriate time for it to be effective. Use of countermeasures is a concern for comparison-question and concealed information polygraph tests (Honts & Amato, 2002), even those that rely on event-related potentials (Rosenfeld, Soskins, Bosh, & Ryan, 2004). Further research is needed to determine if pupil diameter is resistant to countermeasures or if it could be used as a counter-countermeasure, and if it is as effective in the field as it is in the laboratory.

The present study provided evidence of a strong relationship between pupil size and deception that may be partially independent of traditional physiological responses. It suggests that measures of pupil size could increase the diagnostic accuracy of the CQT. Beyond that, the present study links research on the CQT to a broader literature on attention and cognitive effort. The connection to this literature may provide new insights into the psychophysiological processes that underlie the CQT.

Acknowledgements

This research was funded by a grant from the United State's Department of Defense Polygraph Institute, Fort Jackson, SC. All views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Department of Defense or U.S. Government.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*, 1289–1292.

Beatty, J., & Wagoner, B. L. (1978). Pupillometric signs of brain activation vary with level of cognitive processing. *Science*, *199*, 1216–1218.

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

Bernhardt, P. C., Dabbs, J. M., Jr, & Riad, J. K. (1996). Pupillometry system for use in social psychology. *Behavior Research Methods, Instruments, and Computers*, *28*, 61–66.

Bradley, M. T., & Janisse, M. P. (1979). Pupil size and lie detection: The effect of certainty on deception. *Psychology: A Quarterly Journal of Human Behavior*, *16*, 33–39.

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

Dionisio, D. P., Granholm, E., Hillix, W. A., & Perrine, W. F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, *38*, 205–211.

Handler, M. D., & Honts, C. R. (2008a). Psychophysiological mechanisms in deception detection: A theoretical overview. *Polygraph*, *36*, 221–232.

Handler, M. D., & Honts, C. R. (2008b). You can run, but you can't hide: A critical look at the fight or flight response in psychophysiological detection of deception. *European Polygraph*, *2*, 193–207.

Heilveil, I. (1976). Deception and pupil size. *Journal of Clinical Psychology*, *32*, 675–676.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192.

Honts, C. R., & Amato, S. L. (2002). Countermeasures. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 251–264). San Diego, CA: Academic Press.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2005). Scientific status: The case for polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony (Volume 4): Forensics 2005–2006 Edition* (pp. 571–605). Eagan, MN: Thompson West.

Honts, C. R., Reavy, R., Markowski, K., Mcbride, S., Pitman, J., & Pitman, F. (2009). *Variations in Comparison Question Test Methods Have Little Impact*. Paper accepted for presentation at the 2009 meeting of the American Psychology Law Society, San Antonio, Texas, USA.

Hyona, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psychology*, *48A*, 598–612.

Iacono, W. G., & Lykken, D. T. (2005). Scientific status: The case against polygraph tests. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *Modern scientific evidence: The law and science of expert testimony (volume 4): Forensics 2005–2006 edition* (pp. 605–655). Eagan, MN: Thompson West.

Janisse, M. P., & Bradley, M. T. (1980). Deception, information, and the pupillary response. *Perceptual and Motor Skills*, *50*, 748–750.

Just, M. A., & Carpenter, P. A. (1993). The intensity of dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585.

Kircher, J. C. (1981). *Psychophysiological processes in the detection of deception*. unpublished manuscript. Salt Lake City, UT: Department of Psychology, University of Utah.

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior*, *12*, 79–90.

Kircher, J. C., Kristjansson, S. D., Gardner, M. K., & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception*. Final report to the US Department of Defense. (Grant No. DASW01-02-1-0016). Salt Lake City, UT: University of Utah, Department of Educational Psychology.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, *73*, 291–302.

Kircher, J. C., & Raskin, D. C. (2002). Computer methods for the psychophysiological detection of deception. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 287–326). San Diego, CA: Academic Press.

Marchak, F. M. (2006). Eye movement-based assessment of concealed knowledge. *Journal of Credibility Assessment and Witness Psychology*, 7, 149–163.

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

Podlesny, J. A., & Kircher, J. C. (1999). The Finapres (volume clamp) recording method in psychophysiological detection of deception examinations: Experimental comparison with the cardiograph method. *Forensic Science Communications*, *1*(3), 1–17.

Podlesny, J. A., & Raskin, D. C. (1977). Physiological measures and the detection of deception. *Psychological Bulletin*, *84*, 782–799.

Raskin, D. C. (1979). Orienting and defensive reflexes in the detection of deception. In H. D. Kimmel, E. H. van Olst, & J. F. Orlebeke (Eds.), *The orienting reflex in humans* (pp. 587–605). Hillsdale, NJ: Lawrence Erlbaum.

Raskin, D. C., & Honts, C. R. (2002). The comparison question test. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 1–47). San Diego, CA: Academic Press.

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

Schluroff, M. (1982). Pupil responses to grammatical complexity of sentences. *Brain and Language*, *17*, 133–145.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York: Oxford University Press.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Chichester, England: Wiley.

Webb, A. K., Kristjansson, S. D., Osher, D., Cook, A. E., Kircher, J. C., Hacker, D. J., *et al.* (2006). Multimethod assessment of deception on personnel tests: Reading, writing, and response time measures. *Journal of Credibility Assessment and Witness Psychology*, 7, 164–168.

Received 16 March 2008; revised version received 7 November 2008