Effects of Direct and Indirect Questions On The Ocular-Motor Deception Test

Pooja P. Bovard[1]

John C. Kircher[2]

Dan J. Woltz[2]

Doug J. Hacker[2]

and Anne E. Cook[2]

Acknowledgments

We would like to thank an anonymous reviewer from a previous draft of this manuscript for their helpful comments.

Abstract

To discriminate between truthful and deceptive individuals, the ocular-motor deception test (ODT) makes within-subject comparisons of recorded physiological and behavioral response time. In two mock crime experiments, we tested for effects of factors that might improve the efficiency and accuracy of the ODT. In each experiment, half of the participants were guilty of stealing $20 from a secretary's wallet and the other half were innocent. Experiment 1 compared the accuracy of an ODT that directly asks if a person committed illicit acts with accuracy of an ODT that indirectly asks if the person provided false answers on a questionnaire about those illicit acts. Experiment 2 manipulated item presentation, feedback during the practice ODT, and inter-question intervals. In one presentation format, items were sequenced such that no two items of the same type appeared in succession (distributed). In the other condition, items of the same type were presented in succession (blocked).

In Experiment 1, accuracy of classifications as guilty or innocent by logistic regression were significantly higher for participants asked directly about their involvement in the crimes (83%) than for participants asked if they falsified their answers on the pre-test questionnaire (60%). In Experiment 2, 86% and 83% of participants in the distributed and blocked conditions were correctly classified, respectively. Feedback during practice and differences in interval-event intervals had no discernible effects on ocular-motor measures. The results suggest that the ODT should stimulate the individual emotionally with direct questions about illicit behaviors, and cognitively or attentionally with unpredictable transitions between question types.

Keywords: Ocular-motor, deception detection, eye tracking, reading

[1] Draper, 555 Technology Square, Cambridge, MA 02139

Corresponding author and person to whom reprint requests should be addressed: Pooja Bovard, PhD Draper 555 Technology Square, MS 31 Cambridge, MA 02139 Email: ppatnaik@draper.com Phone: 617-258-2091

[2] University of Utah, 1721 Campus Center Drive, Salt Lake City, UT 84112 ppatnaik@draper.com, dan.woltz@utah.edu, doug.hacker@utah.edu, anne.cook@utah.edu, john.kircher@utah.edu

Introduction

Zuckerman, DePaulo, and Rosenthal (1981, 1986) proposed a four factor theory which posits that changes in deceivers' behavior are the result of four physiological processes: physiological arousal, emotional reactions, cognitive effort, and attempted control. Cook et al. (2012) introduced an automated deception detection technique called the ocular-motor deception test (ODT) that derives an index of deception from measures of physiological and emotional arousal, cognitive effort, and attempted control. The ODT is completely automated, and can be administered in approximately 40 minutes without the need for an adversarial interview process. A computer presents voice-synthesized and written instructions, after which examinees will read a series of true/false test statements concerning possible involvement in illicit activities. The instructions inform examinees that if they do not answer quickly and accurately, they will fail the test. The examinee then reads statements presented serially by the computer while a remote eye tracker records eye movements and changes in pupil size. The examinee presses a key on the keyboard to answer true or false. The computer processes the ocular-motor and behavioral data (response time and errors), combines its measurements in a logistic regression equation, and classifies the individual as truthful or deceptive on the test.

The ODT uses a test format known as the Relevant Comparison Test (RCT). The RCT originally was developed as a new polygraph technique for use at ports of entry to screen travelers for trafficking of drugs and transporting explosives (Kircher et al., 2012). The RCT contains questions about two relevant issues (R1 and R2) that are intermixed with neutral questions. The test uses the difference between reactions to the two sets of relevant questions to decide whether the examinee was truthful or deceptive to one or the other relevant issue. Each relevant issue serves as a control for the other. If the examinee reacts more strongly to one set of relevant questions, the computer classifies the individual as deceptive to that issue. In both Experiments 1 and 2, the deceptive issue involved questions about cash. If the examinee responds similarly to the two sets of relevant questions, the computer classifies the person as truthful to both issues. The irrelevant crime questions were about taking an exam from a professor, which was a crime that no one committed. The original RCT covered two relevant issues that were mutually exclusive, such that if the person was deceptive to one issue (transporting drugs), he or she would be truthful to the other (intention to detonate a bomb on an aircraft). The RCT also might compare two relevant issues, where the consequences of failure on one issue, such as espionage, are considerably greater than the consequences of failure to the other issue (e.g., recent drug use).

The ODT is based on the assumption that deception is cognitively more demanding than telling the truth (Johnson, Barnhardt, & Zhu, 2005; Kircher, 1981; Steller, 1989; Vrij, Fisher, Mann, & Leal, 2006). While taking a test for deception, truthful people interpret the questions and then give the appropriate answers. In addition to these tasks, deceptive individuals also must distinguish between questions answered truthfully and deceptively. When they encounter an incriminating statement, they must differentially inhibit the prepotent truthful answer to execute a deceptive one. Deceptive individuals also may attempt to monitor their behavior and the environment during the test to assure themselves that they are not revealing their guilt, for example, by answering too slowly or making too many mistakes. The recruitment of resources to accomplish the additional cognitive and meta-cognitive activities could contribute to the observed effects on autonomic, somatic, and behavioral measures (Hacker et al., 2014; Kahneman, 1973).

The ODT also assumes that deception is associated with emotional arousal. In a personnel screening setting, examinees may believe that they will be subject to adverse decisions or undesirable administrative action not be hired if they fail the deception test. In these contexts, questions answered deceptively might pose threats to the individual and evoke defensive psychophysiological responses. This possibility is consistent with findings that large increases in pupil size are associated with deception during polygraph tests (Bradley & Janisse, 1979; Dionisio et al., 2001; Janisse & Bradley, 1980; Webb et al., 2009).

In the psychology of reading literature, frequent fixations, short inter saccade distances, and long reading times are indications that participants had difficulty processing those items (Rayner, 1998; Rayner, Chace, Slattery, & Ashby, 2006). If deception is more difficult than being truthful, then deception should affect reading patterns. In the Cook et al. experiments, effects were found on reading measures, but they were not the effects that were expected. Within-subject contrasts revealed that deception to questions about one relevant issue (R1) was associated with fewer fixations and shorter reading and rereading times than being truthful to the questions about the other relevant issue (R2). We concluded that guilty participants, to avoid detection, made a concerted effort to spend as little time on the incriminating R1 items as possible. Guilty participants achieved their objective, but in so doing, revealed their deception. This finding is consistent with other evidence that participants can exert some conscious control over their reading behaviors to implement specific reading strategies (Hyona & Nurminen, 2006). We obtained this finding in the two experiments reported by Cook et al., and in a subsequent study by Patnaik et al. (2016).

Experiment 1

In prior studies on the ODT, the test statements directly addressed the participant's possible involvement in each of two crimes (Cook et al., 2012; Patnaik et al., 2016). However, an ODT that asks directly about the person's involvement in a specific incident has limited generalizability. A more general approach would be to administer a short pre-test questionnaire that covers the relevant issues of concern, and then conduct a generic ODT that asks if the participant falsified information on the questionnaire. All of the items on the ODT would remain the same regardless of the particular application; only the pre-test questionnaire would change from one application to another.

In addition to answering a practical question about the possibility of developing a single general-purpose ODT, Experiment 1 also addressed a theoretical question. Since stronger emotions are more likely associated with the commission of a crime than the falsification of an answer on a pre-test questionnaire, we predicted that guilty participants would react more strongly to statements about the crime than to statements about their answers on a pre-test questionnaire.

Method

Design

Participants were randomly assigned to one of six groups: guilt with two levels (guilty or innocent) and protocol with three levels (1. indirect ODT statements with pre-ODT questionnaire, 2. direct ODT statements with pre-ODT questionnaire, or 3. direct ODT statements with no pre-ODT questionnaire). To test whether the pretest affected the accuracy of the ODT independently of the questions included on the ODT itself, the pre-ODT questionnaire was administered to half of the participants who received direct items.

The design also included two within-subject factors: statement type (neutral, cash, and exam) and repetition (5 repetitions of the ODT test items). In some analyses of pupil diameter, time with 40 levels (10 Hz samples x 4 seconds) also was included as a within-subjects variable.

Participants

One hundred nine participants were recruited via flyers on campus from an urban university in the western United States. The flyers offered $30 in pay and an opportunity to earn an additional $30 bonus. Of these 109 participants, five chose not to participate after learning their experimental condition, six did not follow instructions, and two produced inadequate recordings. The remaining 96 participants ranged in age from 18 to 68 years (M=23.79, *SD*=8.88), were predominantly Caucasian (67%), single (80%), full time students at the university (83%) with English as their primary language (87%). Forty-eight participants received indirect statements with a pre-ODT questionnaire, 24 received direct statements with a pre-ODT questionnaire, and 24 received direct statements with no pre-ODT questionnaire.

Apparatus

A ViewPoint EyeFrame Monocular Nystagmus System eye tracker (Arrington Research, Scottsdale, AZ) was used to record eye movements and pupil diameter at 30Hz. The eye tracker was affixed to a pair of lens-less plastic goggles. Viewing was binocular, but eye movement and pupil diameter were recorded from only the right eye. A computer presented instructions and test items to the participant on a 19-inch Dell flat screen LCD monitor with a 5:4 aspect ratio. The monitor was positioned approximately 60 cm from the participant's eyes.

Ocular-motor Deception Test

Test items were presented to the participant in black font on a pale gray background. Participants answered 15 practice items followed by 48 test items, and these same 48 items were presented five times in different orders. Sixteen items pertained to the theft of the $20 (direct- "I had nothing to do with the theft of the $20"; indirect- "I answered truthfully that I was uninvolved in the theft of the $20"), 16 pertained to the theft of the exam (direct- "I took nothing from the professor's office"; indirect- "I correctly reported that I took nothing from the professor's office"), and 16 were neutral items ("I was born prior to the year 2000"). The items were randomized subject to the constraint that no two items from the same category appeared in succession. The correct (non-incriminating) answer was true for 8 of 16 items in a category and false for the remaining 8 items in the category.

Procedures

Participants reported alone to a room in a building on campus. Instructions in an envelope taped to the door instructed the participant to enter the room, read and sign the consent form, and then listen to an audio recording for their instructions. A hard copy of the recorded instructions was included as well. A phone number was provided for participants to call if they did not wish to participate.

Half of the participants were in the guilty condition. Guilty participants were instructed to go to a secretary's office and ask the secretary where Dr. Mitchell's office was located. The secretary (a confederate) informed the participant that there was no Dr. Mitchell in the building, and the participant left. The participant was told to wait inconspicuously for the secretary to leave her office unattended, then enter her office, find her purse, remove $20 from a wallet in the purse, and conceal the money on their person. Participants were told to prepare an alibi in case they were caught and to leave no fingerprints. They were informed that they had no more than 20 min to commit the crime and report to the experimenter (Podlesny & Raskin, 1978). and report to the experimenter (Podlesny & Raskin, 1978).

Half of the participants were in the innocent condition. They were told that some participants had to steal an exam or money, but that they were innocent participants and should not steal anything. Innocent participants were instructed to wait approximately 20 min before reporting to the experimenter.

All participants also were informed that there was another crime in which some participants had to download an exam from a professor's computer onto a disk. In actuality, no one committed that crime.

Participants reported to the experimenter after committing their crime or after the 20 min waiting period. Participants assigned to a pre-ODT questionnaire condition completed the two-question questionnaire that asked (1) if they took the exam, and (2) if they took the money. Guilty participants were instructed to lie on this questionnaire to appear truthful (as if they did not take the money). The participants were fitted with the Arrington eye tracker, calibrated to the eye tracker, and administered the ODT.

After completing the tasks, participants were paid $30 and were given an additional $30 bonus if the computer determined they had passed the test.

Dependent Measures

Behavioral Outcome Measures. *Response time (RT)* was the time in ms from the appearance of the item on the screen to a button press by the participant. To control for differences in item length, RT was divided by the number of characters in the statement.

Proportion wrong for a particular statement type (neutral, cash, exam) was the number of incorrect responses divided by the number of items (16 X 5= 80).

Ocular-Motor Outcome Measures. An area of interest (AOI) was defined for each T/F test item. The AOI began with the first character of the item and ended at the period at the end of the statement. Ocular-motor reading measures were computed for the fixations in each AOI divided by the number of characters in the statement. Fixations were determined from the data files produced by the Arrington eye tracker by identifying a sequence of samples in which the eye shows little movement for at least 100 ms (ASL, 2001). Fixations longer than 1000 ms were considered artifacts and were discarded (Rayner, 1998).

Number of fixations was the number of fixations detected in an AOI.

First pass duration was the sum of all fixation durations in an AOI before the eye fixated outside the AOI.

Reread duration was the sum of fixation durations associated with all leftward eye movements in the AOI, regardless of whether the eye ever fixated outside the AOI.

PD response curve was the change in pupil diameter in mm from statement onset for a period of 4 seconds.

Area under the pupil response curve (PD Area) was obtained by identifying the times and levels of high and low points in the response curve for a 4-second window that began at statement onset. The computer generated a diagonal matrix of differences between each low point and every subsequent high point. Peak amplitude was the greatest obtained difference, and response onset was defined as the low point from which peak amplitude was measured. PD Area was the area under the curve from response onset to the point at which the response returned to the initial level or to the end of the 4-second sampling interval, whichever occurred first.

The 30 Hz PD data samples from the beginning of a block of 48 test items to the end of that block of items were converted to z-scores (standardized) within participants. *PDLevel at T/F response* was the mean of z-scores within +/-1 second of the participant's true or false answer.

Blink rate was the number of blinks per second. *Item blink rate* was computed for each item for 1.5 s immediately preceding the answer. Blink rate also was computed for a period of 1.5 s that began at the participant's answer (*next item blink rate*).

Results

Significance tests involving within-subject factors used Huynh-Feldt corrections to degrees of freedom. An alpha level of .05 was applied for all statistical tests.

Preliminary Test for Effects of the Pretest Questionnaire

For half of the participants, the relevant issue on the ODT was whether the participant had committed the mock crime (direct). For the remaining participants, the relevant issue was whether the participant had falsified answers on the pre-ODT questionnaire (indirect). The primary goal of the experiment was to determine if the type of relevant issue affected the accuracy of the ODT. Prior to testing for effects of relevant issue, we compared groups that received direct statements on the ODT and either did or did not complete the pre-ODT questionnaire. As expected, completion of the pre-ODT questionnaire did not interact with guilt for any of the outcome measures (all $p > .05$). Therefore, the questionnaire/no questionnaire groups that received direct questions were combined, and the presence or absence of pre-ODT questionnaires was dropped as a factor. Pooling groups balanced the cell sizes for subsequent comparisons of direct and indirect question types.

Repeated measures analysis of variance (RMANOVA) was used to analyze each dependent variable. Only main effects of guilt and interactions with guilt are discussed here.

Pupil Diameter. PD was assessed by computing change from the baseline of statement onset. The first data point was subtracted from every subsequent data point in the response curve. A positive value indicated PD increased relative to the initial value, and a negative value indicated PD decreased relative to the initial value.

PD response curves are presented in Figures 1a and 1b for innocent and guilty participants. Innocent participants showed little difference between responses to cash and exam statements (Figure 1a), whereas guilty participants reacted more strongly to statements about the theft of the $20 (Figure 1b). However, neither innocent nor guilty participants who received indirect statements reacted differentially to cash and exam items. The Guilt X Statement type interaction was significant, $F(1.95,179.62) = 11.12$, $p<.05$, partial $\eta^2 =.108$. The effect of the Guilt X Statement type X Relevant issue interaction also was significant, $F(1.95,179.62)= 3.25$, $p<.05$, partial $\eta^2= .034$. The Guilt x Statement type interaction was significant for those who received direct statements, $F(2,92)= 14.45$, $p<.01$, partial $\eta^2= .239$, but not for those who received indirect statements, $F(1.93,88.98)= 2.73$, $p < .08$.

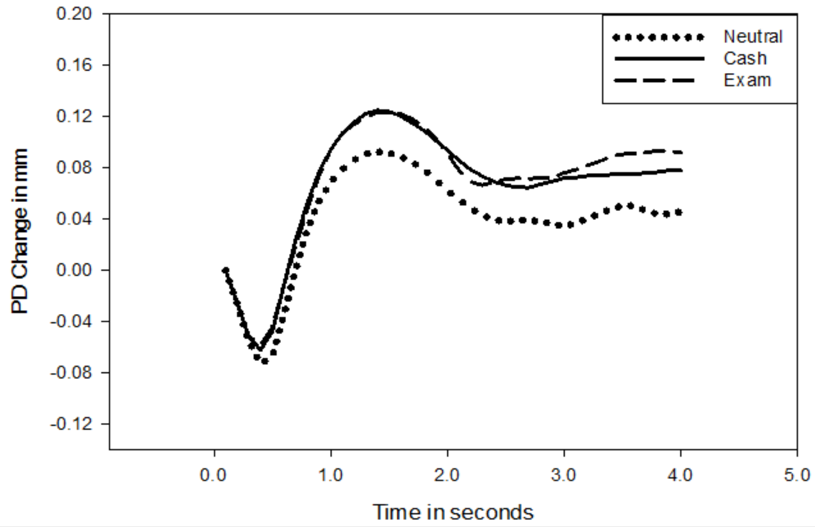
Predictive Validity of Ocular-motor Measures

Between-statement type contrasts were generated to assess the extent to which the ocular-motor measures could be used to distinguish between the groups. CashExam was the difference between the person mean for cash items and the person mean for exam items, which controlled for the perceived relevance of test items. The contrast was derived for each behavioral and ocular-motor variable (Table 1).

To assess the diagnostic validity of an outcome measure, it was correlated with a dichotomous variable that distinguished between innocent (coded 0) and guilty participants (coded 1).

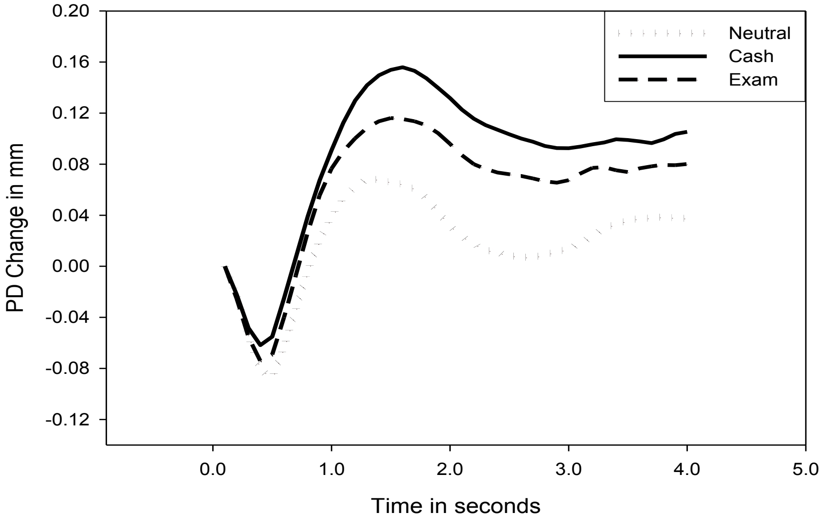
To assess the diagnostic validity of an

Figure 1a. Pupil response to neutral, cash, and exam items for innocent participants.



1a. Pupil response to neutral, cash, and exam items for innocent participants.

Figure 1b. Pupil response to neutral, cash, and exam items for innocent participants.



1b. Pupil response to neutral, cash, and exam items for guilty participants.

outcome measure, it was correlated with a dichotomous variable that distinguished between innocent (coded 0) and guilty participants (coded 1).

Table 1. Point-Biserial Correlations for Direct and Indirect Relevant Issues

Outcome Measure	Relevant Issue	
	Direct	Indirect
RTCashExam	-.311*	-.281
PropWrongCashExam	-.311*	-.281
NfixCashExam	-.402**	-.212
FirstPassCashExam	-.160	-.115
RereadCashExam	-.364*	-.177
PDAreaCashExam[a]	.684**	.268
PDLevelCashExam[a]	.649**	.144
ItemBlinkRateCashExam	.000	.011
NextItemBlinkRateCashExam	.223	-.279

*p < .05, **p <.01. a significant difference between the two correlation coefficients.

Note. RT = response time per character, PropWrong = proportion wrong, NFix = number of fixations per character, FirstPass = time spend reading per character, Reread = time spent rereading per character, PDArea = pupil diameter area under the curve, ItemBlinkRate= number of blinks per second on each item type, NextItemBlinkRate = number of blinks per second on the item following each item type.

The negative point-biserial correlations for RT, proportion wrong, and number of fixations between the relevant crimes indicate that guilty participants took less time to respond, made fewer mistakes, and made fewer fixations on cash items than exam items. The negative correlation for the reread Cash versus Exam contrast indicates that guilty participants did less rereading of cash items than exam items. The correlations for the Cash versus Exam contrasts were positive for PD area and PD level, which indicate that guilty participants showed greater increases in pupil size in response to relevant items than did innocent participants (Table 1).

Previously, ocular-motor data from participants who participated in ODT mock crime experiments in the U.S and Mexico were used to develop a binary logistic regression model to classify participants as truthful or deceptive (Patnaik et al., 2016) . That model included between-question-type differences in RT and PDLevel. In the present study, the Patnaik et al. model correctly classified 83% of direct participants (false positive= 17%; false negative= 17%) and 60% of indirect participants (false positive = 21%; false negative = 58%). Indirect questions produced over three times as many false negatives as did direct questions. The difference in accuracy between direct (83%) and indirect methods (60%) was significant, Yates' $X^2(1) = 5.15$, $p<.05$.

Discussion

The accuracy of an ODT that asks directly if the person committed illicit acts was greater than the accuracy of an ODT that indirectly asks if the person provided false information about those illicit acts on a pre-ODT questionnaire. The differences between cash and exam items were more diagnostic for participants asked about their involvement in the crime than for participants asked about their answers on a questionnaire. The results obtained with direct items were not only stronger than those obtained with indirect items but also more consistent with the rationale that underlies the RCT. Theoretically, the difference between crime-related items should be more diagnostic than the difference between crime-related and neutral items.

Why would indirect items be less effective than direct statements? A participant who lied on the questionnaire wrote "No" to one question on a form. Guilty participants may have been focused on denying culpability about the crime rather than their answer on the questionnaire. Writing "No" on the questionnaire was only the last of a series of illicit behaviors, and it may have been the least emotionally arousing because it posed relatively little risk of discovery. When asked about their answers on the questionnaire during the ODT, guilty participants may have been relieved that they were not asked if they had committed the crime.

The direct statements evoke an episodic memory of stealing with all of the attendant detail and possible emotion of the actual experience, which could account for the observed differences between the groups that received direct and indirect statements. The recall of that episodic memory makes the denial of the truth more difficult and increases cognitive load. Responding to the indirect statement is less likely to evoke a detailed and complex episodic memory, since all they did was mark a question wrong on the questionnaire.

Differences in the semantic complexity of items on the two forms of the ODT also might account for the effects on diagnostic validity. The relevant issue for a direct statement referred to the commission of a particular crime (an action). The relevant issue for an indirect statement referred to falsifying information on a questionnaire (one action) concerning the crime (another action). To answer an indirect statement correctly, the participant had to retain information concerning their possible involvement in the crime and how they responded on the questionnaire. Guilty participants had the added burden of distinguishing between items answered truthfully and items answered deceptively. If there was a ceiling effect for guilty participants, the additional burden of item complexity might raise the load on innocent participants and reduce the difference between guilty and innocent participants. This possibility is consistent with the finding that item difficulty adversely influenced the diagnostic validity of reading measures in an experiment reported by Cook et al. (2012).

There may be greater social stigma associated with lying about committing a theft than lying on a questionnaire. Five participants withdrew from the study upon learning they had to steal $20 from a secretary's wallet, and six participants chose not to steal the money but showed up for the ODT anyway. No one refused to lie on the questionnaire. Although social stigma could account for the difference in withdrawal rates, a selection artifact also could account the difference since only participants who had already agreed to commit the crime had the option to lie on the questionnaire (Shadish, Cook, & Campbell, 2002).

Finally, these findings may have generalizable implications in credibility assessment testing using traditional polygraph instrumentation and test formats. Some polygraph examiners use written statements about a crime as the focus of the polygraph test. On the polygraph test, examinees are not asked directly if they committed some illicit act; rather, they are asked if they falsified their statement about the illicit act. To our knowledge, this is the first research that has addressed this issue in any credibility assessment venue. Although polygraph instrumentation and techniques differ from the ODT, the present findings have implications for polygraph testing to the extent that the same physiological arousal, emotional reactions, cognitive effort, and attempted control underlie the ODT and traditional polygraph approaches.

Experiment 2

The results of Experiment 1 suggest that emotional arousal plays a role in facilitating discrimination between truthful and deceptive individuals. Participants asked about their involvement in a mock crime were more readily identified as truthful or deceptive than participants asked if they had falsified answers on a pre-test questionnaire about the crimes. In Experiment 2, we changed the format of the ODT in an attempt to capitalize on effects of emotion on ocular-motor measures by comparing blocked and distributed presentations of questions concerning the same issue.

The rapid presentation (Experiment 1 inter-event interval was 500 ms) of test items that vary in content may interfere with the development of large pupil responses when the person is deceptive. In the blocked design, all activity that takes place during a series of question of the same type could contribute to a single protracted physiological reaction, whereas the distributed condition may interrupt the development of a sustained response because each item is followed by another item of a different type. One benefit of a blocked design is that phasic reactions to individual questions may be investigated as well as more global activity in the blocked set (Visscher et al., 2003).

Changes in item content for every test item also may counteract attempts by deceptive people to implement reading strategies to defeat the test, and use of strategies may be diagnostic (Hacker et al., 2014). On the other hand, if blocks rather than individual statements serve as the unit of analysis, the number of 'items' on the ODT would be reduced and that could adversely affect the reliability and validity of pupil measures. Experiment 2 tested if the potential benefits of blocking outweigh the cost of reducing the number of items.

Experiment 2 also manipulated the feedback the computer provided to participants following a set of practice items. Although feedback might encourage participants to minimize response errors on the ODT (Adams & Goetz, 1973), the error rates in student samples already are less than 10%. Feedback might not reduce participants' response errors, but it could result in anchoring. Anchoring is the tendency to use initial information to establish a standard against which subsequent performance is evaluated. Response time and accuracy feedback during a practice session should serve to establish high expectations about subsequent performance on the ODT. If anchoring causes participants, especially innocent participants, to respond quickly and consistently, it might reduce variance within and between participants, increase the signal to noise ratio, and improve decision accuracy.

Webb et al. (2009) found that pupil responses during a polygraph examination can last 10 or 12 seconds. During an ODT, a computer presents the next test statement 500 ms following the participant's answer. In light of the Webb et al. results, there is a possibility that the rapid onset of an item soon after the person answers the prior item interrupts a psychophysiological process that attenuates the participant's reactions to test statements. The current brief inter-event interval may not allow sufficient time for the pupil response to reach its maximum and recover. The present study assessed the effects on pupil reactions of longer inter-event intervals.

A longer inter-event interval, during which the participant recovers from the prior event and prepares for the next, also might facilitate efforts to develop a diagnostic measure of eye blink rate. Prior research indicates that deception is associated with fewer eye blinks followed by an increase in blink rate when the deception is complete (Leal & Vrij, 2008; Marchak, 2013). Cook et al. (2012) observed a similar pattern for the ODT, but the effect sizes were small compared to those reported by Leal and Vrij (2008). Lengthening the inter-event interval might improve the reliability and usefulness of post-answer blink rates.

In contrast to prior mock crime studies of the ODT, for Experiment 2, we recruited participants from the general community rather than the university. A community sample may be more heterogeneous with respect to age, intelligence, and educational background and may better represent a more general target population than a sample that consists of only college students.

To summarize, in Experiment 2, we manipulated presentation format (distributed versus blocked), feedback following a pre-ODT practice session, and the interval between the examinee's answer and the presentation of the next test statement, and we recruited participants from the general community.

Methods

Design and Analysis

We used a mixed design with three between-group factors and three within-subject factors. The between-group factors were guilt with two levels (guilty or innocent), presentation format (distributed or blocked), and feedback (practice with or without performance feedback). The within-subject factors were statement type (neutral, cash, credit card), inter-event interval (500 ms, 1500 ms, and 3000 ms), and repetition (2 repetitions of the items at each of the three inter-event intervals). Twenty participants were randomly assigned to each treatment combination of guilt, presentation format, and feedback (N=160). Power analysis indicated that a sample of 160 participants was sufficient to detect medium effects on outcome measures with a probability of at least .80.

Participants

Recruitment ads were posted on KSL (Salt Lake City, Utah), Craigslist, and City Weekly online and print that advertised an opportunity to earn $30 and a possible bonus of $30 for participation in a psychological experiment. Two hundred and eighty-five people were given appointments, and 178 arrived to participate in the study. Of these 178 people, five chose not to participate after learning their experimental condition, three did not follow instructions, and 10 had inadequate data. The mean age of the remaining 160 participants was 33.6 years (SD= 12.99). Males comprised 53% of the sample, and 78% self-identified as Caucasian. Education levels ranged from some high school to graduate degree with some college as the median level of education.

Apparatus

A SensoMotoric Instruments (SMI) RED-m remote eye tracker affixed to a 19-inch 5:4 Dell flat screen monitor recorded eye movements and pupil diameter at 60 Hz. Viewing was binocular, and although the eye tracker allowed for free head movement, a chin rest was used to keep the participant's head still. The computer monitor was 65 centimeters from the participant's eyes. A floor lamp provided 5.57 lumens of light reflected off the ceiling measured at eye level facing the computer monitor.

Presentation Format

For the blocked presentation format, the computer presented four items of the same type in succession. In addition to analyses of individual items, the four statements in a block were treated as a single unit. As a result, for PD Area, PD Level, and blink rates, the onset of the first item of the block was identified as block onset, and PD Area, PD Level, and Blink rate were analyzed from 0 to 12s following block onset.

In the blocked condition, four items of the same type (e.g., neutral) were presented in succession, followed by four items of a different type (e.g., cash). Before each blocked set of four items, a text message appeared for 3500 ms and informed the participants of the issue covered in the next set of items. For each participant, this process was repeated four times for each statement type in each of six sessions (two sessions at each of three inter-event intervals). In the distributed condition, items were distributed randomly with the stipulation that no two items of the same type appeared in succession.

Practice and Feedback

Before the ODT, participants in the no-feedback condition answered 12 practice items twice in different orders. Participants in the feedback condition answered 12 practice statements twice in different orders and were given feedback about their accuracy and response times after each repetition. If the participant took longer than five seconds to answer True or False to a statement, a "Time Out!" screen would appear, and the question was counted as an incorrect answer. The practice items included statements about crimes that were unrelated to the issues covered on the ODT.

Ocular-motor Deception Test (ODT)

The ODT consisted of 48 test statements that were similar to those in the direct condition in Experiment 1, and these same 48 statements were presented six times using either the distributed or blocked presentation format. The statements were presented in the same manner as in Experiment 1, and participants used handheld push buttons to answer True or False.

Procedures

The procedures were the same as Experiment 1 with the following exceptions. Participants were recruited from the community and called in response to ads placed in the community. Participants did not complete a pre-ODT questionnaire. All participants were informed that there was another crime in which some participants had to download credit card information from a professor's computer onto a USB flash drive, but in actuality, no one committed that crime. The comparison crime was changed from questions about stealing an exam in Experiment 1 to stealing credit card information in Experiment 2. Before participants were informed of the decision, they completed a questionnaire to assess their subjective experiences during the experiment. Finally, except for the additional block-level measures of change in pupil size and blink rates, all of the ocular-motor measures in Experiment 2 were the same as those in Experiment 1.

Results

Presentation Format

Of interest were Guilt X Statement type X Presentation format interactions. For reread duration, the Guilt X Statement type X Presentation format was significant, $F(2, 252) = 3.62$, $p<.05$, partial $\eta^2 = .028$. Presentation format had little effect on guilty participants. In contrast, innocent participants spent more time rereading cash and card items than neutral items in the blocked condition as compared to the distributed condition.

For PD waveform, the Guilt X Statement type X Presentation Format interaction was significant, $F(2, 256) = 4.06$, $p<.05$, partial $\eta^2 = .031$ and is illustrated in Figures 2a, 2b, 2c, and 2d. The RCT predicted that guilty participants would react more strongly to statements about the cash than the credit card. The expected difference was observed in the distributed condition but not the blocked condition.

For area under the pupil response curve, the Guilt X Statement type X Presentation format interaction was significant, $F(2, 288) = 5.64$, $p<.05$, partial $\eta^2 = .038$. Consistent with the analysis of the evoked pupil response curve, the guilty distributed group showed stronger pupil responses to cash than credit card statements, whereas guilty blocked participants showed little difference in their pupil responses to cash and credit card statements.

The Guilt X Statement type X Presentation format interaction was significant for PD level, *F*(2, 256) = 5.15, p<.05, partial η^2 = .039. As compared to innocent participants in the distributed condition, innocent participants in the blocked condition reacted less strongly to neutral statements. There was little difference between guilty distributed and guilty blocked participants in their reactions to neutral, cash, and credit card statements.

The Guilt X Statement type X Presentation Format also was significant for item blink rate, *F*(2, 254) = 3.42, p<.05, partial η^2 = .026. As compared to guilty participants in the distributed condition, guilty participants in the blocked condition blinked less often while reading cash statements than neutral and card statements.

Block as the Unit of Analysis

Figures 2c and 2d present the changes in pupil size over the 4 sec interval that began at the onset of the first of four statements of the same type. The pupil dilated in response to cash and card item over the first four seconds by more than 0.10 mm and then slowly recovered. The pupil was more dilated while guilty participants read and responded to cash items than to credit card or neutral items, whereas the opposite pattern was observed for innocent participants. The Guilt X Statement type X Time, F(14.49, 1129.91) = 1.44, p<.05, partial η^2 = .018, and Guilt X Statement Type interactions were significant, F(1.56, 121.80) = 6.35, p<.05, partial η^2 = .075. The observed differences between guilty and innocent groups did not vary significantly by Presentation format (all *p* > .05).

Figure 2a. Pupil response to neutral, cash, and card items for distributed format for innocent participants.

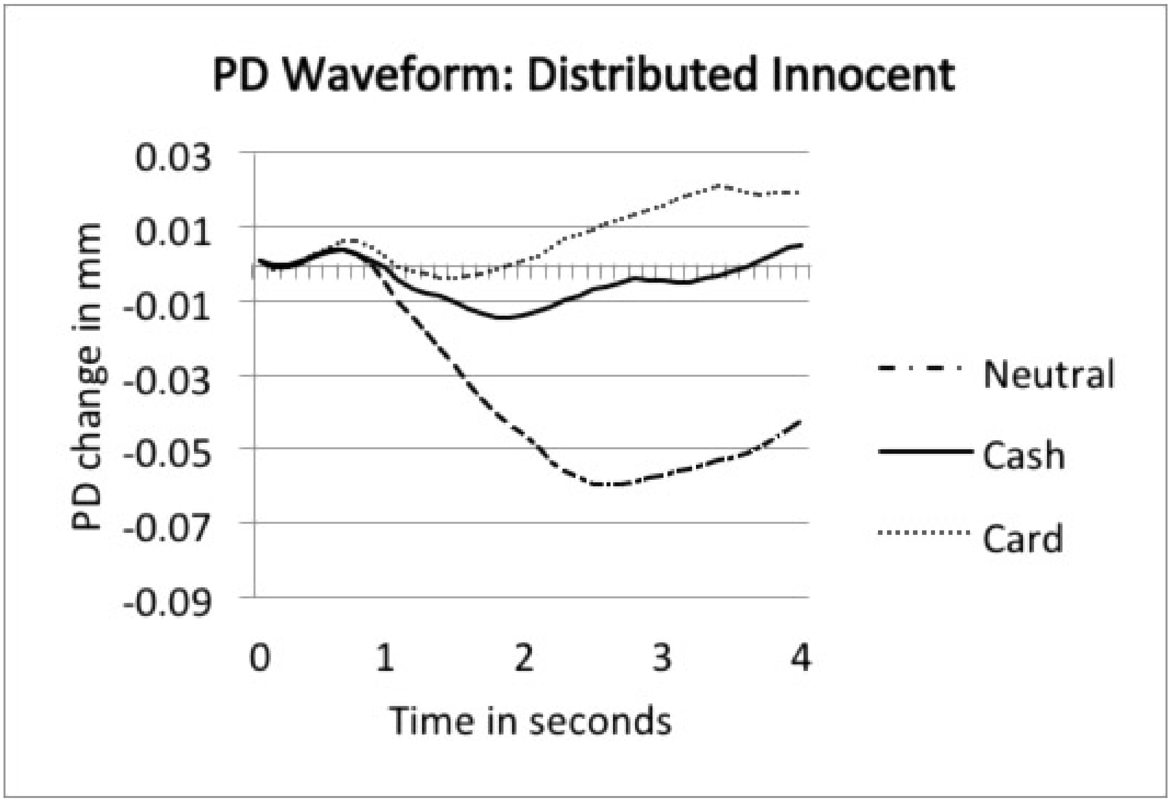


Figure 2b. Pupil response to neutral, cash, and card items for distributed format for guilty participants.

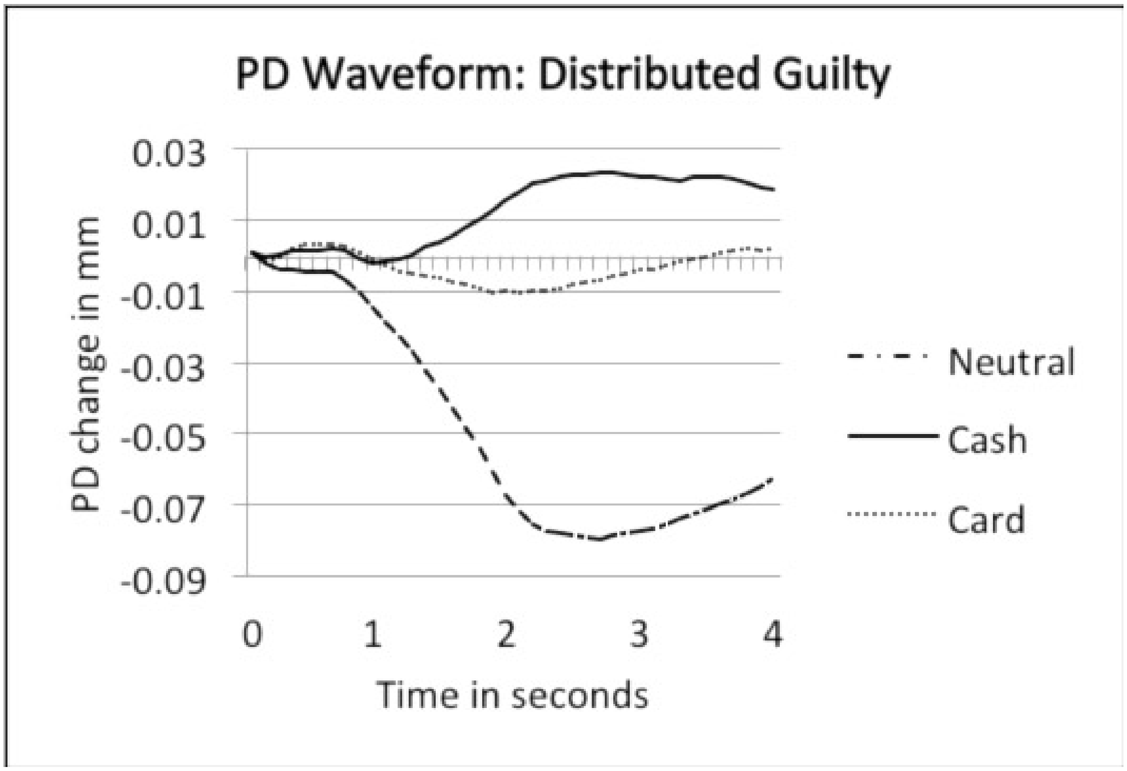


Figure 2c. Pupil response to neutral, cash, and card items for blocked format for innocent participants.

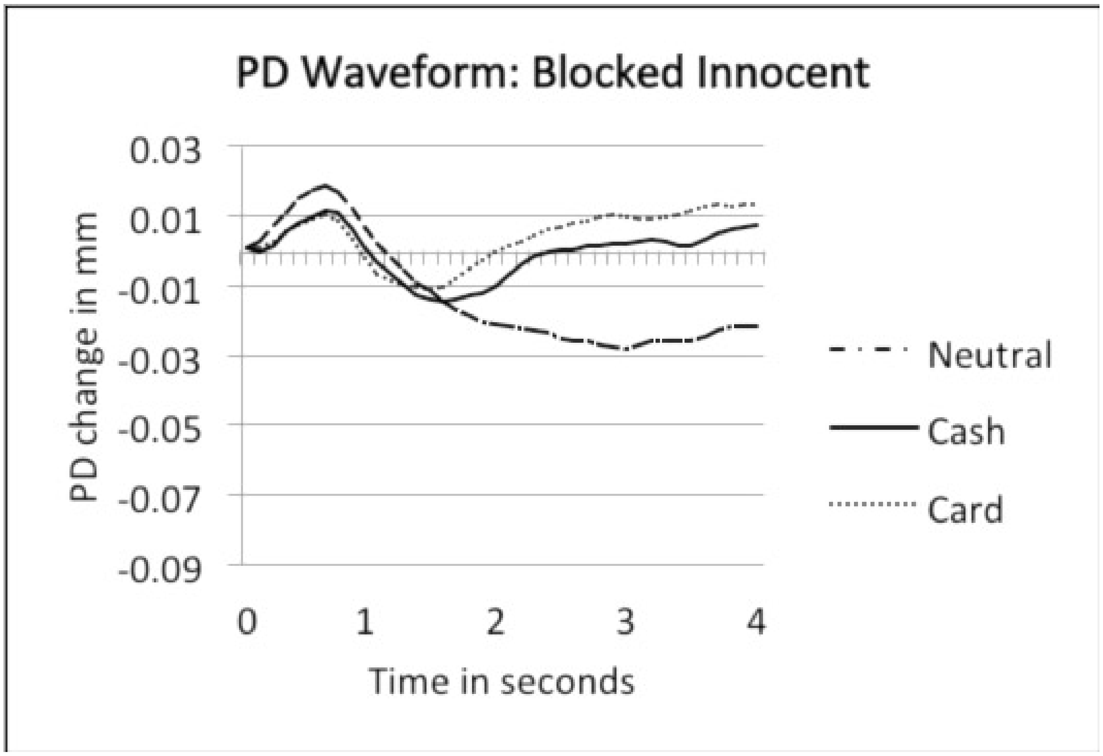


Figure 2d. Pupil response to neutral, cash, and card items for blocked format for guilty participants.

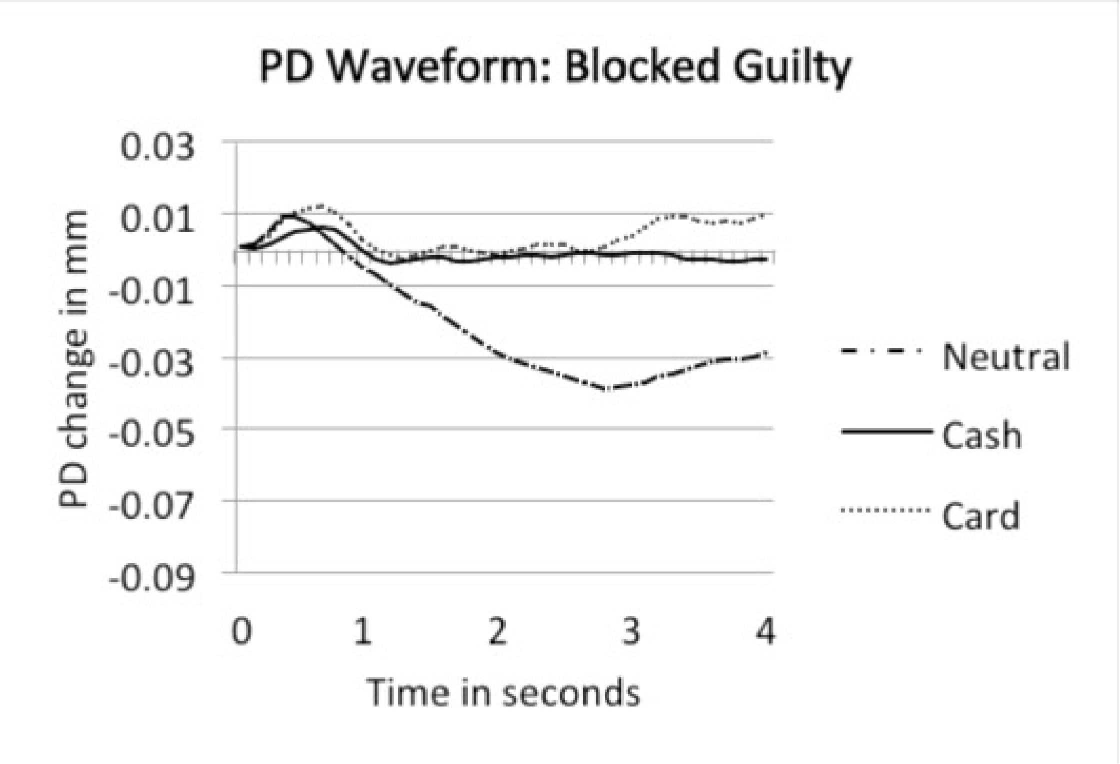


Table 3.3 reports the reliability of ocular-motor measures (coefficient alpha) to determine if reducing the number of items on the ODT adversely affected the reliability of outcome measures. Reliability was measured across the six repetitions of the 48 ODT statements. As a result, the number of 'items' in the coefficient alpha was the number of repetitions. This approach was used for the distributed, blocked, and blocked unit formats. Mean reliability for ocular-motor measures varied little over distributed (M=.61), blocked (M= .54), and blocked unit (M= .56) formats. Mean reliability for ocular-motor measures varied little over distributed (M=.61), blocked (M= .54), and blocked unit (M= .56) formats.

Practice with or without Feedback

There were small Guilt X Feedback, F(1, 144) = 9.124, p<.05, partial n2 = .06, as well as for Guilt X Statement type X Feedback effects, F(2, 288) = 3.151, p<.05, partial n2 = .021, on PD area. Guilty participants had greater increases in pupil size in the feedback condition than in the no feedback condition. Presentation format did not moderate these effects (all *p* > .05).

Interval

The Guilt X Interval interaction was significant for PD area, F(1, 144) = 5.145, p<.05, partial n2 = .021. Although the absolute magnitude of the pupil response increased as the length of the post-response interval increased, F(1, 126) for linear effect = 281.0, *p*<.01, the difference between innocent and guilty groups was greatest at the 500 ms interval. These findings suggest that the 500 ms inter-event interval interrupts the development of the evoked pupil response, but there was no evidence that the length of the interval affected the diagnostic usefulness of this or any other ocular-motor measure.

Measures Based on Longer Inter-event Intervals

We conducted additional analyses to determine if new PD level and blink rate measures that capitalize on longer inter-event intervals are more diagnostic of deception than the traditional measures. A multivariate repeated measures ANOVA compared traditional measures for the two repetitions of test items presented with 500 ms inter-event intervals to the alternative methods for repetitions that used 1500 ms and 3000 ms inter-event intervals but there were no significant interactions if the Guilt X Statement type X Method of measurement interaction (all p>.05).

Post-ODT Questionnaire

A post-ODT questionnaire asked about the participant's perceptions during the ODT. Two questions measured each of eight aspects of subjective experience (Appendix A). The mean of responses to the two items was computed for each participant and group means and standard deviations are reported in Table 2.

As compared to innocent participants, guilty participants rated the experience as more realistic, were more concerned about the cash items, and were more worried about passing the ODT. Presentation format correlated with Concentration, $r(158) = .192$, $p < .05$; participants reported that they were better able to concentrate during the blocked than the distributed format.

Participants were asked to rate their anxiety levels while answering questions about the thefts. As compared to innocent participants, guilty participants were more anxious when answering questions about the $20 than the credit card. However, almost half of both innocent and guilty participants reported being equally anxious when answering questions about the two thefts. The distribution of responses to this item differed for innocent and guilty participants, $x^2(3) = 23.02$.

More than half of the participants in the no feedback and feedback conditions thought that it was just as important to be fast as it was to be accurate. Further analysis revealed that whether or not a participant received feedback did not correlate with their concern about speed or accuracy. There was no relationship between answers to this question and feedback condition, $x^2(3) = 1.54$.

Discriminating Variables

Similar to Experiment 1, contrasts between statement types were correlated with a dichotomous variable that distinguished between guilty (coded 1) and innocent groups (coded 0). In addition to the traditional method for extracting features from evoked pupil responses to individual items, in the case of blocked items, the change in pupil size across the entire block of four items was analyzed as a single evoked response.

PDAreaCashCard and PDLevelCashCard contrasts for the distributed format had validity coefficients that exceeded .55 and

Table 2. Means and SDs of Post-ODT Questionnaire for Innocent and Guilty Participants

	Innocent mean (SD)	Guilty mean (SD)	Eta-Square
Motivation	8.3 (1.75)	7.84 (1.59)	-
Concentration	6.16 (2.11)	5.94 (1.82)	-
Was study realistic	6.60 (1.95)	7.30 (1.65)	.036
Worry about speed	7.16 (2.22)	6.95 (2.00)	-
Worry about accuracy	6.93 (1.81)	6.58 (1.69)	-
Worry about cash items	4.94 (1.65)	5.89 (1.76)	.073
Worry about card items	5.43 (1.81)	5.23 (1.70)	-
Worry about passing ODT	5.15 (2.12)	5.88 (1.61)	.036

Table 3. Point-Biserial Correlations (validity) and Reliability of Outcome Measures for Distributed and Blocked Presentation Formats.

	Distributed		Blocked	
Outcome Measure	Validity	Reliability	Validity	Reliability
RTCashCard	**-.497**	.329	**-.341**	.491
PropWrongCashCard	.093	.209	-.043	.113
NfixCashCard	**-.406**	.627	**-.335**	.318
FirstPassCashCard	**-.253**	.540	-.188	.167
RereadCashCard	**-.342**	.397	-.170	.004
PDAreaCashCard*	**.586**	.615	**.274**	.080
PDLevelCashCard	**.585**	.510	**.604**	.668
ItemBlinkRateCashCard	**-.388**	.182	**-.261**	.130
NextItemBlinkRateCashCard	-.088	.351	-.119	.040

were significantly greater than those obtained from the blocked condition (Table 3). The pupil measures from the distributed format also tended to be more reliable (M = .61) than those from the blocked format (M = .54).

The negative point-biserial correlations for RT, number of fixations, first pass duration, reread duration, and item blink rate between cash and card items indicate that guilty participants were faster to respond, made fewer fixations, spent less time reading and rereading, and blinked fewer times on the cash items than card items. The correlations for the Cash versus Card contrasts were positive for PD area and PD level. As compared to innocent participants, guilty participants showed greater increases in pupil size in response to cash than other items.

For the distributed condition, the decision model correctly classified 90% of the innocent participants and 78% of the guilty participants (M = 84%). For the blocked condition, the accuracy rates for innocent and guilty groups were 74% and 78%, respectively (M=76%). Percent correct decisions was not significantly lower for the blocked condition than for the distributed condition, Yates' $x^2(1)= 1.145$, $p > .05$.

Discussion

The present study evaluated the effects of guilt, blocking, practice with or without feedback, and inter-event intervals on ocular-motor and behavioral measures.

Presentation Format

Mean accuracy for a decision model developed in a prior study (Patnaik et al., 2016) was 84% for the distributed format and 76% for the blocked presentation. That model included response time and relative pupil diameter (PD level) for a 2-second interval surrounding the participant's answer. The decision model achieved good accuracy with distributed and blocked presentations of test items, but there were significant differences between distributed and blocked conditions on measures of reread duration, area under the evoked pupil response, PD level, and blinks per item. In all cases, the distributed format produced superior results. The model performed similarly across formats because only two measures that showed the effects of presentation format were used to make decisions. Examination of evoked pupil responses relative to statement onset revealed that changes in pupil size were diagnostic and consistent with prior research when statement types were distributed, but not when they were presented in blocks.

Participants in the distributed condition reported that they were less able to concentrate when items were distributed than when they were blocked. These findings suggest that participants found it more difficult to read and respond to test items when the items were distributed than when they were blocked. The distributed format appears to be more cognitively demanding than the blocked format.

The magnitude of short-term, phasic increases in pupil size following the onset of test statements (PD area) might be an indication of cognitive effort, whereas pupil size measured the moment participants responded to the statement (PD level) might reflect the emotional impact of the stimulus. For deceptive individuals, the blocked format provided opportunities to anticipate the presentation of incriminating test items. Although these items did not require additional cognitive resources, they did produce large tonic effects on PD level. The possibility that PD area reflects a cognitive response, whereas PD level reflects an emotional response would explain why both measures were diagnostic for the distributed format, but only PD level was diagnostic for the blocked format. If a reduction in the interval from the participant's answer to the onset of the next item contributes to cognitive load, then the hypothesis that PD area reflects mental effort also is consistent with the finding that the difference between guilty and innocent groups was greatest at the shortest inter-event interval. Finally, being indicators of different psychological processes would explain why the two measures make independent contributions to discriminant functions and logistic regressions that form the basis of ODT decision models.

Pre-ODT Performance Feedback

Feedback during the pretest practice session reduced error rates and produced larger phasic pupil reactions to test items for guilty participants and greater differences between pupil responses to cash and credit card items for guilty participants. However, there was little evidence of anchoring because performance feedback did not affect response times.

Post-Answer Intervals

An increase in the length of the inter-event interval had no effect on the diagnostic validity of any ocular-motor measure. Predictably, PD area increased with increased inter-event intervals because the reactions were less truncated by the occurrence of the next stimulus. However, the PD area measures were no more diagnostic for longer inter-event intervals. Likewise, new measures of PD level and blink rates obtained with extended scoring windows for longer inter-event intervals were no more diagnostic than measures previously developed for 500 ms inter-event intervals.

Individual Differences

There were significant differences between innocent and guilty participants on Realism, concern about the cash items, and General Worry. Innocent participants probably did not find the study as realistic as guilty participants, because they could not be sure that someone actually stole $20 or credit card information. The fact that guilty participants were concerned about answering questions about the $20 was reflected in pupil responses and general worry about passing the test. Differences between the guilty and innocent groups' ratings of concern and worry also are consistent with the idea that emotional processes contribute to observed changes in ocular-motor measures.

General Discussion

The primary objective of the present investigation was to explore alternative procedures that might improve the efficiency or effectiveness of the ODT and contribute to our understanding of psychophysiological basis of the ODT. Guilty participants exhibited clear differences from innocent participants in both experiments. Guilty participants responded faster, made fewer fixations, and spent less time reading and rereading statements about the crime they committed than the control crime in both of the Cook et al. studies and in the present study when participants received direct items. In addition, guilty participants showed greater increases in PD for statements answered deceptively than for statements answered truthfully. The observed differences between groups in pupil size are consistent with the idea that deception requires more cognitive effort and greater emotional arousal than truthfulness. The additional investment of cognitive and emotional resources was beneficial to guilty participants, because their error rates were lower than those of innocent participants.

In Experiment 1, we found that the effects of deception were greatest when the items on the ODT directly asked about illicit activities. We attributed the performance gain to the emotional salience of the direct statements, and designed Experiment 2 to capitalize on the presumed emotional aspects of test. To increase arousal, we informed participants about the type of statement they should expect and presented several statements of the same type in sequence. We observed the largest mean effect on pupil measures when the task was made more difficult by changing the type of statement on each trial. The mean effect on pupil measures was greater for the distributed format (mean r-to-z-to-r = .59) than for the blocked format (mean r-to-z-to-r = .38). Together, the results from the two experiments suggest that effects on ocular-motor measures are greatest when the test challenges participants with items that have high arousal value and their occurrence during the test is less predictable.

Conclusions

Results from the present experiments, Patnaik et al. (2016), and Cook et al. (2012) suggest that a combination of behavioral and ocular-motor measures can be used to detect deception. We used a mock crime paradigm that reliably produces large, diagnostic changes in electrodermal, cardiovascular, and respiration reactions during polygraph examinations (Raskin & Kircher, 2014). Although not a comparative study, the magnitude of these observed effects on ocular-motor measures is comparable to that obtained on polygraph measures, as are the accuracy rates obtained for ODT and polygraph examinations. To the extent that ODT and traditional polygraph instrumentation and techniques involve similar underlying cognitive, emotional, memory and control factors, these findings may be of generalizable interest to the field polygraph practitioners and program managers.

References

Adams, J.A., & Goetz, E.T. (1973). Feedback and practice as variables in error detection and correction. *Journal of Motor Behavior, 4,* 217-224.

Applied Sciences Laboratories. (2001). *Eyenal (Eye-Analysis) software manual: for use with ASL series 5000 and ETS-PC eye tracking systems.* Bedford, MA: Applied Science Group, Inc.

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.M., Miccoli, L., Escrig, M.A., & Lang, P.J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45(4),* 602-607.

Cook, A. E., Hacker, D. J., Webb, A., Osher, D., Kristjansson, S., Woltz, D. J., & Kircher, J.C. (2012). Lyin' eyes: Ocular-motor measures of reading reveal deception. *Journal of Experimental Psychology Applied, 18*(3), 301-313.

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.

Hacker, D.J., Kuhlman, B., Kircher, J.C., Cook, A.E., & Woltz, D.J. (2014). Detecting deception using ocular metrics during reading. In D.C. Raskin, C.R. Honts, & J.C. Kircher (Eds.), *Credibility assessment: Scientific research and applications.* Elsevier.

Hyona, J., & Nurminen, A.M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology, 97,* 31-50.

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

Johnson, R., Jr., Barnhardt, J., & Zhu, J. (2005). Differential effects of practice on the executive processes used for truthful and deceptive responses: An event-related brain potential study. *Cognitive Brain Research, 24,* 386-404.

Kahneman, D. (1973). Attention and Effort. Englewood Cliffs, NJ: Prentice-Hall.

Kircher, J. C. (1981, June). *Computerized chart evaluation in the detection of deception.* Master's thesis, University of Utah.

Kircher, J.C., Kristjansson, S., Gardner, M.K., & Webb, A.K. (2012). Human and computer decision making in the psychophysiological detection of deception. *Polygraph, 41(2),* 77-126.

Kircher, J. C., & Raskin, D. C. (2016). Laboratory and field research on the ocular-motor deception test. European Polygraph, 10(4), 159-172.

Leal, S., & Vrij, A. (2008). Blinking during and after lying. *J. of Nonverbal Behavior, 32,* 187-194.

Marchak, F.M. (2013). Detecting false intent using eye blink measures. *Front Psychol, 4,* 1-9.

Patnaik, P., Kircher, J.C., Hacker, D.J., Cook, A.E., Woltz, D.J., Ramm, M.L.F., & Webb, A.K. (2016). Ocular-motor methods for detecting deception: A cross-cultural examination. *International Journal of Applied Psychology, 6*(1), 1-9. doi: 10.5923/j.ijap.20160601.01

Podlesny, J.A. & Raskin, D.C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology, 15,* 344-359.

Raskin, D. C. & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility assessment: Scientific research and applications.* Elsevier.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372-422.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading, 10,* 241-255.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi- Experimental Designs for Generalized Causal Inference*. Unknown Publisher.

Steller, M. (1989). Criteria-based statement analysis. Psychological methods in criminal investigation and evidence. In D.C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217-245). New York, NY: Spring Publishing.

Visscher, K.M., Miezen, F.M., Kelly, J.E., Bucker, R.L., Donaldson, D.I., McAvoy, M.P.,…Petersen, S.E. (2003). Mixed block/event-related designs separate transient and sustained activity in fMRI. *Neuroimage, 19*, 1694-1708.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*, 141–142. doi:10.1016/j.tics.2006.02.001

Webb, A. K, Honts, C. R., Kircher, J. C., Bernhardt, P.C., & Cook, A. E. (2009). Effectiveness of pupil diameter in a probable-lie comparison question test for deception. *Legal and Criminal Psychology, 14(2), 279-292.*

Zuckerman, M., DePaulo, B.M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp 1-59). New York: Academic Press.

Zuckerman, M., DePaulo, B.M., & Rosenthal, R. (1986). Humans as deceivers and lie detectors. In P.S. Blanck, R. Buck, & R. Rosenthal (Eds.), *Nonverbal communication in the clinical context* (pp. 13-35). University Park: Pennsylvania State University Press.