Generalizability of an Ocular-Motor Test for Deception to a Mexican Population

Pooja Patnaik[1,*], Dan J. Woltz[2], Douglas J. Hacker[2], Anne E. Cook[2],
María de Lourdes Francke Ramm[3], Andrea K. Webb[1], John C. Kircher[2]

[1]Draper, Cambridge, MA, U.S.A.
[2]Educational Psychology, University of Utah, Salt Lake City, U.S.A.
[3]Department of Psychology, Tec de Monterrey, Monterrey, Mexico

Abstract   We developed an ocular-motor deception test (ODT) that classifies people as truthful or deceptive based on eye movements and pupillary responses while participants read and respond to true/false statements concerning their possible involvement in illicit activities. The purpose of the present study was to investigate whether the effects of deception on ocular-motor measures generalize to native Spanish speakers in Mexico. One hundred and forty-seven students at a large university in Mexico participated in a mock crime experiment; 83 were guilty of taking $200 pesos from a secretary's purse, and 64 were innocent. On cross-validation, accuracy of classifications based on behavioral and ocular-motor measures exceeded 80% for both truthful and deceptive participants in Mexican and U.S. samples.

Keywords   Deception detection, Ocular-Motor, Reading measures, Cross-Cultural psychology

1. Introduction

Since 2007, the Mexican government, in conjunction with the U. S. State Department, has promoted anti-corruption screening of government employees. The U.S. government has obligated over $1 billion to help eliminate corruption within the ranks of police and security officials in Mexico. Of those funds, Mexican officials have invested tens of millions of dollars in polygraph equipment, training, and examinations of job applicants in law enforcement and employees.

Unfortunately, polygraph tests are time consuming and labor-intensive, requiring 1.5 to 3 hours with a trained examiner. Polygraph testing of federal prosecutors and law enforcement officers is permitted by Mexican law, but the results alone are not sufficient to terminate an employee [1, 2]. The time and labor demands of polygraph testing coupled with the need for additional employee evaluations limit its usefulness for deterring corruption in Mexico.

1.1. Deception Detection

Cook et al. developed a new method for detecting deception called the ocular-motor deception test (ODT) [3]. In contrast to the polygraph, the ODT is automated and can be completed in approximately 40 minutes. A computer presents written instructions and true/false test statements concerning the subject's possible involvement in illicit activities. The subject reads test statements presented individually by the computer and uses a keypad to answer while a remote eye tracker records eye movements and changes in pupil size. The computer processes the ocular-motor data, combines its measurements in a logistic regression, and classifies the individual as truthful or deceptive on the test. With sufficient evidence of validity in a field setting with Spanish readers, the Mexican government could use the test alone or to decide whether or not to conduct a polygraph test.

Most theories of deception detection posit that lying is more cognitively demanding than telling the truth [4-6]. Cognitive resources are needed to inhibit the truth, fabricate the lie, and maintain its consistency, coherence, and believability over time. Inhibiting pre-potent truthful responses, maintaining credibility over time, and self-monitoring for signs of leakage are cognitive processes that require mental effort.

Over the past four decades, psychophysiological studies have shown that the pupil provides a sensitive index of cognitive effort (e.g., [7, 8]). Generally, the greater the cognitive load, the greater is the increase in pupil size. Consistent with the idea that deception requires more mental effort than telling the truth, psychophysiologists also have found that pupil responses discriminate between truthful and deceptive individuals during polygraph examinations (e.g., [9, 10]).

The pupil responds not only to cognitive load but also

emotional stimuli. Several investigators have reported that emotional stimuli evoke pupil responses, the magnitude of which depends on the intensity but not the valence of the emotional stimulus [11-14]. Polygraph tests are based on the idea that deceptive individuals will show stronger emotional responses to some test questions than others. To the extent that emotional reactions to test questions distinguish deceptive from truthful individuals, pupil responses should reflect those differences and be diagnostic of deception.

A reader who has difficulty reading or comprehending text shows more eye fixations, pupil enlargement, and longer reading times [15, 16]. Number of fixations is a count of the number of fixations that fall in a predefined area of interest that surrounds a text. Reading time is the sum of durations for fixations that fall in the area of interest. Investigators sometimes partition reading time into initial reading of the text (first pass duration) and subsequent rereading of the text.

Cook et al. assessed the effects of deception on pupil responses, reading behaviors, and behavioral measures. Participants were assigned to guilty or innocent treatment conditions. Guilty participants committed a mock theft of $20 from a secretary's purse or downloaded credit card information from a graduate student's computer. All participants were instructed to respond quickly and accurately to true or false statements presented serially on a computer monitor while eye movements and changes in pupil size were recorded by an eye tracker. The statements pertained to the theft of the $20, the theft of information, or to neutral content (e.g., Polar bears are native to Mexico.).

As expected, deception was associated with the greatest increases in pupil size, and guilty participants also took longer to read and respond to test statements. Although the main effects of guilt on reading and response time measures across all statement types were consistent with the literature on reading, guilt status interacted with statement type, and those interactions were not consistent with initial expectations based on the reading literature. In both experiments, guilty participants took significantly *less* time to read and answer deceptively to test statements concerning the crime they had committed than when they answered truthfully to the other statements. For guilty participants, deceptive answers were characterized by fewer fixations and shorter reading and rereading times compared to their truthful answers. Cook et al. concluded that when participants were informed they would fail the test if they did not respond quickly and accurately to the test statements, guilty participants made a concerted effort to spend as little time on the incriminating statements as possible to avoid detection.

1.2. Present Study

To date, all laboratory experiments with the ODT have been restricted to English-speaking participants in the United States. Research has shown that members of all cultures believe that liars experience fear, shame, or cognitive difficulties [17, 18], but cultures differ in their norms for interpersonal communication [19], and lying is a form of interpersonal communication [20]. In addition to cultural differences between the U.S. and Mexico, their languages (English and Spanish) differ on semantic, pragmatic, syntactic, orthographic, and morphological dimensions.

A simple theory that relies on cognitive load and emotion does not predict that language will affect behavioral or physiological measures observed during the ODT. Nevertheless, in contrast to results from mock crime experiments in the U.S., we found almost no effect of deception on ocular-motor measures in an unpublished field study conducted in Colombia. Although error in the criterion measure (ground truth) probably contributed to the low accuracy of the ODT in that study, we also believe that the participants did not have adequate reading proficiency to show effects of deception. Deceptive individuals who struggled to understand the meaning of test items do not exhibit the expected patterns changes on reading, pupil, or behavioral measures and are not good candidates for an ODT.

Although reading ability could account for lower accuracy in the field than the laboratory, cultural or linguistic differences between the U.S. and Colombian populations also could explain the failure of the ODT to generalize across settings. The goal of the present study was to test whether cultural or linguistic differences moderate effects of deception on ocular-motor measures. We used the procedures in Cook et al. to conduct a mock crime experiment at a large Mexican university. Following Cook et al., we assessed the effects of deception on individual ocular-motor measures. We then combined the Mexican and U.S. data sets to test whether the effects of deception on ocular-motor measures vary over settings. Failure of the ODT in a population of proficient Mexican readers would indicate that the technology does not generalize well to other cultures or languages.

2. Method

2.1. Participants

One hundred seventy-two participants were recruited from an urban university campus in northeastern Mexico. Recruitment flyers were posted on campus that advertised an opportunity to earn 200 pesos (approximately 15 USD at the time) and a possible bonus of 300 pesos (approximately 23 USD at the time) for participation in a psychological experiment. Interested participants who spoke fluent Spanish, were over the age of 18, and could read a computer screen without glasses were scheduled for a session. Of these 172 individuals, 27 participants were omitted due to problems with calibration, not following directions, and/or inadequate recordings. Two guilty participants were dropped because their answer error rates exceeded 50% (chance). The remaining 145 participants ranged in age from 18-52. They were students at the university, predominantly Hispanic /

Latino Mexican, and Spanish was their primary language. Sixty-two of the 145 participants were innocent of the mock crimes (24 female) and 83 were guilty of the crime (29 female).

2.2. Overview of Design and Procedure

The design was a 2 × (3 × 5) mixed design. The between-subjects factor was guilt with two levels (guilty or innocent). The within-subject factors were statement type (neutral, cash, and exam) and repetition (5 repetitions of the ODT test items). Time with 40 levels (10 Hz samples x 4 seconds) also was included as a within-subjects factor for the analyses of pupil diameter (PD).

2.3. Apparatus

A SensoMotoric Instruments (SMI) RED-m remote eye tracker affixed to a desktop PC monitor recorded eye movements and pupil diameter at 60 Hz. Viewing was binocular and a chin rest was used to keep the participant's head still. Stimuli were presented to the participant on a 19-inch Lenovo flat screen LCD monitor with a 5:4 aspect ratio. The monitor was positioned approximately 65 cm from the participant's eyes.

2.4. Ocular-motor Deception Test

Instructions and practice items were presented in Spanish to the participant in black font with a pale grey 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 200 pesos (e.g., I had nothing to do with the theft of the 200 pesos.), 16 pertained to the theft of the exam (e.g., I took nothing from the professor's office.), and 16 were neutral items (e.g., I am younger than 75 years old.). The items were arranged such that no two items from the same category appeared in succession. Statements were presented one at a time and the characters *V/F* (verdadero/falso) appeared to the far right of the statement to remind participants of their answer choices. The correct (i.e., nonincriminating) answer was true for 8 of 16 items in a category and false for the remaining 8 items in each category.

Between repetitions of the 48 test items, participants completed an intervening task that consisted of 18 T/F general world knowledge questions and was designed to provide a break and clear working memory of ODT test items and answers.

2.5. Dependent Measures

Behavioral Outcome Measures. Behavioral measures consisted of percentage error and response time.

Percentage error for a particular statement type (neutral, cash, exam) was the number of incorrect responses divided by the number of items (16 × 5 = 80) times 100.

Response time (RT) was the time in ms from the appearance of the item on the screen to a button press response by the participant.

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. Fixations were determined from the data files produced by the SMI eye tracker by identifying a sequence of samples in which the eye showed little movement for at least 100 ms. Fixations longer than 1,000 ms were considered artifacts and were discarded [15]. Reading time was the sum of durations for fixations that fell in the area of interest.

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

The PD data samples from the beginning of a block of 48 test items to the end of that block were standardized (converted to z-scores) within participants. For each test item, PD area under the curve (PD AUC) was obtained from the standardized pupil response curve that began the moment the test statement appeared on the computer screen and ended 4 s later. The computer identified high and low points in this 4-second response curve and computed the difference between each low point and every subsequent high point. Peak amplitude was the greatest observed difference, and response onset was defined as the low point from which peak amplitude was measured. In the final step, the area measure was obtained by integrating the area under the pupil response curve from response onset to the point at which the response returned to the level at response onset or to the end of the 4-second sampling interval, whichever occurred first.

PD level at T/F response was the mean of standard scores for a period that began 1 s prior to the participant's T/F response and ended 1 s after the T/F response.

2.6. 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, and read and sign the consent form. The participant then listened to a recording that gave their instructions for the study. All participants were informed that some participants would steal an exam from a professor's computer, other participants would take 200 pesos from a secretary's purse, and a third group of participants would be innocent of both mock crimes. In actuality, there were only two groups; participants were either guilty of taking 200 pesos from a secretary's purse or they were innocent of the crime. All participants were promised a 300-pesos bonus if they could pass the ODT. A phone number was provided for participants to call if they chose not to participate.

Guilty participants went to a secretary's office and asked the secretary where a professor's office was located. The secretary (a female confederate) told the participant that the

professor did not work in the building. The participant thanked the secretary and left her office. The participant then waited inconspicuously for the secretary to leave the office unattended, entered the office, found her purse, removed 200 pesos from a wallet in the purse, and concealed the money on their person. Guilty participants were told to prepare an alibi in case they were caught and not to leave fingerprints. They were informed that they had no more than 20 min to commit the crime and report to the examiner.

Participants in the innocent condition were told that they would be asked about the thefts but should not take anything, and they should wait approximately 20 min before they reported to the examiner.

When participants reported for the ODT, they sat at a computer with their chin in a chin rest, and completed a brief calibration of the eye tracker. After calibrating the eye tracker, the computer administered the ODT. After the test, the examiner debriefed and paid the participant. In addition to their base pay of 200 pesos, participants were paid a 300-pesos bonus if the computer indicated they were innocent.

3. Results

The primary goal of the present study was to determine if the effects of deception on behavioral and ocular-motor measures were similar for students at a Mexican university and students at a university in the U.S. Repeated measures analysis of variance (RMANOVA) was used to analyze each dependent variable. Because the ODT was designed as a relevant comparison test, the evaluation of deception

detection within each measure was based on the difference between person means for the two relevant issues (i.e., exam and cash items). Deception detection was tested by comparing this contrast for guilty versus innocent individuals, and this corresponded to the interaction effect of guilt and the relevant content contrast. Significance level was set at $p < .05$ in all analyses. Supplemental analyses also are reported with the inclusion of 112 participants (56 innocent and 56 guilty) from Experiment 2 who performed under similar experimental conditions in the U.S. [3]. We analyzed the 3-way interaction between guilt condition, the relevant item contrast, and setting (Mexico versus U.S.) for each measure as a test of whether the effects of deception differed in the two settings.

3.1. Behavioral and Reading Measures

Figure 1 presents means for response errors (Panel a) and response time (Panel b). Error bars represent 95% confidence intervals calculated for within-subject comparisons of item type. As can be seen in Figure 1a, guilty participants made more response errors across all three item types compared to those in the innocent condition, $F(1,143) = 10.04$, partial $\eta^2 = .07$. In addition, there was a significant main effect for the contrast between relevant items, $F(1,143) = 17.47$, partial $\eta^2 = .11$. In general, participants in both conditions made more errors in responding to exam compared to cash items. Guilt condition did not interact with the contrast of relevant item content, $F(1,143) < 1, p > .05$. When Cook et al. data were compared with the current data, the 3-way interaction of Setting x Guilt x Relevant Content was not significant, $F(1,253) < 1, p > .05$.
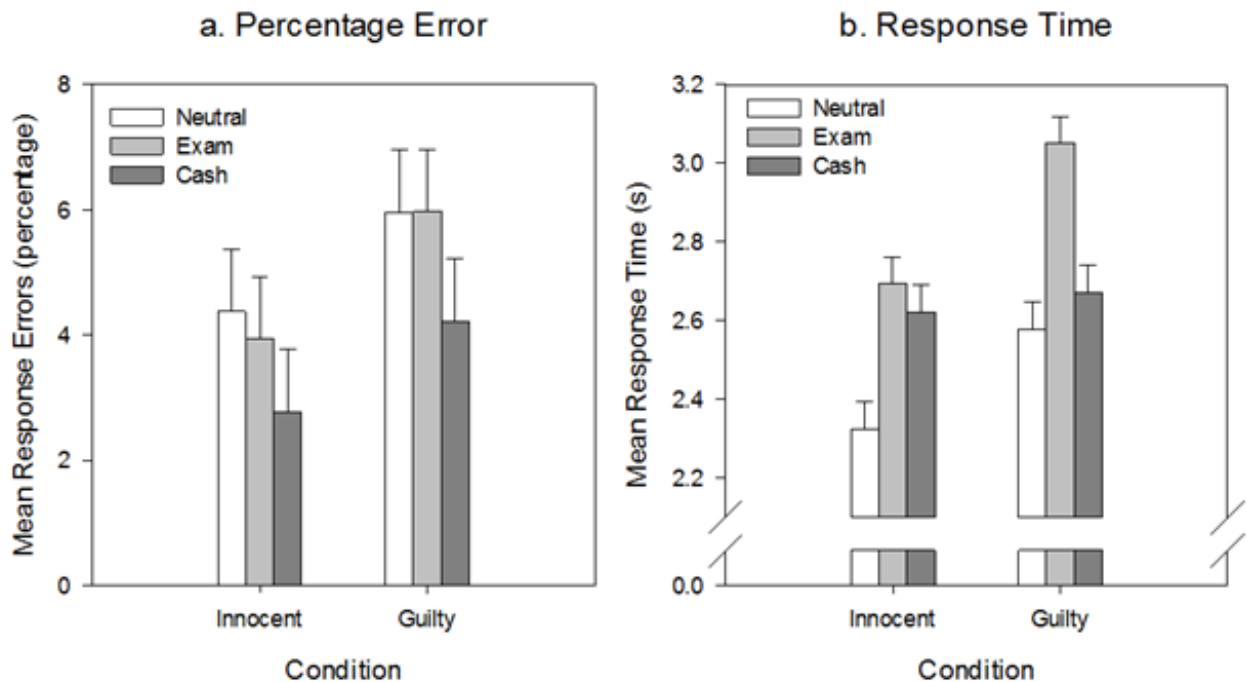
a. Percentage Error

b. Response Time

Figure 1. Mean percentage error and response time by item content and guilty condition

For response time, the main effect of guilt across all items was not significant, F(1,143) = 3.52, p = .06. The main effect of the relevant item contrast was significant, F(1,143) = 63.20, partial η^2 = .31. In general, participants responded more quickly to cash compared to exam items. Of primary importance, guilt interacted the relevant itemcontrast, F(1,143) = 29.37, partial η^2 = .17. As can be seen in Figure 1b, innocent participants responded similarly to exam and cash items, whereas guilty participants responded more quickly to cash relative to exam items. This pattern was similar to that reported by Cook et al. When Cook et al. Experiment 2 data were included with the current data, the 3-way interaction of Setting x Guilt x Relevant Content was not significant, F(1,253) < 1, p > .05.

Figure 2 presents results for first pass reading time per item (Panel a) and rereading time per item (Panel b). As with response time, there was no main effect of guilt on first pass reading duration, F(1, 143) = 2.52, p = .11. There was a main effect for the relevant item contrast, F(1,143) = 13.51, partial η^2 = .09, with shorter first pass reading duration for cash compared to exam items. However, as is evident in Figure 2a, the overall effect of item content was due entirely to the fast first pass reading of guilty participants on cash items. This was reflected in the interaction of relevant item content with guilt, F(1,143) = 37.91, partial η^2 = .21. The pattern found by Cook et al. (2012) was similar in that only guilty participants had shorter first pass durations for cash compared to exam items. The 3-way interaction of Setting x Guilt x Relevant Content was not significant, F(1,253) < 1, p > .05.

For rereading time, there was no main effect of guilt, F(1, 143) = 1.75, p = .19. There was a main effect for the relevant item contrast, F(1,143) = 36.74, partial η^2 = .20. As with first pass time, average rereading duration was shorter for cash compared to exam items. Importantly, guilt condition interacted with the relevant item contrast, F(1,143) = 19.78, partial η^2 = .12. As can be seen in Figure 2b, relative to innocent participants, guilty individuals spent less time rereading cash compared to exam items. This pattern was similar to that reported by Cook et al. (2012). The 3-way interaction of Setting x Guilt x Relevant Content was not significant, F(1,253) = 1.89, p= .28.

Figure 3 presents results for number of fixations per item. There was no main effect of guilt on number of fixations, F(1, 143) < 1 p > .05. Participants did make fewer fixations on cash compared to exam items, F(1,143) = 13.30, partial η^2 = .09, but there was no interaction of relevant item content with guilt, F(1,143) = 2.80, p = .10. When the Mexican and U.S. samples were combined, the 3-way interaction of Setting x Guilt x Relevant Content was not significant, F(1,253) = 1.66, p = .20.the 3-way interaction of Setting x Guilt x Relevant Content was not significant, F(1,253) < 1, p > .05.

3.2. Pupil Measures

Figure 4 shows mean change from baseline in pupil diameter. A positive difference indicated PD increased relative to the initial value, and a negative difference indicated PD decreased relative to the initial value.

One to three seconds after question onset, the pupil response began to distinguish among statement types. Innocent participants showed a slightly stronger pupil response to the exam items than the cash items, and neutral items were associated with the weakest responses. In contrast, guilty participants showed a stronger reaction to cash than exam items. The drop in the curve for neutral items from second 2 to 6 indicates pupil constriction and probably reflects recovery from the previous reaction to a relevant item.
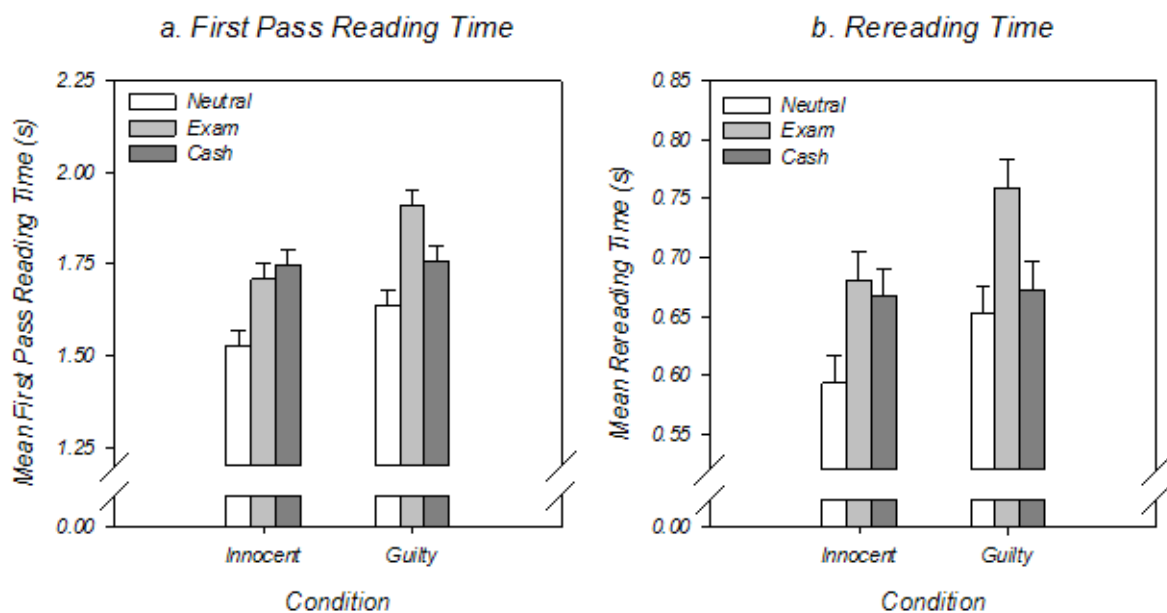

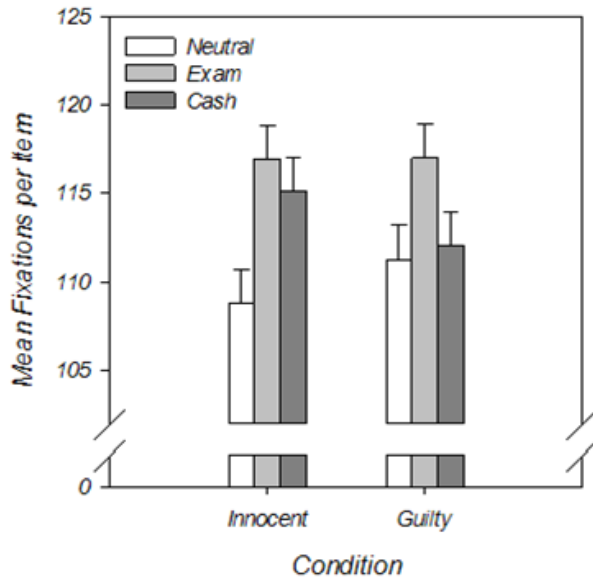**Figure 2.**   Mean first pass and rereading times by item content and guilty condition

Figure 3. Mean number of fixations per item by item content and guilt condition

The magnitude of the pupil response as measured by PD AUC is shown in Figure 5a for each group and item type. Figure 5b shows the PD level at the time the participant answered the item for each group and item type.

There was no main effect of guilt across all item types on either PD AUC, $F(1, 143) < 1, p > .05$, or PD level, $F(1, 143) = 1.53, p = .22$. There was a small main effect for the relevant item contrast in PD AUC with lower values for exam compared to cash items, $F(1,143) = 5.02$, partial $\eta^2 = .03$. There was no main effect for relevant item content in PD level. Of primary importance, there were substantial

interactions between guilt and relevant item content for both PD AUC, $F(1,143) = 60.03$, partial $\eta^2 = .30$, and PD level, $F(1,143) = 73.22$, partial $\eta^2 = .34$. As can be seen in Figures 5a and 5b, compared to innocent participants, guilty individuals showed stronger pupil responses during cash items and weaker responses during exam items. These patterns were similar to those reported by Cook et al. The 3-way interaction including setting was not significant for either PD AUC, $F(1,253) < 1, p > .05$, or PD Level, $F(1,253) < 1, p > .05$.

3.3. Predictive Validity of Ocular-Motor Measures

To assess the accuracy of classifications by the ODT with Spanish speakers, a backward stepwise logistic regression was conducted using contrast measures that had significant interactive effects with guilt. The resulting regression model had three measures that represented behavioral, reading, and pupil responses to test items. This function accurately classified 55 of the 63 innocent participants (87.3%) and 69 of the 82 guilty participants (84.1%), for an overall accuracy of 85.5%. These results were similar to those reported by Cook et al., where overall classification accuracies were 85% for Experiment 1 and 86% for Experiment 2.

To further compare the current Mexico results with those from a U.S. sample, we performed a double cross-validation analysis with data from Experiment 2 of [3]. The Mexico model correctly classified 80.4% of the innocent and 89.3% of the guilty U.S. participants, for a mean accuracy of 84.9%. Using the variables selected in the Mexico model, the U.S. model correctly classified 87.5% of the innocent and 81.9% of the guilty Mexican participants, for a mean accuracy of 84.7%.
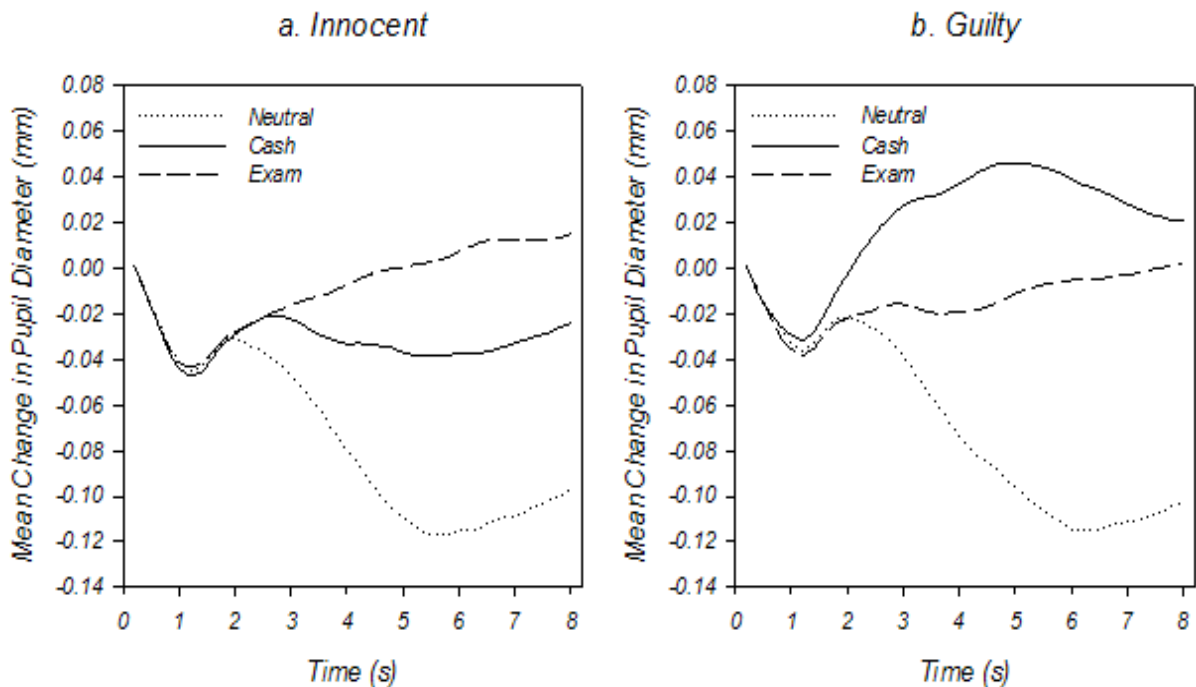

Figure 4. Mean change in pupil diameter for 8 seconds following item onset by item content and guilt condition
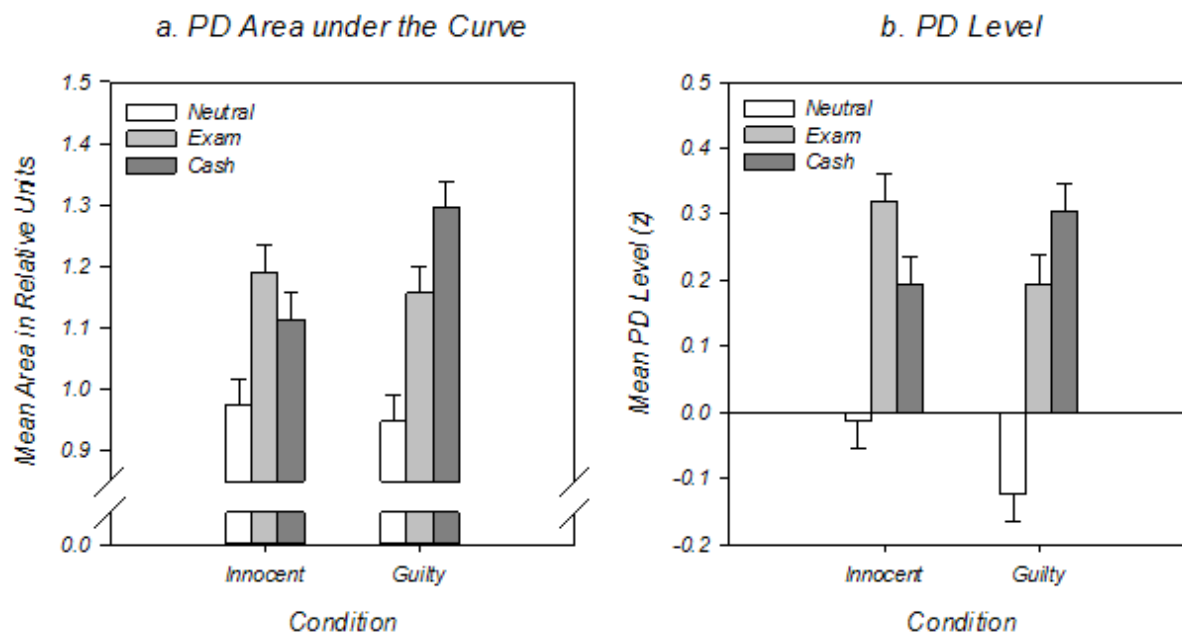
Figure 5. Mean area under the curve and level at response for pupil diameter by item content and guilt condition

4. Discussion

The present mock crime study was conducted at a Mexican university with native Spanish-speaking participants. The study revealed several large effects on behavioral and ocular-motor measures used by the ODT to distinguish between Spanish-speaking truthful and deceptive individuals. Most importantly, the present study failed to reveal any evidence of meaningful differences between Mexican and U.S. participants on any behavioral or ocular-motor measure of deception. Despite differences in culture and language, behavioral measures of response errors and response times did not distinguish between the Mexican and U.S. settings, nor did reading patterns or pupil responses. On cross-validation, the ODT correctly classified approximately 85% of truthful and deceptive participants. The uniformly high levels of decision accuracy achieved by the logistic regression models in Mexican and U.S. samples are indications of the robustness of the effects of deception on ocular-motor measures, as well as the stability of covariance structures among the measures across settings.

Although it is difficult to draw firm conclusions from null results, several aspects of the design supported by a rudimentary theory lessen concerns about the failure to detect effects of settings. For each participant, there were 80 questions per mean for each of three types of questions, and there were 257 participants in the Mexican and U.S. samples. The pooling of measurements over test items of the same type improved the reliability of measures from individual participants, and the large numbers of cases provided sensitive statistical tests for effects of settings. Statistical power also was improved by conducting focused contrasts [21]. Power analysis revealed that our statistical tests had over 80% power to detect an effect that would account for less than 3% of the variance in an outcome measure, and yet

no such test revealed an effect of setting on any measure.

The ODT is based on the idea that it is more difficult to lie than to tell the truth and deception can be associated with a strong emotional response. If these ideas are correct, effects of language on the diagnostic validity of our reading measures may be too small to matter. On the other hand, the proposed theoretical framework would predict effects of culture on the deceptive context created in a mock crime experiment if there were differences between the cultures in moral development. If the theft of money was more reprehensible in one culture than another, commission of the crime might evoke a stronger emotional response that could affect one or more of the outcome measures. However, there was no evidence of such an effect in the present study.

4.1. Future Directions

In the present study, all truthful participants had one common pretest experience and all deceptive individuals had another. The pretest experiences of individuals who undergo an ODT in field settings would be more diverse. Examinees in the field would vary more in age, education, intelligence, and reading ability than did the participants in the present experiment. Variance on any factor that moderates the effects of deception on ocular-motor measures would have to be controlled to maintain high levels of decision accuracy. For instance, unpublished efforts to assess the effectiveness of the ODT with poor readers in Colombia produced only weak effects on ocular-motor measures. Poor readers struggle to comprehend the content of test items, and those efforts appear to outweigh the effects of deception on ocular-motor measures. A reading test might be administered prior to the ODT to determine if the person is a suitable candidate for the test. Alternatively, error rates and response times on the ODT should provide evidence that the examinee is capable of providing diagnostic data.

Another question may be raised about whether an ODT about a specific incident will work as well if the relevant issue is more general. In the present study, we asked participants if they stole 200 pesos from a purse, whereas in a pre-employment screening test, the test might ask if the applicant ever stole from a previous employer. The generality of the relevant questions in a screening context could threaten the validity of an ODT if applicants are uncertain of their guilt [22]. What exactly is meant by stealing from a previous employer? One individual may feel guilty about taking paper from work, whereas another individual might think that taking paper from work is irrelevant. Unless the relevant issues are clearly defined prior to the test, individuals could interpret the meaning of test items differently and accuracy would suffer.

Although there are reasons to expect that field conditions will reduce the accuracy of the ODT, there also is reason to think that field conditions will improve its accuracy. Field subjects are likely to be more motivated to pass the test than participants in a simulated crime experiment. Cook et al. (2012) manipulated monetary incentives to pass the test and found that larger incentives improved the diagnostic validity of some reading measures.

There is additional evidence that the ODT will be effective in the field, despite the generality of issues covered on the test. In 2012, the Narcotics Affairs Section of the U.S. Embassy in Mexico asked us to conduct an evaluation of the ODT for screening attorneys who were applying to work for Procuraduria General de la República, the Mexican equivalent of the U.S. Office of the Attorney General. The issue covered on the ODT was whether or not the applicant had ever used illegal drugs. A day after the ODT, the applicant was given a pre-employment polygraph examination. During an interview with the examiner, 26 of the 332 examinees confessed to having lied about drug use the day before on the ODT. The logistic regression model for Mexican students in the present study correctly classified 23 of the 26 attorneys as deceptive (88.5%), which is slightly higher than the accuracy of the model on deceptive Mexican students (84.1%). Of course, we could not verify the deceptive status of the applicants who did not confess, but the results on cases for those who did suggest that the accuracy could be higher in field settings when the consequences of failing the ODT are more serious than the loss of a cash bonus. An important next step in the validation process is to conduct a field study where ground truth is available on all participants, and there are real-life consequences to failing the test. Until the limits of generalizability are better known, use of the ODT should be limited to individuals who share the common characteristics of our study samples and show similar outcomes on behavioral measures.

REFERENCES

[1]   Byrne, E.V. (2013, February 15). Mexico's supreme court approves polygraph tests for federal prosecutors, but with some limitations. *The Mexico Gulf Reporter.* Retrieved from http://www.mexicogulfreporter.com/2013/02/mexicos-supreme-court-approves.html.

[2]   Byrne, E.V. (2014, March 12). Mexico's supreme court upholds police vetting process. *The Mexico Gulf Reporter.* Retrieved from http://www.mexicogulfreporter.com/2014/03/mexican-supreme-court-upholds-police.html.

[3]   Cook, A. E., Hacker, D. J., Webb, A.K., 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. doi: 10.1037/a0028307.

[4]   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. DOI: 10.1016/j.cogbrainres.2005.02.011.

[5]   Kircher, J. C. (1981, March). *Psycholphysiological processes underlying the detection of deception.* Unpublished manuscript, University of Utah.

[6]   Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitiveload. *Trends in Cognitive Sciences, 10,* 141–142.

[7]   Kahneman, D., Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*, 1583-1585.

[8]   Klingner, Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48,* 323-332.

[9]   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.

[10] 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. doi:10.1348/135532508X398602.

[11] Bradley, M.T., Micolli, L., Escrig, M.A., Lang, P.J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*, 602-607.

[12] Hess, E.H., Polt, J.M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132,* 349-350.

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

[14] Steinhauer S. R., Boller F., Zubin J., Pearlman S. (1983). Pupillary dilation to emotional visual stimuli revisited. *Psychophysiology, 20,* 472.

[15] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372-422. doi: 10.1037/0033-2909.124.3.372Hess, E.H., Polt, J.M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132,* 349-350.

[16] 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. doi**:**10.1207/s1532799xssr1003_3

[17] Bond C.F., Jr, Robinson M.A. (1988) The evolution of deception. *Journal of Nonverbal Behavior, 12,* 295–308. Poulos, H. G., 1971, Behavior of laterally loaded piles-I: Single piles., J. Soil Mech. and Found. Div., 97(5), 711–731.

[18] Ekman P. (2001). *Telling lies: Clues to deceit in the marketplace, politics, and marriage.* New York: Norton.

[19] Anderson, P. A., Hecht, M. L., Hoobler, G. D., & Smallwood, M. (2002). Nonverbal communication across cultures. In W. G. Gudykunst & B. Mody (Eds.), *Handbook of international and intercultural communication* (pp. 89-106). Thousand Oaks, CA: Sage.

[20] Buller, D.B. and J.K. Burgoon (1996). Interpersonal Deception Theory. *Communication Theory, 6(3),* 203–242.

[21] Keppel, G. & Wicken, T. D. (2004). Design & analysis: A researcher's handbook (4rd Ed.). New Jersey: Prentice Hall.

[22] Meijer, E.H., Verschuere, B., Merckelbach, H.L.G.J., & Crombez, G. (2008). Sex offender management using the polygraph: A critical review. *International Journal of Law and Psychiatry, 31(5),* 423-429.