Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

VOLUME 48 **2019** **NUMBER 1**

Contents

Published Semi-annually

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice

Editor-in-Chief: Mark Handler
E-mail: Editor@polygraph.org
Managing Editor: Nayeli Hernandez
E-mail: polygraph.managing.editor@gmail.com

******

APA Officers for 2018 – 2019

Subscription information: *Polygraph* is published semi-annually by the American Polygraph Association. Editorial Address is Editor@polygraph.org. Subscription rates for 2019: One year $150.00 (Domestic). Change of address: APA National Office, P.O. Box 8037 Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

Instructions to Authors

Scope

The journal *Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice* publishes articles about the psychophysiological detection of deception, and related areas. Authors are invited to submit manuscripts of original research, literature reviews, legal briefs, theoretical papers, instructional pieces, case histories, book reviews, short reports, and similar works. Special topics will be considered on an individual basis. A minimum standard for acceptance is that the paper be of general interest to practitioners, instructors and researchers of polygraphy. From time to time there will be a call for papers on specific topics.

Manuscript Submission

Manuscripts must be in English, and may be submitted, along with a cover letter, on electronic media (MS Word). The cover letter should include a telephone number, and e-mail address. All manuscripts will be subject to a formal peer-review. Authors may submit their manuscripts as an e-mail attachment with the cover letter included in the body of the e-mail to:

Editor@polygraph.org

As a condition of publication, authors agree that all text, figures, or other content in the submitted manuscript is correctly cited, and that the work, all or in part, is not under consideration for publication elsewhere. Authors also agree to give reasonable access to their data to APA members upon written request.

Manuscript Organization and Style

All manuscripts must be complete, balanced, and accurate. Authors should follow guidelines in the *Publications Manual of the American Psychological Association*. The manual can be found in most public and university libraries, or it can be ordered from: American Psychological Association Publications, 1200 17th Street, N.W., Washington, DC 20036, USA. Writers may exercise some freedom of style, but they will be held to a standard of clarity, organization, and accuracy. Authors are responsible for assuring their work includes correct citations. Consistent with the ethical standards of the discipline, the American Polygraph Association considers quotation of another's work without proper citation a grievous offense. The standard for nomenclature shall be the *Terminology Reference for the Science of Psycho-physiological Detection of Deception* (2012) which is available from the national office of the American Polygraph Association. Legal case citations should follow the *West* system.

Manuscript Review

An Associate Editor will handle papers, and the author may, at the discretion of the Associate Editor, communicate directly with him or her. For all submissions, every effort will be made to provide the author a review within 4 weeks of receipt of manuscript. Articles submitted for publication are evaluated according to several criteria including significance of the contribution to the polygraph field, clarity, accuracy, and consistency.

Copyright

Authors submitting a paper to the American Polygraph Association (APA) do so with the understanding that the copyright for the paper will be assigned to the American Polygraph Association if the paper is accepted for publication. The APA, however, will not put any limitation on the personal freedom of the author(s) to use material contained in the paper in other works, and request for republication will be granted if the senior author approves.

Polygraph Examiners Unable to Discriminate True and False Juvenile Confessions

Charles R. Honts

Boise State University

Krista Forrest

University of Nebraska Kearney

and

Adela Stepanescu

Boise State University

Abstract

We asked polygraph examiners to assess the credibility of confessions given by incarcerated juveniles. Eighty-three practicing polygraph examiners attending continuing education seminars made 664 true/false judgments and confidence estimates of the credibility of four true and four false confessions. Examiner judgments showed a slight truth bias with 61.8% of the true confessions correctly identified, but with 52% of the false confessions also believed. A believability index developed from judgments and confidence did not show a significant difference between true and false confessions. However, examiners with the Reid interrogation training found false confessions more believable than true confessions. Examiners without Reid training produced the opposite pattern. Our results suggest that Reid training is detrimental when assessing the credibility of juvenile confessions. As with adults, a high degree of caution in evaluating the credibility of confessions given by juveniles is warranted.

Keywords: Confessions, false confessions, juveniles; deception detection

During the past 15 years, a disturbing number of high-profile cases have revealed innocent people who confessed, were convicted at trial, and spent substantial time in jail, only to be exonerated later (Kassin, Drizin, Grisso, Gudjonsson, Leo, & Redlich, 2010). Current data show that false confessions and admissions are present in approximately 25% of all DNA exonerations (Garrett, 2008; Scheck, Neufeld, & Dwyer, 2000; Innocence Project, 2019). Juveniles were involved in a number of those cases. For example, interrogators elicited five false confessions from juveniles in the now infamous New York Central Park Jogger case (*People of the State of New York v. Kharey Wise et al.*, 2002). Jurors convicted all five suspects at trial. After 13 years of incarceration, convicted rapist Matias Reyes confessed to being the sole perpetrator, a confession that was confirmed by DNA and other forensic evidence (Innocence Project, 2019). In December of 2002, the Manhattan District Attorney recommended that the convictions of the five juveniles convicted in the Central Park Jogger case be overturned, and the wrongfully convicted juveniles were released (Drizin & Leo,

Author Note

Correspondence concerning this article should be addressed to Charles R. Honts, Ph.D., Department of Psychological Science, Boise State University, 1910 University Drive, MS-1715, Boise, ID 83725-1715. E-mail: chonts@boisestate.edu

2004).

Juvenile confession and credibility assessment phenomena have received some scientific study and although many of the findings mirror those with college students and other adults, there are some differences. Redlich and Goodman (2003) report that younger and more suggestible teens were more likely to take responsibility for acts they did not commit in a laboratory experiment. Gudjonsson, Sigurdsson, Asgeirsdottir and Sigfusdottir (2006) studied self-reports of actual false confessions in a large sample of students in Iceland and found that 1.6% of more than ten thousand students said that they falsely confessed to police. Among those who reported interrogation by the police, 12% reported giving the police a false confession. Drizin and Leo (2004) also report that, in their sample of criminal cases, about one-third of the confirmed false confessions involved juveniles.

Courts in the United States tend to treat juvenile confessions with the same standards as those for adults and with no additional safeguards (Feld, 2006). Moreover, the trend in the United States is to charge and try juveniles as adults, especially in cases involving serious charges. A survey of 1,828 United States police officers (Reppucci, Meyer, & Kostelnik, 2010) revealed that although police officers recognize that there are developmental differences between juveniles and adults, they fail to apply those differences to the way they interact with and interrogate juveniles. Reppucci et al., concluded that police believe that juveniles can be treated as adults in criminal investigations.

Studies on the development of deception behaviors typically do not find differences in the rates of lie-telling, but they do report that the ability to conceal deception improves with age (Talwar, Gordon, & Lee, 2007). Moreover, it has been suggested that executive functioning skills may be related to lie sophistication (Evans, Xu, & Lee, 2011; Gombos, 2006; Talwar & Lee, 2002; 2008). Evans and Lee (2011) reported that although they found no lie sophistication differences in participants aged 8 to 16, they did find individual differences in the working memory and planning skill of the participants, with both having positive associations with lie sophistication. Evans and Lee note that there is a link between adolescent lying and behavior issues, conduct disorders and delinquency, and they suggest mediation of that relationship by executive function. Specifically, they suggest that due to deficits in adolescents' executive functioning, adolescents fail to construct adequate statements to conceal their transgressions and deception, resulting in high rates of deception detection from others and therefore greater involvement of adolescents in the criminal justice system. One might thus expect that juveniles involved in the criminal justice are less sophisticated liars and that false confessions given by them would be more detectable. It is interesting to note that Craig, Raskin and Kircher (2011) reported an experiment that examined juvenile deception detection with the polygraph. Craig et al., report that although juvenile deception could be discriminated with the polygraph, the polygraph was notably less effective with juveniles than with adults.

Although juveniles are at risk for false confessions during interrogation, one could argue that this would not pose a problem if police, prosecutors, and others could discriminate between false and true confessions. However, there is reason to suspect that the discrimination of juvenile false and true confessions might be difficult. Kassin, Meissner and Norwick (2005) reported a study that examined the ability of laypersons and police officers to discriminate false from true confessions given by adult prison inmates. Kassin et al, found that students were more accurate than police officers, but overall performance was poor for everyone with only 53.9% of the confessions being correctly classified. Moreover, there is a considerable literature on assessing the credibility of adults and children that indicates low accuracy rates for unassisted credibility assessment (Vrij, 2008).

However, to our knowledge, there is one study that specifically examined the validity of credibility assessments of juvenile confessions. Honts, Kassin, and Craig (2014) reported a constructive replication of Kassin et al. (2005) with incarcerated juvenile offenders. They reported two experiments with college student participants who assessed the credibility of confessions given by incarcerated juveniles. The stimulus materials for the Honts et al., study used the same yoked de-

sign and interview methods as in Kassin et al., (2005). The Honts, et al., participants judged juvenile confessions that were presented either as transcripts, audio or video recordings. Judgment accuracy was poor across their two experiments, averaging 52.8% correct, with the participants showing a small truth bias in their judgments. Audio and video presentation modes resulted in more accurate judgments than did transcripts. Participants in Honts et al., were moderately confident in their accuracy judgments and confidence was sometimes weakly associated with accuracy. Assessing the validity and generalizability of the Honts et al., (2014) findings is particularly important as interrogations often follow polygraph examinations.

Polygraph examiners should be particularly well trained and experienced to assess the credibility of confessions and other statements pertaining to wrongdoing, since they do so professionally on a daily basis. Moreover, polygraph examiners are often responsible for interrogating suspects following a failed polygraph examination, and in that context, they frequently elicit confessions. Moreover, the role of polygraph results as a false evidence ploy (either wittingly or unwittingly) was highlighted in some of the false confession reviews (Drizin & Leo, 2004; Kassin, et al., 2010; Kassin et al., 2018) and has been raised in other polygraph research (Honts, 2017). Moreover, in at least two recent Federal cases, expert testimony about poorly conducted and/or improperly evaluated polygraph tests was successfully used as part of the effort to suppress post-polygraph confessions by criminal defendants (*United States of America v. Jamaico Tennison*, 2016; *United States of American v. Tyrone Coriz,* 2018). In this study we replicated the video portion of Honts et al., (2014) with professional polygraph examiners.

Method

Participants

Participants were 83 practicing polygraph examiners (65 male, 18 female) attending continuing education seminars for polygraph examiners. One participant did not provide information concerning law enforcement status or training so that participant's data were eliminated from analyses involving those variables. Forty-nine of the examiners were currently employed by a law enforcement agency. Of the 83 examiners 27 reported that they had attended an interrogation training class from John Reid and Associates.

Procedure

The same confessions used in the Video condition of Experiment 2 in Honts et al, (2014) were used as the stimulus materials in this study. Details about the development of these stimulus materials can be found in Honts et al., (2014). Participants watched videos of 4 true and 4 false confessions and then answered two questions for each confession. Following the methods used by Kassin et al., (2005) participants were told that they would be watching a number of confessions and that some of the confessions were true and that some were false. Participants were not given any information about the base rate of truthfulness. The first question asked for a judgment of whether the confession was true or false. The second asked for a rating of confidence in the decision on a 7-point scale, where 1 = *not confident at all* and 7 = *highly confident*. On a test by test basis we dropped participants who failed to answer one or both of the questions for a confession. The data set for this study consisted of 659 usable judgments regarding the truth of the various confessions and 647 usable confidence ratings.

Results

Overall, 54.7% of the judgments of true versus false confessions were correct. A predictive cross table between true versus false confessions and judgments revealed that 61.8% of the true confessions were judged by the examiners to be true (true negative outcomes), but that 52% of the false confessions were also judged to true (false negative errors). A chi-square analysis revealed that the obtained distribution of judgments was significantly different from chance, $X^2(1, N = 659) = 12.66$, $p < .001$, but the effect size was small, $r(657) = .14$, $p = .04$. This effect size was not different from the video result obtained from the college students who assessed the same confessions in Honts et al., (2014), $z = 0.55$, *ns*. The correlation between judgment accura-

cy and confidence was statistically significant but was small in magnitude, $r(650) = .10$, $p = .01$.

We also examined judgments at the level of employment status. The observed proportions of correct and incorrect judgments of true and false confessions by those polygraph examiners employed by law enforcement were not different from chance, $X^2(1, N = 392) = 2.966$, *ns*. However, the judgments of the civilian polygraph examiners were significantly different from chance, $X^2(1, N = 259) = 12.67$, $p < 0.001$, but the predictive power was low, $r(257) = 0.13$, $p = 0.047$. We also examined the impact of Reid interrogation training on the accuracy of judgments by the polygraph examiners. The judgments of Reid trained polygraph examiners were not different from chance, $X^2(1, N = 216) = 3.63$, *ns*. However, the judgments for polygraph examiners who did not have the Reid training were different from chance, $X^2(1, N = 435) = 9.23$, $p = 0.003$, although the predictive power was low, $r(433) = 0.13$, $p = 0.047$

We also explored the effects of Reid interrogation training and current Law Enforcement status on Confidence with a 2 (participant status: civilian vs. law enforcement) by 2 (Reid trained v. no Reid training) ANOVA of the confidence data. That analysis revealed that law enforcement polygraph examiners were more confident of their judgments ($M = 4.81$, $SD = 1.24$) than were the civilian polygraph examiners ($M = 4.78$, $SD = 1.23$), $F(1, 643) = 8.37$ $p = 0.004$, *partial* $\eta^2 = .013$. Polygraph examiners without Reid training were more confident in their judgments ($M = 4.85$, $SD = 1.22$) than were Reid trained polygraph examiners ($M = 4.70$, $SD = 1.23$) $F(1, 643) = 10.53$ $p = 0.001$, *partial* $\eta^2 = .016$. There was also a significant interaction of Law Enforcement status and Reid interrogation training, $F(1, 643) = 10.53$ $p = 0.001$, *partial* $\eta^2 = .016$, but this interaction should not be interpreted because one cell (Civilian and Reid Trained) contains only the 8 judgments of a single participant.

Following the methods used in Honts et al., (2014) we converted confidence ratings into a predictive believability score by multiplying the confidence value by -1 when the participant concluded that a confession was false. This transformation resulted in a Believability scale where -7 indicated a strong belief that the confession was false and a +7 indicated a strong belief that the confession was true. We treated each estimate of believability as an independent observation and subjected the data to a 2 (confession: true vs. false) by 2 (participant status: civilian vs. law enforcement) by 2 (Reid trained v. no Reid training) ANOVA. Only one significant effect was revealed by that analysis, the interaction of training and confession, $F(1, 639) = 4.33$, $p = .038$, *partial* $\eta^2 = .007$. The analysis also revealed that the ANOVA assumption of homogeneity was violated, *Levine's Statistic* $(7, 639) = 8.267$ $p < .001$. Heteroskedasticity (a test to see if the difference in variability was associated with the independent variable) was significant, *F for heteroskedasticity* $(1, 645) = 4.08$, $p = 0.04$. Figure 1 shows that Reid-trained examiners rated true confessions as less believable ($M = -0.25$) than the false confessions ($M = 2.125$), while examiners without training rated the true confessions as more believable ($M = 1.14$) than false confessions ($M = .01$). Figure 1 also illustrates that the judgments made by Reid trained examiners were much more variable than were judgments made by examiner without the Reid training.

Discussion

Analyses of wrongful convictions and relevant experimental and field psychological research continue to spotlight the indisputable fact that innocent people sometimes confess to crimes they did not commit, either voluntarily or through psychologically coercive interrogation processes (Kassin et al., 2010; Kassin, 2017). Moreover, high profile Innocence Project cases demonstrate that the problem of false confessions is compounded by the fact that police investigators, judges, juries, and others often seem unable to distinguish between true and false confessions, too often accepting the latter at face value. Consistent with decades of research on human failings in deception detection (Vrij, 2008; Vrij, Granhag & Porter, 2011), Kassin et al. (2005) exposed participants to true and false prisoner confessions and found that accuracy rates were generally quite low and that police were no more accurate than laypeople--only more confident and prone to judge confessors guilty. Many of the interrogation risk factors for false con-

fession identified by psychological science are reliable enough for court testimony (Kassin et al., 2018).

As illustrated in several high-profile wrongful convictions, a disproportionate number of false confession cases have involved juveniles (Drizin & Leo, 2004). This pattern is consistent with studies showing that juveniles self-report high false confession rates (Gudjonsson et al., 2006), are more likely to sign false confessions in the laboratory (Redlich & Goodman, 2003), and are prone to compliance effects, suggestibility, and other manifestations of cognitive and emotional immaturity that render them vulnerable to manipulation (Owen-Kostelnik, Reppucci, & Meyer, 2006). Taken together, these literatures have led researchers to identify youth as an important risk factor in the interrogation room (Kassin et al., 2010).

Although juveniles are psychologically at risk, one could argue that the problem could correct itself to the extent that police, polygraph examiners, prosecutors, and others could tell the difference between true and false confessions by juveniles. Kassin et al. (2005) found that people cannot discriminate between true and false confessions given by adult prison inmates. Honts et al. (2014) replicated Kassin et al. with incarcerated juveniles and produced very similar accuracy results with their college student participants.

Here we extended the earlier detection of false confession work by testing polygraph examiners who spend their professional lives detecting deception, conducting interrogations and taking confessions. This population is particularly important as interrogations often follow polygraph examinations and the use of polygraph results as a false evidence ploy (either wittingly or unwittingly) was highlighted in some of the false confession reviews (Drizin & Leo, 2004; Kassin, et al., 2010; Kassin et al., 2018). Using the same confession stimulus materials, we did not find any significant differences between the college students in Honts et al. (2014) and our polygraph examiners. Both college students and the polygraph examiners showed low accuracy and a small truth bias. Within the group of polygraph examiners, there were some individuals who worked in law enforcement and some who were trained in the Reid interrogation technique. The confession veracity judgments of those polygraph examiners who were employed by law enforcement were not better than chance, while the judgments by civilian polygraph examiners were. Judgments by those polygraph examiners who had the Reid training were not different from chance, while the judgments of those without the Reid training were significantly better than chance. The reasons for these differences in judgment accuracy are not clear and need to be examined in further research.

Despite their consistently low accuracy rates, polygraph examiners were generally confident in their judgments showing an average confidence of 4.79 on a 7-point scale where 7 = highly confident. Thus, the average level of confidence by the polygraph examiners was about a standard deviation above the midpoint of the scale. In contrast to the college students in Honts et al., (2014) who were more confident when judging true confessions, our polygraph examiners showed no difference in confidence when judging true and false confessions. However, our analyses of law enforcement status and Reid training revealed that both those variables were associated with significant differences in confidence. Law enforcement officers were more confident than civilian polygraph examiners and participants who were not Reid trained were more confident than those with Reid training. Both of these effects may be due to training histories that instilled a strong belief in the ability to detect deception during interviews and interrogations, despite the lack of scientific evidence that shows such training to be effective. Additional research is needed to follow up on those findings.

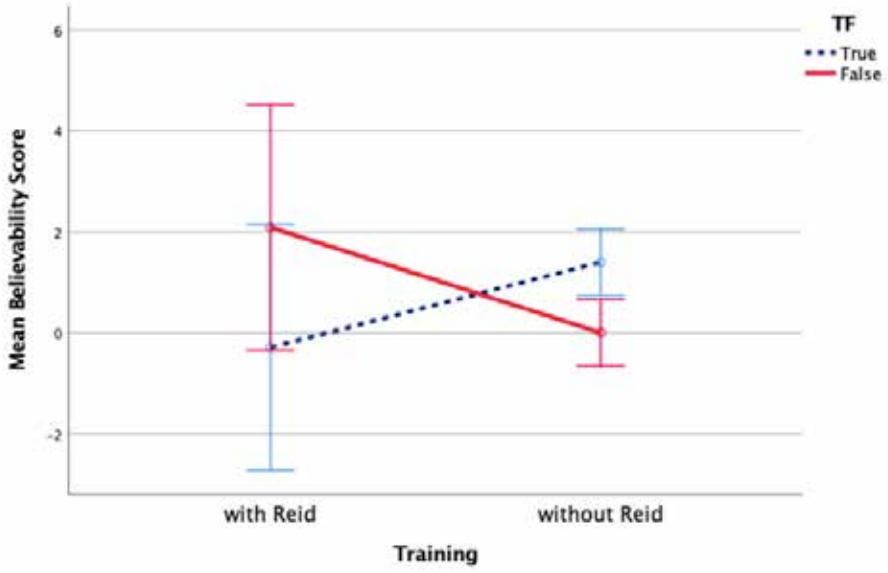
Finally, we examined our data by creating a believability index from the confidence ratings and judgments. When a participant judged a confession to be false, we multiplied their confidence score by -1. This simple transformation created an interval scale to which we could apply more powerful parametric statistics. The results of parametric analyses generally mirrored the effects reported for judgments. However, the analyses of the believability data revealed that the Reid interrogation training not only reduced the accuracy of believability judgments of the confes-

sions to chance, at a more fundamental level it was associated with a complete reversal of reality regarding true and false confessions. Reid trained polygraph examiners found the false confessions to be significantly more believable than the true confessions while those examiners without Reid training found the true confessions more believable than the false confessions. Moreover, the credibility judgments made by the Reid trained examiners were significantly more variable than were those by examiners without the Reid training. Our findings contribute to a growing literature that suggests that not only is Reid training in deception detection not effective, it is significantly counterproductive and dangerous.

In this study, polygraph examiners were asked to make judgments about true and false confession under no threat conditions. Moreover, the polygraph examiners knew they were in a research study and that no real-world consequences were associated with their credibility judgments. Some have questioned the generalizability of findings obtained under similar research circumstances (for example, O'Sullivan, Frank, Hurley, & Tiwana, 2009). However, a meta-analysis (Hartwig & Bond, 2014) failed to find any moderator effects between credibility assessment made under laboratory and field conditions. Hartwig and Bond noted that interpersonal credibility assessments were stable across experimental and field settings for a wide variety of potential moderator variables. As Hartwig and Bond note, their findings are similar to other meta-analytic findings that have compared experimental and field research in a number of other domains (for example, Anderson, Lindsay, and Bushman, 1999). In addition, a recent meta-analysis of the comparison question polygraph test accuracy (138 datasets, 11,474 polygraph examinations) did not find a significant effect of the level of motivation used or of the setting of the research (Honts, Thurber, & Handler, 2018). Although our credibility assessments were set in the context of confessions, we see no reason why these results should be any more or less generalizable than more traditional credibility assessment research.

In summary, our results provide no support for the idea that false confessions by juveniles are more detectable than are those given by adults, even when the judgments are made by professional credibility analysts. Practitioners and triers of fact should be aware that it is unlikely that they will be able to recognize a false confession if one is given to them

Figure 1. Mean polygraph examiner believability scores as a function of true and false confession and Reid Training. Means are illustrated with 95% confidence intervals. Reid training is shown to have negative effects on the evaluation of confession credibility by shifting believability scores in the wrong direction for both true and false confessions and by introducing high variablility.



by a juvenile or an adult. The results of this study highlight the importance of vetting all confessions, either by adults or by juveniles, through independent confirmation of the confession and subsequent confirmation of all new evidence generated in the confession. Confessions that contain only information known to the general public, or known by the interrogators before the interrogation, must be viewed with great suspicion until independent confirmation can be found.

References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3–9. doi: 10.1111/1467-8721.00002

Craig, R. A., Raskin, D. C., & Kircher, J. C. (2011). The use of physiological measures to detect deception in juveniles. *Polygraph*, *40*, 86-99.

Drizin, S. A., & Leo, R. A. (2004). The problem of false confessions in the post-DNA world. *North Carolina Law Review*, *82*, 891–1007.

Evans, A. D., & Lee, K. (2011). Verbal deception from late childhood to middle adolescence and its relation to executive functioning skills. *Developmental Psychology*, *47*, 1108–1116 doi: 10.1037/a0023425

Evans, A. D., Xu, F., & Lee, K. (2011). When all signs point to you: Lies told in the face of evidence. *Developmental Psychology*, *47*, 39–49. doi:10.1037/a0020787

Feld, B. (2006). Police interrogations of juveniles: An empirical study of policy and practice. *Journal of Criminal Law and Criminology*, *97*, 219–316.

Garrett, B. (2008). Judging innocence. *Columbia Law Review*, *108*, 55–142.

Gombos, V. A. (2006). The cognition of deception: The role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs*, *132*, 197–214. doi:10.3200/MONO.132.3.197-214

Gudjonsson, G. H., Sigurdsson, J. F., Asgeirsdottir, B. B., & Sigfusdottir, I. D. (2006). Custodial interrogation, false confession, and individual differences: A national study among Icelandic youth. *Personality and Individuals Differences*, *41*, 49–59.

Hartwig, M. B., & Bond, Jr., C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*, 661-676.

Honts, C. R. (2017, March). *Current FBI Polygraph Practices Put the Innocent at High Risk of Wrongful Accusation, Interrogation, and False Confession.* Paper presented at the American Psychology - Law Society meeting, Seattle, WA.

Honts, C. R., Kassin S. M. & Craig, R. (2014). 'I'd know a false confession if I saw one': A constructive replication with juveniles. *Psychology, Crime and Law*, *20*, 695-704. (published online 15 November 2013). http://dx.doi.org/10.1080/1068316X.2013.854792

Honts, C. R., Thurber, S., & Handler, M., (2018, October). *Polygraph Meta-Analysis. Meta-Análisis de Poligrafía.* Invited lecture series at the annual meeting of the Latin American Polygraph Association (Asociación Latinoamericana de Poligrafistas), Lima, Peru.

Innocence Project (2019). *False Confessions & Recording of Custodial Interrogations*. https://www.innocenceproject.org/causes/false-confessions-admissions/ January13, 2019.

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior, 34*, 3-38.

Kassin, S. M., Meissner, C. A., Norwick, R. J. (2005). 'I'd know a false confession if I saw one': A

comparative study of college students and police investigators. *Law and Human Behavior*, *29*, 211-227.

Kassin, S. M., Redlich, A. D., Alceste, F., & Luke, T. J. (2018). On the general acceptance of confessions research: Opinions of the Scientific Community. American Psychologist, 73, 63-80.

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, *33*, 530–538. doi: 10.1007/s10979-008-9166-4

Owen-Kostelnik, J., Reppucci, N., & Meyer, J. (2006). Testimony and interrogation of minors: Assumptions about maturity and morality. *American Psychologist, 61*, 286-304.

People of the State of New York v. Kharey Wise, Kevin Richardson, Antron McCray, Yusef Salaam, & Raymond Santana: Affirmation in Response to Motion to Vacate Judgment of Conviction (2002). Indictment No. 4762/89, December 5, 2002.

Redlich, A. D., & Goodman, G. S. (2003). Taking responsibility for an act not committed: Influence of age and suggestibility. *Law and Human Behavior*, *27*, 141–156.

Reppucci, N. D., Meyer, J., & Kosteinik, J. (2010). Custodial interrogation of juveniles: Results of a national survey of police. In G. D. Lassiter & C. A. Meissner (Eds.) *Police interrogations and false confessions: Current research, practice, and policy recommendations*, Washington, DC: American Psychological Association, pp. 67-80.

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence*. Garden City, NY: Doubleday.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*, 804–810. doi:10.1037/0012-1649.43.3.804

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behavior during verbal deception. *International Journal of Behavioral Development*, *26*, 436–444. doi:10.1080/01650250143000373

Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children's lying behavior. *Child Development, 79*, 866-881.

(*United States of America v. Jamaico Tennison*, *Crim. No. 15-212 MCA*, 2016. *Order Suppressing Evidence,* United States District Court for the District of New Mexico. (Available from the first author on request.)

United States of American v. Tyrone Coriz, CR. No. 17-1105 JHC, 2018). *Memorandum Opinion and Order*, United States District Court for the District of New Mexico. (Available from the first author on request.)

Vrij, A. (2008). *Detecting lies and Deceit: Pitfalls and opportunities. Second edition*. West Sussex, UK: John Wiley and Sons.

Vrij, A., Granhag, P. A., Porter, S. (2011). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, 11*, 89-121.

A Study of the Validity of Polygraph Examinations in Criminal Investigation

Final Report to the National Institute of Justice
Grant No. 85-IJ-CX-0040

David C. Raskin, Ph. D
Principal Investigator

John C. Kircher, Ph. D
Co-Principal Investigator

Charles R. Honts, Ph. D
Research Associate

Steven W. Horowitz, M.S
Research Assistant

Department of Psychology University of Utah
Salt Lake City, Utah 84112
May 1988

Abstract

This project was designed to answer several major questions concerning the validity of the control question polygraph technique (CQP) for assessing truth and deception in criminal investigations. Confirmed and unconfirmed polygraph charts from examinations by the U. S. Secret Service in criminal investigations were sampled and blindly interpreted by six polygraph examiners from that agency and one psychophysiologist at the University of Utah. They were also subjected to computer interpretation using algorithms developed at the University of Utah.

The accuracy of human and computer interpretations was very high. Original examiners' decisions on individual relevant questions ranged from 91-96% correct on confirmed truthful and 85-95% correct on confirmed deceptive answers. Blind interpretation produced somewhat lower accuracies, from 63-85% on truthful and 84-94% on deceptive answers. The accuracy of computer interpretations was higher than blind interpretations, ranging from 95-96% on confirmed truthful and 83-96% on confirmed deceptive subjects. The results provide considerable support for the accuracy of decisions by the original examiners and for the use of computer interpretations for quality control decisions concerning the outcomes of polygraph tests.

The generalizability of laboratory research on CQP tests was analyzed using computer generated response profiles and double cross-validation of models developed from laboratory and criminal suspects. Results indicated that laboratory findings may provide considerable information about the underlying processes and accuracy of field polygraph examinations. They also indicated a need to improve the choice of relevant questions in multiple issue testing and for modifications to improve the accuracy of field numerical evaluation.

Although the use of polygraph examinations in criminal investigations and security applications by the Federal Government more than tripled during a 10-year period, there appears to be a lack of adequate scientific research on the accuracy of such field applications (Office of Technology Assessment, 1983). The OTA study was mandated by the House Committee on Government Operations, and it provided an extensive review of the existing literature on polygraph research and applications. It concluded that although there is evidence that polygraph accuracy exceeds chance in field applications, there is a strong need for further research.

Every federal investigative agency, including those within the Department of Defense, uses polygraph examinations in criminal investigations (OTA, 1983). State and local law enforcement agencies, courts, and attorneys make extensive use of such techniques to screen suspects, to dispose of cases, to elicit confessions following deceptive results, to generate evidence for court proceedings, to provide information for pre- sentencing investigations, and for various other applications within the criminal justice system. The extent to which these applications provide valid information and the weight that should be accorded to such results in various contexts are hotly debated issues (Lykken, 1981; OTA, 1983; Raskin, 1982, 1986).

The OTA report highlighted the pressing need for additional research on this problem. In response to concerns expressed in their report, this project was designed to provide information that is crucial to enlightened decisions regarding the range of useful applications of polygraph techniques in the criminal justice system, and ways to improve existing techniques.

Objectives of the Research Project

The first objective of this project was to provide a definitive study of the validity of control question polygraph examinations in criminal investigation as well as reliable estimates of the accuracy of truthful and deceptive outcomes. The research was designed to generate important data that will be useful in guiding policy decisions in different settings, such as the extent to which polygraph tests should be used in different contexts and the amount of confidence that can be placed in the outcomes of such tests.

The second objective of this project was to assess the performance of polygraph examiners with different educational backgrounds and different types and amounts of experience with polygraph techniques. The analytic techniques that we applied to the data provided information about the qualitative and quantitative differences in the ability of different polygraph examiners to interpret polygraph recordings accurately.

The third objective of the project was to assess the efficacy of an automatic and objective computer method for interpreting the outcomes of polygraph examinations. At the present time, most federal investigative agencies have quality control procedures that require materials from all polygraph examinations conducted in the field to be sent to a central office for an independent evaluation before the results receive final approval.

Independent evaluations are intended to minimize mistakes in interpretations caused by subjective influences or insufficient skill or experience, and they are also used to identify examiners in the field who are experiencing difficulties in their performance. However, the current procedures are slow and costly and may not solve all of the operational problems.

Computer analysis might perform better than independent human interpreters and be less costly in terms of time and resources. Research has established that there is wide variation in the abilities of polygraph examiners to interpret correctly the physiological recordings obtained in such tests (Raskin, Barland, & Podlesny, 1978), and computer methods have been demonstrated to perform as well as the most experienced and sophisticated human interpreters (Kircher & Raskin, 1988). If a computer method could provide the same information as that obtained from human interpreters, at a significantly lower cost and within minutes instead of days or weeks, problems could be identified more readily and with greater speed. All subjectivity would be removed from the process; more accurate decisions would be available immediately; ex-

aminers in the field would receive immediate feedback that they could consider before the examination is terminated; polygraph examination results could be utilized in a more effective manner; and additional training could be provided on the basis of the computer identification of particular examiner deficiencies. The entire process could benefit from a powerful, rapid, and scientific approach to the diagnosis of truth and deception.

The fourth objective of the project was to assess the extent to which laboratory mock-crime experiments provide information and results that have implications for field applications of polygraph examinations. A large amount of scientific laboratory research has investigated problems such as the influence of personality factors, the effectiveness of countermeasures and drugs, the usefulness of different physiological measures and examination techniques, and the accuracy of control question polygraph examinations (Raskin, 1986). However, the extent to which the results of such studies can be generalized to field applications of polygraph techniques is not entirely clear. The use of computer analytic techniques may provide information concerning the extent to which the findings of laboratory research can be utilized in making decisions and formulating policy regarding applications of polygraph techniques in the criminal justice system.

Methodological Issues

In order to assess the accuracy of control question polygraph tests in criminal investigation, a reliable criterion of ground truth must be available against which the test results may be evaluated. Complete confidence in the criterion can be obtained using laboratory simulations that employ mock crimes and field polygraph techniques (Raskin, 1982). The results of such experiments have frequently produced accuracies in excess of 90% (Bradley & Ainsworth, 1984; Dawson, 1980; Gatchel, et al., 1983; Kircher & Raskin, 1988; Podlesny & Raskin, 1978; Raskin & Hare, 1978; Rovner, Raskin, & Kircher, 1979). However, critics of laboratory research have argued that the motivational structure of the field situation cannot be simulated in the laboratory, and the greater consequences of the test outcomes in criminal investigations produce different physiological reactions and higher rates of error (Lykken, 1981).

If the possible problems of motivation and context inherent in the laboratory simulations are to be overcome, it is necessary to use examinations from actual criminal investigations. On the basis of 10 studies that met minimal criteria for methodological adequacy, OTA concluded that the average accuracy of polygraph tests in the field situation is 90% on guilty subjects and 80% on innocent subjects. However, these studies raise several additional problems.

A criterion for ground truth is more difficult to establish in the field situation than in the laboratory, and it is necessary to develop criteria with a high degree of reliability and accuracy. Three approaches have been taken to that problem. One method is to submit a11 of the case information except the polygraph results to a panel of experts who are asked to make judgments of guilt or innocence using the available information and disregarding legal technicalities (Bersh, 1969; Raskin, Barland, & Podlesny, 1978).

Accuracy of the polygraph tests is then determined by comparing the test outcomes to the composite judgments of the panel. The problem with this method is the fallibility of panel judgments that are based on a vague evaluation of evidence of unknown and variable quality and quantity. Therefore, the findings of panel studies of polygraph accuracy are open to serious question. Similar and more severe problems arise when polygraph accuracy is assessed against a criterion of judicial outcomes (Raskin, 1982, 1987).

It is generally agreed that the best criterion for assessing the accuracy of field polygraph tests is confirmation by means of confessions by guilty persons (Horvath, 1977; Lykken, 1979; Raskin, 1987). In such studies, polygraph charts are obtained from cases in which the guilty person subsequently confessed. Sets of such confirmed deceptive and confirmed truthful polygraph charts are assembled, and they are then submitted to other polygraph examiners for blind interpretation. The accuracy of their interpretations is assessed against the criterion of ground truth

independently established by the confessions. The accuracies reported by such studies range from 64% in the Horvath study (1977) to 98% in the Raskin study (1976).

Only limited conclusions can be drawn from the available field studies that have used a confession criterion. Questions have been raised about the method of selecting the charts to be evaluated, their representativeness with regard to polygraph tests in general, and the training and qualifications of the polygraph interpreters (Raskin, 1987).

For example, the Horvath study included examinations of victims and witnesses as well as suspects (Barland, 1982), which complicates the interpretation of the results. In Barland's re-analysis of Horvath's data, he found that all but one of the false positive errors occurred on victims and witnesses, indicating that the Horvath study cannot be used to estimate the accuracy of polygraph tests on criminal suspects.

Another major problem with the Horvath study and those from the Reid organization (Horvath & Reid, 1971; Hunter & Ash, 1973; Kleinmuntz & Szucko, 1982; Slowik & Buckley, 1975; Wicklander & Hunter, 1975) is the failure to use control question techniques that are accepted by most federal agencies and supported by scientific research (Raskin, 1987). Furthermore, the interpreters in these studies were not adequately trained or experienced in the use of numerical evaluation of polygraph charts, and only the Horvath study even attempted to employ numerical methods. All of the examiners and interpreters in these studies were trained in a method that involves the observation and utilization of so-called "behavior symptoms" in the diagnosis of truth and deception. Such methods have been shown to be useless for diagnosing truth and deception, and they produce lower rates of accuracy than numerical interpretation (Raskin et al., 1978).

The Reid studies also suffer from the additional weakness of having used cases where employers referred their employees for polygraph tests with no option to decline to take the test, and the Kleinmuntz and Szucko study did not even use qualified polygraph examiners or accepted methods of chart interpretation.

Finally, there is the problem of case selection and the generalizability of the results of validity studies based on confession criteria. In addition to the above problems, which indicate that many studies used cases that are not representative of polygraph examinations on criminal suspects, the methods used to select the cases in the Reid studies have not been specified in a manner that permits a definitive evaluation of whether or not the cases were selected in an unbiased manner (Raskin, 1987).

Even if these problems did not exist, there is a more fundamental problem with the use of cases confirmed by confessions. If tests are selected because someone (either the person who took the test or another suspect) confessed to the crime subsequent to the polygraph test, there arises the question of whether or not such tests are representative of the population of tests of suspects who agree to take polygraph tests in connection with a criminal investigation.

In all but one of the studies mentioned above, no data were presented concerning the proportion of cases resolved by confessions in the population of cases from which the data were drawn. Only the Raskin (1976) study provided that information, and it indicated a confession rate of only 17% in the set of tests from which the sample was drawn. It is possible that cases in which confessions are obtained are not representative of polygraph tests of criminal suspects in general, and the generalizability of the results of such studies is thereby limited.

It has been argued that only those subjects whose polygraph charts are most strongly indicative of deception are interrogated (Iacono, in press). Therefore, the resulting confessions may inflate reported accuracy by biasing the selection of charts chosen in field validity studies. Subsequent blind interpretations of those charts are likely to produce correct deceptive decisions more frequently than would occur if subjects who produced weaker deceptive polygraph charts were also interrogated to attempt to obtain confessions. Therefore, we also performed analyses to determine if differences in the strength of physiological results indicative of deception are obtained

from suspects who were considered deceptive and subsequently confessed and from suspects who were considered deceptive and did not confess.

In order to overcome many of the methodological problems cited above, this project investigated the accuracy of the control question test in actual criminal investigations where standard field polygraph examination techniques were used and numerical evaluation was employed by adequately trained interpreters, including blind interpreters. The data also allowed us to assess the effectiveness of the computer methods to analyze polygraph charts for the automatic and objective diagnosis of truth and deception (Kircher & Raskin, 1988).

The use of computer algorithms and software, and of extensive multivariate statistical analyses made it possible to assess the relationships between polygraph recordings obtained in field examinations and the qualitative and quantitative nature of polygraph recordings obtained in mock crime laboratory experiments. These analyses provided the basis for estimating the extent to which laboratory research evokes emotional and physiological responses that are similar to those observed in the field situation. The obtained data allowed us to evaluate how far the results of laboratory research on polygraph techniques may be generalized to the application of polygraph examinations in the criminal investigation context.

Research Methods

Our initial objective was to obtain a sample consisting of the polygraph charts from 200 examinations conducted by a federal law enforcement agency. The U.S. Secret Service agreed to provide the materials from their files and the services of some of their examiners to participate in the study. The Secret Service was a logical choice for this study because they have a very high-quality polygraph program with more than 20 experienced and well-trained examiners (Raskin, 1984). They conduct in excess of 1,000 polygraph examinations per year in the context of criminal investigation, and they utilize standard control question procedures and numerical interpretations of the polygraph charts. Furthermore, OTA (1983) reported that they achieve very high rates of admissions and confessions that provide confirmation of more than 90% of their polygraph diagnoses. That high rate of confirmations would ensure that the results are not dependent on the selection of a small, non representative sample of cases, a common problem in studies that rely on a confession criterion for establishing ground truth.

Cases were to be selected to provide 80 tests of suspects who were confirmed as deceptive at some time after their polygraph test and 80 tests of suspects subsequently confirmed as truthful. An additional sample of 20 unconfirmed deceptive results and 20 unconfirmed truthful results was also sought.

A confirmed deceptive suspect is one who was examined on the polygraph and subsequently admitted having lied to one or more of the relevant questions that pertained to the crime under investigation. A confirmed truthful suspect is one who was examined on the polygraph and was later cleared of the allegation or suspicion by the admission or confession of another person. In this study, we required independent corroboration of the confession in the form of some type of physical evidence. Unconfirmed results are those for which no admission or confession was obtained either to inculpate or exculpate the person who took the test.

By including a sample of unconfirmed polygraph results, we were able to determine if there are qualitative or quantitative differences in the physiological reactions between cases in which deception was indicated by the polygraph charts and the polygraph subject confessed and those in which the charts indicated deception and no confession was obtained.

For each of the categories above, we planned to select tests so as to obtain ha1f from cases where there was on1y one suspect and half where there was more than one suspect. That would have permitted us to determine if there are differences in outcomes when the examiner expects that at least one of the suspects will produce a truthful outcome (multiple-suspect cases) as compared to single-suspect cases where there would be a

higher probability that the suspect is guilty. The resulting design of the sample was to be as follows:

	Confirmed Deceptive	Confirmed Truthful	Unconfirmed Deceptive	Unconfirmed Truthful
Single Suspect	40	40	10	10
Multiple Suspect	40	40	10	10

The polygraph charts were selected only on the basis of the type of case as described above. The decision regarding truth or deception made by the examiner and the quality or characteristics of the polygraph charts themselves played no role in the selection process, with two exceptions. First, the tests must have included at least three charts with a control question format. Tests that were incomplete in those respects were not used. Second, if there was an equipment malfunction or examiner error that rendered the charts technically unusable or incomplete, the examination was not included in the sample. The cases were selected first, and the polygraph charts were then inspected to determine if they were to be retained or discarded for failure to meet the standards of completeness or technical adequacy.

Subject Selection

Three strategies were employed in selecting cases. Initially, the U.S. Secret Service case logs for all 1,757 polygraph examinations conducted during FY1983 and FY1984 were coded for type of case, examiner, pretest admissions, and post-test confessions and entered into a computer file. All of the examinations were then screened by a computer program that selected all cases with post-test admissions or confessions.

The 241 cases selected by the computer program were then requested from the Washington, D. C. Headquarters of the Secret Service. The Secret Service personnel then requested the case files from the field offices where they were located. When they received the files, they removed all identifying information from the polygraph charts and recoded them with new identification numbers that we supplied. These recoded charts were taken from the case files and sent to the University of Utah. The case files without the polygraph charts were sent to the Secret Service Field Office in Salt Lake City, where they were evaluated by members of our University of Utah research team.

Confirmation of truthfulness or deception by the polygraph subject was based on a two-step criterion. The first step required an admission or confession by the subject who took the polygraph test or by another suspect in the case who either inculpated or exculpated the subject who was tested. The second step required that admissions and confessions be supported by independent evidence that corroborated the admission or confession, such as recovering counterfeit notes or printing plates described in the confession, recovering the money stolen from a bank, or an analysis of the handwriting of a forged signature.

A very stringent criterion for confirmation was employed to increase the reliability and validity of the criterion so as to avoid errors in the subsequent analyses of the accuracy of the polygraph results and other types of analyses based on the confirmed polygraph results. The use of such a stringent criterion also made it more difficult to confirm cases that were otherwise confirmed by admissions or confessions. Therefore, it was difficult to fill all of the cells in the planned sample as described above. Although it appeared that the Secret Service had a lower rate of confirmation than that reported to OTA (1983), our stringent requirements for purposes of this research eliminated many cases that can reasonably be assumed to have been confirmed for other purposes.

From this initial sampling, we obtained 127 sets of polygraph charts, 93 from multiple-suspect cases and 34 from single-suspect cases. Of the 93 multiple-suspect cases, 19 polygraph subjects were confirmed as having answered one or more relevant questions truthfully, 32 were confirmed as having answered one or more relevant questions decep-

tively, 7 were confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and 35 were not confirmed on any question. Of the 34 single-suspect cases, 14 polygraph subjects were confirmed as having answered one or more relevant questions deceptively, 4 were confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and the remaining 16 subjects were not confirmed on any question. We obtained no confirmed truthful single-suspect subjects from this initial sampling.

In order to increase the likelihood of obtaining confirmed truthful subjects, we used another approach of requesting all cases with pretest as well as post-test admissions and/or confessions. Cases from the first six months of FY1985 were coded as described above, and cases were then selected by a computer program. Of the 325 cases examined, 95 were selected by the computer program and requested from the Secret Service.

Materials from those cases were sent to Salt Lake City in the same manner as previously described. Only charts from subjects needed to fill incomplete group categories were selected and coded.

From this second sampling we selected 32 multiple-suspect subjects and 5 single-suspect subjects. Of these 32 multiple-suspect subjects, 14 were confirmed as having answered one or more relevant questions truthfully, 11 were confirmed as having answered one or more relevant questions deceptively, one was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and six were not confirmed on any question. Of the five single suspect subjects, four were confirmed as having answered one or more relevant questions deceptively and one was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test. Again, we obtained no confirmed truthful single-suspect subjects.

The third strategy resulted in an exhaustive sample of multiple-suspect cases. By this time, it was clear that it was not possible to fill the confirmed-truthful, single-suspect category, so we concentrated on trying to fill the multiple-suspect cells. We hoped that an exhaustive sample of all multiple-suspect cases would enable us to obtain additional confirmed truthful subjects.

From the 440 cases from the first six months of FY1986, we selected all of the 35 multiple-suspect cases and requested them from the Secret Service. Materials from those cases were sent to Salt Lake City in the same manner as previously described. From this third sample, we obtained 12 subjects, six of whom who were confirmed as having answered one or more relevant questions truthfully, five who were confirmed as having answered one or more relevant questions deceptively, and one who was confirmed as having answered at least one relevant question truthfully and at least one relevant question deceptively in the same test.

The polygraph charts obtained from the total of 176 cases from the three samples consisted of 39 subjects confirmed to have answered one or more relevant questions truthfully, 66 subjects confirmed to have answered one or more relevant questions deceptively, 14 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively on the same test, and 57 subjects who were not confirmed on any questions.

Blind Interpretations

Blind interpretations were conducted by seven interpreters, six of whom were U. S. Secret Service polygraph examiners who had been trained at the U. S. Army Military Police School. Of the Secret Service examiners, two were experienced examiners who performed quality control evaluations at their Washington, D. C. headquarters (quality control), two were stationed at field offices and had more than one year of experience as polygraph examiners (experienced examiners), and two were stationed in field offices and had less than one year of experience as examiners (inexperienced examiners). The other interpreter was a doctoral level psychophysiologist who had been licensed as a polygraph examiner for 10 years.

One hundred of the obtained cases

were selected for scoring by the seven blind interpreters using random processes to fill three categories. Forty deceptive subjects were selected from the total sample of subjects confirmed to have answered at least one relevant question deceptively, but not confirmed to have answered any relevant question truthfully. The 13 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively were coded as truthful subjects and were combined with the other subjects confirmed to have answered at least one question truthfully. Forty subjects were selected at random from this population of truthful subjects. The random procedure resulted in the selection of 13 of the 14 subjects who had been confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively. Twenty subjects were selected randomly from the sample of unconfirmed cases.

After the charts had been blindly interpreted, it was discovered that one confirmed truthful and three confirmed deceptive subjects did not meet the criteria for selection because their polygraph results were from a second test. Therefore, they were discarded from the sample and could not be replaced. That reduced the sample to 26 subjects confirmed to have answered one or more relevant questions truthfully, 37 subjects confirmed to have answered one or more relevant questions deceptively, 13 subjects confirmed to have answered at least one relevant question truthfully and at least one relevant question deceptively in the same test, and 20 unconfirmed subjects.

Division of Cases for Analysis

The cases appeared to belong to three natural categories of verification. Complete Verification occurred when responses to all relevant questions in an examination were confirmed as either truthful or deceptive. Partial Verification occurred when responses to some relevant questions in an examination were confirmed as either truthful or deceptive, but there was also at least one response to a relevant question that remained unconfirmed. Mixed Verification occurred when suspects were confirmed to have answered at least one relevant question truthfully and at least one question deceptively within the same polygraph examination. Subjects were initially separated into these three categories of verification for purposes of data analysis.

Numerical Scoring

The original examiners and the Secret Service interpreters used the numerical scoring system developed and taught to federal polygraph examiners at the U. S. Army Military Police School. The psychophysiologist used the numerical scoring system developed and validated at the University of Utah. Although the psychophysiologist used a different numerical scoring system than the other interpreters, differences in effectiveness of these systems are slight (Weaver, 1985). In general, both numerical scoring systems follow the scoring system described by Raskin and Hare (1978) and Podlesny and Raskin (1978). Differences in physiological reactions to relevant and control questions in electrodermal activity, respiration, peripheral vasomotor activity, and relative blood pressure were evaluated.

The following characteristics were used to assess the strength of the responses: electrodermal response amplitude and duration; decrease in amplitude and rate of respiration, increases in respiration baseline; duration and amplitude of decreases in finger pulse amplitude, and amplitude and duration of baseline increase in relative blood pressure. Reactions were not scored if they began more than 5 seconds following the subject's answer. Minimum latencies of 0.5 second and 2.0 seconds were adopted for skin conductance and finger pulse amplitude responses, respectively, and reactions that began prior to the minimum latencies were not scored. For each physiological system, each pair of control and relevant questions was assigned a score from -3 to +3 (except by the two Secret Service quality control interpreters and one of the inexperienced Secret Service interpreters who elected to assign scores from -1 to +1) depending on the strength of the difference between the reactions to the two question types. Positive scores were assigned when reactions to control questions were stronger, negative scores were assigned when reactions to relevant questions were stronger, and scores of

zero were assigned when the reactions to relevant and control questions were approximately equal in strength.

Computer Scoring

Data Entry. The physiological data had been recorded at 2.5 mm per second on standard polygraph chart paper that was 20 cm in width. Physiological responses to each control and relevant question in the first three repetitions of the question sequence were manually traced on a digital tablet, the output of which was read by a laboratory microcomputer. The laboratory assistants who traced the response wave forms had no knowledge of the subjects' criterion status.

The computer was programmed to sample skin resistance (SR) and thoracic and abdominal respiration(R) channels at 10 Hz for 20 seconds following the onset of each test question. The program also read the times and levels of systolic and diastolic points of the blood pressure (BP) tracings. From the series of systolic and diastolic points for each question, average changes in BP were computed for 2 seconds immediately preceding the onset of question presentation and 20 seconds following question onset. The data for each chart were stored on a floppy disk in a file identified by subject and chart numbers and date.

Data Editing. A second program was written to read the data files from the floppy disks, display the physiological response wave forms on the computer screen, and edit movement artifacts. The editing program also rescaled the data when sensitivity adjustments had been made between charts. Artifacts of approximately 1-3 seconds in duration were replaced with interpolated values. A response containing multiple artifacts or artifacts greater than 3 seconds in duration was considered unusable and was not used.

Data Quantification. The SR and BP response curves were divided into segments, and each segment was tested for positive slope. Approximate times of occurrence of low points in the waveform were identified by changes from zero or negative slope to positive slope. High points in the curve were isolated between successive pairs of low points. The exact times and levels of low points were then isolated between successive pairs of high points.

The procedures for locating high and low points in the SR and BP wave forms differed in two respects. Tests for positive slope were performed between successive samples (seconds) of the BP response curve and between every fifth sample (500 ms segments) of the SR response curve. In addition, a stepwise averaging procedure smoothed the SR response curve prior to testing the 500 ms intervals for positive slope. After the approximate times of low points in the response curve had been identified in those intervals, the exact times and levels of high and low points were isolated in the original sequence of 100 ms time samples.

The times and levels of high and low points in the response curves provided the information needed to quantify all of the physiological variables listed below, except respiration length that was quantified with a separate algorithm (Timm, 1982). The first six of the following seven types of measurements were obtained from the SR and BP response wave forms:

Amplitude. Differences were computed between each low point and every succeeding high point identified in the response curve. Amplitude was defined as the greatest obtained difference.

Rise time. Time to the nearest 100 ms for SR and 1,000 ms for BP was measured between response onset and the occurrence of the maximum.

Half recovery time. Time of occurrence of the maximum was subtracted from the time at which the recovery limb reached a level that was half of the amplitude. When the response did not recover sufficiently to reach the criterion, the interval was measured to the end of the 20-second sampling period.

Rise rate. Amplitude was divided by rise time.

Half recovery rate. Half of amplitude was divided by half recovery time.

Latency to response onset. Time to the nearest 100 ms for SR and 1,000 ms for BP was measured from stimulus onset to response onset.

Respiration length. Linear distance was measured between successive pairs of 1000ms samples from question onset to the 10^{th} post-stimulus second. The 100 measurements were summed to yield a length measure in relative units for each respiration channel. After standardizing the measurements for the two respiration channels as described below, standard scores for the two channels were averaged to obtain a combined index of respiratory suppression (R length) for each control and relevant question.

Variable Generation Procedures

For each subject and each response parameter, repeated measures were obtained across the control and relevant questions for the first three repetitions of the question sequence. The number of measurements depended on the number of control and relevant questions presented, and they ranged from four to eight per chart.

The set of measurements for each response parameter was converted to standard scores. The transformation to standard scores within each subject established a common metric among the various types of response parameters. Since unit variance was partitioned among the repeated observations for each response parameter, it also controlled for the tendency of some individuals to react more strongly in one response system than in another.

The relative magnitudes of reactions to each relevant question were assessed separately for each response parameter. The mean standard score for repetitions of a given relevant question was subtracted from the mean standard score for reactions to all of the control questions on the test. The size of the z-score difference indexed the magnitude of differential reactivity, and its sign indicated if the average response to the relevant question was greater or less than the average response to the control questions.

Variable Selection Procedures

Since it is difficult to obtain a stable prediction model from a large set of redundant measures (McNemar, 1969), three all-possible-subsets regression analyses (Pedhazur, 1982) were performed to identify a reliable subset of variables that was optimal for discriminating between truthful and deceptive responses to relevant questions.

The first regression analysis was performed using only those cases in which all answers to relevant questions had been confirmed as either truthful or deceptive (Complete Verification). The second regression analysis was conducted using only those cases in which some but not all answers to relevant questions were confirmed as either truthful or deceptive (Partial Verification). The Complete and Partial Verification samples were combined for the third analysis. Cases in which some answers to relevant questions had been confirmed as truthful and others had been confirmed as deceptive (Mixed Verification) were not included in these preliminary analyses.

The best subset of variables for discriminating between confirmed truthful and deceptive subjects in the Complete Verification sample consisted of four variables: SR Amplitude, SR Rise Rate, BP Amplitude, and R Length. The same set of four variables was the seventh best subset with four variables for the Partial Verification sample of subjects, but three of the four measures appeared as the best subset of three variables for that sample. When the Complete and Partial Verification samples were combined (Pure Sample), the four-variable model was again selected as optimal for discriminating between the groups. Therefore, the four variable model was adopted for assessing the discriminant validity of the computer method.

Structure of the Probability Model

A probability-generating model was developed to calculate the probability of group membership for each subject. The probability of group membership was defined as the probability of truthfulness for a confirmed Truthful subject or the probability of deception for a

confirmed Deceptive subject. Its complement, one minus the probability of group membership, was the probability that the subject was a member of the wrong criterion group.

The model consisted in part of a discriminant function that was used to calculate a discriminant score for each subject. The discriminant score was a weighted combination of the subjects' scores on the four physiological variables. The weights for the variables were those that maximized the discrimination between confirmed truthful and deceptive individuals in the sample.

The model also incorporated two likelihood functions that were used to calculate the conditional probability of group membership given the obtained discriminant score. The two likelihood functions formed partially overlapping normal curves, the parameters of which were specified by the means and variances of the distributions of discriminant scores for confirmed truthful and deceptive subjects in the sample. To calculate the probability of group membership for a subject, two maximum likelihood estimates were computed using the subject's discriminant score and the equation for the normal probability density function (Winkler & Hays, 1975). The two likelihoods were then combined according to Bayes' Theorem to calculate the probability of group membership for each individual (Kircher & Raskin, 1988).

Results

Numerical Scores

Original Examiners. Differences in the numerical scores assigned by the original examiners for the three verification categories were tested by a 2-way ANOVA comprised of Confirmation (Truthful/Deceptive) and Verification (Complete/Partial/Mixed). The means for the 6 cells of the ANOVA are shown in Table 1.

The analysis indicated a main effect for Confirmation, $F(1, 164) = 247.13$, $p < 0.0001$. Positive numerical scores were associated with questions confirmed to have been answered truthfully, whereas negative numerical scores were associated with questions confirmed to have been answered deceptively.

The analysis also indicated a significant Confirmation X Verification interaction, $E(2, 164) = 5.35$, $Q = 0.006$. An examination of the means indicates that this effect was primarily due to a reduction in numerical scores for confirmed truthful responses in the Mixed Verification Group.

A further ANOVA failed to find differences between the Complete and Partial Verification Groups, so the Complete and Partial Verification Groups were combined to form a Pure Verification Group that was then compared to the Mixed Verification Group. That ANOVA also revealed a similar interaction of Confirmation and Verification, $F(1, 166) = 8.90$, $p = 0.003$.

The extent to which the original examiners' numerical scores predicted the truthful/deceptive criterion was assessed by correlating the numerical scores with the confirmation criterion for individual questions. For Pure Verification subjects, the correlation with the criterion was significant, $r(136) = 0.79$, $r < 0.001$. The correlation with the criterion was also significant for the Mixed Verification sub-

Table 1. Original Examiners' Mean Scores for Confirmed Single Questions

	Complete Verification	Partial	Mixed
Truthful	**4. 1**	**6.0**	**2.7**
Deceptive	**-5.6**	**-4.3**	**-2.8**

jects, r (33) = 0.61, p< 0.01, but the correlation for the Mixed Verification subjects was significantly smaller than the correlation for the Pure Verification subjects, *I*=1.84, P = 0.03 (one-tailed).

Blind Interpretations

Complete, Partial. and Mixed Verifications. Possible differences in numerical scores assigned by various blind interpreters for the three categories of Verification were assessed by a repeated measures ANOVA. An analysis of Interpreters by Confirmation (Truthful/Deceptive) by Verification (Complete/Partial/Mixed) indicated a significant main effect for Confirmation, *F* (1, 162) = 99.40, Q < 0.001. The analysis failed to find a main effect for Verification, but there was a significant Confirmation X Verification interaction, *F* (2, 162) = 6.60, P = 0.002. Inspection of the means indicated that this interaction was primarily due to a reduction in the size of the numerical scores for confirmed truthful responses by subjects in the Mixed Verification Group (M = 0.41) as compared to confirmed truthful responses by subjects in the Complete (M = 2.20) and Partial (M = 2.68) Verification Groups. No interaction of Verification with Interpreters was found. A significant interaction of Interpreters and Confirmation was found, *F* (5, 830) = 3.26, p = 0.006, and is discussed below in the section on Interpreter Characteristics.

An Interpreters by Confirmation by Verification (Complete/Partial) ANOVA was conducted to determine if there were differences between numerical scores for cases with Complete Verification and those with Partial Verification. This analysis indicated a significant main effect for Verification, *F* (1,122) = 4.04, p = 0.047. Inspection of the means indicated that for suspects with Complete Verification the numerical scores for individual questions tended to be more negative (M = -0.72) than the numerical scores to confirmed questions for suspects with only Partial Verification (M = -0.095). There was no significant interaction between Verification and Interpreters or Confirmation.

Since the difference in numerical scores for the Partial and Complete Verification was quite small, these groups were combined (Pure Verification) and compared to the Mixed Verification Group using an Interpreters by Confirmation (Truthful/Deceptive) by Verification (Pure/Mixed) ANOVA. This analysis indicated a strong main effect for Confirmation, *F* (1, 164) = 69.12, p < 0.001, and an interaction of Confirmation and Verification, *F* (1, 164) = p = 0.001. This effect was due to the reduction in the numerical scores for confirmed truthful responses in the Mixed Verification group (M = 0.41) as compared to the Pure Verification group (M = 2.33).

Reliability. All confirmed questions were used to assess inter-rater reliability in the assignments of scores, since ANOVA failed to indicate that Interpreters performed differently on the three Verification groups. A complete pairwise correlation matrix was calculated among the numerical scores assigned by the six Secret Service blind interpreters, and the inter-rater correlations were all significant, ranging from 0.80 to 0.88 *(M* = 0.84). The pairwise correlations between the scores of the psychophysiologist and the Secret Service blind interpreters were also significant, ranging from 0.76 to 0.8 2 (M = 0.79).

Interpreter Characteristics. The numerical scores assigned by the six Secret Service blind interpreters were subjected to a Confirmation (Truthful/ Deceptive) by Interpreter repeated measures ANOVA. This analysis indicated that the main effect of Interpreters was not significant, *F* (5,830) = 1.59, but there was a significant interaction between Interpreters and Confirmation, F (5, 830) = 3.26, p = 0.006. The means for the six Secret Service blind interpreters shown in Table 2 indicate that the interact ion of Interpreter and Confirmation was primarily due to lower scores assigned by the two quality control interpreters on confirmed truthful responses. This may have been a consequence of their use of scores of only +1, 0, and -1.

The performance of the interpreters was further assessed by point bi serial correlations between the interpreters' numerical scores on individual questions and the Truthful/Deceptive criterion. These correlations are also shown in Table 2.

The differences among interpreters appeared to be individual differences not associated with examiner experience. The best

Table 2. Mean Numerical Scores on Individual Questions and Correlations With The Criterion for the Seven Blind Interpreters and The Original Examiners

	Confirmed Truthful	Confirmed Deceptive	Correlation With Criterion
Original Examiners	4.7	- 4.8	0.79
Quality Control Examiner A	1.9	-3.1	0.62
Quality Control Examiner B	2.0	-3.4	0.64
Experienced Examiner A	3.0	-3.4	0.65
Experienced Examiner B	2.3	-3.3	0.57
Inexperienced Examiner A	2.2	-2.7	0.62
Inexperienced Examiner B	2.2	-3.6	0.62
Psychophysiologist	2.6	-4.8	0.66

performance was shown by an experienced field examiner, r = 0.65, and the poorest performance was by the other experienced field examiner, r = 0.57. The difference between these two correlations was significant, 1 (190)= 5.01, p < 0.01. The inexperienced examiners performed at a level similar to that shown by the quality control evaluators, and the performance of the psychophysiologist was approximately midway between the best and poorest performance shown by the Secret Service examiners.

Accuracy of Outcomes

Decisions on individual questions using an inconclusive zone of +2 to -2 are shown in Table 3 for the original examiners and for the average of the six Secret Service blind interpreters. For Pure Verification subjects, the original examiners were 77.6% correct, 3.6% incorrect and 18.8% inconclusive. The blind interpreters averaged 59.1% correct, 5.8% incorrect and 35.1% inconclusive.

The decision accuracy on individual questions for Mixed Verification subjects was poorer than for the Pure Verification subjects. For the original examiners, the overall accuracy was 95.5% for Pure Verification and only 87.5% for the Mixed Verification subjects. The overall accuracy of the blind interpreters averaged 90.5% for Pure Verification subjects and only 74.5% for Mixed Verification.

Table 3. Percent Accuracy on Individual Questions for Original Examiners and Blind Interpreters

Pure Verification										
	Truthful (N=26)					Deceptive (N=37)				
	(n)	C	W	?	Dec	(n)	C	W	?	Dec
Original Examiners	(62)	76	3	21	96	(76)	79	4	17	95
Blind Interpreters	(68)	52	9	39	85	(83)	65	4	31	94
Mixed Verification										
	Truthful (N=13)					Deceptive (N=13)				
	(n)	C	W	?	Dec	(n)	C	W	?	Dec
Original Examiners	(15)	67	7	26	91	(20)	55	10	35	85
Blind Interpreters	(19)	29	17	54	63	(23)	47	9	43	84

It can be seen in Table 3 that the accuracy of decisions on confirmed truthful and deceptive answers differed as a function of type of verification, especially for the blind interpreters. For the original examiners, accuracy on questions answered deceptively was somewhat higher for Pure (95%) as compared to Mixed Verification (85%), and a similar pattern occurred on questions answered truthfully (Pure= 96% and Mixed = 91%). A stronger effect of verification type was observed for the blind interpreters. Again, accuracy of decisions on questions answered deceptively was somewhat higher for Pure (95%) as compared to Mixed Verification (84%). However, for questions answered truthfully there was a large drop in accuracy from 85% for Pure Verification to 63% for Mixed Verification subjects.

Comparison of Strength of Reactions by Confirmed and Unconfirmed Subjects

The magnitudes of numerical scores assigned to individual questions that yielded definite decisions (truthful or deceptive) were tested for possible differences between those decisions that were subsequently confirmed and those that were not confirmed. A 2- way ANOVA of Decision (Truthful/Deceptive) and Confirmation (Confirmed/ Unconfirmed) was performed on the numerical scores that exceeded +2 or -2 assigned by the blind interpreters to the questions from the 100 cases, as described above. The mean numerical scores are shown in Table 4.

Table 4. Mean Numerical Scores for Blind Decisions on Confirmed and Unconfirmed Questions

	Confirmed	Unconfirmed
Truthful	5.9	- 6.0
Deceptive	5.7	4.9

ANOVA showed a significant main effect for Decisions, $F(1, 212) = 1340.26$, $p < 0.0001$. The main effect for Confirmation was not significant, $F(1, 212) = 1.57$, but the interaction of Decision and Confirmation approached significance, $F(1, 212) = 3.84$, $p = 0.051$. That was due to the slightly smaller scores for the Unconfirmed as compared to the Confirmed deceptive questions.

Computer Analyses

Discriminant Validity

The discriminant validity of the computer method was initially assessed separately for the Complete, Mixed, and Partial Verification Groups. Subjects in the Mixed Verification Group had answered some of the relevant questions truthfully and other relevant questions deceptively. For purposes of the analysis of cases with Mixed Verification, it was necessary to split the Mixed Group in half and assign confirmation group membership arbitrarily. When the subject was assigned to the Truthful group, only physiological responses to relevant questions confirmed to have been answered truthfully were included. Conversely, when the subject was assigned to the Deceptive group, only responses to relevant questions confirmed as having been answered deceptively were included in the analysis.

A discriminant function was computed for each verification group and was used to generate a discriminant score for each subject in that group. A subject was defined as correctly classified when the discriminant score yielded a probability of correct group membership that exceeded .50. If the probability was less than .50, the classification by the computer model was considered an error. Since it is known that a small subject-to- variable ratio causes discriminant analysis to capitalize on chance and produce inflated estimates of diagnostic validity (McNemar, 1969), standard

Table 5. Percent Correct Dichotomous Computer Classifications, Magnitude of Effect (R^2), and Tests of Statistical Significance (F) for Complete, Partial, Mixed, and Pure Verification Groups

	Percent Correct Classification				Statistics		
	(n)	Truthful	(n)	Deceptive	R^2	*F*	p
Complete	(17)	88.2	(13)	92.3	.79	24.09	<.0001
Mixed	(7)	85.7	(6)	83.3	.27	.73	ns
Partial	(9)	88.9	(24)	87.5	.56	9.01	<.0001
Pure	(26)	96.2	(37)	83.8	.62	23.91	<.0001

statistical tests were also performed to assess the reliability of the findings. The results obtained for the three verification groups are presented in Table 5.

As shown in Table 5, the accuracy of the computer model was highest for cases with Complete Verification. In those cases, answers to all of the relevant questions had been confirmed as either Truthful or Deceptive. A significant proportion of criterion variance was explained by the optimal linear combination of the four computer variables (R2 = .79). The lowest accuracy was obtained for the Mixed Verification cases. Although the correct classifications in the Mixed Group exceeded 80%, it is clear that the result was unreliable since the F-ratio was not significant.

Complete versus Mixed Verification. A MANOVA with the four physiological parameters as dependent variables was performed to determine if the accuracies obtained for the Complete Verification Group differed significantly from those obtained for the Mixed Verification group. The MANOVA revealed that the Verification (Complete/Mixed) X Confirmation (Truth/Deception) interaction was significant, $F(4,36) = 2.69$, $p < .05$. The discrimination between truthful and deceptive answers was significantly better in Complete Verification cases than in Mixed Verification cases. This finding suggests that there are important differences between Complete and Mixed Verification cases and that the two types of cases should be considered separately.

A within-subjects MANOVA conducted using only Mixed Verification cases revealed that the physiological reactions associated with deceptive answers to relevant questions were not significantly stronger than those associated with truthful answers to relevant questions, $F(4,9) = 2.99$, $p = .08$.

Complete versus Partial Verification. MANOVA revealed no main effect for Complete versus Partial Verification Groups, $F(4,56) = 1.01$, and no evidence of a Verification X Confirmation interaction, $F(4,56) = .89$. Thus, cases in which answers to only some relevant questions were confirmed as either truthful or deceptive were indistinguishable from those in which answers to all relevant questions were confirmed as either truthful or deceptive. Since little would be gained from treating these two subgroups separately, they were pooled to form the Pure Verification sample for all subsequent analyses. The results obtained from the Pure sample are presented in the bottom row of Table 5.

Discriminant Validity in the Pure Verification Sample. Table 6 presents the percentage of correct truthful and deceptive decisions and inconclusive subjects in the Pure Verification sample as a function of various decision criteria. Within the .50 cutoff, a correct decision was defined as a probability of correct group membership greater than .50, and an error occurred if the probability was less than .50. With the .90 cutoff, a correct decision was scored if the probability of correct group membership was .90 or greater; an error was scored if it was equal to or less than .10; and the result was inconclusive if the probability

Table 6. Percent Correct Classifications and Inconclusives for Various Decision Criteria

	Probability Cutoffs for Decisions				
	.50	.60	.70	.80	.90
Truthful (n=26)	96	96	96	95	95
Deceptive (n= 37)	84	83	93	93	96
Inconclusive	0	5	11	21	24

was between .90 and .10.

With the .50 cutoff, 96% of the Truthful and 84% of the Deceptive subjects were correctly classified. There were no inconclusive outcomes since no probability was exactly .50. Predictably, there was a progressive increase in the percentage of inconclusive outcomes as the criterion for a definite truthful or deceptive diagnosis approached unity. Using the .90 criterion, 95% of the Truthful and 96% of the Deceptive subjects were correctly classified, and 15 of the 63 cases (24%) were inconclusive. Examination of the data in Table 6 suggests that an optimal cutoff to maximize the accuracy of decisions and minimize inconclusive outcomes is a probability of approximately .70.

Relative Utility of Physiological Components. The univariate point biserial correlations (rpb) between each of the four physiological variables and the Truth/Deception criterion are presented in Table 7. This statistic provides a measure of the discriminant validity of each physiological parameter. Table 7 also presents the correlations between each of the physiological measures and the discriminant scores (structural coefficients). The structural coefficient for a variable indicates the extent to which the discriminant scores were dependent on changes in that variable.

It may be seen that SR Amplitude was clearly the most diagnostic measure, and it predicted over 53% of the criterion variance (r_{pb}^2). Not surprisingly, SR Amplitude was also correlated most highly with the discriminant scores. BP Amplitude was the next most diagnostic measure, followed by SR Rise Rate and R Length. The relative importance of the variables, as measured by the structural coefficients, followed a similar pattern.

Characteristics of Physiological Responses in Laboratory and Field Examinations. Profile analyses were performed to determine if there were reliable differences between physiological data obtained in laboratory simulations and data obtained from polygraph examinations conducted in the course of actual criminal investigations. The laboratory sample was composed of 26 Truthful and 37 Deceptive adult males randomly selected from a pool of 100 subjects who had participated in a previous mock crime experiment (Kircher & Raskin, 1988). The field subjects were the 26 confirmed Truthful and 37 confirmed Deceptive subjects in the Pure Verification sample. Field cases with Mixed Verification were excluded from the profile analyses because no attempt had been made in the laboratory experiment to represent that condition.

Table 7. Validity and Structural Coefficients for the Physiological Measures

	Validity Coefficient	Structural Coefficient
SR Amplitude	.73	.92
SR Rise Rate	.48	.61
BP Amplitude	.69	.87
R Length	-.39	-.49

The physiological measures for the profile analyses were obtained from subjects' electrodermal, cardiovascular, and respiration responses to control and relevant test questions. Although the procedures for recording blood pressure and respiration data in the laboratory and field settings were similar, different measures of electrodermal activity had been recorded. Specifically, skin conductance (SC) had been recorded in the laboratory, whereas skin resistance (SR) had been recorded in the field examinations. Despite there being a well-defined, nonlinear relationship between SC and SR, the transformation from one to the other requires absolute measures of conductance and resistance that were not available for most of the field cases. Since the original units of measurement in the two data sets were not linearly related and it was not possible to transform the electrodermal measures to a common metric, any observed difference between laboratory and field measures of electrodermal activity was confounded with the method of measurement and should be viewed with caution.

Three physiological variables were selected for the profile analyses: SC or SR Amplitude, BP Amplitude, and R Length. These measures were selected because they comprised the largest subset of measures that had been independently, empirically, and consistently identified as diagnostic in the laboratory (Kircher & Raskin, 1988) and in the Pure Verification sample of field cases.

Parameter Standardization Procedures. In the above analyses, raw measurements of physiological reactions were transformed to z-scores. However, for the profile analyses a z-score transformation is inappropriate since the mean of a set of z- scores is always zero. As a consequence, the z-score for a reaction to one type of test question would necessarily be counterbalanced by a z-score of the same absolute magnitude but of opposite sign for the other type of question. The dependency introduced by use of a z-score transformation would preclude interpretation of differences in physiological response profiles associated with control and relevant test questions.

In order to establish a common metric among the three response variables, within-subject range-adjusted scores were computed separately for each physiological variable according to the following formula:

$$\underline{X}' = 100 * (\underline{X} - \underline{X}_{min}) / (\underline{X}_{max} - \underline{X}_{min})$$

where $\underline{X}$ was a raw score associated with one of the control or relevant questions in the first three repetitions of the question sequence; $\underline{X}_{max}$ was the greatest obtained score in the set of repeated measurements; $\underline{X}_{min}$ was the smallest obtained score in the same set; and $\underline{X}'$ was the range-adjusted value of $\underline{X}$. This transformation produced $\underline{X}$ ' = 0 for the smallest observed score in the original set of raw measurements for the subject (X_{min}) and X'= 100 for the greatest observed score for that subject (X_{max}).

As noted by Nunnally (1978), the levels of response profiles are interpretable only when the variables are "pointed in the same direction" (p. 439). Since relatively strong physiological reactions yielded relatively high scores on the electrodermal and cardiovascular measures but low scores on the respiration measure, all measurements of R Length were reversed in sign prior to projecting the scores onto a standard scale of constant range.

For each subject, the mean of range-adjusted scores associated with each of the two types of test questions was calculated for each physiological measure. A single measure of R Length was obtained for each question by averaging the means of the range-adjusted lengths of thoracic and abdominal respiration tracings. The mean reaction profiles for Truthful and Deceptive subjects in the laboratory and field samples are presented in Figure 1. The order of presentation of the three variables along the abscissa was arbitrary.

To examine possible differences among the response profiles exhibited by laboratory and field subjects, two independent sources of variance were assessed with MANOVA: differences in the levels of response profiles and differences in their shapes (Harris, 1975; Van Egeren, 1973).

The level of a subject's response profile was the mean of the range-adjusted scores for

the three physiological measures that comprised the profile. The level of a response profile may be viewed as a measure of the relative magnitude of generalized arousal associated with control or relevant questions. Observed differences between the shapes of response profiles would suggest qualitative differences in the patterns of physiological responses associated with particular questions (Control or Relevant), criterion status (Truthful or Deceptive), or context for the examination (Laboratory or Field).

Among all comparisons of levels and shapes of response profiles produced by laboratory and field subjects, only one significant effect was observed. This was a significant difference between laboratory and field subjects in the shapes of their response patterns associated with deceptive answers to relevant ques-

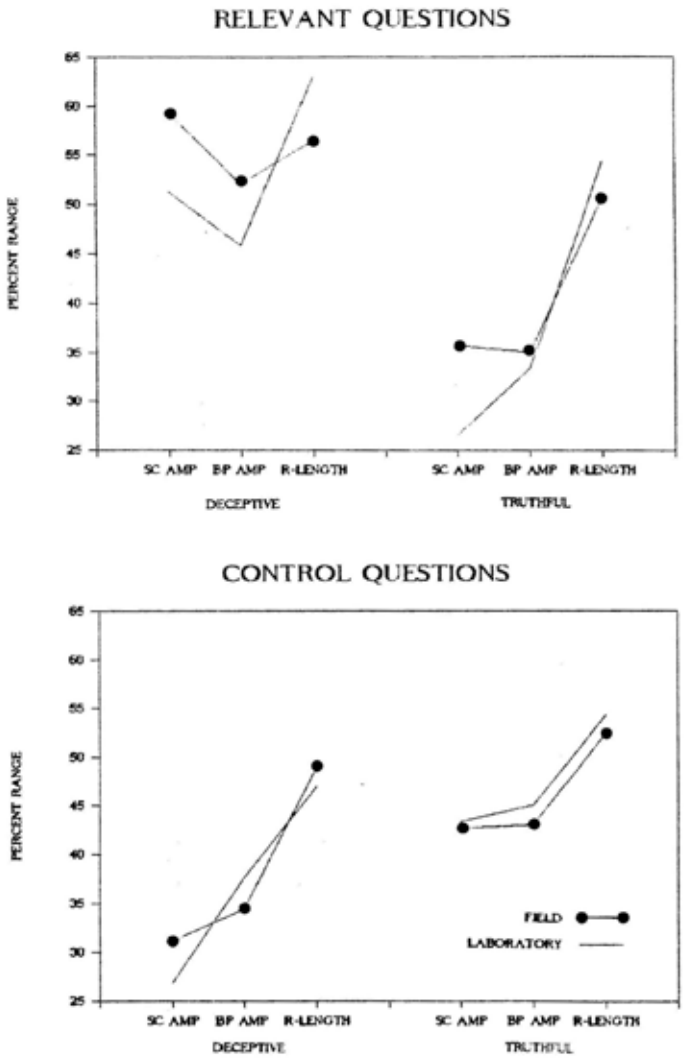


Figure 1. Mean profiles of physiological reactions of Truthful and Deceptive Laboratory and Field Subjects

40

Table 8. Multivariate Comparisons of Response Profiles for Laboratory and Field Subjects

	Control Questions	Relevant Questions
Truthful (n=26)		
Profile level	*F* (1,122) = .53	*F* (1,122) = 1. 12
Profile shape	*F* (2,121) = .05	*F* (2,121) = 2.69
Deceptive (n= 37)		
Profile level	*F* (1,122) = .36	*F* (1,122) = 1. 73
Profile shape	*F* (2,121) = 2.17	*F* (2,121) = 5.71

tions, (p <.01). In order to assess the magnitude of this effect, a discriminant analysis was performed between the laboratory and field samples using the level-adjusted profiles for physiological responses to relevant questions answered deceptively. Level adjusted scores were obtained for each subject and each response variable by subtracting the mean of the three scores that comprised a profile from each variable in that profile.

The differences between laboratory and field subjects accounted for 9.8% of the variance in the shapes of these profiles. By comparison, differences between Truthful and Deceptive subjects accounted for 56.9% of the variance in the physiological measures. In other words, the differences between Truthful and Deceptive subjects accounted for almost six times the amount of variance in physiological responses associated with the differences between the laboratory and field subjects.

The laboratory-field differences between the shapes of subjects' response profiles associated with deceptive answers to relevant questions were examined in greater detail by performing separate univariate tests using level-adjusted scores for the three physiological measures. Univariate tests revealed that the significant effect for profile shape was due to differences in the SR/SC Amplitude, *F* (1,122) = 4.49, p < .04, and R Length measures, F (1,122) = 11.41, p < .001. Level-adjusted scores on BP Amplitude did not distinguish between the groups, F (1,122) = 1.97.

Double Cross-Validation. Separate discriminant functions were developed from the 63 subjects in the Pure Verification sample (37 confirmed Deceptive and 6 confirmed Truthful) and from 50 Guilty and 50 Innocent subjects who had participated in a mock crime experiment (Kircher & Raskin, 1988). Each discriminant function was used to classify the subjects in the sample on which it was developed and also the subjects in the other sample.

The discriminant functions developed from the laboratory and field samples incorporated the same variables, SC or SR Amplitude, BP Amplitude, and R Length.

Generalizability from laboratory to field and vice-versa was first assessed by comparing the accuracy of classification made by each mode1 when applied to the data from laboratory and field samples. Classification accuracies were calculated by comparing the actual status of each subject with the computer-generated probability of group membership using a dichotomous decision rule that defined a correct decision as a probability of correct group membership that exceeded .50, and defined an error as a probability of correct group membership that was less than .50. The results are presented in Table 9.

The results indicated that each model performed similarly when applied to the two samples. Thus, the accuracy of the laboratory model was approximately the same when applied to the original sample of laboratory subjects and to the validation sample of field subjects. Similarly, the accuracy of the field model was approximately the same when applied to the original sample of field subjects and to the validation sample of laboratory subjects. However, it should be noted that the laboratory model showed a drop in performance on Truthful subjects when applied to the field subjects (88% versus 77%), and the field model showed a drop in performance on Deceptive

Table 9 Accuracy of Classifications Based on Laboratory and Field Models

Laboratory Model		Classification		
	Laboratory Sample	Deceptive	Truthful	% Correct
	Deceptive	45	5	90
	Truthful	6	44	88
	Field Sample			
	Deceptive	34	3	92
	Truthful	6	20	77
Field Model		Deceptive	Truthful	% Correct
	Field Sample			
	Deceptive	31	6	84
	Truthful	2	24	92
	Laboratory Sample			
	Deceptive	38	12	76
	Truthful	1	49	98

subjects when applied to the laboratory subjects (84% versus 76%).

The laboratory and field results were also compared by calculating univariate point-biserial correlations with the criterion (validity coefficients) and multivariate structural coefficients for the physiological variables used in the two models. The validity coefficients and structural coefficients for the laboratory and field samples are shown in Table 10.

Table 10. Validity and Structural Coefficients for Laboratory and Field Samples

	Validity Coefficients		Structural Coefficients	
	Laboratory	Field	Laboratory	Field
SC/SR Amplitude	.77	.73	.94	.92
BP Amplitude	.61	.69	.74	.87
R Length	.55	.39	.67	.49

The validity and structural coefficients were similar for laboratory and field samples.

These findings suggest that the relationships among the physiological variables obtained from polygraph tests of subjects in mock crime laboratory experiments are similar to those obtained from suspects in field polygraph tests. However, correlational analyses are not sensitive to differences in the means of the variables obtained from laboratory and field subjects, and the analyses of classification accuracies presented in Table 9 suggest that mean differential physiological reactivity for Deceptive and Truthful subjects may not be symmetrical around zero in both samples. The findings that the laboratory model showed a drop in accuracy on Truthful field subjects and the field model showed a drop in accuracy on Deceptive laboratory subjects may indicate such asymmetry.

In order to examine the possibility of a lack of symmetry in the means of the differential physiological reactivity of laboratory and field subjects, the means of the computer-generated indices of differential physiological reactivity to relevant and control questions were calculated for Truthful and Deceptive laboratory and field subjects and are presented in Table 11.

Table 11. Computer Indices of Differential Reactivity to Control and Relevant Questions for Laboratory and Field Subjects

	Laboratory		Field	
	Truthful	Deceptive	Truthful	Deceptive
SC/SR Amplitude	1.89	-2.41	.67	-2.95
BP Amplitude	1.53	-.93	.88	-2.02
R Length	.25	-1.64	.31	-1.07

Truthful laboratory and field subjects reacted more strongly to control than to relevant questions for all three physiological indices (positive means), and the Deceptive laboratory and field subjects responded more strongly to relevant than to control questions (negative means). However, the means for Truthful and Deceptive laboratory subjects were approximately equidistant from zero, whereas the means for the field sample were generally shifted in the negative direction. Deceptive field subjects showed stronger differential reactivity to relevant questions than did Deceptive laboratory subjects, and Truthful field subjects showed weaker differential reactivity to control questions than did Truthful laboratory subjects.

Since the means for Truthful and Deceptive laboratory subjects are approximately symmetrical around zero, the model derived from those data "expects" that Truthful subjects wi11 produce differential reactions to control questions as strong as those produced by Deceptive subjects to relevant questions. Since Truthful field subjects did not show that pattern to the same degree, there was a fairly high rate of false positive errors when the laboratory model was applied to the field subjects. On the other hand, the laboratory model "expects" only moderately strong reactions to relevant questions from Deceptive subjects.

Since Deceptive field subjects showed much stronger differential reactions to relevant questions than to control questions, the laboratory model produced very few false negative errors when applied to field subjects. These results suggest that computer models developed on laboratory subjects are biased against Truthful field subjects, and they also suggest modifications of the decision cutoffs for numerical scoring based on the results of laboratory experiments. It appears that the cutoffs should be asymmetrical and shifted in the negative direction.

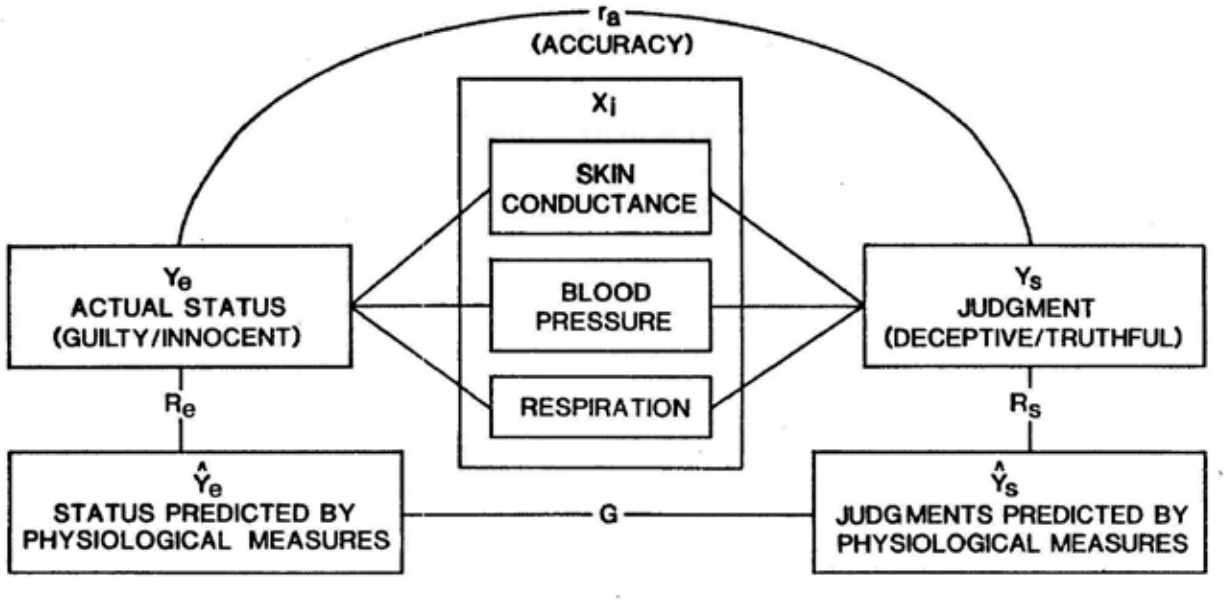
Human Versus Computer Scoring (Lens Model Analyses)

The subjects used in the lens model analyses were the Secret Service examiners who had conducted the polygraph examinations (Original Examiners) the six Secret Service examiners and one psychophysiologist who independently interpreted the polygraph charts.

Only judgments made on examinees in the Pure Verification sample were included in the lens model analyses. To facilitate comparisons among the polygraph interpreters, a forced- choice decision rule was adopted to produce an equal number of decisions for each interpreter. For confirmed relevant questions any positive total numerical score was considered a truthful outcome and any negative total score was considered a deceptive outcome. The physiological measures used to predict the criterion were the four parameters identified by the previous all-possible-subsets regression analyses as the subset that best discriminated between the Truthful and Deceptive subjects in the Pure Verification sample.

Brunswik's lens model (Slavic & Lichtenstein, 1971) was used to compare the performance of the blind numerical interpreters and the computer. The lens model was also used to examine possible differences among the polygraph examiners in their use of information from the polygraph charts to diagnose truth and deception. For the present problem, the lens model organized three sources of information and the relationships among them, as illustrated in Figure 2.

Figure 2. The Lens Model



As shown on the left side of Figure 2, the statistically optimal classification strategy is operationally defined in terms of a multiple regression equation that predicts the actual deceptive status of an individual (Y_e) by means of a linear combination of weighted physiological measures or cues (X_i). The subscript e in the lens model stands for the environment, which is the criterion of truth or deception. The obtained multiple correlation R_e provides a measure of the validity of the combination of physiological measures for predicting group membership.

The decision policy of the polygraph interpreter is represented on the right side of Figure 2 by the regression of diagnoses of truth and deception (Y_s) on the multiple physiological measures (X_i). The subscript s refers to the polygraph interpreter who served as the subject of the lens model analysis. The obtained multiple correlation R_s measures the extent to which the interpreter used information that was contained in the computer-generated physiological variables in making his decisions. The correlation between the interpreter's decisions (Y_s) and the criterion (Y_e) provides a measure of achievement (r_a).

This correlation is the most important component of the lens model since the magnitude of r_a indicates how well the interpreter discriminated between guilty and innocent subjects on the basis of his blind evaluations of the polygraph charts.

According to Tucker (1964), the relationship between achievement (r_a) and other components of the lens model can be represented in terms of the following equation:

$$r_a = G R_e R_s + C \sqrt{(1 - R_e^2)} \sqrt{(1 - R_s^2)}$$

where G is the correlation between the predicted criterion scores (Y_e) and the predicted decisions by the interpreter (Y_s), and R is the correlation between the residuals ($Y_e - Y_e$) and ($Y_s - Y_s$). Since both sets of predictions were made from the same physiological measures, the magnitude of G specifies the degree of similarity between the model used to predict group membership and the model used to predict decisions.

Conceptually, G specifies how closely the interpreter's use of information contained in the physiological measures generated by the computer matched the optimal linear combination of these variables. The C component represents the degree to which errors in predicting the criterion from the physiological measures were correlated with errors in pre-

dicting examiner judgments. The magnitude of C may be taken as a measure of the amount of diagnostic information available in the physiological recordings that was used by the blind interpreter to make valid diagnoses but was not contained in the four features of response wave forms that were quantified by the computer. Therefore, C provides an index of the extent to which the computer failed to use diagnostic information available in the physiological recordings that was effectively used by the human interpreters.

The results of the lens model analysis are presented in Table 12. The interpreters are listed in order of their achievement coefficients (r_a), which ranged between .53 and .87, with a mean of .76. On average, human judgments based on numerical evaluations of the polygraph charts accounted for approximately 58% of the criterion variance. The multiple correlation between the physiological variables and the criterion (B_e) provided an overall estimate of the validity of the combination of the physiological measures for diagnosing truth and deception. The optimal linear combination of physiological measures produced a multiple correlation of . 79 and accounted for 63% of the criterion variance. The average level of discrimination between Truthful and Deceptive subjects achieved by the human interpreters was slightly less than that achieved by the computer model (.76 vs. . 79), but the difference was not significant. The G component is also important for summarizing the performance of a human interpreter (Slavic & Lichtenstein, 1971; Tucker, 1964). The G component, or matching index, exceeded .93 for each of the human interpreters. These findings indicate that most of the human interpreters made optimal use of the information contained in the four computer generated physiological measures.

Table 12. Lens Model Components for the Original Examiners and Seven Blind Interpreters

	R_a	R_e	R_s	G	C
Original Examiners	.87	.79	.74	.99	.70
Experienced Examiner	.87	.79	.81	.99	.64
Quality Control	.84	.79	.76	.99	.62
Psychophysiologist	.77	.79	.77	.99	.47
Inexperienced Examiner	.77	.79	.75	.96	.45
Quality Control	.71	.79	.69	.99	.38
Inexperienced Examiner	.67	.79	.75	.99	.20
Experienced Examiner	.53	.79	.60	.93	.17
Mean (r-to-z-to-r)	.76	.79	.74	.99	.43

Variability in performance was observed among the blind numerical interpreters. Judgments made by the original examiners were highly accurate and were slightly more accurate than those made by the blind interpreters, all of whom used numerical scoring procedures. Since the original examiners interacted with the subjects and had detailed knowledge of the case facts, it is possible that their decisions were influenced by the case facts and the verbal and nonverbal behavior of the subjects during the examinations.

Although the performance of the human interpreters was not clearly related to level of experience, it was directly related to C. This finding may indicate that the major factor that distinguished among the blind numerical interpreters was their ability to extract more diagnostic information from the physiological recordings than was represented by the four response parameters quantified by the computer. The large value for C for the original examiners is another indication that they may have adjusted their numerical scoring of the physiological data by using non-physiologi-

cal, auxiliary sources of information that were available only to them.

The mean C component of the lens model indicated that, on average, the blind evaluators were able to predict 18% (C^2) of the criterion variance that was not predicted by the four computer-generated variables. This finding suggests that significantly more diagnostic information was available in the physiological recordings than was represented in the four parameters quantified by the computer. Some of that variance may be attributed to the human interpreter's ability to make reasonable approximations of the amplitudes of physiological reactions even when the recording pens exceeded the limit of travel because the examiner had set the amplifier sensitivity too high, a common occurrence in the polygraph charts used in the present study. The computer merely quantified the amplitude of the response as it appeared on the chart, and no attempt was made to estimate the true amplitude of the response when the limit of pen travel was exceeded.

Discussion

This study evaluated the accuracy of control question polygraph examinations in criminal investigations conducted by U. S. Secret Service personnel during FY1983 through FY1985. The cases were obtained from their files and were confirmed using a very stringent criterion of admissions and confessions that were independently corroborated by physical evidence. The results of this study clearly indicate that control question polygraph examinations used for purposes of criminal investigation can be highly accurate when conducted by qua lified examiners and numerically evaluated by experienced interpreters or assessed using computer methods developed at the University of Utah.

Accuracy

Human Interpreters. The overall accuracy of decisions made by the Secret Service examiners on individual relevant questions was 96% for confirmed truthful answers and 95% for confirmed deceptive answers in those cases where suspects were either truthful to all confirmed relevant questions or deceptive to all confirmed relevant questions (Pure Verification). When suspects were confirmed as deceptive to at least one relevant question and also truthful to at least one relevant question in the same test (mixed verification), the accuracy of the decisions made by the original examiners dropped to 91% on confirmed truthful answers and 85% on confirmed deceptive answers. It should be noted that this high level of accuracy was achieved even though the level of analysis at individual questions would be expected to produce lower reliability and accuracy than analyses of all relevant questions combined.

The results also indicated that the accuracy of decisions by examiners who made blind interpretations of the polygraph charts was also high, but not quite as high as the original examiners. The accuracy of blind interpreters on Pure Verification subjects was 85% on truthful answers and 94% on deceptive answers. However, when there was mixed verification, their accuracy dropped to 63% on truthful answers and 84% on deceptive answers. From these results, it appears that control question polygraph tests perform best when the relevant questions deal with issues that elicit either all truthful or all deceptive answers from the subject. It should also be noted that the blind interpreters made more false positive than false negative errors, a result that consistently appears in the data from laboratory and field studies (Raskin, 1986). However, the original examiners did not show that pattern.

The effects of context of the interpretation (original or blind) and interpreter experience or type of training on the accuracy of chart interpretations were assessed by comparisons of the performance of the original examiners, highly experienced quality control interpreters, experienced and inexperienced field examiners, and an experienced field examiner - psychophysiologist. Analyses of the numerical scores and lens model analyses were used for these purposes, and the results produced two somewhat unexpected findings.

There was no demonstrable effect on accuracy as a function of experience or type of training among all of the blind interpreters. However, the original examiners clearly outperformed all of the blind interpreters and the

computer model. The lens mode1 analyses indicated that level of performance of the human interpreters was directly related to the extent to which they either extracted more diagnostic information from the polygraph charts than did the computer model or used nonphysiological information to adjust their numerical scoring to increase their accuracy. The original examiners, one quality control, and one experienced blind interpreter outperformed the computer, but the computer outperformed the remaining five blind interpreters. The superior performance of the original examiners suggests that they used their knowledge of the case facts and their interactions with the subjects to achieve more effective use of the physiological information contained in the polygraph charts.

Computer Interpretations. The computer interpretations of the polygraph recordings also produced a high degree of accuracy. Using the discriminant function generated from these data and various probabilities to define truthful and deceptive decisions, the accuracies ranged between 95% and 96% on confirmed truthful suspects and between 83% and 96% on confirmed deceptive suspects. As the probability required for a decision was increased, the accuracies and the rate of inconclusive outcomes increased. The optimal cutoffs of .70 probability of truthfulness for truthful decisions and .30 probability of truthfulness for deceptive decisions yielded accuracies of 96% on Truthful suspects and 93% on Deceptive suspects, with only 11% inconclusive outcomes. These analyses seem to indicate that the use of cutoffs of approximately .70 and 30 for probabilities of truthfulness yield the best results in field applications.

Comparisons of the computer-generated decisions and those produced by the human interpreters indicated that the computer was generally more accurate than the blind interpreters, but not as accurate as the original examiners. These findings are consistent with a recent review of the literature concerning clinical versus statistical prediction (Wiggins, 1981), indicating that statistical methods are frequently, but not always, superior to clinical judgments. If the computer could take advantage of the case information and observations of the suspect's behavior that were available to the original examiners, computer models might equal or exceed the performance of the original examiners. Achievement of that goal would require additional research to determine the factors that account for the increment in performance of the original examiners and how to incorporate that information in the computer decision models. Toward that end, research that explores relationships between individual differences in expressive behavior, case information, and truthfulness seems feasible and desirable.

Research Issues

Validity of the Confession Criterion. Questions have been raised with respect to the validity of results obtained in field studies; that selecting polygraph examinations for analysis using a criterion of ground truth based on confessions (Iacono, in press; Raskin, 1987). Iacono argued that such studies overestimate accuracy because they do not include the polygraph charts of innocent suspects who failed tests and did not confess and guilty suspects who passed tests and were not interrogated or failed to confess. Iacono also argued that guilty suspects selected for confession studies were only those who produced charts that were strong enough to cause the examiner to elicit a confession. The latter argument seems specious since it implicitly recognizes the accuracy of polygraph charts that are strongly indicative of deception. It also implies that the test results of suspects who failed the test and did not confess are weaker than those who failed the test and did confess. These arguments were addressed by the methods and results of this study.

The manner of selecting cases prevented the problem of not selecting innocent suspects who failed tests (false positive errors) because all of the confirmed truthful suspects were obtained from multiple-suspect cases. Since the truthfulness of these suspects was established by corroborated confessions of other suspects, all truthful suspects who might have failed the tests were included in the sample and would have contributed to the observed error rate. Similarly, the large majority of confirmed deceptive suspects were obtained from multiple-suspect cases in which there was usually more than one deceptive person who could, and often did, confess and incriminate one or more of the other suspects

who were tested. Thus, the potential problems of false positives and false negatives proposed by Iacono were reduced or eliminated by the methodology of this study.

This study also evaluated the suggestion that suspects who failed the tests and confessed produced stronger deceptive charts than those who failed the tests and did not confess. In order to answer that question, we compared the strengths of the deceptive results produced by suspects who confessed to the original examiners and deceptive results produced by suspects who were scored as deceptive by the original examiners but did not confess. The analyses indicated a difference of approximately 20% between the magnitude of negative scores assigned to confirmed and unconfirmed deceptive results. However, the mean scores for unconfirmed deceptive results were 63% higher than the minimum score required for a conclusive deceptive decision. Therefore, it appears that the success or failure in eliciting a confession was unrelated to the strength of the physiological reactions to relevant questions. These results provide little support for Iacono's argument concerning the lack of validity of confession-based field polygraph studies.

Generalizability of Laboratory Results

Two types of analyses were conducted to assess the extent to which the results of laboratory experiments can be used to make inferences about the accuracy and processes that underly control question polygraph examinations of criminal suspects.

The first compared profiles of physiological responses of confirmed truthful and deceptive laboratory subjects and criminal suspects. The results indicated that although there was a small but significant difference in the shape of the profiles of deceptive laboratory and field subjects, the size of the effect was very small in comparison to the differences between the physiological responses to control and relevant questions produced by truthful and deceptive laboratory and field subjects. Since the latter is the basis for rendering decisions in the field as well in realistic simulations of the field situation (Kircher, Horowitz, & Raskin, 1987), the findings lend support to the generalizability of the results of such laboratory studies to applications of polygraph examinations in criminal investigation.

The second type of analysis used a double cross-validation procedure to determine the accuracy of computer classifications of criminal suspects based on a discriminant function derived from laboratory data and the accuracy of computer classifications of laboratory subjects based on a discriminant function developed on criminal suspects. The results indicated that the accuracies of each model were similar when applied to laboratory and field data. However, the laboratory model produced an increase in false positive errors when applied to field suspects and the field model showed an increase in false negative errors when applied to laboratory subjects. The structural coefficients and univariate validity coefficients also were consistent with the principle of generalizability.

The suggestion of asymmetry in false positive and false negative errors produced by the laboratory and field models was further assessed by a comparison of the means of the computer-generated indices of differential reactivity to control and relevant questions by laboratory subjects and criminal suspects. The differential reactivity indices for laboratory subjects were symmetrical around zero, but the means for the field suspects were shifted in the negative direction. These results reinforce an interpretation that compared to deceptive laboratory subjects, deceptive field suspects show stronger differential reactions to relevant questions than to control questions; and compared to truthful laboratory subjects, truthful field suspects showed much weaker differential reactions to control than to relevant questions. Although it appears that the underlying structure of physiological responses in laboratory subjects is similar to that obtained in polygraph examinations of criminal suspects, the obtained differences suggest using somewhat different numerical cutoffs for decision-making in the two situations.

Implications of the Results for Investigative Applications

Three major conclusions for applications and procedures for control question polygraph examinations of criminal suspects

are suggested by the results of this study. They concern the accuracy of such tests, the optimal composition of relevant questions to be used in such tests, and the optimal methods for interpreting the outcomes of such tests. The overall pattern of results indicates that properly conducted and interpreted examinations have a high degree of accuracy and can be of considerable benefit in evaluations of the credibility of criminal suspects. However, certain changes in current practices should be considered.

The results suggest that blind numerical scoring procedures using cutoffs that are symmetrical around zero may be biased against truthful criminal suspects. Although the scores assigned by the original examiners did not show this effect, the blind interpreters made relatively more errors on confirmed truthful responses. Apparently, the original examiners used other information to compensate for the inherent bias of the test against truthful suspects.

Even though the six U. S. Secret Service blind interpreters scored the charts using the federal system that compares the reactions to relevant questions to the control questions that evoke stronger physiological responses (Weaver, 1980, 1985), they still made more false positive errors than did the original examiners and the computer model. Thus, it appears that blind numerical interpretation would be more accurate if stronger negative scores were required for deceptive decisions and somewhat weaker positive scores were required for truthful decisions. The present data seem to suggest cutoffs of -3 and +2 for individual questions and -7 and +4 for overall decisions. However, additional analyses are required in order to establish definitive cutoffs for decisions based on blind numerical evaluations.

A related problem is raised by the finding of higher false positive rates for questions answered truthfully by suspects who were also deceptive to at least one relevant question in the same test. It appears that answering deceptively to at least one relevant question in the test tends to weaken the reactions to the control questions, thereby making it difficult for them to produce reactions that are larger than those to relevant questions that are answered truthfully.

Therefore, field polygraph examiners should attempt to devise sets of relevant questions that the suspect can be expected to answer all truthfully or all deceptively. The case information and the importance of each relevant question should be carefully considered in formulating the set of relevant questions to be asked, and separate question series should be used whenever it seems likely that the suspect might answer some of the relevant questions truthfully and some of them deceptively.

Finally, the results of this research clearly support the utility of computer models for the analysis and interpretation of polygraph test outcomes. The results obtained with computer models derived from the data on criminal suspects demonstrated higher accuracy than blind numerical interpretations. Computer evaluations have the additional virtues of being objective and providing a rapid and readily available form of quality control for field examiners. Computer analyses would be especially useful when performing examinations in important cases and another examiner is not available for independent interpretation when decisions must be made on the spot. In most cases, decisions must be made in order to determine if the suspect is to be excused, interrogated, or administered additional examinations. Under such circumstances, an independent computer analysis may be increase confidence in the decisions and guide the course of further testing.

References

Barland, G. H. (1982). On the accuracy of the polygraph: An evaluative review of Lykken's Tremor in the Blood. *Polygraph, 11*, 258–272.

Bersh, P. J. (1969). A validation study of polygraph examiner judgments. J*ournal of Applied Psychology, 53,* 399-403.

Bradley, M. T., & Ainsworth, D. (1984). Alcohol and the psychophysiological detection of deception. *Psychophysiology, 21*, 63-71.

Dawson, M. E. (1980). Physiological detection of deception: Measurement of responses to questions and answers during countermeasure maneuvers. *Psychophysiology, 17*, 8- 17.

Gatchel, R. J., Smith, J. E., & Kaplan, N. M. (1983). The effect of propanolol on polygraphic detection of deception. Unpublished manuscript, University of Texas.

Harris, R. J. (1975). *A primer of multivariate statistics.* New York: Academic Press. Horvath, F. S. (1977). The effect of selected variables on interpretation of polygraph records.

Journal of Applied Psychology, 62, 127-136.

Horvath, F. S., & Reid, J. E. (1971). The reliability of polygraph examiner diagnosis of truth and deception. *Journal of Criminal Law, Criminology and Police Science, 62*, 276- 281.

Hunter, F. L., & Ash, P. (1973). The accuracy and consistency of polygraph examiners' diagnoses. *Journal of Police Science and Administration, 1*, 370-375.

Iacono, W. G. (in press). Can we determine the accuracy of polygraph tests? In P. K. Ackles,

J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 4). Greenwich, CT: JAI Press.

Kircher, J. C., Horowitz, S. W., & Raskin, D. C. (1987). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior, 12*, 79-90.

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

Kleinmuntz, B. J., & Szucko, J. J. (1982) On the fallibility of lie detection. *Law & Society Review, 17(1),* 85-104.

Lykken, D. T. (1979), The detection of deception. *Psychological Bulletin, 86,* 47-53. Lykken, D. T. (1981). *A tremor in the blood.* New York: McGraw-Hill.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley. Nunnally, J. C. (1978) *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Office of Technology Assessment (1983). *Scientific validity of polygraph testing: A research review and evaluation.* Washington, D. C. : U. S. Government Printing Office.

Pedhazur, E. (1982). *Multiple regression in behavioral research: Explanation and prediction.* (2nd ed.). New York: Holt.

Podlesny, J. A. & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology, 125*, 344-358.

Raskin, D. C. (1976). *Reliability of chart interpretation and sources of errors in polygraph examinations.* (Report 76-3, Contract 75-NI-99-0001, U.S. Department of Justice.) Salt Lake City, Utah: Department of Psychology, University of Utah.

Raskin, D. C. (1982). The scientific basis of polygraph techniques and their uses in the judicial process. In A. Trankell (Ed.), *Reconstructing the past: the role of psychologists in criminal trials.* Stockholm: Norstedt and Soners.

Raskin, D. C. (1984). *An evaluation of the polygraph policies and programs of U. S. Department of the Treasury.* Unpublished report to the U. S. Department of the Treasury.

Raskin, D. C. (1986) The polygraph in 1986: Scientific, professional and legal issues surrounding applications and acceptance of polygraph evidence. *Utah Law Review, 1986*, 29-74.

Raskin, D. C. (1987). Methodological issues in estimating polygraph accuracy in field applications. *Canadian Journal of Behavioural Science, 19*, 389-404.

Raskin, D. C., Barland, G. H., & Podlesny, J. A. (1978). *Validity and reliability of detection of deception.* (Contract 75-NI-99-0001, U.S. Department of Justice). Washington, D.C.: U.S. Government Printing Office.

Raskin, D. C., & Hare, R. D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15*, 126-136.

Rovner, L. I., Raskin, D. C., & Kircher, J. C. (1979). Effects of information and practice on detection of deception. *Psychophysiology, 16*, 197-198. (Abstract)

Slovik, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 3*, 305-309.

Slowik, S. M., & Buckley, J. P. (1975). Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure and GSR recordings. *Journal of Police Science and Administration, 3,* 305-309.

Timm, H. W. (1982). Effect of altered outcome expectancies stemming from placebo and feedback treatments on the validity of the guilty knowledge technique. *Journal of Applied Psychology, 67,* 391-400.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch and by Hammond, Hursch and Todd. *Psychological Review, 71*, 528-530.

Van Egeren, L. F. (1973). Multivariate statistical analysis. *Psychophysiology, 10*, 517-532.

Weaver, R. S. (1985). Effects of differing numerical chart evaluation systems on polygraph examination results. *Polygraph, 14*, 34-41.

Wicklander, D. E., & Hunter, F. L. (1975). The influence of auxiliary sources of information in polygraph diagnoses. *Journal of Police Science and Administration, 3*, 405-409.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where do we go from here?

Clinical Psychology Review, 1, 3-18.

Winkler, R. L., & Hays, W. L. (1975). *Statistics: Probability, inference and decisions.* New York: Holt.

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 sub-

tracted 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, $F(1.93,88.98)= 2.73$, $p < .08$.

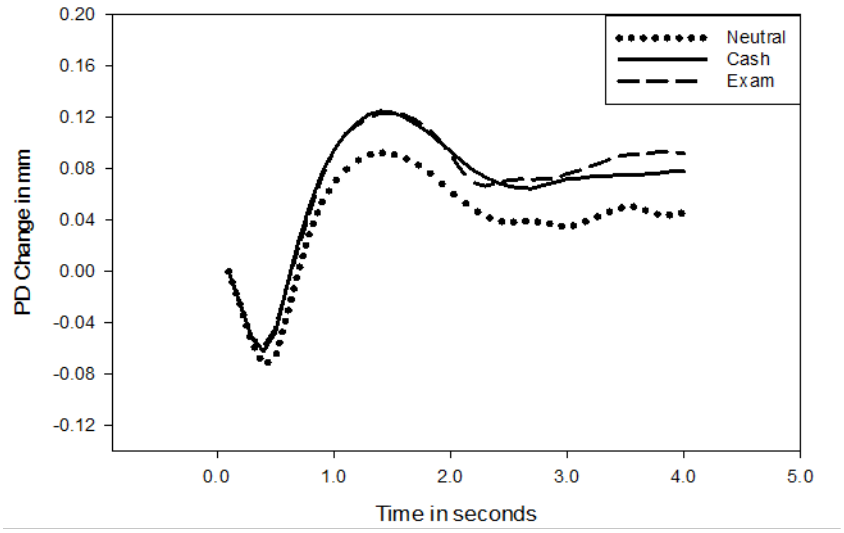
Predictive Validity of Ocular-motor Measures

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).

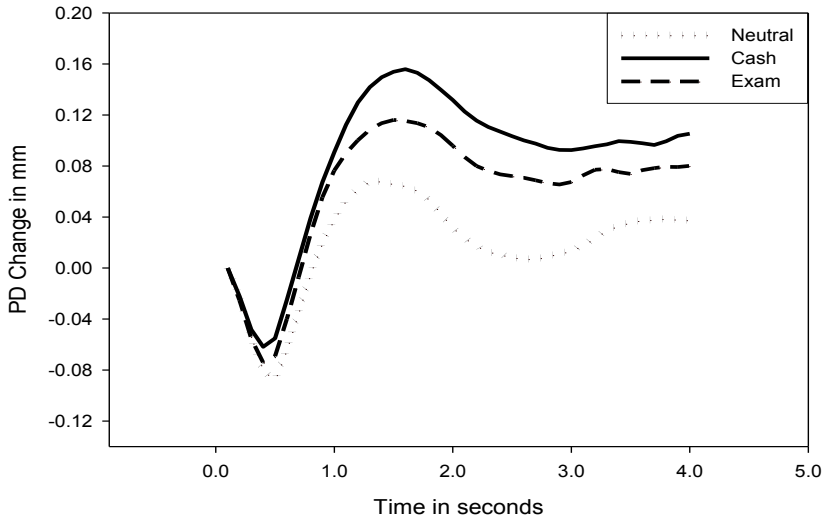
To assess the diagnostic validity of an

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 $\eta^2 = .039$. 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 $\eta^2 = .026$. 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 $\eta^2 = .018$, and Guilt X Statement Type interactions were significant, $F(1.56, 121.80) = 6.35$, $p<.05$, partial $\eta^2 = .075$. 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.

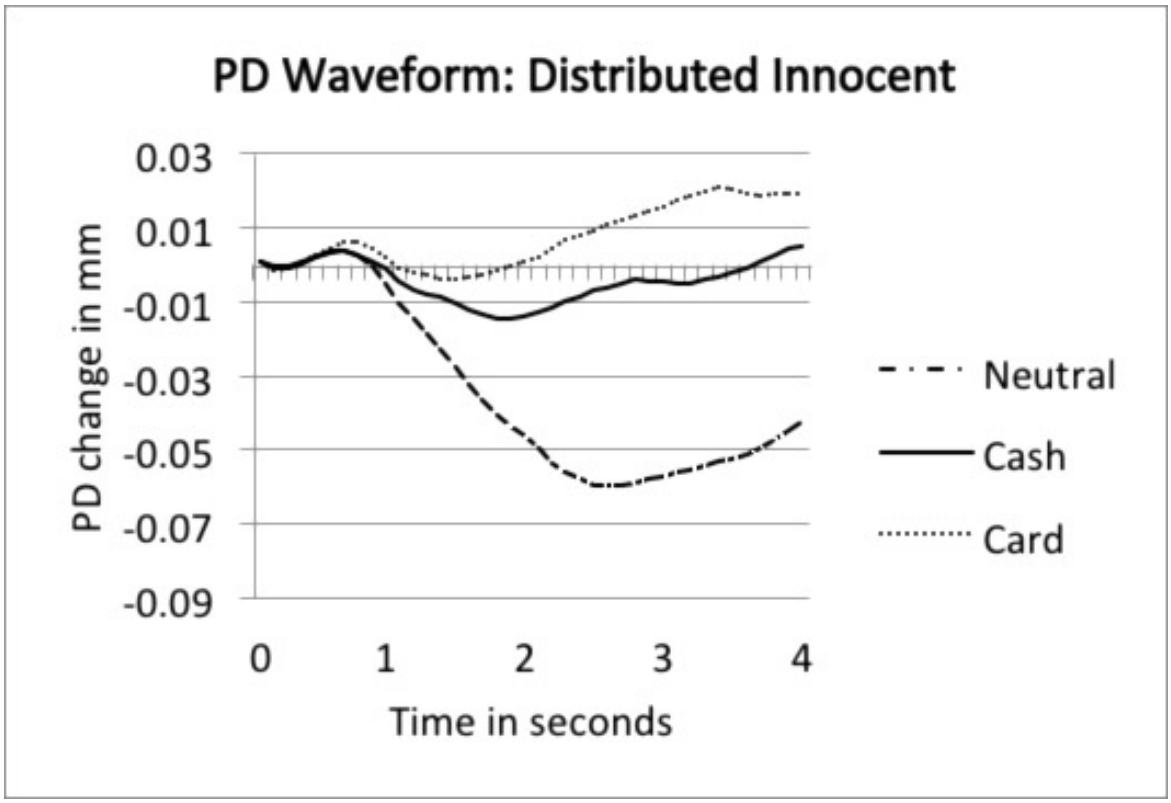


Figure 2b. Pupil response to neutral, cash, and card items for distributed format for guilty participants.

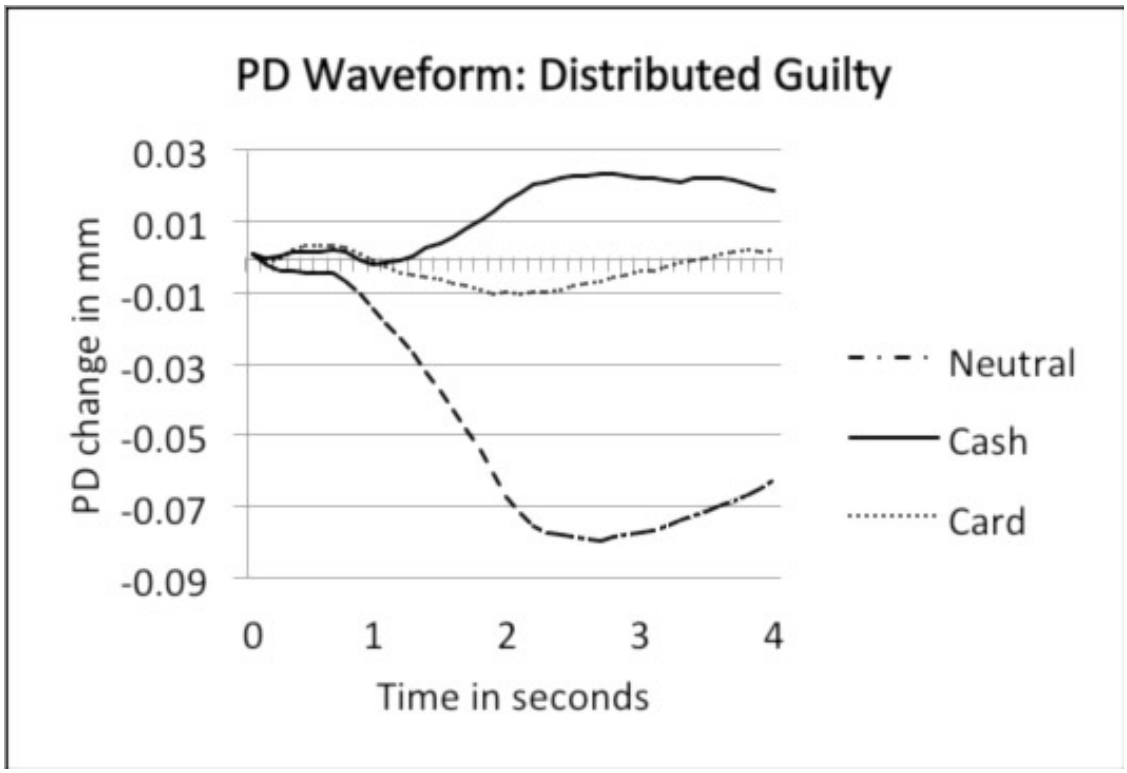


Figure 2c. Pupil response to neutral, cash, and card items for blocked format for innocent participants.

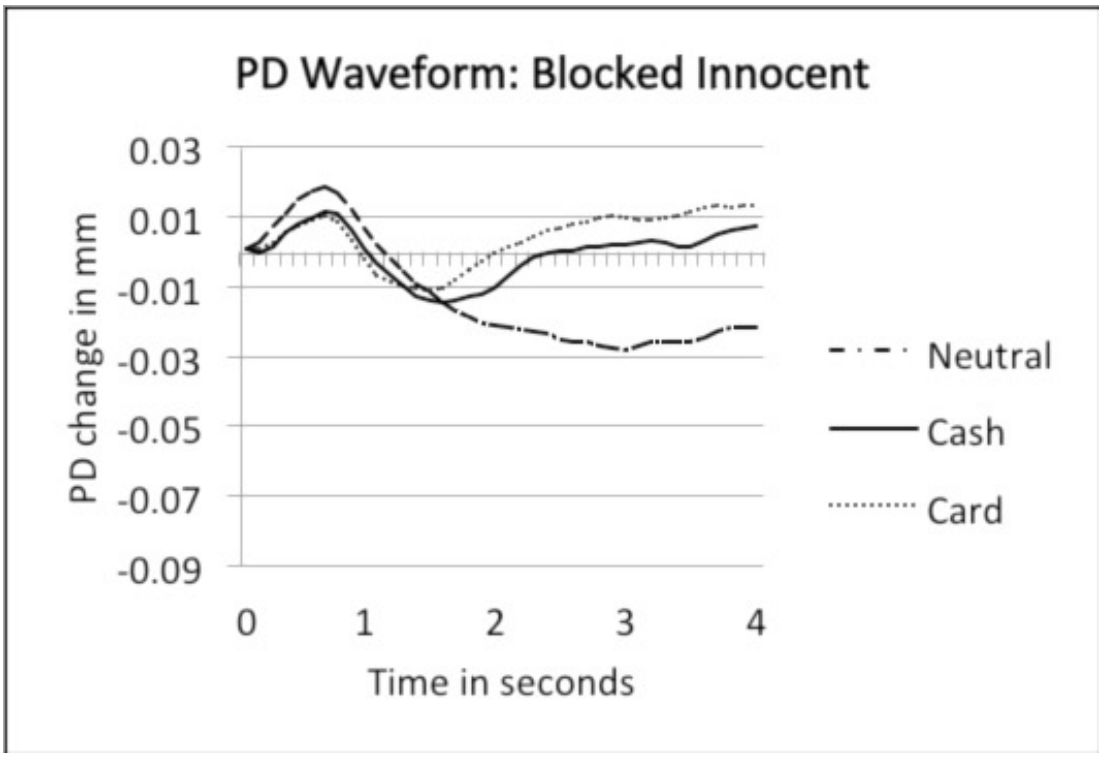


Figure 2d. Pupil response to neutral, cash, and card items for blocked format for guilty participants.

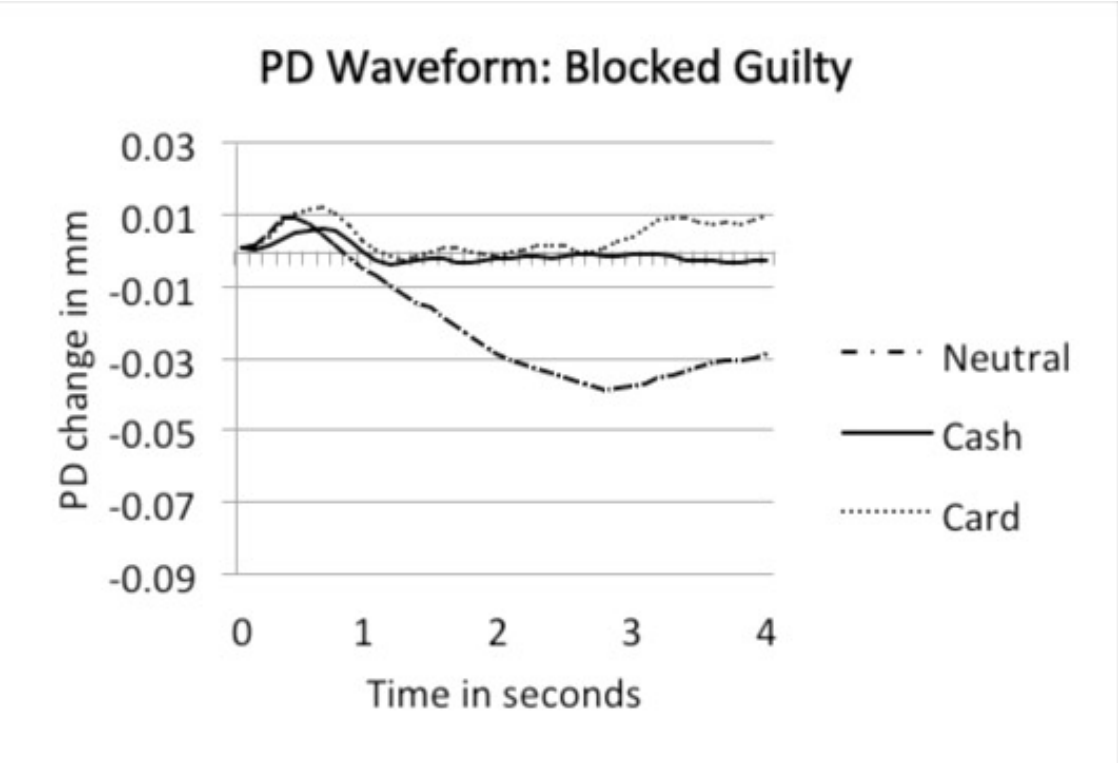


Table 3.3 reports the reliability of ocular-motor measures (coefficient alpha) to determine if reducing the number of items on the ODT adversely affected the reliability of outcome measures. Reliability was measured across the six repetitions of the 48 ODT statements. As a result, the number of 'items' 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 er-

ror 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.

How To:

A Step-by-Step Worksheet for the Multinomial ESS

Raymond Nelson[1], Mark Handler[2], Tom Coffey[3], Rodolfo Prado[4]

and Ben Blalock[5] [6]†

Abstract

We describe the use of a structured analytic process for the Multinomial update to the Empirical Scoring System (ESS-M). Polygraph data analysis begins with feature extraction, numerical transformation, and data reduction. Later stages of analysis include the calculation of a statistical classifier and the parsing of a categorical test result from the numerical and probabilistic test data. This work-flow is designed to capture and organize the information of interest to polygraph field practitioners and other professionals who may need to work with, interpret, and understand the result of polygraphic credibility assessment test data. Appendices include a structured worksheet, multinomial reference tables for diagnostic and screening polygraphs with three to five iterations of two to four relevant questions, and a short glossary of terms that are foundational to Bayesian analysis. Step-by-step procedures describe the use of the ESS-M Analysis Worksheet and the multinomial reference tables. They also include the determination of numerical cut-scores and the calculation of both posterior odds and the lower limit of the 95% credible interval. These materials may be a useful resource for field practice, education and training, and can provide insight into the process automation issues for both manual and automated analysis of polygraphic credibility assessment test data.

Introduction

Nelson (2017a) described the development of a multinomial likelihood function for the Empirical Scoring System (ESS; Nelson, Krapohl & Handler, 2008; Nelson et al., 2011) and provided reference tables for polygraph exams with three to five iterations of two, three, and four relevant questions. The multinomial likelihood function is calculated under the analytic theory of the polygraph test: that greater changes in physiological activity are loaded at different types of test stimuli as a function of deception and truth-telling in response to investigation target stimuli. [Refer to Nelson (2016a) for more information about the analytic theory of polygraphic credibility assessment testing.]

1 Lafayette Instrument Company

2 Converus, Inc.

3 Detective, Chicago Police Department

4 International Polygraph Studies Center

5 PEAK Credibility Assessment Training Center

6 There are no proprietary interests associated with this manuscript. The views and opinions expressed herein are those of the authors.

Multinomial reference tables were subsequently incorporated into a Bayesian classifier for the Empirical Scoring System – Multinomial (ESS-M; Nelson, 2017b). Feature extraction, numerical transformation and data reduction with the ESS-M are unchanged from the ESS. Decision rules are also unchanged. Field practitioners will notice some change in numerical cut-scores, and this reflects the change to a multinomial reference distribution Bayesian classifier. This current project involves the development of a work-flow for computation and documentation of the Bayesian ESS-M result. It describes a simplified multinomial reference model that can be used for Bayesian analysis of polygraph exams consisting of three to five iterations of two, three or four relevant questions.

Bayesian analyses (Berger, 1985; 2006a; Bernardo & Smith, 1994; Box & Tiao, 1973; Cohen, 1994; Efron, 1986; Gelman et al., 2014; Stone, 2013; Winkler, 1972) are advantageous because the statistical results can provide a more intuitively useful estimate of the effect size of interest. This change in information is often reported in the form of odds – versus decimal probabilities that are common to frequentist inference. Probabilistic information in the form of odds can be easily understood by persons with a wide range of professional and educational backgrounds. [For more information about p-values and potential misunderstanding, refer to the published statement from the American Statistical Association (Wasserstein & Lazar 2016), and Nelson (2018a) for a discussion of p-values in the classification of polygraph test results.] Polygraph effect sizes can be thought of as the posterior likelihood of deception or truth-telling, or as a Bayes Factor (Berger, 2006a; Morey & Rouder, 2011; Rouder et al., 2009). A Bayes Factor tells us the strength of the posterior information relative to the prior information.

Bayesian analyses are fairly straightforward and polygraph examiners can benefit from using a carefully structured process when using the ESS-M in field practice settings. Structured work-flows can increase both reliability and efficiency. Automation of an analysis process is one obvious way to increase reliability and efficiency, but it does little to advance knowledge and skill levels. Well-structured procedures offer the advantage of organizing the information of interest while also providing insight as to how the information is used. It is our hope that by using the worksheet, examiners will become more knowledgeable about Bayesian analysis and the ESS-M classifier and will have a better understanding of what the test result can signify. This will contribute to improved communication when conveying information to professional consumers of polygraph test results (supervisors, attorneys, judges, juries, therapists, probation/parole officers, adjudicators, etc.)

How to use the ESS-M Analysis Worksheet.

Appendix A is a structured worksheet to assist in the learning and execution of the foundational concepts of Bayesian analysis and the ESS-M. Appendices B1 through B3 show the simplified ESS-M reference tables that can be applied to the variety of diagnostic and screening polygraph settings with three, four or five iterations of two, three or four relevant questions, using the traditional array of polygraph sensors with or without the optional vasomotor sensor. Tables in Appendices B1 through 3 are limited to calculations with an equal prior and alpha = .05. If a practitioner chooses to use an unequal prior or alpha other than .05, then the calculations will be different and the tables in Appendices B1 through B3 will not be applicable. The tables for subtotal scores include a statistical correction to account for the deleterious effects of multiplicity which are often not considered in other test data analysis models. Appendix C shows a short terminology list that will orient readers to the foundational concepts of Bayesian analysis.

The ESS-M Analysis Worksheet (Appendix A) is used after obtaining all numerical scores for all iterations of all relevant questions. This structured worksheet helps orient professionals to the conceptual vocabulary and analytic process of the Bayesian ESS-M. It helps to organize information for competent documentation and reporting of test results. Finally, it can foster greater understanding of automated computer algorithms that can expedite test data analyses and improve their reliability. In field practice and training, the ESS-M Analysis Worksheet can be printed or

converted to a spreadsheet. Numbered line items on the ESS-M Analysis Worksheet are explained below.

1. Examiners should clearly identify the examination data and/or examinee to which the analysis pertains by recording the name of the examination or examinee on line 1

2. Record the date of the examination on line 2. You may also record the date of the analysis on this line if the analysis is completed at a later date.

3. Record on Line 3 the name or ID of the professional who completed this analysis.

4. Circle one item on Line 4 to indicate whether the test is a *diagnostic* exam (in response to a known allegation or known incident) or *screening* exam (conducted in the absence of any known allegation or incident).

5. Circle the item on Line 5 to indicate the decision rule that will be used to parse the categorical test result from the numerical and probabilistic information. Options include the grand total rule (GTR; Bell, Raskin, Honts & Kircher, 1999; Kircher & Raskin, 1988; Senter, 2003; Weaver, 1980), the subtotal score rule (SSR; Department of Defense, 2006a, 2006b, Capps & Ansley 1992, Senter Waller & Krapohl, 2008) and the two-stage rules (TSR; Senter, 2003; Senter & Dollins, 2003; Krapohl, 2005; Krapohl & Cushman, 2006). These options are aligned with the options on item 4. Diagnostic exams and single-issue screening exams should be evaluated with the GTR or TSR, while multiple-issue exams are most often evaluated with the SSR. [Refer to Nelson (2018b) for more information about polygraph decision rules.]

6. Circle one item to indicate the use of any mathematical correction for statistical multiplicity when using subtotal scores. These options are aligned with those on line 5. Use of the GTR does not require statistical correction. The TSR uses a correction for deceptive subtotal scores at stage 2. The SSR employs a statistical correction for truthful subtotal scores. [Refer to Nelson (2015) for more general information about the use of statistical corrections in polygraph test data analysis.]

7. Enter the *alpha* level for *statistical significance* on Line 7. Two alpha boundaries are required because polygraph testing involves two possible classifications, deception or truth-telling. The alpha level will be used later to determine the coverage level of the Bayesian credible interval and the ESS-M numerical cut-scores. ESS-M reference tables, shown in Appendix B, are calculated with α = .05 for deception and α = .05 for truth-telling. ESS-M alphas are one-tailed unless otherwise specified.

8. Enter the *a priori* odds of deception on Line 8. Without reliable information, our objective knowledge is often limited to 2 possibilities – the examinee is possibly deceptive/guilty, or the examinee is possibly innocent/truthful – with no objective basis for concluding that one possibility is more likely than the other. In these cases, the likelihood of deception can be argued as objectively equal to the likelihood of truth-telling – and the prior odds of deception should be entered as 1 to 1. Of course, some situations may warrant the use of a different prior. In some circumstances, however, objective information might obviate the need for polygraph testing. When the prior information exists in the form of a decimal probability, the prior odds can be calculated using this formula: odds = p / (1 - p).

9. Indicate the prior *probability* of deception on Line 9. The prior probability can be calculated from the prior odds using the following formula: p = odds / (1 + odds). Because the prior odds of deception are often 1 to 1, unless objective information is available that would suggest changing the prior odds, the prior probability is often .5.

10.Use the tables in Appendices B1 through 3 to determine the numerical cut-scores for deception and truth-telling and enter them on Line 10. ESS-M numerical cut-scores are a function of both the prior odds of deception and the lower limit of the *credible interval* (Bayesian confidence interval). When alpha = .05, the credible interval can be referred to as the 95% credible interval. This interval tells us the range in which the test score is likely to be observed upon repetitive testing, given the potential for some random error variation in the test data and test results.

To determine the cut-scores for grand total scores, use the table shown in Appendix B-1. Use the *oddsLL05* column that shows the lower limit of the 95% credible interval for the posterior odds. Locate the smallest lower limit odds that exceeds the prior odds for deception and truth-telling (usually 1:1), then locate the numerical cut-scores in the corresponding row using *score* column, as shown in Figure 1. ESS-M cut-scores for grand total scores are: grand total = +3 or greater for truthful classifications and grand total = -3 or lower for deceptive classifications. Grand total scores from -2 to +2 are not statistically significant, and no opinion is supported by these inconclusive values.

Figure 1. ESS-M cut-scores for grand total scores (from Appendix B-1).

-6	14000	.0369	.2471	.2383	3.2 (.762)	2
-5	14086	.0398	.2847	.2766	2.62 (.723)	1.67
-4	14155	.0424	.3247	.3177	2.15 (.682)	1.39
-3	14210	.0446	.3667	.3611	1.77 (.639)	1.16
-2	14248	.0461	.4102	.4064	1.46 (.594)	0.97
-1	14272	.0471	.4548	.4529	1.21 (.547)	0.8
0	14279	.0475	.5000	.5000	1 (.500)	0.67
1	14272	.0471	.5452	.5471	1.21 (.547)	0.8
2	14248	.0461	.5898	.5936	1.46 (.594)	0.97
3	14210	.0446	.6333	.6389	1.77 (.639)	1.16
4	14155	.0424	.6753	.6823	2.15 (.682)	1.39
5	14086	.0398	.7153	.7234	2.62 (.723)	1.67
6	14000	.0369	.7529	.7617	3.2 (.762)	2

When using the TSR the corrected subtotal cut-score can be determined using the reference table shown in Appendix B-2. Use the *odds234LL05* column to locate the smallest lower limit odds that exceeds the prior odds (usually 1:1). Then locate the cut-score in the same row using the *score* column. Appendix B-2 includes a statistical correction is used to prevent an increase in false positive errors when using TSR with subtotal scores. Figure 2 shows the process.

The ESS-M subtotal cut-score is -7 or lower for deceptive classifications when using the TSR. The overall test result is classified as deceptive when any subtotal has equaled or exceeded these cut-scores. A test result is not-statistically significant (and is therefore inconclusive) if neither the grand total nor any subtotal score has equaled or exceeded the ESS-M cut-scores. *Notice that subtotal scores are not used to make truthful classifications when using the TSR.*

Figure 2. ESS-M cut-scores for deceptive results with subtotal scores using the TSR (Appendix B-2).

-9	423	.0150	.0383	.0315	30.72	3.13 (.758)	4.84	1.37
-8	465	.0216	.0592	.0500	19.01	2.67 (.728)	4.11	1.19
-7	505	.0297	.0875	.0758	12.19	2.3 (.697)	3.3	1.05
-6	540	.0389	.1242	.1104	8.06	2.01 (.668)	2.66	0.93
-5	571	.0489	.1697	.1546	5.47	1.76 (.638)	2.06	0.83
-4	595	.0588	.2236	.2087	3.79	1.56 (.609)	1.58	0.74
-3	615	.0678	.2852	.2720	2.68	1.39 (.582)	1.19	0.66
-2	628	.0750	.3531	.3432	1.91	1.24 (.554)	0.89	0.59
-1	637	.0797	.4254	.4201	1.38	1.11 (.526)	0.65	0.53
0	639	.0814	.5000	.5000	1	1 (.5)	0.48	0.48

Use the reference table shown in Appendix B-3 when using the SSR. For deceptive classifications of subtotal scores use the *oddsLL05* column to determine the smallest lower limit odds that exceed the prior odds of deception (usually 1:1). Then locate the subtotal cut-score for deception in the same row of the *score* column. Figure 3 shows the process. For truthful classifications of subtotal scores use the *odds234LL05* column to determine the smallest lower limit odds that exceed the prior odds of truth-telling (usually 1:1). Then locate the subtotal cut-score for truth-telling in the same row using the *score* column.

When using the SSR, ESS-M cut-scores are -3 or lower at any subtotal for deceptive classifications and +1 or greater at all subtotals for truthful classifications. Subtotal scores between these values are not statistically significant and are therefore inconclusive. When using the SSR, the overall test result is classified as truthful when all subtotal scores have equaled or exceeded the truthful cut-score. A statistical correction is used for truthful classifications to prevent a loss of screening specificity that could result from the SSR requirement that all subtotals scores are statistically significant for truthful classification.

The overall test result is classified as deceptive when one or more subtotal score has equaled or exceeded the cut-score for deception. To prevent loss of screening sensitivity, no statistical correction is used for the deceptive ESS-M subtotal cut-scores for multiple-issue screening polygraphs.

Note that when using the SSR, deceptive and truthful classifications are not made within the same examination. This prevents an increased potential for false-negative error results in the same exam. If any subtotal score is statistically significant for deception, all subtotals that are not significant for deception are meaningless and uninterpretable and are therefore inconclusive.

Figure 3. ESS-M cut-scores for deceptive and truthful classifications using the SSR (Appendix B-3).

-6	540	.0389	.1242	.1104	8.06	2.01 (.668)	2.66	0.93
-5	571	.0489	.1697	.1546	5.47	1.76 (.638)	2.06	0.83
-4	595	.0588	.2236	.2087	3.79	1.56 (.609)	1.58	0.74
-3	615	.0678	.2852	.2720	2.68	1.39 (.582)	1.19	0.66
-2	628	.0750	.3531	.3432	1.91	1.24 (.554)	0.89	0.59
-1	637	.0797	.4254	.4201	1.38	1.11 (.526)	0.65	0.53
0	639	.0814	.5000	.5000	1	1 (.5)	0.48	0.48
1	637	.0797	.5746	.5799	1.38	2.63 (.725)	0.65	1.18
2	628	.0750	.6469	.6568	1.91	7.01 (.875)	0.89	2.45
3	615	.0678	.7148	.7280	2.68	19.17 (.95)	1.19	4.13
4	595	.0588	.7764	.7913	3.79	54.52 (.982)	1.58	5.31
5	571	.0489	.8303	.8454	5.47	163.4 (.994)	2.06	6.77
6	540	.0389	.8758	.8896	8.06	522.8 (.998)	2.66	7.47

11. Circle the item on Line 11 to indicate the number of relevant questions used in this examination.

12-15. Use Lines 12 through 15 to record the information for each relevant question. Enter both the question label and subtotal score. Do not leave blank lines. When less than four relevant questions are used, enter NA or line-out all blank items.

16. Enter the grand total score on Line 16 when using the GTR or TSR. Enter N/A or line-out this block when using the SSR.

17. Locate and transfer to Line 17 the question label and subtotal score for the *lowest* relevant question subtotal. When subtotal scores are used for classification, only the lowest subtotal is used for classification and statistical inference. Enter N/A or line-out this block when using the GTR.

18. Circle the categorical result on

Line 18 that is supported by the numerical scores. Determine the overall test result using the GTR, TSR, or SSR – as indicated in item 5.

19. Circle the either *grand total score* or *lowest subtotal score* on Line 19 to indicate which value was used to classify the overall test result.

20. Enter the name or identifier for the ESS-M reference table (Appendix name) on Line 20.

21. Determine the posterior odds of deception or truth-telling for that score using the tables in Appendix B and enter the value on Line 21.

Use Appendix B-1 if the overall test result is determined by the grand total score, as would occur when using the GTR or TSR and the grand total score has equaled or exceeded a numerical cut-score. Locate the grand total score in the *score* column. Then locate the posterior odds of deception or truth-telling in the corresponding row of the odds column.

Figure 4 shows an example for which a grand total score of +8 produces a posterior odds of 4.9 for truth-telling. When reporting ESS-M posterior odds it is preferable to round the results to one decimal place for values less than 10 and round results to the nearest integer when the posterior odds are 10 or greater.

Figure 4. Locate the posterior odds using the grand total score (Appendix B-1).

-5	14086	.0398	.2847	.2766	2.62 (.723)	1.67
-4	14155	.0424	.3247	.3177	2.15 (.682)	1.39
-3	14210	.0446	.3667	.3611	1.77 (.639)	1.16
-2	14248	.0461	.4102	.4064	1.46 (.594)	0.97
-1	14272	.0471	.4548	.4529	1.21 (.547)	0.8
0	14279	.0475	.5000	.5000	1 (.500)	0.67
1	14272	.0471	.5452	.5471	1.21 (.547)	0.8
2	14248	.0461	.5898	.5936	1.46 (.594)	0.97
3	14210	.0446	.6333	.6389	1.77 (.639)	1.16
4	14155	.0424	.6753	.6823	2.15 (.682)	1.39
5	14086	.0398	.7153	.7234	2.62 (.723)	1.67
6	14000	.0369	.7529	.7617	3.2 (.762)	2
7	13900	.0336	.7878	.7970	3.93 (.797)	2.39
8	13783	.0303	.8197	.8290	4.85 (.829)	2.85
9	13652	.0269	.8486	.8576	6.02 (.858)	3.41

Use the reference table in Appendix B-2 for subtotal scores during the second stage of the TSR, when the grand total score is inconclusive. Figure 5 shows an example of the second stage of TSR, with a subtotal score of -9, for which the posterior odds of deception are 3.1 to 1.

Figure 5. Locate the posterior odds using a subtotal score when using the TSR (Appendix B-2).

-10	378	.0099	.0236	.0190	51.67	3.73 (.789)	5.22	1.56
-9	423	.0150	.0383	.0315	30.72	3.13 (.758)	4.84	1.37
-8	465	.0216	.0592	.0500	19.01	2.67 (.728)	4.11	1.19
-7	505	.0297	.0875	.0758	12.19	2.3 (.697)	3.3	1.05
-6	540	.0389	.1242	.1104	8.06	2.01 (.668)	2.66	0.93
-5	571	.0489	.1697	.1546	5.47	1.76 (.638)	2.06	0.83
-4	595	.0588	.2236	.2087	3.79	1.56 (.609)	1.58	0.74
-3	615	.0678	.2852	.2720	2.68	1.39 (.582)	1.19	0.66
-2	628	.0750	.3531	.3432	1.91	1.24 (.554)	0.89	0.59
-1	637	.0797	.4254	.4201	1.38	1.11 (.526)	0.65	0.53
0	639	.0814	.5000	.5000	1	1 (.5)	0.48	0.48
1	637	.0797	.5746	.5799	1.38	2.63 (.725)	0.65	1.18

Use Appendix B-3 when using the SSR. Figure 6 shows the results of a polygraph exam for which the lowest subtotal score is +2. In this example, using the SSR, the multiplicity corrected posterior odds of truth-telling are 7 to 1.

Figure 6. Posterior odds for truthful subtotal scores of multiple issue screening exams.

-5	571	.0489	.1697	.1546	5.47 (.845)	1.76	2.06	0.83
-4	595	.0588	.2236	.2087	3.79 (.791)	1.56	1.58	0.74
-3	615	.0678	.2852	.2720	2.68 (.728)	1.39	1.19	0.66
-2	628	.0750	.3531	.3432	1.91 (.656)	1.24	0.89	0.59
-1	637	.0797	.4254	.4201	1.38 (.580)	1.11	0.65	0.53
0	639	.0814	.5000	.5000	1	1	0.48	0.48
1	637	.0797	.5746	.5799	1.38	2.63 (.725)	0.65	1.18
2	628	.0750	.6469	.6568	1.91	7.01 (.875)	0.89	2.45
3	615	.0678	.7148	.7280	2.68	19.17 (.95)	1.19	4.13
4	595	.0588	.7764	.7913	3.79	54.52 (.982)	1.58	5.31
5	571	.0489	.8303	.8454	5.47	163.4 (.994)	2.06	6.77

22. Enter the posterior probability of deception or truth-telling on Line 22. This can be calculated manually from the posterior odds using this formula: p = odds / (1 + odds). For convenience, the *odds* columns in Appendices B1-3 show the posterior probability in parenthesis along with the posterior odds. The posterior odds can also be seen in Figures 4, 5, and 6.

23. Enter the *lower limit* of the credible interval (Bayesian confidence interval) for the posterior odds of deception or truth-telling on Line 23. This value can be obtained by locating the grand total score or lowest subtotal scores – depending on which was indicated on line 19. The lower limit can be found in the columns *oddsLL05* column or the *odds234LL05* column using the reference tables in Appendices B1-3. For convenience, the lower limit columns are shaded along with the odds and score columns in Appendices B1-3.

24. Calculate the Bayes Factor (Berger, 2006b) and enter it on Line 24. Bayes Factor is a statistic that tells us the relative strength of the posterior information when compared to the prior. For example, a Bayes factor of 7 indicates that the posterior information supporting a conclusion of deception or truth-telling is 7 times greater than the prior information. Bayes Factor is easily calculated using this formula: posterior odds / prior odds. As a convenience, Bayes Factor will be equal to the posterior odds whenever the prior odds are 1 to 1 (because any number divided by 1 is equal to the same number).

25. Transfer the information to line 25 using the information in Line 7 (alpha for deception | alpha for truth-telling). Select the value that was used to classify the test result as deceptive or truthful. Two alpha values were entered on line 7, and only one alpha should be entered here on Line 25. Circle DI/SR or NDI/NSR to clearly indicate which alpha value was used to classify the test result. This may seem unimportant when the two alpha values are symmetrical, but it will become important whenever the alphas are asymmetrical.

26. Finally, calculate the coverage interval for the Posterior Credible Interval (Bayesian confidence interval) and enter the result on Line 26. This is easily calculated using this formula: (1 - *alpha*) x 100%. For example: with *a* = .05, the coverage area is (1 – .05) x 100% = 95%. This value is not intended for use as a practical effect size – for which posterior odds are better intended. The credible interval is only an estimate of the degree of certainty that a test result is *indicative* of deception or truth-telling. This can also be thought of as an estimate of the random error potential or the likelihood of obtaining another similar test result under similar testing conditions.

Algorithmic approaches to test data analysis.

In the most rudimentary sense, an algorithm is merely a structured procedure used to solve a problem. Mathematics is full of algorithms for basic operations such as addition, subtraction, multiplication, division, and other more challenging problems. In a more general sense, an algorithm can be thought of as somewhat similar to a recipe, involving both ingredients and a procedure [See Nelson (2016b) for a more complete discussion]. A structured worksheet, such as the one described in this manuscript, is an analog or manual form of an algorithm. Use of algorithms is abundant in data analytics, and some will immediately associate the word algorithm as meaning *computer algorithm* or *computer scoring algorithm* when discussing software applications developed to analyze test or experimental data. For example, polygraph test data can be analyzed with an automated computer scoring algorithm such as the OSS-3 (Nelson, Krapohl & Handler, 2008). Any procedure or rubric can be thought of as an algorithm – including procedures such as the ESS-M – whether executed manually or via automated computer. The structured analysis worksheet shown in Appendix A, can also be thought of as an algorithm – used to organize the information pertaining to the analysis.

Conclusion

Test data analysis models of all types consist of similar parts or similar operations. These include: feature development and extraction, numerical transformation and data reduction, the use of a likelihood function to calculate a statistical value for the test data, and structured procedures to interpret the probabilistic and categorical test results.

Polygraphic feature extraction has historically been a somewhat subjective process of visual inspection – though the use of automated feature extraction algorithms has increased during recent years. Numerical transformations by automated computer scoring algorithms can employ a variety of approaches. These include ratios, proportions, z-scores, log transformations, probit transformations, and other methods. Manual scoring procedures have traditionally used integer level numerical transformations such as Likert-type scores (Likert, 1932) to transform polygraph response features into numerical values – in both 7-point and 3-point variants. Rank order transformations have also been described and applied to polygraph testing.

Numerical transformations can be executed either manually or automatically. The traditional 7-position Likert scale – originally used to quantify subjective opinion data – is an example of a transformation that may have limited potential for automation due to the use of subjective and arbitrary differences within the scale values. Unlike the 7-position model, 3-position scale transformations are easily automated because they are more readily made objective and more easily subject to statistical optimization. Moreover, no theoretical distribution exists for 7-position scores, whereas 3-position scores can be characterized using a multinomial distribution. ESS-M scores are a simple variation of the 3-point scoring method.

Data reduction for manual scoring systems is a straightforward matter of addition with positive and negative integer scores. Computer algorithms will often employ more advanced methods of data aggregation and reduction, including the use of averaging, weighted averaging, and other structural functions to aggregate data. Likelihood functions can take a variety of forms, from mathematical formulae, to empirical sampling distributions. They can also include statistical/mathematical reference distributions such as the multinomial distribution shown in Appendices B1 through 3.

Similarly, decision rules – used to parse a categorical result from numerical and statistical data – can take many forms. For example, the worksheet included in Appendix A captures information about the selection of one of three polygraphic decision rules: the grand total rule, the two-stage rule, and the subtotal-score rule.

The ESS-M is a simple, yet powerful application of Bayes theorem and the principles of Bayesian analysis with polygraphic credibility assessment test data. Whereas the original ESS introduced the potential for

a statistical classifier to manual scoring procedures, the ESS-M provides a platform for the convenient use of Bayesian analysis. The structured work-flow described in this manuscript can assist polygraph examiners and others to better understand the ESS-M Bayesian classifier, and can be easily implemented in both manual and automated processes.

Automated processes offer the advantages of ease of use, increased reliability, and the potential for more sophisticated analytic methods that can meet the requirements for statistical classifications in 21st century forensic science. We anticipate the increased use of automation as software applications become more widely available. We caution against any notion that it is acceptable for polygraph professionals to assert that their understanding of data analysis should stop when they click a button to run an algorithm. We further caution against the belief that polygraph analytics has remained static or has failed to develop beyond the manual scoring innovations from the mid-century (pre-computer) epoch. As a final caution, use of automated analysis algorithms should not be limited to circumstances in which an examiner desires merely to strengthen the *impression* of advancement in the science of polygraphic credibility assessment testing. Continuing to rely on subjective integer scoring methods without statistical quantification as the actual basis for decision seems ill-advised.

Our hope is that all polygraph professionals would endeavor to become proficient with the details and procedures of manual test data analysis and the manual calculation of statistical classifiers for their polygraph test result. By doing so, they will be better equipped to make effective use of computer software tools, and better equipped to account for their conclusions when discussing them with other professionals. We hope that this manuscript and the accompanying ESS-M Analysis Worksheet will be a useful educational and field-practice resource to polygraph examiners and other professionals.

References

Bell, B. G., Raskin, D. C., Honts, C. R. & Kircher, J.C. (1999). The Utah numerical scoring system. *Polygraph, 28(1)*, 1-9.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Second edition. New York: Springer Verlag.

Berger, J. (2006a). The Case for Objective Bayesian Analysis. Bayesian Analysis, 1(3), 385–402.

Berger, J. O. (2006b). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory.* John Wiley.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Wiley.

Capps, M. H. & Ansley, N. (1992). Analysis of federal polygraph charts by spot and chart total. *Polygraph, 21,* 110-131.

Cohen, J. (1994). The earth is round (p<.05). *Psychological Bulletin, 112,* 155-159.

Department of Defense (2006a). *Federal Psychophysiological Detection of Deception Examiner Handbook.* Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007. Reprinted in Polygraph, 40(1), 2-66.

Department of Defense (2006b). *Psychophysiological Detection of Deception Analysis II -- Course #503. Test data analysis: DoDPI numerical evaluation scoring system.* Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007).

Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician. 40(1).* 1-5.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtary, A., & Rubin, D. B. (2014). *Bayesian Data Analysis.* CRC Press.

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

Krapohl, D. J. (2005). Polygraph decision rules for evidentiary and paired-testing (Marin Protocol) applications. *Polygraph, 34(3)*, 184-192.

Krapohl, D. J. & Cushman, B. (2006). Comparison of evidentiary and investigative decision rules: a replication. *Polygraph, 35(1),* 55-63.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5-55.

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16(4)*, 406–419.

Nelson, R. (2015). Bonferroni and Šidák corrections for multiplicity effects with subtotal score of

comparison question polygraph tests. *Polygraph, 44(2)*, 162-167.

Nelson, R. (2016a). Scientific (analytic) theory of polygraph testing. *APA Magazine, 49(5)*, 69-82.

Nelson, R. (2016b). What Is An Algorithm? And Why Don't We Use 'Em? (Bonus Recipe Included: Habanero-Cranberry Cookies). *APA Magazine, 49(2),* 49-68.

Nelson, R. (2017a). Multinomial reference distributions for the Empirical Scoring System. *Polygraph & Forensic Credibility Assessment, 46(2)*, 81-115.

Nelson, R. (2017b). Updated numerical distributions for the Empirical Scoring System: An accuracy demonstration with archival datasets with and without the Vasomotor Sensor. *Polygraph & Forensic Credibility Assessment, 46 (2),* 116-131.

Nelson, R. (2018a). Five-minute science lesson: review and discussion of the American Statistical Association's "Statement on Statistical Significance and P-Values" and polygraph test results. *APA Magazine, 51(1)*, 60-64.

Nelson, R. (2018b). Practical polygraph: a survey and description of decision rules. *APA Magazine, 51(2),* 127-133.

Nelson, R., Handler, M., Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B. & Oelrich, M. (2011). Using the Empirical Scoring System. *Polygraph, 40,* 67-78.

Nelson, R., Krapohl, D. & Handler, M. (2008). Brute force comparison: A Monte Carlo study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers. *Polygraph, 37,* 185-215.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16(2)*, 225–237.

Senter, S. (2003). Modified general question test decision rule exploration. *Polygraph, 32*, 251- 263.

Senter, S., Waller, J. & Krapohl, D. (2008). Air Force Modified General Question Test Validation Study. *Polygraph, 37(3)*, 174-184.

Senter, S. M. & Dollins, A. B. (2003). *New Decision Rule Development: Exploration of a two-stage approach.* Report number DoDPI00-R-0001. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC. Reprinted in Polygraph, 37(2), 149-164.

Stone, J. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press.

Weaver, R. S. (1980). The numerical evaluation of polygraph charts: Evolution and comparison of three major systems. *Polygraph, 9*, 94-108.

Wasserstein, R. L. & Lazar, N. A. (2016) The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70(2)*, 129-133.

Winkler, R. L. (1972). *An Introduction to Bayesian Inference and Decision.* Holt McDougal.

Appendix A: ESS-M Analysis Worksheet

1.	**Exam ID / Examinee Name**	
2.	**Exam date (analysis date)**	
3.	**Examiner / analyst**	
4.	Diagnostic or Screening Exam	Diagnostic | Screening
5.	Decision Rule	GTR | TSR | SSR
6.	Statistical correction for subtotals	NA | Deceptive subtotals | Truthful subtotals
7.	Alpha levels (1-tailed level of sig.): Truth | Deception	|
8.	Prior odds of deception = p / (1 - p)	
9.	Prior probability of deception = odds / (1 + odds)	
10.	Cut-scores: Truth / Deception (Subtotal)	| ( )
11.	Number of RQs	2RQs 3RQs 4RQs
12.	RQ: Question ID / Score	|
13.	RQ: Question ID / Score	|
14.	RQ: Question ID / Score	|
15.	RQ: Question ID / Score	|
16.	Grand total (enter NA when using the SSR)	
17.	Lowest Subtotal Score: Question ID | Score	|
18.	**Result**	**DI/SR NDI/NSR INC/NO**
19.	**Classified by**	**Grand Total Lowest Subtotal**
20.	**ESS-M Reference Table (circle one)**	**B-1 (GTR & TSR) | B-2 (TSR) | B-3 (SSR)**
21.	**Posterior odds = p / (1 – p) (systematic error est.)**	
22.	**Posterior probability = odds / (1 + odds)**	
23.	**Lower-limit Posterior Credible Interval (oddsLL05 or odds234LL05)**	
24.	**Bayes Factor = Posterior Odds / Prior Odds**	
25.	**Alpha tolerance/sig. level (random error est.)**	**DI/SR NDI/NSR**
26.	**(1 - *alpha*) x 100% Bayesian credible interval**	

Appendix B-1: Simple ESS-M Cut-scores for Grand Total Scores with 3, 4 or 5 presentations of 2, 3, or 4 Relevant Questions with or without the Vasomotor Sensor

Prior = .5 (1 to 1), Alpha = .05 / .05 (truth / deception)

score	ways	*pmf*	*cdf*	cdfContCor	odds	oddsLL05
-24	9915	.0008*	.0023	.0019	518.7 (.998)	21.4
-23	10248	.0011	.0034	.0028	352.2 (.997)	20.18
-22	10572	.0015	.0048	.0041	242.7 (.996)	18.69
-21	10888	.0020	.0069	.0059	169.7 (.994)	16.95
-20	11193	.0027	.0096	.0082	120.4 (.992)	17.25
-19	11488	.0036	.0132	.0114	86.55 (.989)	14.98
-18	11770	.0047	.0179	.0156	63.05 (.984)	13.98
-17	12040	.0061	.0239	.0210	46.52 (.979)	12.51
-16	12295	.0077	.0316	.0280	34.75 (.972)	10.89
-15	12536	.0097	.0411	.0367	26.26 (.963)	9.29
-14	12760	.0119	.0527	.0475	20.06 (.953)	8.05
-13	12970	.0144	.0668	.0607	15.49 (.939)	6.83
-12	13163	.0172	.0835	.0765	12.07 (.924)	5.74
-11	13342	.0202	.1031	.0953	9.5 (.905)	4.85
-10	13504	.0235	.1257	.1172	7.53 (.883)	4.06
-9	13652	.0269	.1514	.1424	6.02 (.858)	3.41
-8	13783	.0303	.1803	.1710	4.85 (.829)	2.85
-7	13900	.0336	.2122	.2030	3.93 (.797)	2.39
-6	14000	.0369	.2471	.2383	3.2 (.762)	2
-5	14086	.0398	.2847	.2766	2.62 (.723)	1.67
-4	14155	.0424	.3247	.3177	2.15 (.682)	1.39
-3	14210	.0446	.3667	.3611	1.77 (.639)	1.16
-2	14248	.0461	.4102	.4064	1.46 (.594)	0.97
-1	14272	.0471	.4548	.4529	1.21 (.547)	0.8
0	14279	.0475	.5000	.5000	1 (.500)	0.67
1	14272	.0471	.5452	.5471	1.21 (.547)	0.8
2	14248	.0461	.5898	.5936	1.46 (.594)	0.97
3	14210	.0446	.6333	.6389	1.77 (.639)	1.16
4	14155	.0424	.6753	.6823	2.15 (.682)	1.39
5	14086	.0398	.7153	.7234	2.62 (.723)	1.67
6	14000	.0369	.7529	.7617	3.2 (.762)	2
7	13900	.0336	.7878	.7970	3.93 (.797)	2.39
8	13783	.0303	.8197	.8290	4.85 (.829)	2.85
9	13652	.0269	.8486	.8576	6.02 (.858)	3.41
10	13504	.0235	.8743	.8828	7.53 (.883)	4.06
11	13342	.0202	.8969	.9047	9.5 (.905)	4.85
12	13163	.0172	.9165	.9235	12.07 (.924)	5.74
13	12970	.0144	.9332	.9393	15.49 (.939)	6.83
14	12760	.0119	.9473	.9525	20.06 (.953)	8.05
15	12536	.0097	.9590	.9633	26.26 (.963)	9.29
16	12295	.0077	.9685	.9720	34.75 (.972)	10.89
17	12040	.0061	.9761	.9790	46.52 (.979)	12.51
18	11770	.0047	.9821	.9844	63.05 (.984)	13.98
19	11488	.0036	.9868	.9886	86.55 (.989)	14.98
20	11193	.0027	.9904	.9918	120.4 (.992)	17.25
21	10888	.0020	.9931	.9941	169.7 (.994)	16.95
22	10572	.0015	.9952	.9959	242.7 (.996)	18.69
23	10248	.0011	.9966	.9972	352.2 (.997)	20.18
24	9915	.0008*	.9977	.9981	518.7 (.998)	21.4
* extreme values omitted						

Score is the grand total score. Ways is the number of sensor-score combinations that can achieve each total score. *pmf* is the probability mass for each score. *cdf* is the cumulative sum of the *pmf*. *cdfContCor* is the continuity corrected *cdf,* so that the statistical estimate always exceeds the actual statistical value – also used as the posterior probability. *odds* are the posterior odds of truth or deception - calculated from the *cdfContCor* using *p*/(1-*p*). Also, the *cdfContCor* can be calculated from the odds using *odds*/(1+*odds*). The *oddsLL05* are the lower limits of the credible interval (Bayesian confidence interval) for prior=.5 and alpha/2 = .05 for truth and deception.

Appendix B-2: Simple ESS-M Cut-scores for Sub-total Scores of Single- Issue Exams with 3, 4 or 5 presentations of 2, 3 or 4 RQs with or without the Vasomotor Sensor

Prior = .5 (1 to 1), Alpha = .05 / .05 (truth / deception) – all statistical corrections are included

score	ways	*pmf*	*cdf*	Cdf ContCor	odds	Odds234RQs	oddsLL05	odds234LL05
-15	161	.0005*	.0009	.0007	1517	11.49 (.92)	7.71	3.32
-14	200	.0011	.0020	.0015	682.2	8.8 (.898)	7.56	2.84
-13	243	.0021	.0041	.0030	328.4	6.9 (.873)	7.27	2.42
-12	287	.0037	.0077	.0059	168	5.52 (.847)	6.79	2.07
-11	333	.0062	.0139	.0109	90.88	4.5 (.818)	6.1	1.81
-10	378	.0099	.0236	.0190	51.67	3.73 (.789)	5.22	1.56
-9	423	.0150	.0383	.0315	30.72	3.13 (.758)	4.84	1.37
-8	465	.0216	.0592	.0500	19.01	2.67 (.728)	4.11	1.19
-7	505	.0297	.0875	.0758	12.19	2.3 (.697)	3.3	1.05
-6	540	.0389	.1242	.1104	8.06	2.01 (.668)	2.66	0.93
-5	571	.0489	.1697	.1546	5.47	1.76 (.638)	2.06	0.83
-4	595	.0588	.2236	.2087	3.79	1.56 (.609)	1.58	0.74
-3	615	.0678	.2852	.2720	2.68	1.39 (.582)	1.19	0.66
-2	628	.0750	.3531	.3432	1.91	1.24 (.554)	0.89	0.59
-1	637	.0797	.4254	.4201	1.38	1.11 (.526)	0.65	0.53
0	639	.0814	.5000	.5000	1	1	0.48	0.48
1	637	.0797	.5746	.5799	1.38	2.63	0.65	1.18
2	628	.0750	.6469	.6568	1.91	7.01	0.89	2.45
3	615	.0678	.7148	.7280	2.68	19.17	1.19	4.13
4	595	.0588	.7764	.7913	3.79	54.52	1.58	5.31
5	571	.0489	.8303	.8454	5.47	163.4	2.06	6.77
6	540	.0389	.8758	.8896	8.06	522.8	2.66	7.47
7	505	.0297	.9125	.9242	12.19	1810	3.3	7.73
8	465	.0216	.9408	.9500	19.01	6870	4.11	7.82
9	423	.0150	.9617	.9685	30.72	28990	4.84	7.84
10	378	.0099	.9764	.9810	51.67	137900	5.22	7.84
11	333	.0062	.9861	.9891	90.88	750600	6.1	7.85
12	287	.0037	.9923	.9941	168	4745000	6.79	7.85
13	243	.0021	.9959	.9970	328.4	3.54E+07	7.27	7.85
14	200	.0011	.9980	.9985	682.2	3.17E+08	7.56	7.85
15	161	.0005*	.9991	.9993	1517	3.49E+09	7.71	7.85

* extreme values omitted

Score is the lowest subtotal score. Ways is the number of sensor-score combinations that can achieve each subtotal score. *pmf* is the probability mass for each score. *cdf* is the cumulative sum of the *pmf*. *cdfContCor* is the continuity corrected *cdf,* so that the statistical estimate always exceeds the actual statistical value – also used as the posterior probability.

Odds are the posterior odds of truth or deception for a single subtotal (without statistical correction) - calculated from the *cdfContCor* using *p*/(1-*p*). Also, the *cdfContCor* can be calculated from the odds using *odds*/(1+*odds*). *oddsLL05* is the lower limit of the credible interval (Bayesian confidence interval) for prior=.5 and alpha/2 = .05 for truth and deception.

Odds 234RQ median odds of deception with statistical correction for 2, 3 and 4 RQs. *Odds234LL05* are the median lower limits of the credible interval (Bayesian confidence interval) for prior=.5 and alpha/2 = .05 for truth and deception with 2, 3, or 4 RQs.

Appendix B-3: Simple ESS-M Cut-scores for Sub-total Scores of Multiple-Issue Exams with 3, 4 5 presentations of 2, 3 or 4 RQs with or without the Vasomotor Sensor

Prior = .5 (1 to 1), Alpha = .05 / .05 (truth / deception) – all statistical corrections are included

score	ways	*pmf*	*cdf*	Cdf ContCor	odds	Odds234RQs	oddsLL05	odds234LL05
-15	161	.0005*	.0009	.0007	1517 (>.999)	11.49	7.71	3.32
-14	200	.0011	.0020	.0015	682.2 (.999)	8.8	7.56	2.84
-13	243	.0021	.0041	.0030	328.4 (.997)	6.9	7.27	2.42
-12	287	.0037	.0077	.0059	168 (.994)	5.52	6.79	2.07
-11	333	.0062	.0139	.0109	90.88 (.989)	4.5	6.1	1.81
-10	378	.0099	.0236	.0190	51.67 (.981)	3.73	5.22	1.56
-9	423	.0150	.0383	.0315	30.72 (.968)	3.13	4.84	1.37
-8	465	.0216	.0592	.0500	19.01 (.950)	2.67	4.11	1.19
-7	505	.0297	.0875	.0758	12.19 (.924)	2.3	3.3	1.05
-6	540	.0389	.1242	.1104	8.06 (.890)	2.01	2.66	0.93
-5	571	.0489	.1697	.1546	5.47 (.845)	1.76	2.06	0.83
-4	595	.0588	.2236	.2087	3.79 (.791)	1.56	1.58	0.74
-3	615	.0678	.2852	.2720	2.68 (.728)	1.39	1.19	0.66
-2	628	.0750	.3531	.3432	1.91 (.656)	1.24	0.89	0.59
-1	637	.0797	.4254	.4201	1.38 (.580)	1.11	0.65	0.53
0	639	.0814	.5000	.5000	1	1	0.48	0.48
1	637	.0797	.5746	.5799	1.38	2.63 (.725)	0.65	1.18
2	628	.0750	.6469	.6568	1.91	7.01 (.875)	0.89	2.45
3	615	.0678	.7148	.7280	2.68	19.17 (.95)	1.19	4.13
4	595	.0588	.7764	.7913	3.79	54.52 (.982)	1.58	5.31
5	571	.0489	.8303	.8454	5.47	163.4 (.994)	2.06	6.77
6	540	.0389	.8758	.8896	8.06	522.8 (.998)	2.66	7.47
7	505	.0297	.9125	.9242	12.19	1810 (>.999)	3.3	7.73
8	465	.0216	.9408	.9500	19.01	6870 (>.999)	4.11	7.82
9	423	.0150	.9617	.9685	30.72	28990 (>.999)	4.84	7.84
10	378	.0099	.9764	.9810	51.67	137900 (>.999)	5.22	7.84
11	333	.0062	.9861	.9891	90.88	750600 (>.999)	6.1	7.85
12	287	.0037	.9923	.9941	168	4745000 (>.999)	6.79	7.85
13	243	.0021	.9959	.9970	328.4	3.54E+07 (>.999)	7.27	7.85
14	200	.0011	.9980	.9985	682.2	3.17E+08 (>.999)	7.56	7.85
15	161	.0005*	.9991	.9993	1517	3.49E+09 (>.999)	7.71	7.85

* extreme values omitted

Score is the lowest subtotal score. Ways is the number of sensor-score combinations that can achieve each subtotal score. *pmf* is the probability mass for each score. *cdf* is the cumulative sum of the *pmf*. *cdfContCor* is the continuity corrected *cdf,* so that the statistical estimate always exceeds the actual statistical value – also used as the posterior probability.

Odds are the posterior odds of truth or deception for a single subtotal (without statistical correction) - calculated from the *cdfContCor* using *p*/(1-*p*). Also, the *cdfContCor* can be calculated from the odds using *odds*/(1+*odds*). *oddsLL05* is the lower limit of the credible interval (Bayesian confidence interval) for prior=.5 and alpha/2 = .05 for truth and deception.

Odds 234RQ median odds of deception with statistical correction for 2, 3 and 4 RQs. *Odds234LL05* are the median lower limits of the credible interval (Bayesian confidence interval) for prior=.5 and alpha/2 = .05 for truth and deception with 2, 3, or 4 RQs.

Appendix C: Vocabulary Primer for Bayesian Analysis

Bayesian inference Inference is the process of using data to make an estimate of an unknown quantity of interest, such as the likelihood that a person has been deceptive or truthful. Inference is necessary when neither deterministic observation nor physical measurement are possible. Bayesian inference is the use of Bayes' theorem for this purpose.

Bayes theorem Bayes theorem is a mathematical idea that is used to calculate a posterior probability by using evidence from a test or experiment to update a prior probability. Theorems are mathematical ideas that have been subject to exhaustive mathematical proof.

Probability Bayesian probability refers to the degree of belief, or degree of certainty, that can be attributed to some knowledge or conclusion. This is in contrast to the frequentist definition of probability, which refers to the number of observed occurrences of something compared to a number of possible occurrences. Whereas the frequentist definition of probability can be applied only to phenomena that are both observable and repeatable, Bayesian probability has wide-ranging application in medicine, psychology, forensics, epidemiology, business, sports, and other fields.

Prior probability Sometimes referred to as simply "prior" and sometimes using the Latin a priori, this refers to what is known before a test or experiment about the likelihood of different possible outcomes. Prior probability refers to the degree of belief in some knowledge or conclusion before more evidence is obtained from a test, experiment, or other investigation. Objective knowledge is often unavailable, and in this case the prior probability may be considered equal for the different possible outcomes.

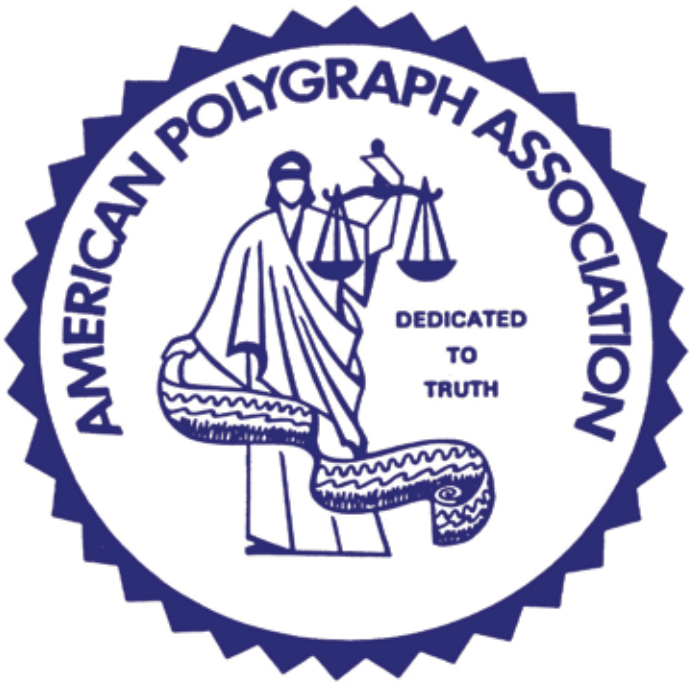
Likelihood function A mathematical device used to obtain a statistical value for some data. A likelihood function can take the form of a mathematical formula, a reference distribution, or a reference table, or even a small computer program to execute the formula to provide a statistical value for the data.

Posterior probability Refers to the probability or likelihood associated with some knowledge or conclusion after the evidence is taken into consideration – through the use of Bayes theorem.

Odds A way of expressing probabilistic information using whole numbers instead of decimal values, and are therefore more intuitively understood by some persons. Odds convey clearly that all probabilities are a comparison of some possibility compared to some other possibility. Odds are easily calculated from decimal probabilities using: odds = p / (1 – p). Also, decimal probabilities may be obtained from odds by: p = odds / (1 + odds).

Bayes Factor Bayes Factor tells us the relative change in the strength of information before and after a test or experiment. Bayes Factor is the ratio of: probability / prior. Bayes Factor will be equal to the posterior whenever the prior is equal to 1.

Credible interval A credible interval is the Bayesian analog for a frequentist confidence interval, and tells us the expected range of variability for a posterior probability. For example, a 95% credible interval tells us the range in which we expect to observe a similar result if a test or experiment is repeated. Whereas a frequentist confidence interval regards data as variable and reality as fixed, Bayesian analysis regards the available data as a fixed quantity with which we can estimate the likelihood of an unobservable though real phenomena of interest.


AMERICAN POLYGRAPH ASSOCIATION
DEDICATED
TO
TRUTH