Testing the Truth-Teller Who Was There

Avital Ginton[1]

Abstract

The danger of having a higher false positive (FP) error rate in testing victims has been acknowledged over the years (Ginton, 1993; Ginton, 1997; Horvath, 1977; Raskin, 1986), and calls for extra caution and specific steps to be taken (Ginton, 1993; Ginton, 1997). Based on a recently published new concept - Relevant Issue Gravity (RIG) (Ginton, 2009) - the present paper claims that this danger exists also in non-victim situations when testing truth-teller examinees that have vivid memories related to the event under investigation (the relevant issue). A recommended preventive remedy in the way of conducting the test is suggested.

Keywords: Comparison Question Technique, CQT, Relevant Issue Gravity, RIG, Truth-teller, False Positive, Elastic Cover, Adaptive Polygraphy.

The physiological reactions we are looking for in psychophysiological detection of deception are by no means “lie reactions.” Yes, we do see such reactions accompany the act of deception quite a lot, but not always, and of course, they are known to occur in the absence of any actual or intended lie. Whether they indicate the arising of an emotional response(s) accompanying the perception of stimuli that threaten the examinee’s safety or well being (Fight, Flight or Freeze notion) or the cognitive activity reflecting the perceived importance of the stimuli presented to the examinee by the question (salience hypothesis), both of them, or none of them (e.g. internal cognitive conflict or even mere physiological activity), they might look the same. Our task is to make sure to detect and measure their appearance and safely relate them to the occurrence of specific acts of deception. The first task is taken care mainly by the instrument, but the latter has to do with the way we conduct the test, and in particular, choosing the appropriate testing technique preceded by a proper pre-test interview. While it is up to the instrument to detect and measure the physiological reactions when they occur, it is the examiner’s responsibility to make sure the examinee is reacting. Most examinees in most instances will react to the questions spontaneously, but some might need to be stimulated to do so. Having taken care of that part of our mission (i.e., making sure the physiological reactions we expect are actually induced), we still face our most important and difficult task, namely, relating them or some of them to specific acts of deception. This is what the examination is all about, and this is the main reason and justification for developing various testing techniques and formats.

The most common technique in modern polygraphy is the Comparison Question Technique (CQT), previously known as the Control Question Technique. The Comparison Question Technique appears in quite a few different formats originating mostly from the pioneering works of John Reid (1947) or Cleve Backster (1963), but always based on a common denominator. The common denominator – which is the essence of the CQT – is the need to compare the physiological reactions between two types of questions: the Relevant and the Comparison (Control) questions. The most basic decision

[1] Avital Ginton, Ph.D is a Psychologist and a retired Commander (Colonel) of Israel Police, in which he served for 22 years (1977-1999). Dr. Ginton has been a full APA member since 2002, and the President of the Israel Polygraph Examiners Association from 1995 – 2001. He has authored several articles appearing in this and other publications.

rule in any of the CQT formats is very straightforward: if the detected reactions to the relevant questions are stronger, on average, than the reactions to the comparison questions, then the examinee is considered to be deceptive with regard to the relevant questions; when they are weaker, the examinee is deemed "non-deceptive."

The most common theoretical basis for this decision rule, until lately, was laid down some fifty years ago by Cleve Backster, who used the term "Psychological Set" to explain this differential reactivity between deceptive and non-deceptive examinees. By that, I believe, he meant: 1) The examinee concentrates on the aspects posing the greatest threat to his or her well being and automatically reacts to this danger with the emotional physiological fight or flight mechanism in an effort to protect himself, and 2) Due to a kind of differential attention process, while deceptive examinees identify the relevant questions as posing the greatest threat for them, the truth-tellers find the comparison questions to posses this quality.[2]

In recent years another approach has been introduced, namely the Differential Salience Hypothesis that puts the emphasis on cognition rather than emotion (Honts, 2004; Handler & Nelson, 2007; Senter, Weatherman, Krapohl & Horvath, 2010). According to this perspective, the physiological reactions reflect the salience value of the stimuli impinging upon the examinees, and the reason we can see differential reactivity is due to the difference between the truth-tellers and the deceptive examinees in the perceived relative salience of the two types of questions. While for the deceptive examinees the relevant questions are more salient than the comparison questions, the opposite is right for the truth-tellers. Unfortunately the presenters of the Differential Salience Hypothesis in their efforts to stay away from the "Psychological Set" term in its prevailing meaning in the field, have not yet provided a good and sufficient reasoning that can explain the cause or the origin of this differentiation in the states of mind of the deceptive versus the truth-teller examinees, that in turn results in the aforementioned differential salience of the two question types. It should be clear that the differential salience occurs in the minds of the examinees and unless explaining the reason or the dynamic of the build up of this difference between the liars and the truth-tellers states of mind, it seems that what is left is not more than the assumption that "by nature" the two types of questions possess different subjective qualities for the guilty versus the innocent examinees.

Lately, this has been addressed by Avital Ginton (2009), who has introduced a new concept into the polygraph arena, namely the Relevant Issue Gravity (RIG) strength. It is assumed that in order to perceive attentively a stimulus (a must for acknowledging its salience), one has to first be relatively free from other attention-attracting-stimuli. Whenever one's attention is focused heavily on a certain stimulus, it is very hard to distract his/her attention from it and divert it to other stimuli.

[2] "Psychological Set" with different qualifiers for prefixes, is a concept widely used in psychology between the 1950s and the 1980s, describing a psychological state of mind of having predisposition to perceive, interpret, and/or to react to stimuli in a particular way, while relatively ignoring other stimuli, interpretations, or various possible reactions. This tendency or readiness, which might be situational or context bounded, is caused by specific prior experiences, instructions or biases towards a particular interpretation of the target stimuli. (McKeachie & Doyle, 1966; Hilgard & Atkinson, 1967; Marx, 1976; Myers, 1986; Reber, 1995).

Unfortunately, the concept of "Psychological Set" has been used or understood in our field in somewhat erroneous ways, that gives the impression that "Psychological Set" is a term describing specifically the tendency of examinees to respond physiologically with a Fight, Flight or Freeze (FFF) autonomic pattern, to stimuli that pose the greatest threat to their well-being or interests at the moment. Responding to stimuli that pose a threat is not a Psychological Set. However, the reason that an examinee identifies certain stimuli as posing a threat to him and reacts accordingly, is highly influenced by his Psychological Set. Thus, the differentiation found between liars and truth-tellers in responding more to the relevant or comparison questions might be related to different Psychological Sets that they hold.

Upon arrival at the examination room to take a specific issue[3] CQT, whether guilty or innocent, the examinees' state of minds are focused on the relevant issue because they know they are about to be tested on their veracity in this regard. Any stimulus that stems from this issue is preloaded with salience, and the more the examinees' minds are preoccupied with that issue, giving questions related to it more signal value, the more difficult it is to divert their attention to other stimuli and make those stimuli (i.e. other question types) salient for them.

Several factors might contribute to the tendency of the examinees' minds to be bound to and preoccupied by the relevant issue(s), and the overall bounding force that leads to this preoccupation of the mind with the relevant issue(s) was termed by Ginton "Relevant Issue Gravity" (RIG).

According to the RIG strength hypothesis, it is hypothesized that deceptive examinees, as a whole, are more preoccupied with the relevant issue to begin with, relative to the truth-tellers and that results in relatively higher resistance to diverting their attention to the comparison questions' domain when they are presented during the pretest interview and later in the test phase. That brings about a mirror image kind of differential salience of the two question types between the deceptive and the truth-tellers. This means that while for the deceptive examinees the relevant questions are more salient than the comparison questions, it is the comparison questions that are more salient for the truth-tellers. This differential salience in turn leads to the differential emotional reactivity.

One of the main factors contributing to the RIG is the very fact that in most "classic" cases, deceptive examinees actually carry real experiences and memories of the issue probed in the relevant questions - unlike the innocent truth-tellers who have more of an abstract understanding of event with no exact memory traces. These emotional and cognitive traces of memory hold a psychological bounding force towards the relevant issue and strengthen the Relevant Issue Gravity for the deceptive examinees. The RIG strength theory suggests that the success or failure in maneuvering the focus of the examinee's attention from the relevant issues' domain to the domain of the comparison questions, which is manifested in his/her relative strength of reactions to the relevant versus the comparison questions, indicates the strength of the RIG for that particular examinee on that specific occasion. A strong RIG indicates a deceptive examinee while a weaker RIG that results in shifting the attention towards the comparison questions, indicates a truth-teller.

However, if a main factor in strengthening the RIG for the deceptive is the existence of memory traces from the relevant event, then we might also expect to detect a relatively strong RIG in truth-tellers who have actually been through that event and carry with them emotional and cognitive traces of memory of what has happened to them from their perspective. Hence, they do possess a strong bounding force that ties their attention to the relevant issue and interferes with the diversion of their attention towards the comparison questions' domain, resulting in a higher rate of false positives. The risk of having a higher rate of false positive has been acknowledged in the field for many years with regard to various kinds of victims such as victims of sexual abuse, sexual assault, or other kinds of violence, victims of fraud and so forth (Ginton, 1993; Ginton, 1997; Horvath, 1977; Raskin, 1986). However, from the RIG strength theory perspective it also applies to non-victim situations such as a case in which a person who claims to be an eyewitness to a crime, becomes a suspect of the very same crime and ought to take the polygraph examination to refute the suspicion. Similarly, but probably to a lesser degree, an examinee who is suspected of killing a person claims that when he arrived at the scene the person was already dead. Thus, there are

[3] A specific issue CQT is a test that covers one specific issue that is under investigation, aimed to diagnose whether the examinee's version about the case is a lie or is he telling the truth. The examinee knows in advance that the test is targeting that issue.

cases in which the RIG strength theory predicts that the existence of the mirror image like differential salience of the relevant versus the comparison questions between the deceptive examinees and the truth-tellers is somewhat questionable not only with regard to victims but in non-victim situations as well. How should we deal with such cases to prevent false positive outcomes?

When testing an examinee who might have been through the relevant event(s) one way or another, and probably has vivid memories, but denies the specific allegations, it is recommended to opt for a pre-test that starts by discussing the relevant event(s) but very quickly diverts from asking whether the examinee did or did not do the alleged acts, towards whether or not he/she is lying now when denying it? The relevant issue becomes not the alleged acts in the investigated event but the issue of whether or not he or she is now lying in that regard. That kind of professional recommendation has been in the field, for many years for testing alleged victims, at least by Israel Police (Ginton, 1993; Ginton, 1997). But to the best of my knowledge it has never been suggested in the professional literature with regard to other allegations.

Turning the relevant issue away from the original event while keeping the effort to detect deception about it, is expected to result in reducing the RIG strength for all examinees, but it should weaken the truth-tellers' RIG to a higher degree, improving the chances that their attention could be diverted from the relevant to the comparison questions. The deceptive examinee, when asked about lying in his/her version of the event, usually couldn't help thinking and experiencing what had actually happened because, to answer the question, he would have to process it, which we would expect would cue the original incident. For the truth-teller, on the other hand, it is easier to dissociate the relevant questions about lying from the original event's memory traces that he may carry because these traces have nothing to do with lying and because the interaction with the examiner: discussing the relevant questions (about lying), tends to avoid cueing these traces. By so doing the bounding effects of the memories from the original event on the RIG strength will still exert their influence in the deceivers' minds, interfering with diverting their attention towards the comparison questions' realm, while relatively reducing their influence on the mind of the truth-tellers, and it will be easier to divert the attention of the truth-teller examinee towards the comparison domain. It is important to say that it is not expected to totally eliminate the impact of the memory traces on the RIG of the truth-teller rather it is only expected to weaken their effect, so, whenever the traces of memory carry a very heavy load, traumatic or sensitive, this remedy won't be enough to prevent false positive outcome.

In terms of technique, the best way to follow the recommendation is to ask the examinee to write, in the presence of the examiner, a short statement in which he or she denies the allegations, and then to ask whether he or she was lying in that written statement. Unfortunately due to lack of relevant empirical research, this recommendation could not be supported by clear cut data. However, it does gain support, for what it is worth, from a lot of personal experience with both the probable false positive outcomes in such cases, if the tests are conducted in the regular direct manner and with the success of the recommended remedy to cope with such situations.

Finally, some people might think that the kind of recommendation given above contradicts the important, and in a way the "bon ton" tendency to pursue greater standardization in our field because it introduces the notion of state-dependent variations in the way the CQT examinations should be conducted. While this claim seems to be true at first glance, it is still for the benefit of our profession. To put it in a wider perspective, it is the opinion of the present author that the extreme striving for rigid standardization in the name of science is based in a way on a simplistic and limited concept of what science is. It is true that behavioral and biological sciences should deal with the central tendencies of phenomena which are formalized in general rules that concern most of the existing variance while treating the individual differences or the variation between existing situations as irrelevant noise. However, this is only the first step, and probably the easiest one, the next steps must deal with the individual and specific situational variance not as a noise, but as part of the phenomenon that needs to be systematically addressed and explained. It is therefore that nowadays in the field of medicine there is a clear trend to shift from the simple standardization of diagnoses and treatment to individualized or personalized medicine, which is leaned heavily on individual differences found between the patients in biological, psychological and environmental aspects, and applies tailor-made diagnostic yardsticks and treatments based on the specific variations found in that specific patient at the time. This medical philosophy and practice says that modern medicine should be Personalized Medicine, meaning "Different Things to Different People," as has been stated by a leading international organization of medicine the Personalized Medicine Coalition Organization, in its Mission and Principles chapter (2010).

It is the belief of the present author that we should not, in the name of science, throw away the tailor-made approach in conducting polygraph examinations that for years has characterized the work of the best polygraph examiners and shift into the standardized "scientific" mediocre kind of work. We should adopt the scientific methods not only in favor of standardizing our profession but also to improve our understanding of the "art" quality found in our work rather than suppress it in the name of science and standardization. Thus, I call to keep in mind that modern polygraphy means understanding and conducting "Different Things to Different People and Different Situations." In other words I call for developing an adaptive approach or adaptive polygraphy.

Polygraph or the Psychophysiological Detection of Deception, is a short blanket that can not cover everything without paying in errors. A clever polygraph examiner and a wise usage of polygraph must make a choice whether to cover the feet or the head with this short blanket and conduct the examination accordingly (Ginton & Ber, 1992). That seemed to be recognized lately more and more, at least with regard to the scoring (e.g. Krapohl, Stern & Bronkema, 2003), but a wiser approach should look to turn the short blanket into an elastic cover that can deal differently with different people and different situations (Ginton & Ber, 1992). That is the only way that can improve our performance beyond the glass ceiling of 90% accuracy.

This doesn't mean to abandon the attempt to formulate standard rules but rather to try to formulate second or third generation rules, which should be applied differentially in accordance with the differences between the cases, the kind of examinees and the specific situation, sometimes unique, that characterize the particular examination. The case of testing "Truth-tellers who were there," presented in the article, is but one example of this adaptive polygraphy approach.

References

Backster, C., (1963). *Backster Standardized Polygraph Notepack and Technique Guide*. New York, NY: Backster School of Lie Detection.

Ginton, A., (1993). *Usage of polygraphic detection of deception test, in verifying complaints about violence; analysis, and policy recommendations*. Unpublished manuscript. Kennedy School of Government, Harvard University.

Ginton, A., (1997). Polygraph examination for verifying the complaint of an alleged victim. (In Hebrew). *Polygraph - The Journal of Israeli Polygraph Examiners Association (IPEA)*, October, 1997.

Ginton, A., (2009). Relevant Issue Gravity (RIG) Strength – A new concept in PDD that reframes the notion of Psychological Set and the role of attention in CQT polygraph examinations. *Polygraph*, 38(3), 204-217

Ginton, A. & Ber, Y. (1992). *Polygraph Doctrine – Understanding Polygraphy - Theory and Practice.* Unpublished internal manuscript (in Hebrew). Scientific Interrogation Lab, Israel Police.

Hilgard, E. R., & Atkinson, R. C., (1967). *Introduction to Psychology*. 4th edition, Harcourt, Brace & World, Inc. (Hebrew edition, (1975) Tel-Aviv, Israel)

Handler, M., & Nelson, R., (2007). Polygraph terms for the 21st Century. *Polygraph*, 36(3), 157-162.

Honts, C. R., (2004). The psychophysiological detection of deception. In: Granhag,P., & Stromwall,, L., (Eds) *The Detection of Deception in Forensic Contexts*. (pp. 103-123) New York, NY:Cambridge University Press.

Horvath, F. S., (1977). The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology*, 62, 127-136.

Krapohl, D. J., Stern,B. A., & Bronkema, Y., (2003). Numerical evaluation and wise decisions. *Polygraph*, 32(1), 2-14.

Marx, M. H., (1976). *Introduction to Psychology*. New-York, NY : Collier Macmillan, Inc..

McKeachie, W. J. & Doyle, C. L. (1966). *Psychology*. Reading, MA : Addison-Wesley Publishing Company Inc.

Myers, D. G., (1986). *Psychology*. New-York, NY : Worth Publishers, Inc.

Personalized Medicine Coalition Organization (2010). http://www.personalizedmedicinecoalition.org /PMC Mission and Principles.

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

Reber, A. S., (1995). *The Penguin Dictionary of Psychology*. London, England: Penguin Books Ltd.

Reid, J. E. (1947). A revised questioning technique in lie detection tests. *Journal of Criminal Law and Criminology of Northwestern University*, 37(6), 542-547.

Senter, S., Weatherman, D., Krapohl, D., & Horvath, F. (2010). Psychological set or differential salience: A proposal for reconcilling theory and terminology in polygraph testing. *Polygraph*, 39(2), 109-117.

Empirical Scoring System: A Cross-cultural Replication and Extension Study of Manual Scoring and Decision Policies

Mark Handler, Raymond Nelson, Walt Goodson and Matt Hicks

"The path of least resistance and least trouble is a mental rut already made. It requires troublesome work to undertake the alteration of old beliefs." — John Dewey

Abstract

A cohort of 19 international polygraph examiner trainees at the Texas Department of Public Safety Polygraph School used the Empirical Scoring System (Blalock, Cushman & Nelson, 2009; Krapohl, 2010; Nelson, Krapohl & Handler, 2008) to evaluate 100 confirmed event-specific criminal investigation polygraph examinations. Bootstrap analytic procedures were used to calculate accuracy profiles and statistical confidence intervals for test results comparing decision rules, including; the Grand Total Rule, Two-Stage Rules (Senter, 2003; Senter & Dollins, 2002 & 2004), Spot Scoring Rules, and traditional ZCT decision rules (Department of Defense Polygraph Institute, 2006). Bootstrap analysis of the distribution of trainee scores with the Empirical Scoring System resulted in a mean accuracy rate of 90.1% (95% CI = 83.8% to 95.8%), excluding 3.3% inconclusives (95% CI = 1.0% to 7.0%). A second bootstrap analysis of decision agreement showed that these inexperienced examiners demonstrated an average rate of agreement of 85% (95% CI = 65 – 97%). Evaluation of the distribution of sub-total scores revealed that 61% (95% CI = 51% to 70%), of the sub-total scores of truthful cases produced a non-positive score (a zero or negative value). Results from this study are consistent with those from previous studies (Blalock, Cushman & Nelson, 2009; Krapohl, 2010; Nelson, Krapohl & Handler, 2008), and provide further support for the validity of the principles inherent to the ESS, including the bigger-is-better rule, three position scoring, electrodermal weighting, two-stage decision rules, and the use of optimal cut-scores. The authors recommend continued interest in and additional research on the ESS as an expedient, valid and reliable method for manually scoring PDD examination data using statistical decision theory.

Acknowledgements

We are extremely grateful to Sabino Martinez, the Policia Federal in Mexico, Eva del Carmen Acua Tabares, Andres Aguilar Leon, Ivan Israel Albarran Santacruz, Miguel Angel Barcenas Monsivais, Lourdes Castro Garcia, Diego Armando Galindo Trujillo, Jose Manel Hurtado Garduño, Beatriz Irene Lopez Hernandez, Leslie Gabriela Montoya Arreola, Ana Silvia Von Schmeling Arechiga, Patricia Morales Luna, Jorge Zepeda Gonzalez, Claudia Anel Peña Alvarado, Juan Carranza Valencia, Zaida Ruth Alonzo Magaña, Miralva Garcia Ornelas, Omar Espino Lumbreras, Karina Barrios Arzate, and Isaac Ricardo Garcia Palacios. Without the commitment of these professionals none of this work would have been accomplished.

The authors thank Mr. Donald Krapohl, Mr. John Schwartz, Mr. Don Imbordino, Mr. Toby McSwain, Ms. Pam Shaw, Mr. Rick Kurtz, Mr. Geoffrey Flohr, Sgt. Robert Gilford, Mr. Chris Fausett, Mr. Mike Gougler, Dr. Stuart Senter, Dr. George Deitchman, and Dr. Charles Honts for their thoughtful reviews and comments to earlier drafts of this paper. The authors would grant unlimited use and duplication rights to any polygraph school accredited by the American Polygraph Association or the American Association of Police Polygraphists. Questions and comments are welcome at polygraphmark@gmail.com or raymond.nelson@gmail.com.

Introduction

The Empirical Scoring System (ESS) is a manual scoring model, first described by Nelson, Handler and Krapohl (2008). The developmental intent was to anchor every procedure and assumption used in the analysis of psycho-physiological detection of deception (PDD) examination data to empirical evidence and published scientific studies. A unique aspect of the ESS is that while it makes a strict demand for scientific proof and evidence for procedures and assumptions, the operational steps are quite simple compared to other manual scoring methods. The ESS employs a pattern-recognition approach using the on-screen data, and is completed visually, without the use of printing or any mechanical or automated measurements.

Psychologically, the ESS is based on the construct of salience (Handler & Nelson, 2007), which assumes that the magnitude of physiological responses to psychological stimuli are a function of the salience of those stimuli, and are mediated by a combination of emotional, cognitive, and behaviorally conditioned factors (Khan, Nelson & Handler, 2009). Salience, and the ESS make no assumptions about which exact emotion, articulate cognitions, or finite set of behavioral events do or do not serve as a basis for response to test stimuli. Instead all responses to stimuli are regarded as inclusive of some unknown proportion of each of these dimensional aspects of psychological response potential.

Physiologically, the ESS method of Test Data Analysis (TDA) requires analysis of only the primary reaction patterns derived from numerous studies on polygraph feature extraction (Dutton, 2000; Harris, Horner & McQuarrie, 2000; Honts & Driscoll, 1987, 1988; Kircher, Kristjansson, Gardner & Webb, 2005; Kircher & Raskin, 1988; Krapohl, 2002; Krapohl & McManus, 1999; MacLaren & Krapohl, 2003; Raskin, Kircher, Honts, & Horowitz, 1988; Nelson et al., 2008). Those reactions include phasic increases in skin conductance (or decreased resistance), increased relative blood pressure, and patterns of breathing movement often associated with movement suppression. Visual recognition of breathing movement suppression is accomplished through the evaluation of: 1) a suppression of waveform amplitude of three or more respiratory cycles following stimulus onset, or 2) a slowing of breathing rate for three or more respiratory cycles from a consistent pre-stimulus level, or 3) an increase in respiratory waveform baseline following stimulus onset and containing three or more breathing cycles before return to pre-stimulus baseline. This last pattern may or may not result from breathing movement suppression but has been shown to be a valid evaluation criterion (Kircher, Kristjansson, Gardner & Webb, 2005). Breathing apnea is regarded as the ultimate form of suppression (Department of Defense Polygraph Institute, 2006), but is easily faked and therefore scored only when it occurs at the relevant question.

The ESS does not employ rigid measurement periods or scoring windows, but requires that scores be assigned to reactions that are timely with, and caused by, the test stimuli. Early onset reactions are not scored nor are those with latencies that are atypically long for the examinee. Reactions that are obviously altered by movement, deep breath, or other voluntary or involuntary artifact event are also not scored.

The ESS makes no assumptions about, and places no requirements on, the linearity, scale or parametric shape of physiological response data. Instead, the ESS is based on the simple and robust assumption that larger reactions tend to occur in response to stimuli that are more salient due to dimensional factors that may include emotion, cognition, and/or behavioral conditioning. The ESS is based on the assumption that all observations or measurements of responses to test stimuli are estimates or approximations of the actual value of the response, and include elements of both systematic variance (i.e., data indicative of response to test stimuli) and uncontrolled variance (i.e., random measurement noise due to uncontrolled physiological, psychological, environmental or statistical measurement factors). The ESS further assumes that a more robust observation or measurement of responses to test stimuli can be achieved through the aggregation of multiple observations and measurements from several presentations of the test stimuli (e.g., *measure-twice cut-once* procedures in construction and woodworking). Existing polygraph techniques have been developed around these assumptions, with several presentations of several versions of test stimulus questions which query the examinee's involvement in an allegation or issue of concern.

The ESS uses on-screen visual analysis to assign 3-position, nonparametric scores whenever a visibly perceptible difference in response magnitude is observed between pairs of relevant and comparison questions. Numerical scores are assigned for each component sensor, and a single composite score is assigned to the upper and lower pneumograph sensors as redundant measures of the same physiological response activity. The ESS does not make complex assumptions that the upper and lower pneumograph sensors somehow cancel, balance, or enhance each other. Instead, strength of reaction in pneumograph data is interpreted as a function of the frequency of occurrence of the scorable reaction patterns for the two pneumograph sensors. Several previous studies have shown electrodermal activity to be the most powerful and effective contributor to PDD examination results (Ansley & Krapohl, 2000; Capps & Ansley, 1992; Kircher & Raskin, 1988; Kircher et al., 2005; Krapohl & McManus, 1999; Nelson et al., 2008; Olsen, Harris & Chiu, 1994; Raskin et al., 1988). For this reason, all electrodermal scores are doubled before calculating the sums for sub-total and total scores. ESS then uses simple addition to achieve weighted aggregate sub-total and grand-total scores for the several presentations of the test stimuli.

Categorical decisions of truthfulness or deception are made through statistical inference, using an equivariance-Gaussian decision model described by Barland (1985). This is accomplished by subjecting the sub-total and total scores to two-stage decision rules (Senter, 2003; Senter & Dollins, 2002; 2004). ESS scores are compared to cut-scores that are selected for a desired alpha boundary which represents a stated tolerance for error and required level of statistical significance and probability of error, based on normative data from Nelson et al. (2008). Decision alpha (cut-score) for deceptive results was set at .05, meaning that a test result would be considered statistically significant when the observed p-value (probability of error) is less than or equal to this level. Decision alpha (cut-score) for non-deceptive results was set at .1, meaning that a test result would be considered statistically significant when the observed p-value (probability of error) is less than or equal to this level.

Decisions based on sub-total scores present the well-known problem of inflated alpha, and corresponding increase in the potential for false-positive (FP) or type 1 errors when basing categorical decisions on multiple statistical comparisons regarding a single allegation or incident. Therefore, a Bonferroni corrected alpha of .017 (desired alpha of .05, divided by the 3 relevant questions) was used to reduce FP errors when basing decisions on any of the sub-total scores from the 3-question zone comparison tests (ZCT). If the number of sub-totals is different, then the correction factor is adjusted accordingly (for example, the correction factor for a 2-question ZCT would be .025). Because most inconclusive results tend to be truthful (Honts & Schweinle, 2009), the alpha for truthful decision was set at .1 in an attempt to balance sensitivity and specificity. Those interested in a more restrictive alpha for truthful case resolution may review the table in Analysis 4 to appreciate the changes in the accuracy profile when the alpha is adjusted via the cut-scores from .1 to .05 for truthful cases.

There are several advantages to selecting cut-scores based on normative data and statistical p-values, including the ability to select a cut-score that provides a desired level of decision accuracy or specified tolerance for error or risk, which is lacking in most current PDD hand scoring models in use today. Another important advantage of a decision theoretic and statistical approach to the selection of the PDD cut-score is that calculations of sensitivity and specificity levels, and their corresponding error rates, will be robust against difference in base-rates, such as in field settings where it is probably impossible to calculate the actual prior probability of deception; for example, a criminal suspect or examinee subject to polygraph screening. Knowing this ahead of time gives utility to the test result. In contrast, statistical metrics based only on Bayesian statistics or simple frequency calculations from sample data will be inherently non-resistant to differences in base-rates and subject to legitimate criticism that they have poor generalizability to field situations in which the base rate of deception is unknown or expected to differ substantially from the circumstances of the research study.

Previous studies on the ESS (Blalock et al, 2009; Nelson et al, 2008) showed that inexperienced examiners, using a simplified empirically based manual scoring system of TDA were able to perform blind scoring tasks with decision accuracy, inconclusive rates, and interrater reliability that were statistically equivalent to those of experienced scorers (Krapohl & Cushman, 2006) using the prevailing and more complex 7-position TDA methods.

The present study is a cross-cultural replication of the original ESS experiment (Nelson et al, 2008), with a cohort of inexperienced polygraph examiner trainees from Mexico, who participated in training in the United States. In this study we:

1. Compared the accuracy profiles achieved by these international trainees to those achieved in previous studies (Blalock et al, 2009; Krapohl, 2010 ; Nelson et al, 2008);

2. Explored the level of interrater agreement among the participants in this study;

3. Investigated the use of two-stage decision rules (Senter, 2003; Senter & Dollins, 2002 & 2004) as compared to traditional, grand-total and sub-total (aka "spot score") rules;

4. Looked at the trade-offs of symmetric versus asymmetric alpha decision thresholds for truthful and deceptive cut-scores; and

5. Evaluated the prevalence of non-positive sub-total (spot) scores among truthful cases.

Method

Participants

A cohort of 19 police polygraph trainees in their eighth week of polygraph training at the Texas Department of Public Safety Law Enforcement Polygraph School, an American Polygraph Association accredited school, participated in the study. Participants were employees of the Policía Federal in Mexico, and were training to deploy in field environments in which PDD exams are used in the context of criminal investigations, and for the purpose of integrity and background screening of municipal police officers and police applicants. The ten female and nine male trainees all possessed a four-year college degree, equivalent with undergraduate education in the United States, in subjects including law, psychology, criminology and forensics. All were native Spanish speaking, and instruction was provided in Spanish by bilingual instructors from the United States.

Data Collection

Participants were instructed, in Spanish, in the use of the ESS, and then requested to evaluate an archival matched sample of 100 confirmed polygraph examinations that were randomly selected from the Department of Defense confirmed case archive. Fifty of the cases were conducted on examinees that were later confirmed as deceptive; the other 50 examinations were conducted on examinees that were later confirmed as non-deceptive to the investigative issue of concern. The same sample was previously used by Krapohl and Cushman (2006). All examinations were conducted using the Federal ZCT format (Department of Defense Research Staff, 2006), with three relevant questions and three test charts. Participants had received prior instruction in the current TDA procedures used by the National Center for Credibility Assessment (Department of Defense Polygraph Institute, 2006), and were asked to score the cases using the ESS, after approximately one hour of instruction on using the ESS model. Participants were asked to provide numerical scores only, and to refrain from making categorical decisions about the test results. Decision rules and cut-scores were established via normative data reported by Nelson et al (2008). Instructors who proctored the data collection phase were blind regarding the guilty status of each case.

Analysis 1 – Accuracy Profile

A Bootstrap Monte Carlo experiment was constructed to calculate the accuracy profile achieved by the study participants using statistically optimal cut scores (alpha <0.1 for truthful and <.05 for deceptive, with Bonferroni corrected sub-totals) and using two-stage decision rules.

Results: Analysis 1 – Accuracy Profile.

Table 1 shows the mean and confidence intervals for the accuracy profile developed from a bootstrap resampling experiment of 1,000 iterations of the resample space of N = 100 sets of scores from the study participants. Bootstrap mean unweighted decision accuracy was 90.1% (95% CI = 83.8% to 95.8%), excluding 3.3% inconclusives (95% CI = 1.0% to 7.0%).

Table 1. Bootstrap Mean and Confidence Intervals for the ESS Accuracy Profile

	Result	95% Confidence Range
Proportion Correct	.901	(.838 to .958)
Inconclusives	.033	(.010 to .070)
Inconclusive Deceptive	.040	(.018 to .093)
Inconclusive Truthful	.039	(.017 to .091)
Sensitivity	.865	(.762 to .955)
Specificity	.881	(.782 to .961)
False Negative Errors	.103	(.024 to .192)
False Positive Errors	.089	(.021 to .174)
Positive Predictive Value	.906	(.818 to .978)
Negative Predictive Value	.895	(.800 to .976)
Unweighted Mean Accuracy	.901	(.839 to .954)

Discussion: Analysis 1 – Accuracy Profile

Test accuracy is a complex phenomenon composed of the interaction of several factors including among other things; construct validity, decision threshold and incidence rate. For this reason, it is not realistic to expect a single numerical index to adequately represent all of the dimensional variations that encompass the accuracy profile of a test or classification method. Instead, accuracy is most accurately understood through the evaluation of the various dimensions which determine the capability of a test to contribute incremental validity to practical decision making. Evaluation of multiple dimensional characteristics of test accuracy will allow developers to adjust testing protocols to optimize their testing objectives, and allows testing professionals, program administrators, and test consumers to make more effective use of the capabilities and advantages of the results from the PDD test.

Participants in this study produced results that were statistically equivalent to those achieved by previous studies on the ESS. Sensitivity and specificity rates were relatively balanced, as were inconclusive rates for truthful and deceptive examinations.

False positive and false negative errors were also found to be closely balanced in this experiment. Inconclusive rates observed during this experiment require additional explanation. Because of the randomization inherent to bootstrapping and Monte Carlo experiments, it is possible that some bootstrap or Monte Carlo distributions will result in zero inconclusives for some distribution. Inconclusive rates were calculated both within and between the truthful and deceptive groups. It is possible, under some randomized iterations, that there are zero inconclusives in one of the groups and not the other. When this occurs under exhaustive repetitions, the resulting between-group zero-inconclusive rate will be lower than the unweighted average of the within-group mean inconclusive rate, and will be more generalizable to field settings than the average of within-group inconclusive rates.

Analysis 2 – Interrater Reliability

We calculated the Fleiss Kappa statistic as a measurement of interrater agreement among the participants in the study. A two-dimensional double-bootstrap was calculated, for which both cases and scorers were selected randomly to construct 100 x 100 resampled sets of the participant scores (N = 100). Statistical confidence intervals were then constructed from the bootstrap distribution of scores.

To further illustrate the profile of interrater agreement achieved by the 19 study participants, we calculated the bootstrap distribution, including mean and 95% confidence range, for the proportion of agreement between decisions made by the study participants, using 1,000 iterations of the bootstrap resample space of 19 x 100 decisions.

Results: Analysis 2 – Interrater Reliability.

A moderate to substantial level of scoring agreement was achieved by the study participants, with $k = 0.59$ (95% CI = .52 to .65). However, the proportion of decision agreement observed among the participants was .84 (95% CI = .73 to .95). A bootstrap of the Pearson correlation coefficient among numerical scores was .84 (95% CI =.71 to .96), which was statistically significantly better than chance ($p < .01$).

Discussion: Analysis 2 – Interrater Reliability.

Interrater agreement among the inexperienced participants in this study was moderate to high, and was not statistically different from those observed in previous studies on the ESS (Blalock et al, 2009; Krapohl, 2010; Nelson et al, 2008). The ESS, using on-line evaluation, without mechanical measurements, outperformed previous reports of interrater agreement for experienced examiners (Blackwell, 1999) by a non-significant margin. These results were consistent with previous studies on the ESS (Krapohl, 2010; Nelson et al., 2008).

Analysis 3 – Decision Rules

To further investigate the influence of decision rules on ESS accuracy, additional analyses were conducted using 1,000 iterations of a bootstrap Monte Carlo model that was seeded with the scores from the study participants. Using statistically optimal thresholds (alpha < .1 for truthful and < .05 for deceptive, including Bonferroni correction to alpha for decisions based on sub-total scores), means and statistical confidence intervals were calculated for the accuracy profiles of ESS scores that were interpreted using different decision rules, including: the Grand Total Rule (GTR), Spot Scoring Rules (SSR), and traditional ZCT rules (TZR) (which involve the simultaneous use of the Grand Total and sub-total scores). Those results were then compared to ESS results using Two-Stage Rules (TSR).

Results: Analysis 3 – Decision Rules.

Table 2 shows the mean and confidence intervals for the different decision rules. The GTR produced the highest level of decision accuracy, however, differences in decision accuracy compared to the other rules was not significant. The GTR resulted in a significant increase in inconclusive results ($p = .03$) compared to the TSR. This difference loaded primarily on deceptive cases, but the overall change in inconclusives within the deceptive cases was not significant, nor was the corresponding reduction in test sensitivity to deception.

Table 2. Mean and confidence intervals for ESS2, TZR (Federal), GTR, & SSR

	ESS 2-stage	TZR	GTR	SSR
Proportion Correct	.901 {.837 to .958}	.870 (.17) {.789 to .942}	.914 (.36) {.853 to .968}	.875 (.22) {.792 to .946}
Inconclusives	.033 {.010 to .070}	.256 (<.01)** {.170 to .340}	.071 (.03)* {.030 to .130}	.285 (<.01)** {.200 to .370}
Inconclusive Deceptive	.040 {.018 to .093}	.091 (.05)* {.021 to .179}	.082 (.08) {.020 to .167}	.134 (<.01)** {.048 to .236}
Inconclusive Truthful	.039 {.017 to .091}	.426 (<.01)** {.259 to .625}	.065 (.16) {.018 to .140}	.441 (<.01)** {.275 to .628}
Sensitivity	.865 {.762 to .955}	.897 (.480) {.804 to .976}	.817 (.46) {.704 to .917}	.854 (.49) {.746 to .942}
Specificity	.881 {.782 to .961}	.398 (<.01)** {.262 to .536}	.880 (.48) {.783 to .961}	.397 (<.01)** {.260 to .539}
False Negative Errors	.103 {.024 to .192}	.027 (<.01)** {.017 to .063}	.103 (.49) {.023 to .196}	.027 (<.01)** {.017 to .060}
False Positive Errors	.089 {.021 to .174}	.181 (.03)* {.080 to .292}	.061 (.22) {.018 to .132}	.166 (.05)* {.067 to .260}
Positive Predictive Value	.906 {.818 to .978}	.832 (.05)* {.726 to .924}	.931 (.28) {.848 to .891}	.837 (.07) {.727 to .934}
Negative Predictive Value	.895 {.800 to .976}	.935 (.13) {.851 to .965}	.896 (.49) {.810 to .975}	.935 (.14) {.854 to .966}
Unweighted Mean Accuracy	.901 {.839 to .954}	.883 (.25) {.819 to .935}	.913 (.36) {.852 to .963}	.885 (.29) {.815 to .941}
*p < .05 **p < .01				

Compared to the ESS with TSR, the TZR, which include the simultaneous use of grand-total and spot-scoring rules, and require a positive score for each sub-total, resulted in statistically significant differences among several accuracy dimensions, including: increased inconclusives (*p* < .01) for both deceptive (*p* = .05) and truthful cases (*p* < .01), decreased specificity with truthful cases (*p* < .01) and increased false-positive errors (*p* = .03). Also, a statistically significant decrease was observed in positive-predictive-value (*p* = .05) when using the TZR. While most of the changes in accuracy resulting from the TZR were undesirable, one desirable change was observed in the form of a decrease in false-negative errors (*p* < .01).

The observed effect size for NPV for the (.935 - .895 = .040) was approaching, but did not achieve, statistical significance at the .05 level for the TZR. A post-hoc power analysis showed the power of the dimensional comparison to be (β = .813). A minimum statistical effect of .059 would be significant at the .05 level. A similar post-hoc power analysis on the percent correct achieved by the TSR and TZR indicated that the observed effect of .301 was achieved with (β = .859), while a minimum effect of .058, for decision accuracy, could achieve statistical significance at the .05 level. Post-hoc analysis of the effect size for unweighted accuracy (.018) revealed the power of the present experiment to be (β = .907), while a minimum effect size of .049 would be significant at the .05 level.

Spot Scoring Rules (SSR), using statistically optimal alpha cut-scores that were corrected for multiple within-test comparisons of deceptive and truthful scores, produced decreases in decision accuracy that were similar to the TZR and not significantly different from the other scoring conditions. The SSR resulted in statistically significant increases in inconclusives (*p* < .01) for both deceptive (*p* < .01) and truthful cases (*p* < .01), along with decreased specificity with truthful cases (*p* < .01) and increased false-positive errors (*p* = .05). The overall change in positive-predictive-value (PPV) was not significant (*p* = .07) at the .05 level, but was approaching statistical significance. A post-hoc power analysis indicates the power of the experimental dimension to be (β = .690). Like the TZR, the SSR did result in one desirable change, a decrease in false-negative errors (*p* < .01), likely a result of the requirement for all positive subtotals.

Discussion: Analysis 3 – Decision Rules.

Unweighted decision accuracy rates did not differ significantly among the four scoring conditions, and none of the scoring conditions produced a statistically significant difference in terms of test sensitivity to deception. The TSR produced a significant decrease in inconclusives compared to the GTR and TZR, along with significantly fewer FP errors and significantly greater PPV. The TZR produced significantly fewer FN errors than the TSR and GTR, at the cost of statistically significant increases in FP errors, and a very large significant effect for increase inconclusive results among truthful cases.

Different decision rules offer different advantages, constrain inconclusives, minimize certain types of error, or optimize specific dimensions of decision accuracy. Most of the differences in inconclusives appear to be due to the requirement for positive scores at all sub-totals in order to achieve a truthful result with the TZR. A significant increase in inconclusive results for deceptive cases for the TZR may be interpreted as a desirable change, because this dimensional change is related to the decrease in FN errors. A reduction in false negatives may be attainable with the TSR through the selection of a more conservative alpha decision threshold for truthful cases, though this is likely to result in an increase in inconclusives. This should be the focus of some future research.

Operationally, the difference between the TSR and TZR is that the TSR prioritizes the grand-total score first, regardless of the sub-total scores, and then only if inconclusive, proceeds to make deceptive classifications based sub-total scores. The TSR can be considered to emphasize balanced test sensitivity and test specificity, by making sequential use of the GTR and SSR, while the TZR prioritizes test sensitivity over test specificity, and amounts to the simultaneous use of the GTR and SSR. The TZR will permit a deceptive sub-total score to “trump” a truthful grand-total score, while the TSR regards the grand-total as more important than the sub-totals and will not allow a sub-total to “trump” the grand-total. The TZR does not, however produce any increase in test sensitivity compared to the TSR, and the observed effect was limited to the reduction of FN errors at a cost of a loss of specificity and increased inconclusives. As always, practical decisions such as this are a matter of policies and operational priorities, just as much as they are a matter of science and decision theory.

These results show that the TZR is not more effective at catching liars than other decision rules. These results were obtained while using statistically optimal cutscores for all scoring conditions, so that any observed effect is not due to differences in decision cutscores and can be attributed to the decision rules. Readers should note that most, if not all, of the presently available and widely used scoring methods lack normative data and lack the ability to make inferential calculations of the probability of a test error. Field examiners, quality assurance reviewers, and program managers should be cautioned that using the ESS cutscores with other scoring methods is not recommended.

Based on these data, the TSR appears to be the optimal solution, with decreased inconclusives compared to the GTR. Use of the TZR should be restricted to circumstances that warrant a need for reduced false negatives, with a risk of a corresponding significant increase in inconclusives and a decrease in test specificity and positive predictive value and increased false-positive results. There appears to be no advantages to the use of the SSR with the ESS. Also, the SSR data reported here were calculated accounting for the deflation of alpha occurring with multiple within-test comparisons and optimal alpha cutscores. These precautions are not typically done in field settings and we predict uncorrected field-practice results will not improve the balance of test results. This too should be explored in future research.

Analysis 4 – Alpha Cut-scores

Using the ESS rules, we varied the decision alpha thresholds (cut-scores) to the effect of using a more conservative alpha for truthful cases. This should be of interest to those examiners who are concerned with risk-aversion and interested in a lower rate of false-negative (FN) results. As is common in many forms of testing, efforts to reduce errors may result in an increase in inconclusive results. We show the changes in the accuracy profiles for alpha held at < .05 for deceptive and varying alpha for the truthful from < .1 to < .05 in Table 3.

Results: Analysis 4 – Alpha Cut-scores.

FN error rates were reduced from the expected overall rate of ~.10 to ~.05 when we changed the alpha cut-score from .1 to .05. The difference in the rate of inconclusive results was significant (p < .01) and this difference was loaded on truthful cases, for which the difference was also significant (p < .05). Loss of test specificity within the truthful cases was statistically significant (p < .01) and Table 3 shows the results.

Discussion: Analysis 4 – Alpha Cut-scores.

This analysis compares decision thresholds in an effort to demonstrate the trade-offs encountered when a more stringent alpha is observed for the truthful cases, (equivalent to requiring a higher positive score to achieve a “No Significant Response” result). As can be seen, proportion correct, deceptive inconclusives, sensitivity, false positive and false negative results, positive and negative predictive value and unweighted accuracies do not differ significantly. However, imposing the more stringent threshold, results in increased inconclusive results for overall cases and especially for the truthful cases and a decrease in specificity. While the selection of alpha decision cut-scores is ultimately a matter of administrative policy as much as it is a matter of science, these results indicate that the current balanced approach of observing alpha at <.05 for deceptive cases and <.1 for truthful cases maintains a relatively high level of sensitivity and specificity, while holding the inconclusive rate low and constraining errors to tolerable proportions.

Analysis 5 – Proportion of Non-positive Sub-total Scores

To further evaluate the assumptions of the TZR, which require a positive sub-total score (spot scores) for all investigation target questions, bootstrap analytic procedures were used to calculate frequency, proportion and confidence intervals for the presence of non-positive sub-totals (i.e., sub-totals that are zero or negative scores) among the confirmed truthful cases.

Table 3. Results of varying the truthful decision threshold (alpha) with ESS rules

	ESS rules *truthful alpha < .1* deceptive alpha < .05	ESS rules *truthful alpha < .05* deceptive alpha < .05	Sig.
Proportion Correct	.901 (.837 to .958)	.904 (.839 to .957)	.488
Inconclusives	.033 (.010 to .070)	.095 (.040 to .160)	.005**
Inconclusive Deceptive	.040 (.018 to .093)	.072 (.019 to .149)	.141
Inconclusive Truthful	.039 (.017 to .091)	.121 (.038 to .229)	.016*
Sensitivity	.865 (.762 to .955)	.888 (.791 to .963)	.331
Specificity	.881 (.782 to .961)	.747 (.627 to .867)	.007**
False Negative Errors	.103 (.024 to .192)	.048 (.018 to .104)	.072
False Positive Errors	.089 (.021 to .174)	.131 (.041 to .234)	.166
Positive Predictive Value	.906 (.818 to .978)	.871 (.762 to .961)	.207
Negative Predictive Value	.895 (.800 to .976)	.940 (.870 to .978)	.134
Unweighted Mean Accuracy	.901 (.839 to .954)	.905 (.841 to .959)	.467
*p < .05 **p < .01			

Table 4. Frequency of non-positive sub-totals.

Questions	Proportion	95% CI
R5	13%	(71% to 91%)
R7	29%	(21% to 39%)
R10	37%	(27% to 46%)
Any RQ	61%	(51% to 70%)

Results: Analysis 5 – Proportion of Non-position Sub-total Scores.

Bootstrap analysis revealed that 61% (95% CI = 51% to 70%) of the truthful cases can be expected to result in at least one or more sub-total scores that are non-positive (i.e., zero [0] or negative scores).

Discussion: Analysis 5 – Proportion of Non-position Sub-total Scores.

Results of this experiment suggest that a large proportion of truthful persons will produce at least one non-positive sub-total score. This requirement results in a condition in which more than one half of truthful cases are regarded as incapable of being correctly classified, and the value of this rule (requirement for positive scores in all sub-totals) is therefore questionable. Some may assume this rule increases decision accuracy with deceptive cases, in terms of increased sensitivity to deception or decreased false negative errors. While it would be procedurally and mathematically impossible for this requirement to produce an increase in test sensitivity, this procedural requirement does result in a statistically significant reduction in false negative results, at the cost of a statistically significant increase in inconclusive results among truthful persons. A more practical, and precise, solution to the need for low false-negative error rates might be achieved through the selection of an alpha decision cut-score that assures the required level of precision with greater ability to constrain error rates. This should become the focus of a future study.

General Discussion

The trainees from the Mexico Federal Police demonstrated that ESS can produce balanced sensitivity and specificity, with no significant differences from results achieved during previous studies on the ESS (Blalock, Cushman & Nelson, 2009; Blalock, Nelson, Cushman, & Oelrich, 2010; Krapohl, 2010; Nelson et al., 2008). The inexperienced examiners (trainees) in this study scored polygraph charts at accuracy and reliability rates consistent with those of the experienced examiners reported by Krapohl and Cushman (2006) which should be of interest to trainers, field examiners and program managers. It seems reasonable to assume that field experience is valuable and contributes to increased skill and performance in test data analysis. Therefore, the performance of the inexperienced scorers might be attributable to an improved emphasis on empirically sound principles in their scoring method. Additionally, historical scoring exercises may have involved evaluators with considerable experience and expertise. Using these experts to test a scoring model is less likely to generalize to what will happen in the field. It is perhaps more informative to test the "weakest link in the chain" to estimate how well a model will work for the many, as opposed to the few. A final consideration is this system was taught to the students via a translator suggesting this simplified system is easy to communicate across language barriers.

Grand total decision rules were the simplest solution and provided the highest level of decision accuracy, though the difference was not significant. Two stage rules outperformed the grand total decision rules in terms of inconclusive results, and sensitivity to deception, with no significant difference in false-positive or false-negative errors. Traditional decision rules produced a significant increase in the rate of false - positive errors and inconclusive results for the truthful sample cases – a result that is consistent with previously published studies (Krapohl & Cushman, 2006).

Two-stage decision rules seem to provide more balanced test sensitivity and test specificity than traditional rules. While the traditional rules have served the profession well through the years, they may be sub-optimal. In addition it is becoming increasingly clear that traditional decision cutscores have not been studied in the context of normative data or correspondence with decision alpha levels. On the surface the traditional rules seem to benefit a "risk-aversive" testing program. Examiners with an inherent fear of a false-negative error may become convinced they would rather "interrogate and apologize" than allow themselves to be beaten. A closer consideration of this attitude reveals that in the end it may actually be detrimental.

Polygraph examiners and consumers of polygraph rely on the test's ability to differentiate the truthful from the deceptive. A test with high sensitivity but poor specificity will have a low positive predictive value because of the high false-positive rate. The mathematical reality of this is that a lot of truthful subjects are classified as deceptive, and logic dictates, will be interrogated. Interrogating a truthful subject offers the opportunity for a number of negative outcomes, not the least of which could be a false confession. Also, field examiners traditionally pride themselves on their ability to secure a posttest admission, and examiners, their peers and their supervisors use this as a metric of success. Indeed, a number of organizations keep statistics on confession rates! Being unable to separate the truthful from the deceptive because of a test that is heavy on sensitivity and light on specificity is a set up for disappointment for the examiner, his or her supervisor, or consumer. While it may seem initially convenient to ensure test sensitivity at the cost of imbalance specificity, the long term effect will be corrosive of confidence among consumers. These consumers of polygraph rely on the diagnostic value of the test result to add incremental validity to the process in which the polygraph has been applied. Without diagnostic value, polygraph will be of no more value than computer voice stress analyzers.

The ESS model applies the principle of weighting the contribution of the EDA component most heavily, employs empirically supported two-stage decision policies, and uses statistically optimal thresholds (cut scores) that allow for error estimations and the dispensing of “lore-based” decision rules. The ESS is straightforward to use and easy to explain to polygraph examiners and non-examiners such as department administrators or adjudicators. ESS offers promising potential for gaining increased understanding and increased credibility among consumers of polygraph test results, with good consistently high criterion validity and interrater reliability that is as good or better than other scoring models.

A major advantage of the ESS, compared to current hand-scoring systems, is the existence of normative data that can be used to provide an understanding of the level of statistical significance achieved by various decision cut-scores (see Appendix A). In an era that emphasizes theoretically sound decision models, mathematically defensible results, and known methods for calculating the likelihood of an erroneous test result, all investigators involved in development and research of polygraph scoring systems should feel an obligation to publish normative data and significance tables for all polygraph scoring systems in present use.

No study is without limitations, and we note several limitations in this study. First, the number of evaluators contributing to this study is small and while, the sample itself is not small, it is also not large. This study addresses only Zone Comparison Technique polygraph results and does not attempt to address data collected using other polygraph techniques. We also realize that balance of sensitivity and specificity may not be the goal in all testing situations and there may be times when more strict or lenient tolerance exists for one type of error over another. This point is precisely why we advocate moving the profession towards results based on p-values, normative data, and the ability to compare a calculated probability of error to a stated tolerance for error. Polygraph professionals should strive to study and understand the normative data to the point where we may make reasonable estimates of our errors on individual cases. In this way we can more precisely predict the scientific strength of confidence in the results. We suggest others replicate this experiment to support or refute our findings in hopes that we can collectively improve the quality of polygraph for all.

As the late great social psychologist Leon Festinger (1987) stated in his remarks during the symposium Reflections on Cognitive Dissonance: 30 Years Later at the 95th Annual Convention of the American Psychology Association:

> No theory is going to be inviolate. Let me put it clearly. The only kind of theory that can be proposed and ever will be proposed that absolutely will remain inviolate for decades, certainly centuries, is a theory that is not testable. If a theory is at all testable, it will not remain unchanged. It has to change. All theories are wrong. One does not ask about theories, can I show that they are wrong or can I show that they are right, but rather

> one asks, how much of the empirical realm can it handle and how must it be modified and changed as it matures?

One thing is for sure: if we presently consider the polygraph to be either perfect or just as good as it need be, then there is no reason to study data or pursue improved methods of hand-scoring. If, however, we think of the polygraph as imperfect and capable of being improved upon, we must rise to the challenge of studying our theories and assumptions and be willing to release our grasp of any procedures which are arcane and suboptimal. Holding on for the sake of posterity, in the face of evidence and data that informs of ways to improve the accuracy profile of the polygraph examination will not only hold the profession back, it will be considered irresponsible and unethical by others with whom we share the social sciences. The authors recommend continued interest in, and additional research on, the ESS as an expedient, valid and reliable evidence based method for manually scoring PDD examination data using statistical decision theory.

References

Ansley, N. & Krapohl, D.J. (2000). The Frequency of appearance of evaluative criteria in field polygraph charts. *Polygraph*, 29 (2), 169-176.

Barland, G.H. (1985). A method for estimating the accuracy of individual control question tests. *Proceedings of Identa-85*, 142-147.

Blackwell, N.J. (1999). *PolyScore 3.3 and psychophysiological detection of deception examiner rates of accuracy when scoring examination from actual criminal investigations*. Department of Defense Polygraph Institute Report DoDPI97-R-006. Ft. McClellan, AL. Available at the Defense Technical Information Center. DTIC AD Number A355504/PAA. Reprinted in Polygraph, 28, (2) 149-175.

Blalock, B., Cushman, B. & Nelson, R. (2009). A replication and validation study on an empirically based manual scoring system. *Polygraph*, 38(4), 281-288.

Blalock, B., Nelson, R., Cushman, B. & Oelrich, M. (submitted 2010). Reliability of the Empirical Scoring System with Expert Examiners. A manuscript submitted to *Polygraph*, for consideration.

Capps, M. H. & Ansley, N. (1992). Analysis of private industry polygraph charts by spot and chart control. *Polygraph*, 21, 132-142.

Department of Defense Polygraph Institute (2006). Test Data Analysis: DoDPI numerical evaluation scoring system. Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007.

Department of Defense Research Staff (2006). Federal Psychophysiological Detection of Deception Examiner Handbook. available online: Retrieved 1-10-2007 from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf.

Dutton, D. (2000). Guide for performing the Objective Scoring System. *Polygraph*, 29, 177-184.

Festinger, L. (1987). Appendix B: Reflections on cognitive dissonance: 30 years later. In E. Harmon-Jones & J. Mills (Eds.), *Cognitive Dissonance-Progress on a Pivotal Theory in Social Psychology* (1999). Washington, DC: American Psychological Association.

Handler, M. & Nelson, R. (2007). Polygraph terms for the 21st century. *Polygraph*, 36, 157-164.

Harris, J., Horner, A. & McQuarrie, D. (2000). *An evaluation of the criteria taught by the Department of Defense Polygraph Institute for interpreting polygraph examinations*. Johns Hopkins University, Applied Physics Laboratory. SSD-POR-POR-00-7272

Honts, C. R. & Driscoll, L. N. (1987). An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. *Polygraph*, 16, 241-257.

Honts, C. R. & Driscoll, L. N. (1988). A field validity study of rank order scoring system (ROSS) in multiple issue control question tests. *Polygraph*, 17, p. 1-15.

Honts, C. R., & Schweinle, W. (2009). Information gain of psychophysiological detection of deception in forensic and screening settings. Manuscript accepted for publication pending revision, *Applied Psychophysiology and Biofeedback*.

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

Kircher, J. C., Kristjansson, S. D., Gardner, M. K. & Webb, A. (2005). *Human and computer decision-making in the psychophysiological detection of deception*. University of Utah. Final Report

Khan, J., Nelson, R., & Handler, M. (2009). An exploration of emotion and cognition during polygraph testing. *Polygraph*, 38 (3), 184-197.

Krapohl, D. (2002). The polygraph in personnel screening. In M. Kleiner (Ed.), *Handbook of Polygraph Testing* (2002). San Diego, CA: Academic Press.

Krapohl, D. & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph*, 28, 209-222.

Krapohl, D.J. (2010). Short Report: A test of the ESS with two-question field cases. *Polygraph*, 39(2), 124-126.

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

MacLaren, V. & Krapohl, D. (2003). Objective assessment of comparison question polygraphy. *Polygraph*, 32, 107-126.

Nelson, R., Krapohl, D. J. & 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.

Olsen, D. E., Harris, J. C. & Chiu, W. W. (1994). The development of a physiological detection of deception scoring algorithm. *Psychophysiology*, 31, p. S11.

Raskin, D. C., Kircher, J. C., Honts, C. R. & Horowitz, S. W. (1988). *A study of the validity of polygraph examinations in criminal investigations*. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040. Final Report, National Institute of Justice, Grant No. 85-IJ-CX-0040

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

Senter, S. M. & Dollins, A. B. (2002). *New decision rule development: Exploration of a two-stage approach*. Department of Defense Polygraph Institute Research Division, Fort Jackson, SC.

Senter, S. & Dollins, A. (2004). Comparison of question series and decision rules: A replication. *Polygraph*, 33, 223-233.

Appendix A

ESS – Monte Carlo Normative Data for Event-Specific ZCT Exams with 3 RQs
Mean deceptive score = -9.14 (SD = 8.74) Mean truthful score = 8.35 (SD = 7.89)

Truthful (NSR) Cut-scores	
Total NSR Cut-score	p-value (alpha)
-1	0.159
0	0.130
1	0.106
2	***0.085***
3	0.067
4	0.052
5	***0.040***
6	0.030
7	0.023
8	0.017
9	0.012
10	***0.008***
11	0.006
12	0.004
13	0.003
14	0.002
15	***0.001***
Deceptive (SR) Cut-scores	
Total SR Cut-score	p-value (alpha)
1	0.159
0	0.127
-1	***0.099***
-2	0.077
-3	0.058
-4	***0.043***
-5	***0.032***
-6	0.023
-7	***0.016***
-8	0.011
-9	***0.008***
-10	0.005
-11	0.003
-12	***0.002***
-13	***0.001***

The Effects of Augmented Physiological Feedback on Detection of Deception

Robert M. Stern[1] and John C. Kircher[2]

Abstract

The purpose of this study was to assess the effects of two types of augmented physiological feedback (APF) on the reliability and accuracy of probable-lie comparison question tests (CQT). Two hundred and ten college students participated in the study, half of whom were guilty of a mock crime and half innocent. During questioning, one group received skin conductance feedback, a second group received composite feedback (skin conductance, cardiograph, and respiration), and a third group received no feedback. The results indicated that APF did not increase detection rates above that of the no-feedback group in this study. However, APF did decrease the rate of habituation during repetition of the question sequences thus allowing for greater discrimination between innocent and guilty participants as the CQT progressed.

The purpose of this study was to assess the relative effects of two types of physiological feedback on the reliability and accuracy of probable-lie comparison question tests (CQT). The CQT is the most widely used method of psychophysiological detection of deception by field polygraph examiners (Ben-Shakhar, 1991). Decisions about a suspect's involvement in criminal activities are based upon within-participant comparisons of physiological reactions to questions relevant to the criminal investigation (e.g., theft of a determined amount of money) and to probable-lie comparison questions. Comparison questions address a general content area that is related to, but excludes the specific criminal activity in question (Reid & Inbau, 1977). For example, if the criminal activity under investigation pertained to the theft of money, a comparison question might be, "Before the age of 21, did you ever take something that didn't belong to you?" Comparison questions are intentionally vague and are nearly impossible to answer truthfully with an unqualified "No." During a pretest interview, suspects are embarrassed or intimidated into answering "No." If an affirmative response is given to a probable-lie question, the question is reworded so that suspect will ultimately answer in the negative, which is probably a lie.

The CQT assumes that the suspect's degree of involvement with each type of question and the relative amount of concern that each type of question would evoke is diagnostic (Stern, Breen, Watanabe, & Perry, 1981). For guilty suspects, polygraph procedures are designed to reinforce their concern that deceptive answers (i.e., a "No" response) to crime relevant questions will be detected. Even though guilty suspects would answer "No" to probable-lie comparison questions, the crime-relevant questions are expected to cause more concern since relevant questions deal specifically with the matter under investigation. For innocent suspects, only the probable-lie comparison questions are answered deceptively. Because innocent suspects had no involvement in the criminal activity in question, the probable-lie questions are expected to elicit greater concern about being (in their opinion, falsely) detected than the crime-relevant questions. If these goals are

[1]The Pennsylvania State University

[2]University of Utah

This study was conducted for the DoD Polygraph Institute, and the report was submitted March 26, 2002. DTIC No. ADA400823. ONR Grant N00014-99-1-1083.

achieved, guilty suspects should show stronger physiological responses to the relevant questions than to the probable-lie questions, whereas innocent suspects should show stronger reactions to the probable-lie questions.

Although these predictions have been confirmed in laboratory and field settings (Office of Technology Assessment, 1983; Raskin, Honts, & Kircher, 1997), it is equally important to understand the psycho-physiological processes that underlie these findings and to devise techniques that would exploit these processes in order to increase detection accuracy rates.

Kircher (1981) offered a theoretical framework to explain the differential reactivity to probable-lie and relevant questions seen in CQTs. He suggested that when a suspect intends to answer a question deceptively during a polygraph test, the presentation of the question signals the occurrence of an involuntary physiological reaction. The participants' expectation that their bodies will reveal deception with large physiological reactions is established and reinforced during the pretest phase of the polygraph examination. According to this view, the participant's expectation that a large involuntary reaction will accompany deception is an essential component of a valid polygraph outcome.

If participants expect large autonomic reactions when they lie, it is reasonable to assume that when they lie, they will attempt to monitor and suppress these internal changes. Borrowing from Kahneman's (1973) theory of attention and effort, Kircher (1981) hypothesized that mental effort is required to suppress and monitor the leakage of incriminating information. That is, the participant must mobilize and expend energy to perform these cognitive activities. According to this hypothesis, the physiological changes recorded by the polygraph are indicators of the amount of mental effort or attention required by participants to monitor and suppress their autonomic responses to test questions. This hypothesis predicts that the perception of increased autonomic reactions will create a positive feedback loop that requires additional mental effort and prolongs the participant's cognitive appraisal of yet another threatening event. The perception of physiological arousal that occurs after the presentation of the test question may explain observed increases in the duration of physiological responses associated with deception (e.g., Kircher & Raskin, 1988; Raskin et al., 1988).

The theory also predicts that the proposed use of augmented physiological feedback (APF) will increase the differences between physiological reactions to comparison and relevant questions and thereby improve discrimination between truthful and deceptive participants. When the participant is deceptive, feedback that a strong autonomic response has occurred would be viewed as an aversive event. Like the test question that initiated the response, the feedback would threaten the participant. If the feedback is public, such that the participant knows the polygraph examiner is also hearing it, it should increase the threat to the participant, or in other words, it should signal to the participant that he/she is revealing him/herself.

Previous evidence suggests that guilty and innocent suspects respond differentially to probable-lie and crime-relevant questions (e.g., OTA, 1983). Guilty suspects react more strongly to crime-relevant questions than to comparison questions, whereas innocent suspects react more strongly to comparison questions than to crime-relevant questions. It is hypothesized that if APF increases the amount of involvement or attention allotted to questions that already pose the greatest threat to the suspect, then guilty suspects should appear more deceptive on their polygraph tests by showing an even greater response to crime-relevant questions, whereas innocent suspects should appear more truthful by showing a greater response to probable-lie questions versus crime-relevant questions. Hence, it should be easier to distinguish between truthful and deceptive suspects, thereby increasing the accuracy of the CQT.

Using the Guilty Knowledge Test, Stern, Breen, Watanabe, and Perry (1981) tested for the hypothesized beneficial effects of APF on detection rates. Participants in the APF condition received an auditory signal that varied with changes in their heart rate or skin resistance (SR), whereas control participants received no feedback. All participants were given two GKT polygraph tests: the first test was based on a geometric figure chosen by the participant from a list of five (low personal relevance test), and the second test concerned the participant's Social Security Number (SS#) (high personal relevance) that was embedded among a list of four other SS#s. Participants answered "No" to each of the four alternatives for both tests. Stern et al. found that discrimination between critical and noncritical items, based on participants' SR responses, was statistically significantly greater for the SR feedback group than the no feedback group. An effect for personal relevance was also found, such that accuracy for tests concerning SS# was statistically significantly greater than the accuracy for tests about the geometric figures.

The second experiment reported by Stern et al. (1981) assessed the effects of APF on innocent and guilty participants involved in a simulated murder plot. Participants in the guilty condition studied a document that contained several details about their role in the murder plot. Innocent participants studied a document that contained the same details, but the details were totally unrelated to any murderous activity. Half of each group was assigned to a SR Feedback condition, and the remaining participants served as a No-Feedback control. Although no statistically significant effect of feedback was found, participant mean SR response to critical items was greater in the feedback condition than the no-feedback condition for both guilty and innocent participants. The lack of statistical significance is probably due to a ceiling effect seen in the No-Feedback Group, such that detection rates in this control condition were high enough that any added benefit of an experimental procedure would be very difficult to detect without a very large sample size.

The results of the Stern et al. (1981) experiments are consistent with the prediction that feedback will enhance physiological responses to items of greater relative importance to the suspect. If this hypothesis is correct, then APF should differentially affect the responses of guilty and innocent suspects to relevant and probable-lie questions, respectively, using the CQT. The present experiment was designed to test that prediction.

Another investigation of the effects of auditory biofeedback on the Guilty Knowledge Test was conducted by Timm (1987). He found that electrodermal feedback statistically significantly enhanced detection efficiency associated with respiration amplitude changes, but that skin conductance detection efficiency was not statistically significantly affected. These results were similar to the results found in the Stern mock murder experiment. The null results found in the studies by Stern, et al. and Timm may be due to a ceiling effect for the No Feedback condition, as they suggest; however, the null results might also be due to low power, as the Stern et al. study employed only 52 participants, and the Timm study employed 68 participants. In the present study, the sample size was increased to 210 participants, which provided an 80% probability of detecting moderate (i.e., .4 - .6) differences between feedback conditions.

In addition to the techniques employed by Stern et al. (1981) and Timm (1987), this study explored alternative methods of providing feedback to participants, as well as alternative analysis procedures used to identify and classify innocent or guilty suspects. Specifically, the study also addressed the question of whether or not feedback based on electrodermal activity alone results in a more reliable index of guilt than a composite of several physiological measures. Stern et al. (1981) had greater success with electrodermal feedback than heart rate feedback. However, with the current state of computer technology, a composite index of arousal based on electrodermal, cardiovascular, and respiration measures may be generated and presented to the participant in real-time. Since some examinees show little or no electrodermal activity or their electrodermal responses quickly habituate (defined as a decrease in amplitude as a result of repeated exposure to the polygraph questions), the use of a composite index should allow investigators to provide those individuals with variable feedback even in the absence of changes in electrodermal activity.

Stern et al. (1981) classified participants as truthful or deceptive and assessed the number of correct hits and correct rejections. In addition to reporting decision accuracy, the present study tested for effects of guilt and APF on discrete measures of electrodermal, cardiovascular, and respiratory activity.

Method

Participants

Two hundred-ten college students (males = 71; age range 18-60) from the Pennsylvania State University volunteered to participate in this study. Participants were in good health, free of psychotropic medication and had not previously taken a polygraph test. Participants received extra credit for their psychology courses; and, if found innocent on the polygraph test, they were given $20. The participants were randomly assigned to one of six conditions in a balanced 2 X 3 between-groups factorial design. Specifically, there were two levels of Guilt (guilty and innocent) and three levels of Feedback (no feedback, skin conductance, and composite.) The university's Institutional Review Board approved the study protocol and informed consent document prior to participant recruitment.

The polygraph examiner was a graduate student who had been trained to use standard polygraph procedures; the examiner did not have any previous academic interactions with any of the participants.

Apparatus

Physiological Data Collection: The CPSLAB system (Scientific Assessment Technologies, SLC, UT) was used to configure the data collection hardware, specify storage rates for the physiological signals, and build automated data collection protocols. CPSLAB was also used to collect, edit, and score the physiological data.

The physiological data acquisition subsystem (PDAS) of CPSLAB generated analog signals for thoracic and abdominal respiration, skin conductance, finger pulse amplitude, and EKG. Each of the five analog signals was digitized at 1000 Hz with a Metrabyte DAS 16F analog-to-digital converted installed in the CPSLAB computer. The CPSLAB computer collected and stored the polygraph charts.

Respiration was recorded from two Hg strain gauges secured with Velcro straps around the upper chest and the abdomen just below the rib cage. Resistance changes were recorded DC-coupled with a 2-pole, low-pass filter, fc = 13Hz.

Skin conductance was obtained by applying a constant voltage of .5V to two UFI 10mm Ag-AgCl electrodes filled with .075M NaCl in a Unibase medium. The electrodes were strapped with adhesive to the middle phalanx of the fourth and fifth fingers of the left hand. The signal was recorded DC-coupled with a 2-pole, lowpass filter, fc = 6 Hz.

Finger pulse amplitude was obtained from a UFI photoplethysmograph attached to the index finger of the left hand with a Velcro strap. The signal from the photocell was AC-coupled with a 0.2-second time constant and a 2-pole, low-pass, fc = 10 Hz.

The electrocardiogram was obtained from the limb Lead II configuration of Einthoven's Triangle using disposable, pre-gelled Ag-AgCl snap electrodes. The PDAS generated a 20 ms square wave pulse that coincided with the R-wave in the electrocardiogram. The square wave from the PDAS was routed to the analog-to-digital converter, and the CPSLAB software measured and stored the time between successive pulses (interbeat intervals).

The 1000 Hz samples for each channel were reduced prior to storing them on the hard disk by computing the mean of samples for successive data points. Respiration and electrodermal channels were stored at 10 Hz. Cardiograph and finger pulse signals were stored at 100 Hz. The cardiotachometer produced an interbeat interval measured to the nearest ms for each heartbeat.

Feedback. The analog respiration, skin conductance, and cardiograph signals along with event marks were routed to a second computer equipped with a Metrabyte DAS 8 analog-to-digital converter. The second computer was programmed to provide the appropriate type of feedback (if any) to the participant. Auditory feedback was produced by the speaker in the PC. The auditory feedback was a tone that varied in pitch with changes in skin conductance or the composite index of arousal. The composite feedback was based on a weighted combination of changes in skin conductance (50%), cardiograph (25%), and respiration (25%). The feedback began at question onset, lasted for 20 seconds, and then was turned off until the next question was presented.

Procedure

Prospective participants registered on the Internet for an appointment to participate in the experiment. When the participant arrived, an envelope addressed to the participant was taped to the door of the meeting room. Instructions within the envelope directed the participant to enter the room, close the door, read and sign an informed consent form, complete the polygraph accuracy questionnaires, and then play a cassette recorder that presented instructions over earphones.

Guilty participants received tape-recorded instructions to commit a mock theft of $20 from a purse in a desk drawer in the room where they received their instructions. The participant searched the desk for the purse, took the $20, concealed it on his/her person, and then reported to the laboratory where the polygraph test was administered. Innocent participants listened to a general description of the crime and then reported to the laboratory for their polygraph examination.

All participants were told that a polygraph expert who didn't know if they had committed the theft would give them a polygraph test. They were told that the examiner would use a computer to assist in the analysis of their polygraph charts. They were also told that they would receive course credit and would be paid $20 if they passed the polygraph test; but they would receive only the course credit and not be paid if they failed the test or the test was inconclusive. Thus, innocent participants were paid the bonus if they were correctly classified by the computer as innocent, whereas guilty participants were paid $20 if the computer incorrectly classified them as innocent. Throughout the polygraph procedure, the polygraph examiners remained blinded to the participant's condition. Once the computer decision was revealed to the participant, the polygraph examiners received documentation from an assistant who assigned the condition to the participant. The assistant did not participate in any aspect of the polygraph test, other than determining the participant's guilt and feedback status.

Pretest. When the participant arrived at the laboratory, the polygraph examiner introduced herself, obtained some demographical data, and reviewed the test questions with the participant. Standard field polygraph procedures were used. Relevant questions that pertain to the theft and the sacrifice relevant were reviewed first, probable lie questions were reviewed next, and the neutral questions were reviewed last. The test questions were as follows:

Test Questions

(Sacrifice Relevant) 1. Do you intend to answer truthfully all of the questions about the theft of the $20?

(Neutral) 2. Do you live in the United States?

(Probable lie) 3. Before the age of _____, did you ever take something that didn't belong to you?

(Relevant) 4. Did you take that $20 from the purse?

(Neutral) 5. Is today _____?

(Probable lie) 6. During the first _____ years of your life, did you ever do anything that was dishonest or illegal?

(Relevant) 7. Did you take that $20?

(Neutral) 8. Is your first name _____?

(Probable lie) 9. Between the ages of _____ and _____, did you ever lie to get out of trouble?

(Relevant) 10. Do you have that $20 with you now?

After reviewing the test questions, sensors were attached to the participant. The polygraph examiner then described the role of the autonomic nervous system in the detection of deception and administered a standard numbers test. Consistent with field practice, participants were informed that they showed their strongest reaction when they lied about the number they chose and showed smaller reactions when they were truthful.

No APF was given during the numbers test. Since the numbers test is a relatively weak manipulation, a high percentage of participants did not actually show their strongest reaction to the chosen number. If participants were to receive APF that revealed that they showed a relative weak reaction to the chosen number, it would not be possible for the polygraph examiner to claim that they did. Moreover, if participants learned from the APF that the technique failed to detect their deception during the numbers test, the accuracy of the subsequent CQT might suffer (Bradley & Janisse, 1981).

Following the numbers test, participants in the APF conditions were informed about the nature of the feedback they would receive during the CQT. Participants in the skin conductance and the composite feedback condition were told that a tone would be presented during the polygraph test. They were told that this tone would rise in pitch as a function of the magnitude of their physiological response to each question.

Interrogation. The probable-lie test was then administered. The question sequence was presented five times, and the interval between repetitions of the question sequence was from one to three minutes. The order of neutral and probable-lie questions varied over repetitions of the question sequence such that each neutral and each probable-lie question at least once preceded each relevant question. The interval between question onsets was a minimum of 35 s.

At the conclusion of the test, the sensors were removed, and the participant was asked to complete the post-test questionnaire. The probability that the participant answered the relevant questions truthfully was then computed using the CPS algorithms developed at the University of Utah (Kirchner & Raskin, 1988; 1994). According to the CPS algorithm, if the probability that the participant was truthful exceeded 0.70, the participant was classified as innocent and was awarded the $20 and course credit. Otherwise, the participant was given only course credit. The participant was then debriefed and informed that the study was finished.

Data Analysis

Dependent Variables

Dependent measures consisted of computer measurements and computer decisions. The CPSLAB software provided the computer measurements and the CPS program provided computer decisions.

Computer Measurements. The CPSLAB software measured skin conductance amplitude (SC amplitude), cardiograph amplitude, and respiration excursions as follows:

SC Amplitude. A SC response curve was defined as the series of samples taken at 10 Hz from question onset to the 20th post-stimulus second. The computer identified points of inflection in the response curve and measured the difference between each low point and every succeeding high point. SC Amplitude was quantified as the greatest observed difference between a low and high point.

Cardiograph Amplitude. CPSLAB identified the time and level of each systolic point in the cardiograph. The systolic points were used to create a second-by-second systolic curve from question onset to 20 seconds post-question onset. Another second-by-second curve was computed from the diastolic points. The mean of the systolic and diastolic points was then compared for each second. Cardiograph amplitude was extracted from the mean response curve in the manner described above for SC amplitude.

Respiration Excursion. Excursion was operationalized as the sum of absolute values of differences between successive 10 Hz samples of respiration obtained from question onset to 20 seconds post stimulus (100 discrete samples). A separate sum of absolute values (excursion) was obtained for thoracic and abdominal respiration. The mean of thoracic and abdominal excursion was computed for each test question. The repeated measurements of thoracic and abdominal respiration excursions, taken separately, were transformed to standard scores. Respiration excursion was defined as the mean of the standard measurements of thoracic and abdominal excursions.

For each physiological measure, an index of differential reactivity to relevant and comparison questions were computed in the manner described by Kircher and Raskin (1988). Briefly, the three probable-lie and three relevant questions on each of the first three charts provided 18 repeated measures of a physiological component. The 18 measurements for a physiological measure (e.g., SC amplitude) were converted to standard scores.

Mean standard scores for relevant questions were subtracted from mean standard scores for comparison questions. The sign of the computer index indicated which question produced the stronger reaction, and the magnitude of the score provided a precise measure of the difference between reactions to the two types of questions.

For all measures except respiratory excursion, a large measured response was indicative of a strong physiological response to a question. However, relatively small measured responses for respiration indicated greater respiratory suppression, which is associated with deception (Kircher & Raskin, 1988; Timm, 1982). Therefore, the sign of the standardized scores for respiration was reversed so that higher scores indicated stronger reactions, consistent with the other physiological measures.

Computer Decisions

The procedures used for making computer decisions paralleled those used by field polygraph examiners who perform numerical evaluations of polygraph charts. If the computer analysis of the first three charts yielded a probability of truthfulness of .70 or greater, or .30 or less, the participant was classified as innocent or guilty, respectively. If the computer analysis was inconclusive after three charts, the final two charts were included in the computer analysis. Participants were classified as inconclusive only if after five charts, their probability of truthfulness remained between .30 and .70.

Reliability of Computer Measurements. Coefficient alpha assessed the internal consistency of computer indices of differential reactivity. To compute coefficient alpha, an index of differential reactivity was computed for each of the 15 comparison-relevant question pairs obtained from the five charts. The 15 difference scores were treated as responses to 15 items on a test (Kircher & Raskin, 1988). If APF captured the attention of participants and reduced random variation in how they processed test questions, the reliability of physiological measures should be greater for participants who received APF than for those in the no-feedback control conditions.

Statistical Tests and Power. A series of univariate comparisons were performed to determine if there were statistically significant effects for Guilt, Feedback, and Sex on each computer index of differential reactivity. For the proposed analyses, the power to detect a medium effect (0.5 within-group standard deviation) exceeded .90 with 210 participants; hence this design had sufficient power to determine if feedback was statistically significantly better or worse than no-feedback. Planned Guilt X Type of Feedback interaction contrasts (Keppel, 1991) were performed to determine if APF affected discrimination between guilty and innocent participants.

Based on the results of those statistical tests, the type of APF that maximized discrimination between truthful and deceptive participants was identified; and the data from that condition were used to test if APF reduced habituation of physiological responses to comparison and relevant questions. A Guilt X Feedback X Charts split-plot ANOVA was performed for each index of differential reactivity. Feedback had two levels (the selected APF and No Feedback); and Charts consisted of a repeated measure with three levels. A more rapid decline in the discrimination between guilty and innocent participants across charts was expected in the no-feedback condition. The selected APF condition was expected to reduce habituation and improve discrimination between guilty and innocent participants across charts.

Analyses of Computer Decisions. Yate's corrected chi-square tests were used to test if there existed differences in accuracy between feedback and no-feedback conditions and between types of feedback. For each of these analyses, a dichotomous decision rule ensured that all participants were classified as truthful or deceptive. "Truthfulness" was defined as having a probability of .50 or higher. These chi-square analyses were performed separately for guilty and innocent participants.

Analyses of Physiological Waveforms. To explore the possibility that APF affects the duration of a physiological response, rather than its amplitude, traditional split-plot ANOVA was used to test for differences in shapes of physiological responses obtained for comparison and relevant questions over the 20-second interval that followed question onset (Kircher, Woltz, Bell & Bernhardt, 1998; Podlesny & Raskin, 1978). These analyses included second-by-second skin conductance, cardiograph, respiration, finger pulse amplitude, and heart rate measures. The between-groups factors consisted of Guilt (2 levels), Feedback (3 levels), and Sex (2 levels); within-participants factors consisted of Question Type (Comparison and Relevant), Charts, and Time (20 seconds). Vagal tone was measured via the Porges-Bohrer algorithm every five seconds during the 20 seconds that followed question onset. Therefore, the time factor in the ANOVA for vagal tone had four levels rather than 20.

Results

Missing Values

Forty-eight of the 1050 charts for the 210 participants (210 X 5) were missing due to participants' reports of fatigue (~20) or due to data collection malfunction (~20); but there was no statistically significant relationship between the loss of charts and group assignment. The first three charts were available for every participant but one. That participant's missing first chart was replaced with her second chart. Charts 4 and 5 were used only to make decisions and only in the event that the outcome based on the analysis of the first three charts was inconclusive. In two cases, the fourth and fifth charts were unavailable for participants with inconclusive outcomes after the first three charts. For those participants, the test was considered inconclusive.

Computer Decisions and Reliability

Table 1 presents the percentage of cases classified correctly, incorrectly, and as inconclusive. Table 1 also presents the percent correct decisions including inconclusive outcomes for each group of participants. Table 2 shows the reliability of differential reactivity measured across the 15 probable-lie/relevant question pairs in the five repetitions of the question sequence (charts). Mean reliability as measured by coefficient alpha was slightly higher for the APF groups than for the control group.

Effects of Gender on Dependent Measures

Preliminary Guilt X Feedback X Gender ANOVAs revealed no main or interaction effects on SC, cardiograph, or respiration measures that involved Gender. Therefore, Gender was dropped as a factor from all subsequent analyses.

Effects of SC Feedback on Physiological Measures

To determine if continuous auditory feedback of SC activity increased discrimination between guilty and innocent participants, a separate Guilt X Feedback interaction comparison was performed for each of the three computer indices of differential reactivity. Guilt had two levels (guilty and innocent) and Feedback had two levels (no-feedback and SC-feedback). The means for SC amplitude, cardiograph amplitude, and respiration excursion are plotted in Figure 1.

Figure 1 suggests that discrimination between guilty and innocent participants tended to be greater in the SC-feedback condition than in the no-feedback condition for measures of SC amplitude and cardiograph amplitude and less for respiration excursion. However, the interaction comparisons for SC amplitude, $t(204) = 1.51$, $p < .14$, cardiograph amplitude, $t(204) = 1.09$, and respiration excursion, $t(204) = -1.31$, were not statistically significant.

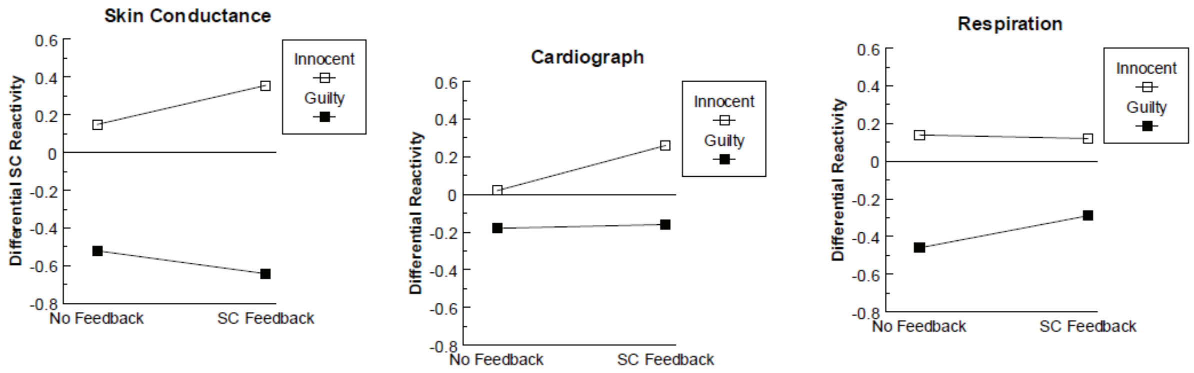
Table 1. Percent computer outcomes and percent correct decisions excluding inconclusive outcomes (n = 35 per group)

		Correct	Wrong	Inconclusive	Correct Decisions
No Feedback	Innocent	69	23	9	75
	Guilty	74	14	11	84
SC Feedback	Innocent	74	17	9	81
	Guilty	77	17	6	82
Composite Feedback	Innocent	69	17	14	80
	Guilty	60	23	17	72

Table 2. Coefficient alphas (internal consistency reliability) for computer indices of differential reactivity across five polygraph charts

Physiological Measure	No Feedback (n = 61)	Skin Conductance Feedback (n = 61)	Composite Feedback (n = 60)
Skin Conductance Amplitude	.85	.91	.80
Cardiograph Amplitude	.40	.53	.71
Respiration Excursion	.81	.79	.71
Mean	.67	.74	.74

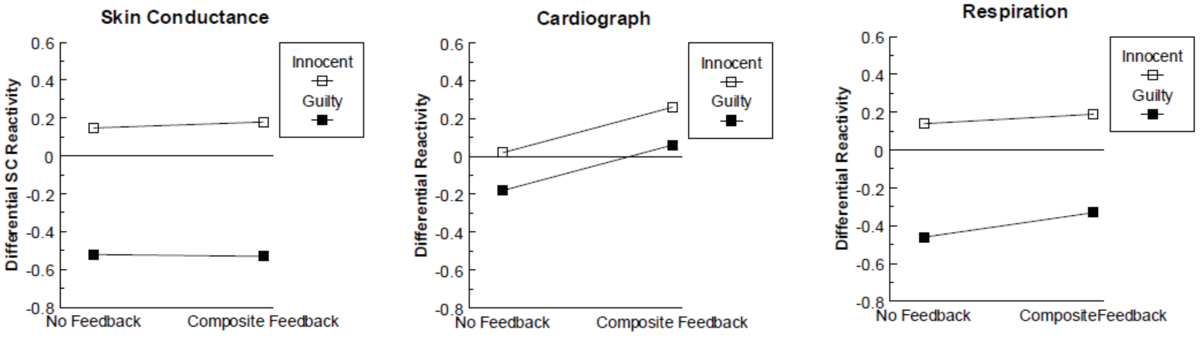
Figure 1. Skin conductance, cardiograph, and respiration indices of differential reactivity for no-feedback and SC feedback groups



Effects of Composite Feedback on Physiological Measures

To determine if composite feedback improved discrimination between guilty and innocent participants, separate Guilt X Feedback interaction comparisons of no-feedback and composite feedback conditions were performed for SC amplitude, cardiograph amplitude, and respiration measures. The means for the three physiological measures are plotted in Figure 2. Again, none of the interaction comparisons was statistically significant.

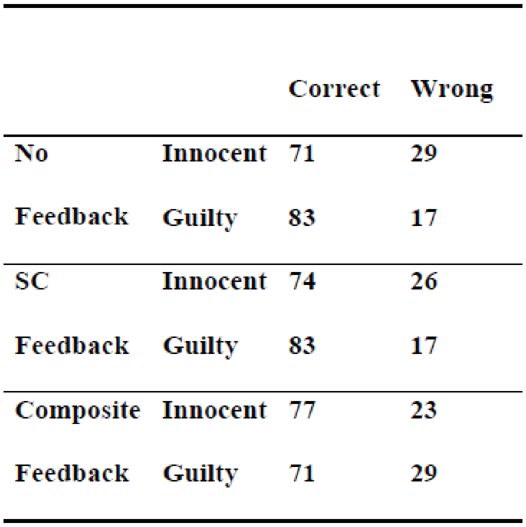
Figure 2. Skin conductance, cardiograph, and respiration indices of differential reactivity for no-feedback and composite feedback groups



Effects of Feedback on Dichotomous Computer Decisions

Table 3 presents the percentage of cases classified correctly and incorrectly when participants were considered truthful if the probability of truthfulness exceeded 0.50 and were considered deceptive if the probability of truthfulness was less than .50. Consistent with the results reported above for individual physiological measures, chi-square analyses revealed no statistically significant differences between no-feedback and SC feedback conditions, between no-feedback and composite-feedback conditions, or between the SC-feedback and composite feedback for either innocent or guilty groups.

Table 3. Percent dichotomous computer outcomes based on first three charts (n = 35 per group)



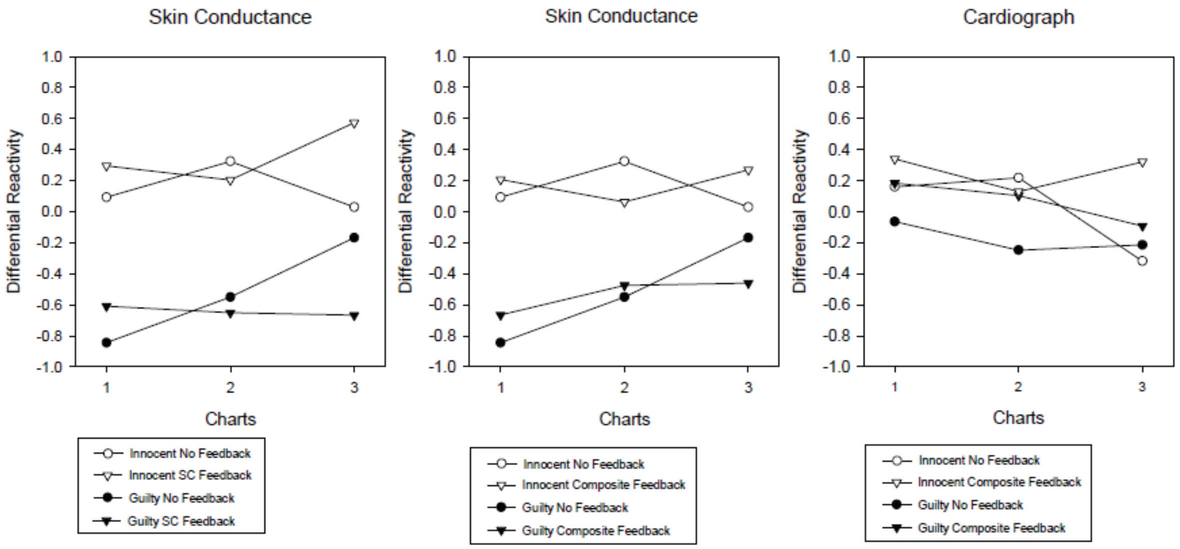
		Correct	Wrong
No Feedback	Innocent	71	29
	Guilty	83	17
SC Feedback	Innocent	74	26
	Guilty	83	17
Composite Feedback	Innocent	77	23
	Guilty	71	29

Effects of Feedback on Habituation Rates

Guilt X Feedback X Charts split-plot ANOVAs were conducted to test the prediction that APF would reduce the habituation of SC, cardiograph, and respiration responses to repeated presentations of comparison and relevant questions. One ANOVA was performed to compare the no-feedback and SC feedback conditions, and another ANOVA compared the no-feedback and composite-feedback conditions. The first set of analyses, displayed graphically in Figure 3, was limited to the first three polygraph charts. This three-chart analysis was conducted independent of the full five-chart analysis to determine APF effects on habituation in a situation more similar to a field polygraph test, where only three charts are collected. P-values based on Huynd-Feldt corrected degrees of freedom were used to decide if an effect was statistically significant. Results suggest the Guilt X Feedback X Charts interaction effect on SC amplitude was statistically significant when participants who received SC feedback were compared to those who received no feedback, F(2, 272) = 6.84, p < .01, η^2 = .05. The means for the first three charts are presented in the left panel of Figure 3. Figure 3 reveals a clear difference between guilty and innocent groups on the first two charts whether or not the participants received auditory SC feedback. However, on the third chart, discrimination between guilty and innocent participants was statistically significantly greater for participants who received APF than no-feedback. A similar effect on SC amplitude emerged when participants who received composite feedback were compared to those who received no feedback, F(2, 272) = 3.70, p < .05, η^2 = .03. The means for the no-feedback and composite feedback conditions are presented in the center panel of Figure 3. Similar to the SC feedback condition, discrimination between guilty and innocent participants by the third chart was greater for participants who received composite APF than for those who received no feedback.

To determine whether or not the effects of APF persisted in further chart presentation, a second analysis which included the fourth and fifth charts was conducted. The Guilt X Feedback X Charts effect was still significant for the No-Feedback versus SC Feedback comparison, p < .02 with Geisser-Greenhouse corrected df. As predicted, guilty feedback participants had more negative SC differential reactivity scores (appeared more deceptive) than guilty no-feedback participants.

Figure 3. Effects of feedback on habituation of skin conductance and cardiograph responses



However, the difference between innocent feedback participants and innocent no-feedback participants that was found for chart 3 attenuated in charts 4 and 5. Thus, for charts 4 and 5, the effect of SC feedback in detecting deception in guilty participants remained; but by the fifth chart, the beneficial effects seen in the previous analysis for innocent participants was not statistically significant.

Composite APF also affected cardiograph responses, F(2, 272) = 3.25, $p < .05$, $\eta^2 = .02$. However, in this case the effects were relatively small and not consistently beneficial. Examination of the right panel of Figure 3 reveals that there was greater discrimination between guilty and innocent participants on the second chart for participants who received no feedback (circles) than for participants who received APF (triangles). In contrast, discrimination between guilty and innocent participants was greater on the third chart for those who received AFP than for those who did not. Indeed, innocent participants who received no feedback (open circles) evidenced slightly more negative cardiograph scores than did guilty participants who received no feedback (closed circles). This same trend remained for analyses conducted on charts 4 and 5.

There were no statistically significant effects of SC-feedback on habituation rates of respiration or cardiograph responses, nor were there effects of composite-feedback on the habituation rates of respiration responses.

Waveform Analysis

Diagnoses of truth and deception by the computer and by polygraph examiners are often based on increases in electrodermal and cardiovascular activity and respiration suppression. In the presence of APF, measures other than SC amplitude, cardiograph amplitude, and respiration excursion may be more diagnostic of truth and deception. To explore this possibility, split-plot ANOVA was used to test for differences in the shapes of various physiological responses to probable-lie and relevant questions over the 20-second interval that followed question onset. ANOVA was performed separately for SC, cardiograph, thoracic and abdominal respiration excursion, finger pulse amplitude, heart rate, and vagal tone. Between-group factors were Guilt (guilty and innocent) and Feedback (no-feedback, SC-feedback, and composite-feedback). Within-participant factors were Question Type (comparison and relevant) and Time (e.g., seconds).

Twenty second-by-second measurements were analyzed for all physiological measures except vagal tone (Podlesny & Raskin, 1978). Vagal tone was measured for each of four successive 5-second intervals. Of interest were Guilt X Question Type X Feedback, and Guilt X Question Type X Feedback X Time interactions. If the differences between comparison and relevant questions for guilty and innocent participants do not depend on the presence or type of APF, then measures found useful for individuals who do not receive APF also should be useful for individuals who do receive APF.

The Guilt X Question Type X Feedback interaction was statistically significant for thoracic respiration excursion, F(2, 203) = 3.37, p < . 05, η^2 = .03. The means for comparison and relevant questions are presented in Figure 4. Baseline respiration measurements for neutral questions are included in Figure 4, but they were not included in the ANOVA. As predicted, innocent participants generally evidenced less respiration activity in response to comparison questions than to relevant questions, whereas guilty participants showed less respiration activity in response to relevant questions.

Figure 4. Thoracic respiration excursion for neutral, comparison, and relevant questions under no-feedback, SC feedback, and composite

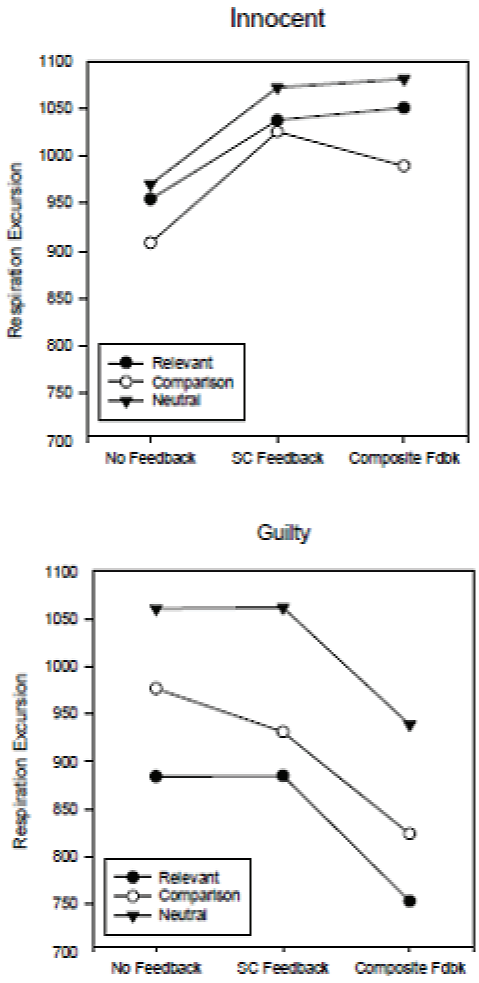
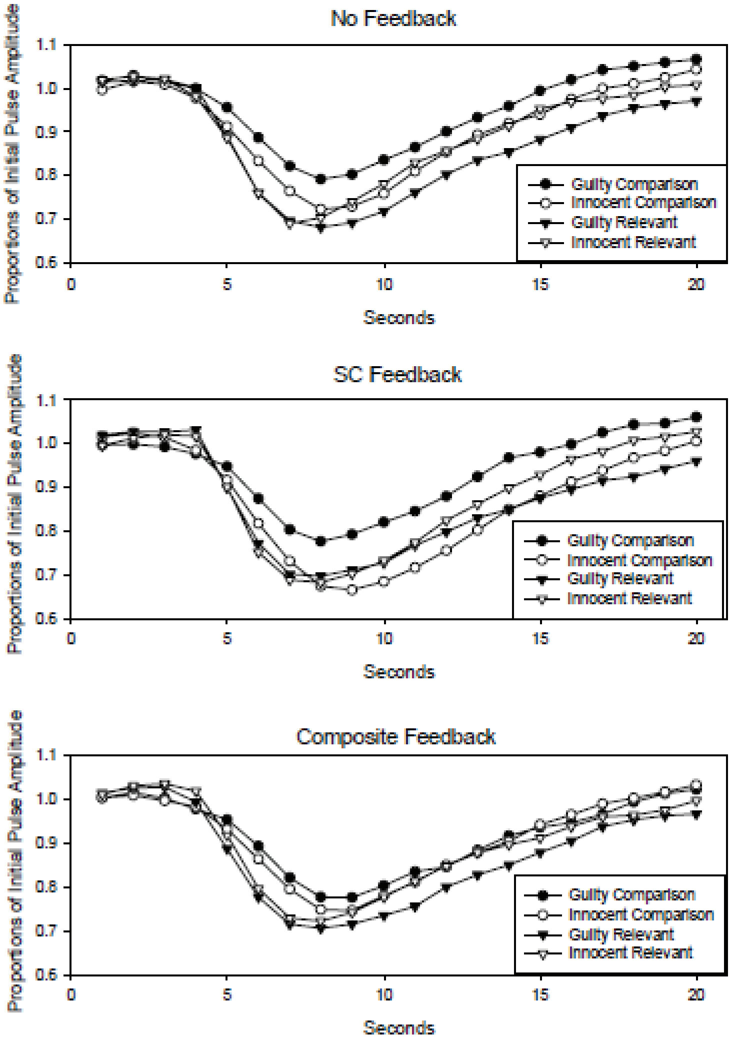


Figure 4 suggests that the interaction was due to an atypical pattern of responses by innocent participants who received SC feedback. As can be seen, the difference between comparison and relevant questions for innocent participants who received SC feedback was less than the difference for the participants in other groups.

A statistically significant effect was also found for the Guilt X Question Type X Feedback X Time interaction for finger pulse amplitude (FPA), F(38, 3876) = 2.53, *p* < .02, η^2 = .02. Figure 5 displays the FPA curves for the three feedback conditions. In general, a strong vasomotor response was indicated by a large reduction in the amplitude of finger pulses. The results given in Figure 5 indicate that guilty participants responded as predicted; they evidenced stronger vasomotor responses to relevant questions than to probable-lie questions across all feedback conditions. In contrast, innocent participants in the no-feedback and composite-feedback conditions showed little difference in their vasomotor responses to probable-lie and relevant questions. Only innocent participants in the SC-feedback condition responded more strongly to comparison questions than to relevant questions.

Figure 5. Finger pulse amplitude responses to comparison and relevant questions for guilty and innocent participants in three feedback conditions



Discussion

The goals of this study were, through the use of APF, to attempt to increase the reliability and validity of the physiological measures obtained during a conventional CQT polygraph test, and to reduce habituation to repeated exposures to polygraph questions. Although not all of our hypotheses were substantiated, the results of the study that did confirm our hypotheses offer useful information to those conducting polygraph tests in the field.

Our first hypothesis, that the use of APF would increase the reliability of physiological measures was not statistically significantly substantiated. Although the use of both composite and skin conductance auditory feedback did increase the magnitude of the coefficient alpha index by five percentage points relative to the no-feedback condition, this increase in reliability probably is not "clinically" statistically significant, in that noticeable improvements in decision accuracy by reducing random variation in the way participants processed questions probably would not result from using APF. The data in Table 1, Figure 1, Figure 2 and Table 3, suggest the difference in percent correct decisions for both innocent and guilty participants in the SC Feedback condition was improved, but not statistically significantly so.

Although results suggest that APF may not enhance the reliability of CQT polygraph tests, an aspect of the protocol implemented in this study may account for the null effects observed from these data. Specifically, it may be the case that the lack of time elapsed from the "mock" crime committed by participants to the actual polygraph test, or the reward offered for an innocent verdict, caused the participants to experience enhanced intrinsic motivation to "perform" well on test and receive the cash bonus. Such motivation to be classified as innocent may not differ substantially from a suspect in a criminal investigation; however, rarely is an individual offered cash in exchange for an innocent verdict or given a polygraph test concerning alleged criminal involvement immediately following the actual crime. Because the guilty participant was given the test immediately after committing the theft, his/her memory of the crime, and involvement with that criminal activity was probability more easily accessible affectively and difficult to suppress physiologically than the criminal who committed a theft in the days or even weeks preceding the polygraph test. Hence the effectiveness of APF might have been confounded by the degree of involvement of the participants with the recency of the crime, such that added benefits of detection over the No-feedback condition were lost.

Perhaps future investigations of APF on detection should include a mock crime that is committed three or more days preceding the polygraph investigation to more adequately represent the typical field polygraph investigation into alleged criminal activity. Using a smaller cash bonus may also serve to reduce the "ceiling effect" observed in this study. This idea, that the more "ego-involving" and relevant the participants perceive the testing situation to be, the less effective is the use of APF, was first mentioned in the Timm (1987) study of biofeedback effectiveness in assessing guilt as measured from the Guilty Knowledge Test (GKT); and the results of the first experiment of relevant (SS#) and non-relevant (geometric figures) items observed in the Stern, et al. (1981) study support such a claim. Thus, these studies, coupled with the results found in this investigation, indicate that further investigation into perceived participant involvement with the test is necessary.

Our hypothesis that APF would decrease habituation rates as participants completed successive polygraph charts was substantiated. Specifically, SC amplitude for those in the SC Feedback condition did not evidence the decrease typically seen as suspects complete the second and third polygraph charts. In fact, by the completion of the third chart, there still existed a greater delineation between innocent and guilty participants who received SC feedback or composite feedback than those who received no feedback. This effect persisted even after including two additional charts of data for the guilty participants. These two charts are not part of a standard polygraph test and further support the strength of APF in detecting guilty participants who may be required to complete a longer version of a standard polygraph test. Thus, APF may serve to decrease fatigue effects or lack of involvement in the test commonly observed after repeated exposures to the test questions for guilty participants and increase the usefulness of latter charts for detecting deception.

Because diagnoses of truth and deception by polygraph examiners are often based on increases in electrodermal and cardiovascular activity and respiratory suppression, another goal of this investigation was to attempt to examine alternative methods of interpreting physiological responses in the presence of APF that may be more diagnostic of truth or deception. As predicted, investigations into second-by-second measurements of thoracic respiration excisions for innocent participants showed that their responses were more suppressed to comparison questions than to relevant questions, whereas guilty participants evidenced more suppression to relevant questions than to comparison questions.

A statistically significant effect of finger pulse amplitude was also found for guilty participants in all feedback conditions. As expected, a stronger vasomotor response was observed in guilty participants for relevant questions than comparison questions. Results for innocent participants suggest that SC Feedback enhances the predicted increase in vasomotor responding to comparison questions. These promising results obtained for SC Feedback are also consistent with the findings of the Stern, et al. (1981) study that found greater success with SC Feedback than the heart rate feedback.

Overall, the results suggest a number of implications concerning the use of APF during CQT polygraph tests. Although detection rates did not appear to be enhanced by APF in this study, further investigation into the benefits of APF are needed. The use of APF in this investigation was shown to offer at least one clear benefit for the CQT polygraph test. APF decreases the rate of habituation over repetitions of the question sequence and allows for greater discrimination between innocent and guilty participants as the polygraph test progresses.

References

Ben-Shakhar, G. (1991). Clinical judgment and decision-making in CQT polygraphy: A comparison with other pseudoscientific applications in psychology. *Integrative Physiological and Behavioral Science*, 26, 232-240.

Bradley, M. T. & Janisse, M. P. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal and pupillary measures. *Psychophysiology*, 18, 307-314.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, N. J.: Prentice Hall.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook* (3rd Ed.). Engelwood Cliffs, NJ: Prentice Hall.

Kircher, J. C. (1981). *Psychological and psychophysiological processes underlying the detection of deception.* Unpublished manuscript. Department of Psychology, University of Utah, Salt Lake City, UT

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

Kircher, J. C. & Raskin, D. C. (1994). *The Computerized Polygraph System.* Version 2.00 Manual. Salt Lake City, UT: Scientific Assessment Technologies.

Kircher, J. C., Woltz, D. J., Bell, B. G., & Bernhardt, P. C. (1998). *Effects of audiovisual presentation of test questions during relevant-irrelevant polygraph examinations and new measures.* Final report to the U. S. Government. University of Utah, Salt Lake City, UT.

Podlesny, J. A., & Raskin, D. C. (1978). Effectiveness of techniques and physiological measures in the detection of deception. *Psychophysiology*, 15, 344-358.

Porges, S. W., Bohrer, R. E., Cheung, M. N., Drasgow, F., McCabe, P. M., & Keren, G. (1980). New time series statistic for detecting rhythmic co-occurrence in the frequency domain: The weighted coherence and its application to psychophysiological research. *Psychological Bulletin*, 88, 580-587.

Raskin, D. C., Kircher, J. C., Honts, C. R , & Horowitz, S. W. (1988). *A study of the validity of polygraph examinations in criminal investigation* (Grant No. 85-IJ-CX0040). Salt Lake City: University of Utah, Department of Psychology.

Raskin, D. C., Honts, C. R., & Kircher, J. C. (1997). Polygraph techniques: Theory, research and applications from the perspective of scientists-practitioners. In D. L. Faigman, D. Kaye, M. J. Saks, & J. Sanders (Eds.), *Scientific Evidence Reference Manual* (pp.551-582). St. Paul, MN: West Publishing Co.

Reid, J. E., & Inbau, F. E. (1977). *Truth and Deception: The Polygraph ("Lie Detector") Technique.* Baltimore: Williams & Wilkins.

Stern, R. M., Breen, J. P., Watanabe, T., & Perry, B. S. (1981). Effects of feedback of physiological information on responses to innocent associations and guilty knowledge. *Journal of Applied Psychology*, 66, 677-681.

Timm, H. W. (1987). Effect of biofeedback on the detection of deception. *Journal of Forensic Sciences*, 32, 736-746.

The Effects of Aural Versus Visual Presentations of Questions during a Detection of Deception Task[1]

Barbara L. Carlton and Brenda J. Smith

Abstract

The purpose of this research was to investigate the relationship between accuracy of a detection of deception task and the stimulus mode of the question presentation. That is, will the presentation of questions on a computer screen change the accuracy rate when compared to exams conducted, more traditionally, in a verbal mode? Eighty subjects were assigned to either a guilty or innocent condition. Guilty subjects were shown a video of a mock crime scenario, while innocent subjects viewed a clip from a training video. Half of the innocent and half of the guilty groups were given the exams aurally using a tape recorder, and the other half shown the questions on a computer terminal. Subjects were then given a guilty knowledge test by the experimenter using a Coulbourn polygraph.

While the polygraph exam was being administered, a second experimenter sat across from the subject. This second experimenter was responsible for programming the subject, while the experimenter running the exam was blind to the subject's guilt/innocent status. During the exam, the subject was required to respond to the experimenter with "no" to every item. The charts were scored by the following: (1) the original examiner; (2) a blind evaluator; and (3) using a scoring system introduced by Lykken. Overall accuracy of the decisions of the original examiner was 78%, 74% for the blind examiner, and 76% for the Lykken system. Accuracy rates for subjects in the visual condition were 83% for the original examiner, 78% for the blind evaluator, and 70% for the Lykken system. The decisions for the aural condition were 73% accurate for the original examiner, 70% accurate for the blind evaluator, and 83% accurate for Lykken scoring system. There was no significant association between an accurate decision and the stimulus mode condition for the original examiner, the blind evaluator or the Lykken scoring decision. (χ^2 = .6091; p < .4351 and χ^2 = 2.0378; p < .1534; χ^2 = 1.065, p < .3020). There was no significant association between the type of error and the stimulus mode for the original examiner (Fisher's exact p < .14) or the decision rendered by the Lykken system (Fisher's exact p < .25) whereas the type of error was associated with stimulus mode for the blind examiner (Fisher's exact p < .0075). This may be due to an artifact associated with the use of the experimenter as a confronter during the exam.

Introduction

The method of presenting questions in field polygraph exams has remained relatively unchanged since 1917. Examiners are taught to ask questions in an unemotional tone of voice to be sure it is the content of the question and not the delivery that is associated with any physiological reaction.

The advent of television and personal computers has made presentation of written material on a video screen rather common. There is, however, a dearth of research on the application of this common technology to polygraph testing. Application of visual technology in physiological detection of deception (PDD) has both certain advantages and disadvantages.

[1] Original citation for this reprinted article is: Carlton, B., & Smith, B.J. (1991). *The effects of aural versus visual presentations of questions during a detection of deception task* (DoDPI91-R-0002). Fort McClellan, AL: Department of Defense Polygraph Institute. DTIC# ADA304657. Portions of the original report were redacted to bring the paper into the publication style of *Polygraph.*

No doubt it would increase the cost of apparatus in the field and, until perfected, might be more awkward to use than verbal presentations. However, using a computer to deliver the questions might be a good way of ensuring that physiological responses are associated with the content of the questions, and not any intentional or unintentional verbal or nonverbal behavior on the part of the examiner. If this is true, the use of visually presented techniques would take the field of PDD a long way toward standardization. Also, there is little research that examines the accuracy of a polygraph test given to someone with impaired hearing, where visual presentation of the questions may be a necessity.

Lacking conclusive research support, there has been no temptation to adopt visual presentation methods. To date, only one investigation can be found in current literature which compared the effects of the type of stimulus mode in which the questions are presented.

An investigation by Beijk (1980) attempted to evaluate potential differences found in skin resistance responses as a function of mode of stimulus presentation on a numbers test. A prior experiment found a significant 'hit' rate on a numbers test. A follow-up experiment was conducted to examine different modes of presentation (auditory versus visual) and found no significant difference between visual and auditory presentation of the stimuli. The authors "conclude that a small difference in experimental procedure, be it an attempt to change motivation of the mode of stimulus presentation, did not significantly change the results found in Experiment 1." (p 276).

Beijk used a type of information test (Podlesny & Raskin, 1977). There are several types of information tests. One type of information test that might prove to be useful in the field is the guilty knowledge test or GKT.[2] An information test presumes that a guilty person possesses knowledge or information that an innocent person would not. It is the exposure of this knowledge or information that is associated with the response made during the polygraph exam.

According to Andreassi, the GKT is superior to the more typically used control question technique[3] (CQT), because it is standardized, error rates can be specified with GKT, and researchers believe that it is less vulnerable to faking or the use of countermeasures (Andreassi, 1989).

The purpose of this research is to compare the distributions of decisions obtained when the questions are presented verbally to those rendered when the questions are presented visually on a GKT. Does one mode of presentation result in more accurate decision concerning deception?

Method

Subjects

Twenty-two female and 60 male basic trainees at Fort McClellan, Alabama participated as subjects in this investigation. Due to excessive movement, the data for two of the male subjects were not included in the final analyses. Subjects were, for the most part, in average to excellent health. The age of the subjects ranged from 17 and 33.

Equipment/Apparatus

Subjects' physiological data was recorded using a Coulbourn Skin Conductance Coupler and preamplifier (S71-22). The coupler was set on AC coupling, sensitivity on 1000 mV/micromho, using silver-silver chloride electrodes attached to the palmer side of the index and middle fingers of the subject right hand. The data was collected on a PC Brand 286 with an NEC Multisync monitor using CODAS Software by DATACQ. CODAS is a data acquisition program which digitizes analog information and stores it in a file in the computer, no hard copy is made.

[2] The current terminology is Concealed Information Test, or CIT. The original language has been retained for this publication.

[3] The current terminology is "comparison question technique." The original language has been retained for this publication.

After the data has been digitized and stored, the data was printed out on hard copy using a HP LaserJet Series II printer.

The questions presented in the visual condition were presented on a Zenith IBM PC Compatible using Harvard Graphics Software. The questions presented in the aural condition were delivered via a Marantz PMD 221 Portable 3-head Cassette Recorder.

Procedure/Method

Upon arrival at the Institute, subjects were met and briefed on the purpose of this investigation. The purpose and procedure of the study was fully explained to all subjects. Subjects were also given a copy of a justification and explanation sheet. At this time, subjects were asked to read and sign a volunteer affidavit or participation consent form. Copies of the justification/explanation sheet and the volunteer affidavit can be found in Appendix A.[4] The volunteer affidavit informed the subject that his/her participation is solely voluntary. The form specifies that if the subject wishes to discontinue their participation, she/he may do so at any time and no penalty will be assessed. Due to the specific nature of the exam, no personal or biographical information was required; therefore, the subjects were not asked questions of a personal nature.

All subjects were given a guilty knowledge test. There were five questions and each question had six alternatives or possible answers. The specific questions and alternatives, with the critical item identified, can be found in Appendix B. The questions were presented in the same sequence, as were the alternatives, for all subjects. Subjects were informed that one of the six alternatives was the correct alternative, however, only a guilty person would know which alternative was correct. Prior to each question, the experimenter told the subject what the question would be, but did not go over the alternatives. Subjects were then told that if they were innocent, none of the alternatives would be any more meaningful than the rest, however, if they were guilty then they would know exactly what the correct alternative was and they could expect it to be presented at some point during the recording of the question.

Subjects were instructed not to respond to the question itself, but to wait until they were presented with an alternative. The required response to each alternative was "NO". Since the question began with "Do you know .. ," an innocent person would never be forced to lie since they would not know which of the alternatives was true.

Subjects were randomly assigned to one of the following four conditions: (1) Aural-Guilty, (2) Aural-Innocent, (3) Visual-Guilty, and (4) Visual-Innocent.

Subjects were randomly programmed innocent or guilty individually. All subjects viewed a short video. Subjects who were programmed guilty viewed a video of a mock crime. The video depicted the theft of a gun and some money. The video was shot from the criminal's perspective, meaning as if the camera person was committing the crime. The criminal's face was never shown. However, the arms were visible at times during the crime and they were easily identified as a man's arms. During the theft, an unwitting victim came upon the crime scene. At this point, the criminal pointed the gun at the victim and fired twice. While making sure the victim was dead, the criminal also stole the victim's wrist watch. After viewing the video, the subjects were then questioned by the investigator concerning the critical elements to ensure that the guilty subject indeed had the guilty knowledge prior to the polygraph examination.

Subjects who were programmed innocent were shown a brief training film and asked questions concerning the content afterwards. All subjects were told that the purpose of the polygraph exam is to determine if a polygraph examiner could tell whether or not a subject witnessed the crime based solely on their physiological activity. All subjects

[4] Appendix A is not included here. It is available with the original report (DTIC# ADA304657) which can be ordered from the Defense Technical Information Center: www.dtic.mil/

were strictly warned not to inadvertently alert the examiner to which video they viewed. The subjects were told that if they did allow the examiner to ‘guess’ their condition prior to running charts, either verbally (admission) or nonverbally, they would be released from the investigation and returned immediately to their unit.

Once the subjects were programmed, they were taken to the polygraph room and introduced to the examiner. Only one examiner was used to run the polygraph exams. The investigator who programmed the subject remained in the room. Once in the polygraph room, the subject was briefed on what measures were being taken, how a polygraph works, what kind of question would be asked, and how they were expected to respond. The subjects were informed that they would be taking a polygraph, because they were suspected of having been an accomplice during a crime. The components were attached.

Subjects were seated in a typical polygraph chair, outfitted with the elongated arm rests. The subject was seated approximately 1 meter from a computer monitor and 30 cm in front of the Coulbourn equipment. The examiner sat at a computer terminal located next to the Coulbourn equipment and was therefore approximately 1 meter to the left of, but slightly behind, the subject. The arrangement was designed so that movement of the examiner would occur outside of the subjects' peripheral vision.

Subjects in the visual condition were told that the questions would be presented on the screen in front of them, while subjects in the aural condition were told that the questions would be presented via a tape recording. The subjects were given an example question (presented either visually or verbally on the tape recorder) to make sure they understood the instructions.

An example question can be found in Appendix B. The example question was unrelated to the crime and the subjects were told this prior to the presentation of the example. The subjects were fully aware that the purpose of the example was to give them a chance to see what the actual testing would be like and to make sure they understood what they were supposed to do.

The visual stimuli were created and presented using Harvard Graphics version 2.0. Each character presented visually was approximately 2 cm in height. Subjects in the visual condition were questioned concerning clarity and those requiring reading glasses were requested to use them if necessary. After the presentation of the last question, subjects were required to read the last alternative out loud to ensure that the subject could see and read the word clearly. Subjects in the aural condition were asked if the volume was acceptable.

Each question was presented once. There were three cases in which a question was interrupted during recording by the telephone or someone at the lab door. In these cases, the question was stopped immediately and the question was asked a second time. There were approximately three minutes between each question, while the examiner informed the examinee what the next question would be. Prior to the presentation of each question, the examiner said, "Please remain still, the test is about to begin." At this point, the data collection program was started and physiological recording began. Simultaneously, either the tape recorder was turned on (aural condition) or the program for the specific question (visual condition) was initiated. After a 20-second pause, the question was presented.

In both the aural and the visual conditions, there were 15 seconds between the presentation of the question and the first alternative, as well as between each subsequent alternative. In the visual condition, the question remained on the computer screen until the first alternative was presented. Each alternative also remained on the screen until the next alternative was presented. After the last alternative was presented in both the aural and visual conditions, there was a 15 second pause until the examiner said, "Now you can relax, this portion of the test is complete." During the aural condition the recorder was then turned off. The program in Harvard Graphics terminates automatically using the slideshow option of presentation.

Upon completion of the polygraph examination, the subjects were taken to another room and asked to fill out a questionnaire. The questionnaire was simply a copy of the GKT questions. A copy of this questionnaire can be found in Appendix B. In the questionnaire subjects were asked to identify the critical items for each question. The purpose of this task was to ensure the following: (a) that no programming mistakes were made; (b) the guilty subjects did remember what the critical items were; and (c) innocent subjects did not identify what the critical items were at a better than chance rate. Since all of the subjects were told not to discuss the nature of the study with anybody, the questionnaire might also reveal an innocent subject who had been given information about the crime from a buddy who served earlier.

The Confronter

A confronter was used to increase the accuracy of the examination. For all subjects, the computer screen was approximately three feet directly in front of the subject. All the subjects were told that during the recording they should focus on the computer in front of them. Subjects in the visual condition were told to watch the computer screen so they would not miss the presentation of the questions or alternatives while those in the aural condition were told to focus on the screen to prevent them from becoming distracted and looking about the room. The investigator who programmed the subjects acted as the confronter. The confronter sat next to the computer screen. Subjects were told to focus on the computer screen while the questions and alternatives were presented but when they had to respond they were to look directly in the eyes of the confronter and say 'NO' just as if the confronter had asked the question.

The rationale behind the use of the confronter was to increase physiological responsivity. By increasing physiological responsivity, one would be more likely to observe differential responding which should, in turn, increase the overall accuracy.

Basically, this strategy should serve to make the guilty subjects more uncomfortable about lying. Perhaps lying to someone who knows you are lying is potentially far more disturbing than the simple act of lying alone. Requiring the examinee to look directly into the eyes of the confronter was designed to make the act of 'lying' a little more uncomfortable for the guilty person.

It may be true that simply looking at a stranger during this process would be uncomfortable for the innocent subjects as well; however, the guilty person also has to lie to a strange person who knows they are lying. It was hoped that this differential anxiogenic procedure would increase the accuracy of detecting the guilty subjects. If this did indeed increase the accuracy for detecting guilt then accuracy in establishing innocence would increase as well.

Previous piloting of this study, using field instruments and regular field polygraph examiners, rendered very poor accuracy, statistically around chance levels. Since the purpose of this study was to compare the accuracy between aural versus visual presentation of questions, it was decided to duplicate the conditions of a previous study conducted earlier in this lab (Richardson, Carlton & Dutton, 1990). This previous study used the same video, virtually identical questions and used a confronter. Since this earlier study obtained a high accuracy rate (76% - 80% for the original examiners) it was decided to include the confronter on this study.

Scoring

The skin conductance data were scored in following fashion: (1) by the original examiner upon completion of the polygraph examination; (2) by a blind evaluator; and (3) using a scoring system introduced by Lykken (1959) devised exclusively for scoring guilty knowledge tests.

(1) Original Examiner. The first author of this report served as the examiner who ran the polygraph test and, therefore, was the original examiner. After the subject was run, the data files were printed out to get the hard copy. There were five questions and each question was called a "chart." Scoring of the charts was subjective. A call of Deception Indicated (DI) or No Deception Indicated (NDI) was made based on these five charts alone. The original examiner used information derived from the electrodermal responses. The following physiological indices were used: (1)

amplitude; (2) rise-time; (3) latency changes; (4) changes in frequency of responding.

Of the four indices, the examiner generally placed more weight on the amplitude information. If the largest response on a chart occurred after the presentation of the critical item, the chart was scored a 'hit'. A subject could be called DI if they hit 3/5 keys or more. However, on a few occasions only 2/5 keys were given a 'hit' designation if any or all of the following occurred: (a) rapid decrease in rise-time for response occurring at the key, but not at the other alternatives; (b) shorter latencies for responses occurring at the key and not elsewhere; and (c) the electrodermal activity diminished after the presentation of the key.

(2) Blind Evaluator. A blind evaluator was given information about how the guilty knowledge test was conducted and simply asked to render a decision.

(3) The Lykken Scoring System. The Lykken scoring system uses only the amplitude of the electrodermal responses for scoring purposes. For a given question, the subject's electrodermal responses for the first alternative are discarded while the remaining responses are ranked according to amplitude. If the largest response occurs at the key, the question is given a score of '2.' If the response is the second largest response on the question, the score of '1' is given. Since there are 5 questions, the largest score possible is 10. A subject was classified as deceptive if the total score was 6 or higher. The total score is referred to as a Lykken score.

Results

All of the statistical calculations were conducted using Crunch statistical software.

Questionnaire Results

Analyses were conducted on the questionnaires to address two issues. The first issue was concerned with the accuracy of guilty subjects, that is, to determine if the guilty subjects knew and remembered all of the critical items to each question. The results of the questionnaire showed that all of the guilty subjects correctly identified all of the critical items.

The second issue was to determine if the innocent subjects could correctly identify the critical items. This could occur if the incorrect alternatives were not adequate and the critical item was too obvious or if the subject was given information about the crime by a buddy who served as a subject earlier in the study. Table 1 shows the probability distribution of correctly guessing the critical items, and the number of the innocent subjects who correctly guessed the specified number of critical items.

Table 1. Probability, frequency and expected frequency distributions of innocent subject currently identifying critical items.

# Correct	p	N Observed	N Expected
0	0.328	16	13
1	0.410	13	16
2	0.205	10	8
3	0.051	1	2
4	0.006	0	0
5	0.000	0	0

Table 1 provides probability, observed, and expected frequency distribution for the number of critical items identified by innocent subjects. The "N Expected" is the number of subjects that would correctly identify that number of critical items by chance alone (out of 40).

The table shows that 16 or 40% of the innocent subjects could not correctly identify any of the critical items, 13 subjects (32.5%) correctly identified one critical item, 10 subjects (25%) could correctly identify two critical items and 1 subject (.025%) correctly identified three of the critical items. These two frequency distributions (observed and expected based on chance) are not statistically significantly different (χ^2 = 2.25, p < .05).

A partial item analysis on the correctly chosen critical item for innocent subjects showed that of the 35 correct answers given by innocent subjects, 26% (9) occurred on question 1, 26% (9) occurred on question 2, 31% (11) occurred on question 3, 6% (2) occurred on question 4, and 6% (4) occurred on question 5. These figures can be found in Table 2.

Polygraph Examination Results

The decisions of the two examiners and the Lykken scores were all highly correlated, Table 3 shows the correlation matrix between the three evaluations.

The correlations between the original examiner and the blind evaluator and Lykken scores were .68 and .67, respectively. The correlation between the blind examiner and the Lykken scores was .67. All of the correlations were statistically significant with p < .0001.

Table 2. Frequency of correctly identified critical items for each question

Question #	# of Correct
1	9
2	9
3	11
4	2
5	5

Table 3. Inter-scoring system/evaluator matrix.

	Original Examiner	Blind Evaluator	Lykken Scores
Original Examiner	1.00	0.68	0.67
Blind Evaluator		1.00	0.67
Lykken Scores			1.00

Overall Accuracy

The accuracy levels for the original examiner, blind evaluator, and the Lykken scores are found in Figure 1.

Overall accuracy for the original examiner was 78%. This level of accuracy is highly statistically significant (χ^2 =24.2; p < .0001). The blind evaluator obtained an accuracy of 74%, also highly statistically significant (χ^2 = 18.05; p < .0001). The Lykken scores showed an overall accuracy rate of 76%, again highly statistically significant (χ^2 = 22.05, p < .0001).

Role

Figure 2 shows the accuracy levels of the original examiner, blind evaluator, and the Lykken scoring system for both guilty and innocent subjects. It shows that accuracy for the guilty subjects was 80% for the original examiner, 73% for the blind examiner, and 63% for the Lykken scores. Accuracy for the innocent subjects was 75% for both the original examiner and the blind evaluator and 90% for the Lykken scores.

Tables 4, 5 and 6 provide the χ^2 contingency tables for the decision of the original examiner, blind evaluator and Lykken scores.

Table 4 indicates that there is a significant association between role and the decision of the original examiner (χ^2 = 22.1, p < 0.0001) in that 32 of the 40 guilty subjects were correctly identified as DI with only 8 false negative errors (guilty subjects called NDI), while 30 of the 40 innocent subjects were correctly identified as NDI with 10 false positive errors (innocent subjects called DI).

Table 5 indicates that there is a significant association between role and the decision on the blind examiner (χ^2 = 16.21, p < 0.0001) in that 29 of the guilty subjects were correctly identified as DI with 11 false negatives and 30 innocent subjects were correctly identified as NDI with 10 false positive errors.

Table 6 indicates that there is a significant association between role and the decision made using the Lykken scoring system (χ^2 = 21.64, p < 0.0001) in that 25 of the guilty subjects were correctly identified as DI with 15 false negative errors and 36 of the innocent subjects were correctly identified as NDI with 4 false positive errors.

Figure 1. Percent correct decisions for three scorers

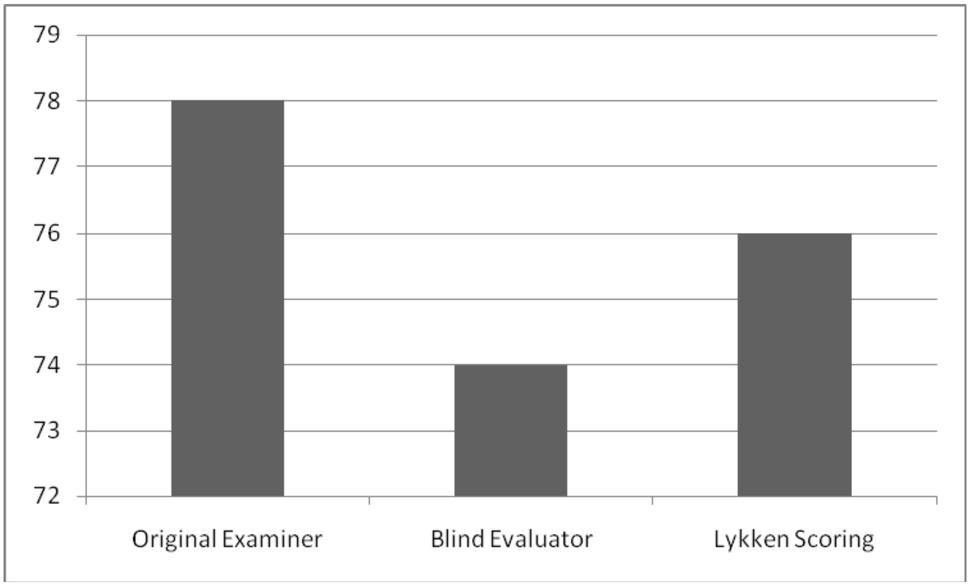


Figure 2. Percent correct decisions by scorer and guilt status

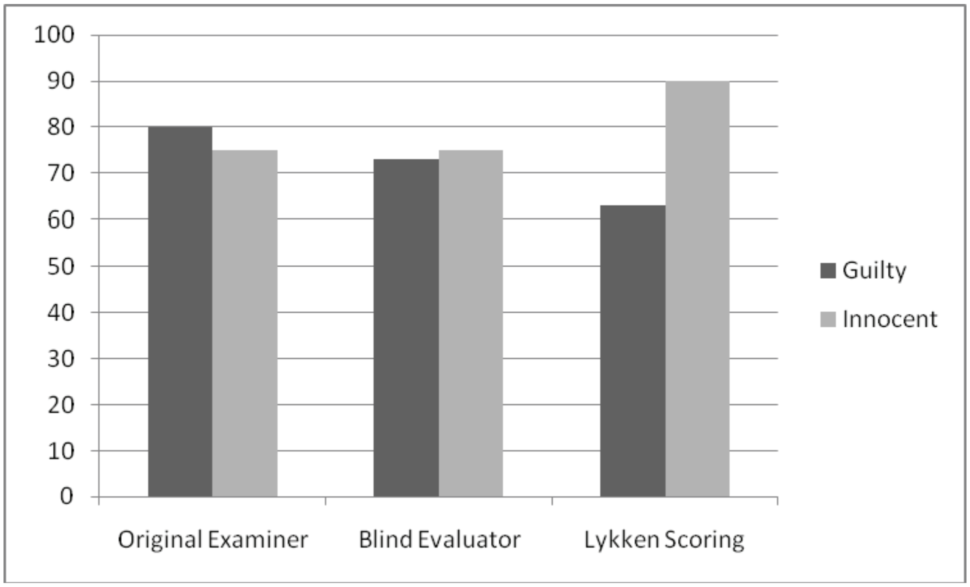


Table 4. Contingency table for role versus decision of the original examiner.

	DI	NDI	Total
Guilty	32	8	40
Innocent	10	30	40
Total	42	38	80

Table 5. Contingency table for role versus decision of the blind evaluator.

	DI	NDI	Total
Guilty	29	11	40
Innocent	10	30	40
Total	39	41	80

Table 6. Contingency table for role versus decision of the Lykken scoring system.

	DI	NDI	Total
Guilty	25	15	40
Innocent	4	36	40
Total	29	51	80

Stimulus Mode

The accuracy levels for the original examiner, blind evaluator, and the Lykken system for the visual and the aural conditions are found in Figure 3.

Accuracy for subjects in the visual condition was 83% for the original examiner, 78% for the blind evaluator, and 70% for the Lykken scores. In the aural condition, accuracy rates were 73%, 70%, and 83% for the original examiner, blind evaluator, and Lykken scores, respectively.

To compare the stimulus modes, one way to organize such a comparison is compare stimulus mode on correct decisions and stimulus mode on errors. The first analysis indicates whether or not the types of correct calls are distributed differently by stimulus mode. The second analysis examines whether or not the types of errors are distributed differently for the two stimulus modes.

Distribution of correct calls as a function of stimulus mode

A decision x stimulus mode chi-square statistic was calculated on correct decisions for the original examiner, blind evaluator and Lykken score. The χ^2 contingency tables for these analyses can be found in Tables 7, 8, and 9, respectively. No significant associations were found between the type of correct decision and the stimulus mode of question presentation for either the original examiner or the blind evaluator, or the Lykken scores on accuracy of decision (χ^2= 0.6091, p < 0.4351); χ^2= 2.0378; p < 0.1534; χ^2= 1.0651, p < .3020).

Figure 3. Percent of Correct Decisions

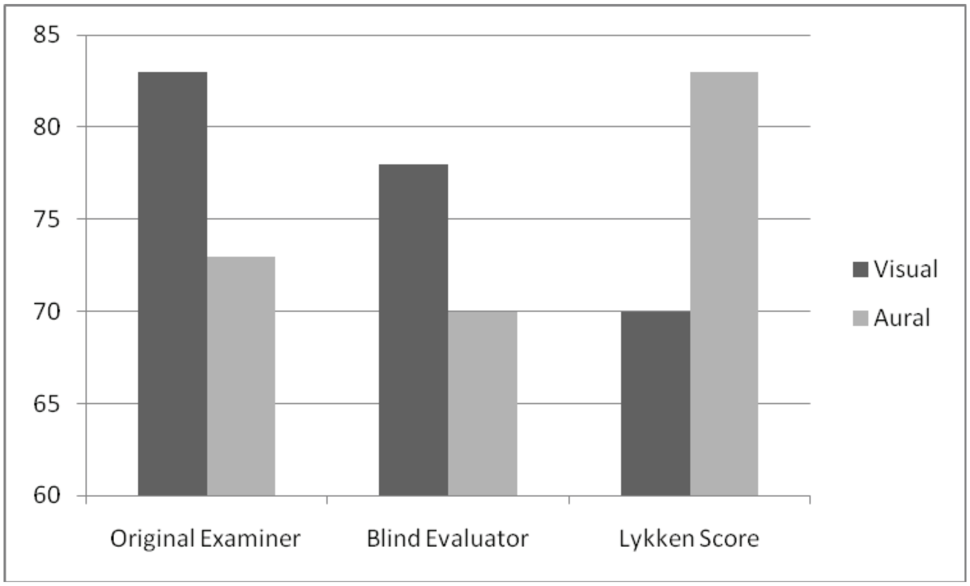


Table 7. Distribution of the correct original examiner decisions as a function of stimulus mode.

	True Negative	True Positive	Total
Aural	12	17	29
Visual	18	15	33
Total	30	32	62

Table 8. Distribution of the correct blind evaluator decisions as a function of stimulus mode.

	True Negative	True Positive	Total
Aural	11	17	28
Visual	19	12	31
Total	30	29	59

Table 9. Distribution of the correct Lykken scoring decisions as a function of stimulus mode.

	True Negative	True Positive	Total
Aural	17	16	33
Visual	19	9	28
Total	36	25	61

Distribution of error-type as a function of stimulus mode

Due to much smaller expected frequencies per cell, the association between the type of error in decisions and the stimulus mode was calculated using a Fisher's exact test. The contingency tables for error-type by stimulus mode for original examiner, blind evaluator, and the Lykken scores are found in Tables 10, 11 and 12, respectively.

No significant association was found between the role of subject and the stimulus mode of presentation for the original examiner or the Lykken scores on type of error (Fisher's exact test, two-tailed, p_2 = 0.1448; p_2 =.2451, respectively). There was a significant association found between the role of the subject and stimulus mode on error type for the blind evaluator (Fisher's exact test, two-tailed, p <.001).

Table 10. Distribution of the errors made by the original examiner as a function of stimulus mode.

	False Negative	False Positive	Total
Aural	3	8	11
Visual	5	2	7
Total	8	10	18

Table 11. Distribution of the errors made by the blind examiner as a function of stimulus mode.

	False Negative	False Positive	Total
Aural	3	9	12
Visual	8	1	9
Total	11	10	21

Table 12. Distribution of the errors made using the Lykken scoring system as a function of stimulus mode.

	False Negative	False Positive	Total
Aural	4	3	7
Visual	11	1	12
Total	15	4	19

Discussion

The stimulus mode in which the questions are presented appears to have very little influence on the rate of detection of the GKT. This was true for both subjective decisions of the original examiner and the blind evaluator as well as the more objective scoring system described by Lykken, when examining the accuracy of the decisions. These results support the earlier finding of Beijk, (1980).

It appears that the stimulus mode in which the question is presented also has little effect on the type of error in decision that is made at least for the original examiner and the more objective Lykken scoring system. The finding of a significant association between the type of error in decision and stimulus mode for the blind evaluator is somewhat puzzling. It is interesting to note that more false positive errors were made for subjects in the aural condition than in the visual condition. This relationship is reversed for false negative errors. More false negative errors were made for subjects in the visual condition than in the aural condition (See Tables 7, 8 and 9). This distribution of errors was found for all of the scores from the original examiner, the blind evaluator and the Lykken scores; however, the association was significant for the blind evaluator alone. Perhaps with a larger sample size this distribution might be significant for the original examiner and the Lykken system. There are a couple of possible explanations for this result.

It is possible that there is a type of confronter effect. The confronter sat next to the computer during the polygraph examination. Therefore, she could not see each alternative as it was presented. She was aware of the presentation of each alternative by the click sound of the event marker used by the examiner, but she could not see which alternative was presented. However, in the aural condition, the confronter could hear each alternative as it was presented. It is possible that the confronter inadvertently reacted when the critical item was presented. If the confronter did react strongly enough for the subject to respond this would only have affected innocent people in the aural condition as the confronter would not have known (for all subjects and all questions) when the critical item was presented. One possibility is that the confronter somehow elicited a larger response from innocent subjects when the critical item was presented in the aural condition.

This does not explain why there are more false negatives in the visual condition than in the aural condition, unless one makes a couple of assumptions about how the confronter affects the subjects. Perhaps the important element is that the confronter must know the following to have any effect: (a) that the subject is lying and (b) exactly when the subject is lying.

During the visual condition even though the confronter knew the subject would be lying, she was unaware of the exact moment that the subject was lying.

Another possible explanation for the higher false positive rate in the aural condition could be that the inflection in the voice of the person asking the questions could have caused the reactions. The tape of the questions was made by the examiner who ran the polygraph examination. Therefore, when the questions were being recorded, the examiner may have accidently, through some tone or inflection, made the critical item more salient such that an innocent person could detect the difference. However, this is not supported by the questionnaire data.

The results of the questionnaire data indicate that innocent subjects were not aware of the critical items at the time the questionnaire was given to them after the exam. The distribution of correctly guessed critical items was not statistically different from what would be predicted from chance alone. This would mean if the confronter has any effect on the innocent subjects in the aural condition, the subject was unaware or not conscious of the effect. The innocent subjects in the aural condition did not know or learn what the critical item was in the questions, and, therefore, the reasons underlying false positive errors are unknown.

Table 2 provides a distribution of the number of times innocent subjects correctly chose the critical item for all of the questions. Although it is apparent that questions 1 through 3 were more often correctly guessed than were 4 and 5, this does not provide much insight to the problem. To examine whether or not this distribution is unusual would require a complete item analysis of the questionnaire data. The purpose of the questionnaire data was to ensure that the guilty subjects could correctly identify the critical items and that the innocent subjects could not do so at a better than chance level. Both of these assurances were maintained.

The question of intonation is an empirical question. However, it is a question that this investigation was not designed to answer. Given that accuracy was not significantly better for aural versus visual presentation, clearly a way to negate the debate is to rely on more visual presentations during polygraph exams.

In spite of the results concerning the types of errors found in this study, the fact remains that there was no significant association between the stimulus mode of question presentation on accuracy. This interpretation does support a greater role for visual stimuli in the polygraph test. In spite of this, subsequent research must address the potential differences found in error type before questions may be presented visually during a polygraph exam.

An interesting observation gleaned from the results is the difference between the two subjective scoring systems and the more objective scoring system proposed by Lykken. It should be pointed out that the Lykken system is objective only in that it uses amplitude as the scoring criterion and attempts to apply a numerical scoring system. However, the cut-off point is arbitrary. Perhaps manipulating different cut-offs for the scores would prove to be a very informative exercise and should be done in subsequent research.

In this study the cut-off score of 6 resulted in a very high false negative rate. This is consistent with what is generally assumed about the GKT. Due to the probabilities involved, it is reasonable to assume that most of the errors should be false negatives. It should be very difficult to reach a false positive result due to chance alone. This investigation would support this notion as there were only 4 false positives and 15 false negatives when using the scoring system developed by Lykken.

Lastly, another interesting result of this investigation is the confronter issue. Although no firm conclusions may be stated, it is curious that the pilot studies for this investigation rendered very poor results (around chance) when using field polygraphs and field polygraph examiners. The decision was then made to use the Coulbourn equipment with one examiner and the confronter. After this decision was made, the accuracy for the investigation increased dramatically with overall accuracies ranging between 74% and 78%.

It is difficult to maintain that the equipment alone is responsible for this increase in accuracy. It is possible that the conductance recordings from the Coulbourn coupler were superior to the resistance recordings on the field polygraphs. Since this variable was not included in the design or even manipulated, no conclusion on this issue may be reached.

It is also possible that changing from multiple examiners to one examiner also played some role in the increase in accuracy. Even though the base rate of 50/50 was common knowledge to all four examiners, that did not necessarily relate to the base rate for any one examiner. There was no attempt to ensure that all of the examiners were given equal numbers of innocent and guilty subjects. This would have violated the random assignment to conditions since the schedules of the examiners varied from day-to-day and week to week.

Another consideration related to multiple examiners is that the examiners used during the pilot phase were all federally licensed polygraph examiners with no experience running GKTs in the field. The examiner who ran the GKT for this study is not a polygraph examiner, but does have some experience with a GKT in laboratory situations. Perhaps the more experienced examiners maintained a peak of tension bias as that is a technique they are familiar with and is most similar to the GKT that somehow interfered with the running of the GKT.

A related possibility is that even though all of the examiners were given scripts to follow for the pre-test and testing, simply by virtue of differences in experience in the field, the examiners would not necessarily handle the subjects in the same way. The switch from several examiners to one examiner would eliminate any differences due to variability between examiners. However, if this is true, there are certain implications on accuracy in the field, where there is no attempt to require examiners to treat all suspects the same and the base rates also vary by examiner. This would mean that overall accuracy in the field would suffer simply due to differences between examiners.

That would leave the confronter issue as a primary candidate for explaining the differences in accuracy rates. How the use of the confronter increases accuracy is an empirical question. One possible explanation is that it increases the accuracy of detecting the guilty subjects simply by making the subject more uncomfortable during a lie. It is logical that if accuracy improves for the guilty subjects, the accuracy for the innocent would also improve.

The confronter issue is certainly one that should be addressed in subsequent research. This issue could affect many aspects of physiological detection of deception. It has ramifications on future research, both theoretical and applied, as well as on how examinations may be conducted in the future.

References

Andreassi, J. L. (1989). *Psychophysiology*, 2nd Edition, Lawrence Erlbaum Associates: New Jersey.

Beijk, J., (1980). Experimental and procedural influences on differential electrodermal activity. *Psychophysiology*, 17(3), 274 - 278.

CODAS Software by DATACQ Inc., Release Level 4, 825 Sweitzer Ave., Akron, Ohio 44311.

Harvard Graphics. SPC Software Publishing Corp., Version 2.0. 1901 Landings Drive, Mountain View, CA 94039-7210.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43(6), 385 - 388.

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

Richardson, D. C., Carlton, B. L., & Dutton, D. (1990). Assessing systolic time intervals. Presented at the annual meeting of American Psychological Society, June 1990, Dallas, TX.

Appendix B

Examination Questions and Alternatives

GKT QUESTIONS

1. Do you know how entry was gained into the building? Was it …

   a. Climbing through an open window?
   b. Entering an unlocked door?
   c. Crowbarring the door?
   d. Breaking the window?
   e. Cutting the padlock on the door?
   f. Climbing through an attic vent?

2. Do you know what the sign read on the door to the room that was entered? Was it …

   a. Cashier?
   b. Receptionist?
   c. Director?
   d. Paymaster?
   e. Supply?
   f. Secretary?

3. Do you know how the victim was killed? Was it…

   a. Choked with a scarf?
   b. Shot with a pistol?
   c. Stabbed with a knife?
   d. Struck over the head?
   e. Drowned in the bath tub?
   f. Hit with a car?

4. Do you know what was removed from the body? Was it…

   a. Money?
   b. Dog Tags?
   c. Watch?
   d. Pocket knife?
   e. Ring?
   f. Keys?

5. In the room entered, there were two boxes with names on them. Do you know what name was on the bottom box? Was it…

   a. William?
   b. Raymond?
   c. Gordon?
   d. Charles?
   e. Matthew?
   f. Steve?

The Critical Item is in **bold** print.

Example of GKT Question Given to All Subjects

Do you know what kind of shoes that man was wearing? Were they....

a. Tennis Shoes?
b. Combat Boots?
c. Loafers?
d. Hiking Shoes?
e. Dress Shoes?

The question was given to subjects via tape recording (aural condition) or a computer monitor (visual condition). Subjects were requested to respond to the alternatives just as if it was an actual test question. This question was not significant to any of the subjects. The question was not related to the mock crime witnessed by the guilty subjects and the subjects were informed of this fact.