Polygraph

VOLUME 30 2001 NUMBER 4

Contents

Published Quarterly

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

Polygraph

Subscription Information: *Polygraph* is published quarterly by the American Polygraph Association. Advertising and Editorial Address is P.O. Box 10342, Ft. Jackson, SC 29207 (USA). Subscription Rates: One year $80.00 (domestic), $100.00 (foreign). Change of address: APA National Office, P.O. Box 8037, Chattanooga, TN 37414-0037.
The Sheridan Press

Editorial

Dean A. Pollina

I would like to begin my tenure as editor of *Polygraph* and the American Polygraph Association publications with an expression of gratitude to Donald Krapohl, the previous editor-in-chief. Throughout his extremely illustrious career, Don has devoted his time and his various efforts to improve the field, specifically in the area of polygraph research. I'm very confident that the journal will continue to benefit from Don's expertise in his new role as associate editor. Don has profoundly influenced the polygraph profession and polygraph researchers. I'm truly grateful that I will continue to benefit from Don's skill and ability in the years to come. I would also like to express similar gratitude to the other prominent polygraph professionals on the editorial board. I look forward to working with you all.

The *Polygraph* Journal has been very unique in scope, uniting disparate areas of human endeavor including scientific research and technology, philosophy, the legal profession, law enforcement, and government. It is also evident that this diverse readership is the journal's strength, drawing together creative and thoughtful individuals from these various professions. We will continue this tradition of diversity and encourage the submission of research and theoretical papers, literature reviews and meta-analyses, bibliographies, legal briefs, technical reports, and case histories. Any serious scholarly work that could influence theoretical or practical aspects of the psychophysiological detection of deception, the polygraph profession, or readership will be given consideration. We would also welcome questions from prospective contributors at any point during their manuscript's production. In keeping with previous policies regarding editorial decisions, the most important considerations as to the acceptance of a manuscript include its interest to a segment of the readership, clarity, and accuracy.

We will continue to encourage writers to send electronic versions of their manuscripts in addition to paper copies to facilitate the editorial and peer review process. We will also be available for assistance via email correspondence to offer suggestions and assistance. We request that all manuscripts submitted to the editorial staff conform to the standards of style recommended by the *Publication Manual of the American Psychological Association.*

What will the *Polygraph* of the future look like? Questions about humankind's representations of reality and truth have formed a cornerstone of intellectual preoccupation and thought throughout the ages. Despite the derision of the polygraph profession's most ardent critics, history will show that the polygraph profession and the journal it created have contributed to the understanding of these representations and to a better society. I look forward to the continued success of this journal and its readership, and I am honored to serve as your editor. I also look forward to any comments, issues, or concerns that you might have.

My thoughts and most fervent prayers go out to the victims of the September 11, 2001 terrorist attacks, and the brave men and women who continue to work so hard in the recovery effort. I've heard many comments about this being "the first attack on the mainland by a foreign power." This statement is, of course, not true. In fact, in August 1776, a foreign power scored a major victory against the newly independent forces of the United States. In this Battle of Long Island, American troops were defeated. On the night of August 29-30, George Washington took command and evacuated his force to Manhattan – to the very same streets where the world trade center once stood. Later, he would defeat these foreign invaders in a war no other world power thought he could possibly win. But they had all underestimated the Americans' ingenuity, resolve, and belief in the righteousness of their cause. In George Washington's century, an amazing revolution was sparked. On September 11, we learned that we must and we will continue to "fight the good fight."

An Assessment of the Total Chart Minutes Concept with Field Data

Donald J. Krapohl[1, 2]

Abstract

The Total Chart Minutes Concept (Backster, 1963a) is a specialized habituation model for polygraph testing. According to the model, each of the three traditional channels of polygraph data are more diagnostic at certain periods during polygraph testing than in others, and no two channels have the same exact periods of good and poor productivity. Backster created what he called "dependability curves" to represent the diagnosticity of each of the channels as a function of time. In the present project, data from six separate studies were examined for evidence of conformity with the dependability curves. No evidence was found. It was concluded that the Total Chart Minutes Concept is not predictive of habituation, and provides no benefit to numerical scorers.

Introduction

Cleve Backster (1963a) introduced the polygraph community to the Total Chart Minutes Concept (TCMC). The TCMC was a notion intended to reconcile the differences in emphasis placed on individual tracings by different schools of thought. There had been disagreements among the various writers and practitioners of that time whether the respiration, electrodermal, or cardiovascular channel contained the most diagnostic information, a minor controversy that continues even today. According to Backster, the TCMC referred "to the accumulation of the units of time that a suspect has been actually undergoing the series of tests making up his total examination on one particular day" (p77). Based on five years of study, he reported that the respiration, electrodermal, and cardiovascular tracings of the traditional polygraph recording are more diagnostic at different periods of the testing time, and he published what he called "relative tracing dependability curves" to characterize this relationship between time and tracing usefulness. Backster asserted that the TCMC was responsible for the differences in viewpoints among polygraph practitioners, since at some points in time each would actually be correct about which channel was best.

Backster's dependability curves are very interesting for a number of reasons, not the least of which is they did not follow habituation patterns expected for physiological data. For at least two of the traditional polygraph channels, strongest reactivity did not occur until several minutes of testing had passed. If the dependability curves are accurate, they would reveal a heretofore unknown psychophysiological response habituation patterns, and as such, they warrant special attention.

While no data were presented in the Backster concept paper, he produced dependability curves with time plotted on the x axis, and effectiveness on the y axis. The time dimension is graduated in minutes, from 0 to 32. The effectiveness axis is divided into four sectors, which from bottom to top of the figure, were marked as poor, fair, good, and excellent. The dependability curves took the shape of graceful arcs, and purportedly revealed how the polygraph data behaved for each channel

[1] Department of Defense Polygraph Institute
[2] To whom correspondence should be addressed at Department of Defense Polygraph Institute, Fort Jackson, SC 29207

over time. The following is a description of the dependability curves for the three channels.

Respiration begins in the good range immediately, peaking at the height of the excellent range at about 4 or 5 minutes of chart time, and falling off into the fair range at the 12- to 16-minute range. Deceptive and nondeceptive curves are plotted separately, and the deceptive curves remain higher than the nondeceptive curves at all times. Backster attributes the shape of the respiration dependability curve to the examinee's gradual accommodation to the polygraph setting and procedures, with a corresponding reduction in differential responding after the optimum period between 2 and 10 minutes.

The cardiograph dependability curve begins near the junction of poor and fair ranges, rising up to the good range after three or four minutes, hovering in excellent range in the 7- to 14-minute period, and thereafter falling off. As with the respiration tracing, the deceptive and nondeceptive curves are plotted separately. The nondeceptive cardiograph tracing is slightly better than the deceptive cardiograph tracing from 0 to 7 minutes, with the deceptive tracing being much more useful from there to the 32-minute line. The reason for the late entry for the cardiograph usefulness, according to Backster, is that the examinee's heightened emotional state during the early phases must settle down in order to allow the polygraph examiner to distinguish extraneous responses from diagnostic responses. In other words, the signal of interest becomes more apparent as the noise decreases with examinee accommodation to the testing environment. This explanation is in contrast to that proffered for the respiration dependability curve, where the heightened emotional state early in the testing period purportedly improves the diagnosticity of that channel.

The dependability curve of the electrodermal activity starts near the bottom of the poor range, is considered fair after three to five minutes, is good at 6 or 7 minutes, excellent at about 10 minutes, and falls off after 16 to 20 minutes. Electrodermal tracings of nondeceptive examinees are better than those of deceptive examinees for the first 11 minutes, with the opposite relationship after that time. Again, Backster reports that the poorer early performance of the EDA is because it is during this time that the subject's adjustment to the setting is taking place.

If the phenomena Backster describes with his dependability curves are reliable, his concept could be used to improve accuracy of manual or automated chart scoring. Tracings that are expected to be poor for some section of time could be ignored in scoring systems, or the tracing might not be recorded at all. Weighting of scores, by time and channel, could exploit the temporal usefulness of each tracing. Testing procedures would be modified, to capture the optimum recording periods for the data. No published scoring system has yet taken advantage of the TCMC.

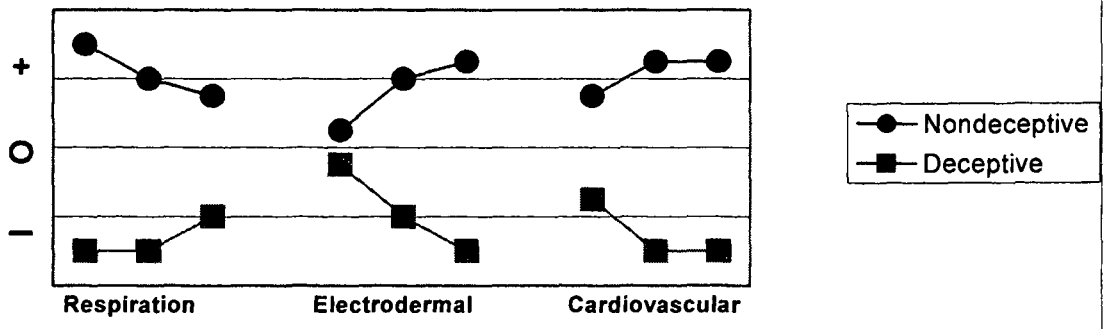
Unfortunately, little in known about how the dependability curves were developed. The source and descriptions of the sample cases were not provided in Backster's concept paper, nor the statistical treatments. Equally problematic for the TCMC is that no one since 1963 has provided convincing data to support the TCMC. This has not prevented the TCMC from being widely taught in polygraph schools for 40 years, however, nor from being cited in texts (Matte, 1996).

The TCMC's staying power in the absence of data is noteworthy. The present writer set out to replicate the Backster methodology in hopes of confirming his findings, but given the insufficient detail in the original report, a new paradigm was necessarily developed. It began with this assumption: if the TCMC has some discernible affect on the interpretability of polygraph channels over time, one could reasonably expect the TCMC to influence manual 7-position numerical scoring of polygraph cases. For example, the TCMC dependability curve for the electrodermal channel indicates that it is least useful in the first several minutes of physiological recording. Therefore, the TCMC dependability curve would predict that the sum of the scores assigned to the electrodermal channel on the first chart would be much closer to zero than those in the final chart. Similarly, the pneumograph would perform well in the first chart, and do less well as time passes, if the dependability curve were

accurate. The trend for the cardiograph is less clear, showing itself to be somewhere between the respiration and electrodermal activity in usefulness (See Figure 1 for a profile suggested by the TCMC for charts 1 – 3). If these are the projected trends according to the TCMC, it would be possible to test those predictions against the scores from live cases.

The average chart using conventional polygraph examination techniques lasts about four or five minutes. A typical examination entails the collection of three charts, and perhaps a stimulation test of about 2 minutes. These charts would produce from 12 to 17 minutes of chart time, adequate to look for the trends characterized by the dependability curves. (See Fig 1.) While Backster suggests that there may be as much as four minutes of variability in the dependability curves, over a sufficiently large sample this variability would average so that the trends would still be apparent, especially for those tracings that pass through the poor to excellent stages within the testing time frame. The goal of the present paper is to compare the profile of channel diagnosticity indicated by the TCMC dependability curves against various samples of polygraph case scorings.

Figure 1. Suggested profile of average scores for each channel for nondeceptive and deceptive cases for charts 1, 2 and 3 according to the Backster Total Chart Minutes Concept (1963a)



Method

Cases

Six separate sets of data were used in the present undertaking. The data consisted of 7-position numerical scorings of confirmed field Comparison Question Technique (CQT) polygraph cases. Scores were summed within channel and chart, then averaged by group, deceptive or nondeceptive. The resulting values were then plotted. Below are the sources of the data.

Sample 1. From Capps & Ansley (1992). One hundred cases were randomly selected from an archive of cases at a US Department of Defense agency. The field cases had been collected from an APA accredited polygraph school that offered polygraph services to the private and law enforcement sector. Forty-eight cases were nondeceptive, and 52 were deceptive. One experienced polygraph examiner performed all of the scorings, using the 7-position scoring system, and US federal scoring rules.

Sample 2. From Blackwell (1999). One hundred confirmed field cases conducted in a Zone Comparison Technique format were scored blindly by three federal polygraph examiners using the 7-position scoring system and federal scoring rules. Case sampling was stratified due to a limited number of confirmed nondeceptive cases. All available nondeceptive cases were used, totaling 35, and the remaining 65 deceptive cases were randomly selected from a database of 400 cases.

Sample 3. From Blackwell (1999). One hundred confirmed field cases conducted in the Modified General Question Technique (MGQT) format were scored blindly by three federal polygraph examiners using the 7-position scoring system and federal scoring rules. Case sampling was stratified due to a limited number of confirmed nondeceptive cases. All available nondeceptive cases were used, totaling 20, and the remaining 80 deceptive cases were randomly selected from the larger database.

Sample 4. From Capps & Ansley (1992). Forty field cases were independently scored blindly by 11 experienced polygraph examiners. Twenty-three of the cases were confirmed deceptive, and the remaining 17 were confirmed nondeceptive. The strategy for case selection was to take the first 20 male and first 20 female cases from a larger sample of 100 cases used by Franz (1989) in another study. Franz had selected 50 confirmed deceptive and 50 confirmed nondeceptive cases from the files of Argenbright Polygraph, Inc, of Atlanta, Georgia. The charts had been collected on analog polygraphs by 10 examiners between 1982 and 1989. For Sample 4, score sheets were not available. Channel scores were derived from recalculations of the data in Table 4 of the Capps & Ansley report (pg 307). However, there was no breakout for deceptive and nondeceptive data in that report. Therefore, for Sample 4 the average of the absolute values were combined to produce a single value for each chart by channel.

Sample 5. From Matte (paper in progress). Using 123 confirmed deceptive cases provided by the Commonwealth of Virginia Department of State Police, Matte had six experienced law enforcement examiners score the cases using three different scoring methods. However, there was only one scoring for each method for each case. The scorings from the Backster scoring system were chosen for this paper. Criteria for selection of the cases were that they be confirmed deceptive, conducted between January 1, 1998 and September 1, 1999, and be in the Backster "You Phase" single-issue ZCT format. There were no nondeceptive cases included in this sample.

Sample 6. Krapohl & McManus (1999). Three hundred confirmed ZCT cases were drawn at random from the DoDPI confirmed case database such that half of the cases were deceptive, and the other half nondeceptive. Using features described by Kircher & Raskin (1988), ratios were calculated for measurements of relevant question features divided by comparison question features. Those ratios were ordered by size, and mapped over the traditional 7-position scoring system such that each 7-position score (from –3 to +3) received one-seventh of the ratios. See the original paper for greater detail on this scoring system.

Results

Respiration

The TCMC suggests that the respiration channel provides more diagnostic information at the early stages of testing, and that its productivity begins to decline after 4-6 minutes. Figure 2 below is the profile suggested by the TCMC, followed by the profiles of respiration pattern produced from the averages of scores for charts 1, 2 and 3 (Figures 3 – 8). The profiles generated from the six samples show considerable variety, not corresponding with any particular profile, and match the TCMC model no more often than any other profile. This suggests that, if there were any habituation underpinnings corresponding with the TCMC, they are not dramatic or robust. Moreover, the average variation of totals for the respiration channel across charts is but a fraction of a single point, making any effect of the TCMC on respiration scores negligible. Therefore, even if the TCMC were a true effect for the respiration channel, it has virtually no practical value to human scorers of polygraph data.

Figure 2. Suggested Profile of Average Respiration Scores by Channel for Nondeceptive and Deceptive Cases for Charts 1, 2 and 3 from the Backster Total Chart Minutes Concept (1963a)

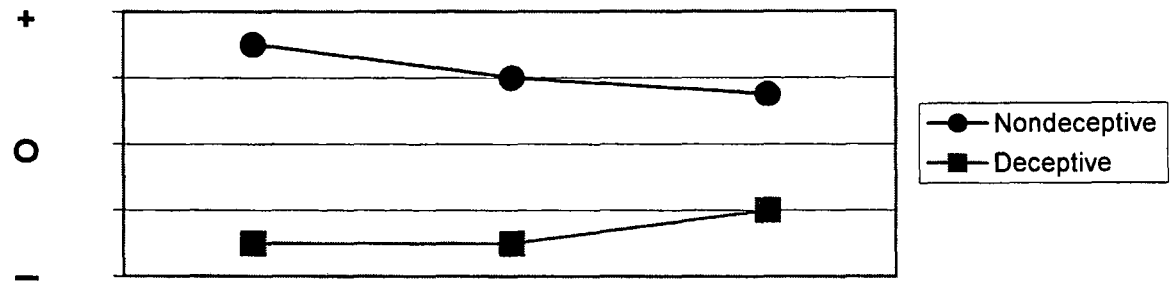


Figure 3. Sample 1: Average Respiration Scores for Deceptive and Nondeceptive Cases by Chart with the ZCT Format (from Capps & Ansley, 1992)

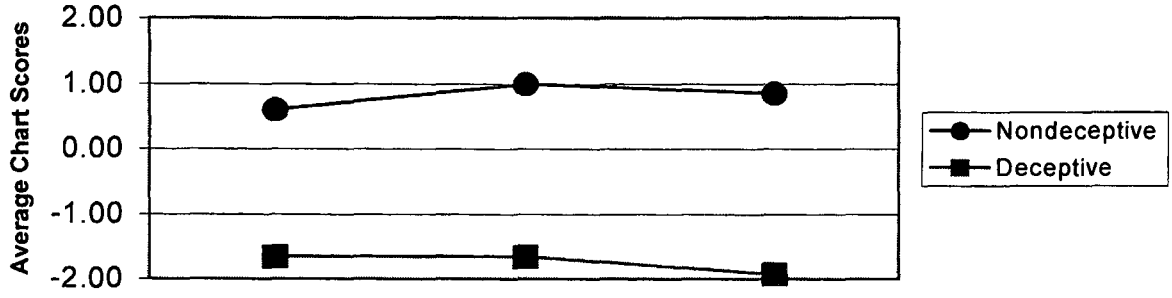


Figure 4. Sample 2: Average Respiration Scores for Deceptive and Nondeceptive Cases by Chart for the ZCT Format (from Blackwell, 1999)

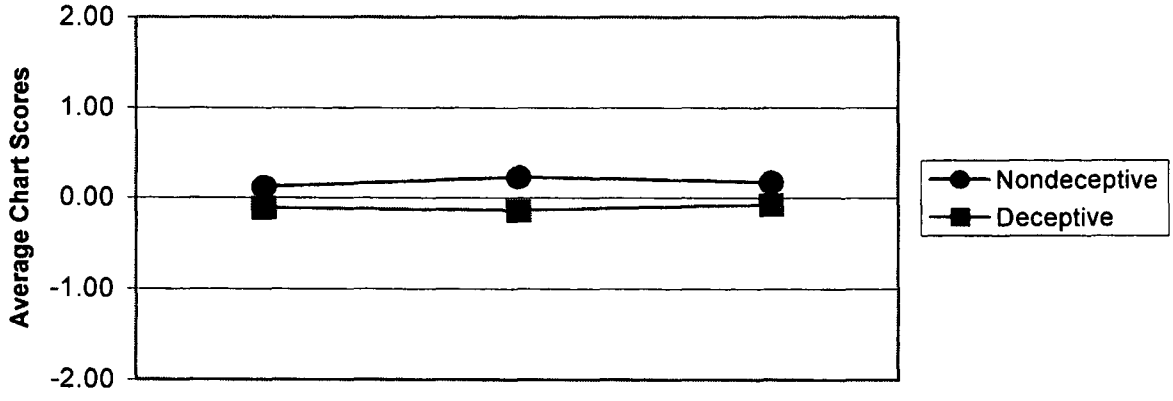


Figure 5. Sample 3: Average Respiration Scores for Deceptive and Nondeceptive Cases by Chart for the MGQT Format (from Blackwell, 1999)

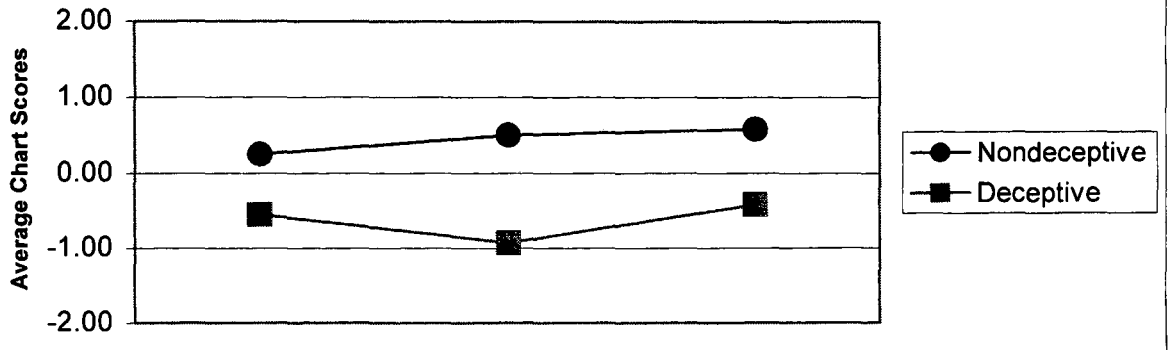


Figure 6. Sample 4: Average Absolute Respiration Scores for Deceptive and Nondeceptive Cases Combined, by Chart, for the ZCT Format (from Capps & Ansley, 1992)

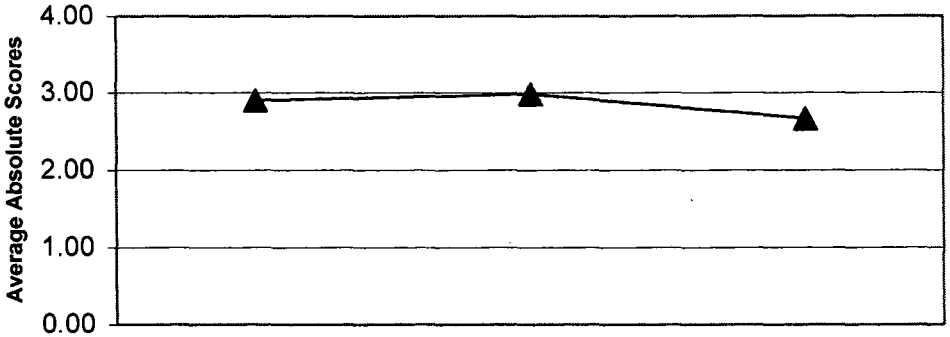


Figure 7. Sample 5: Average Respiration Scores for Deceptive Cases by Chart for the ZCT Format (Matte, unpublished data)

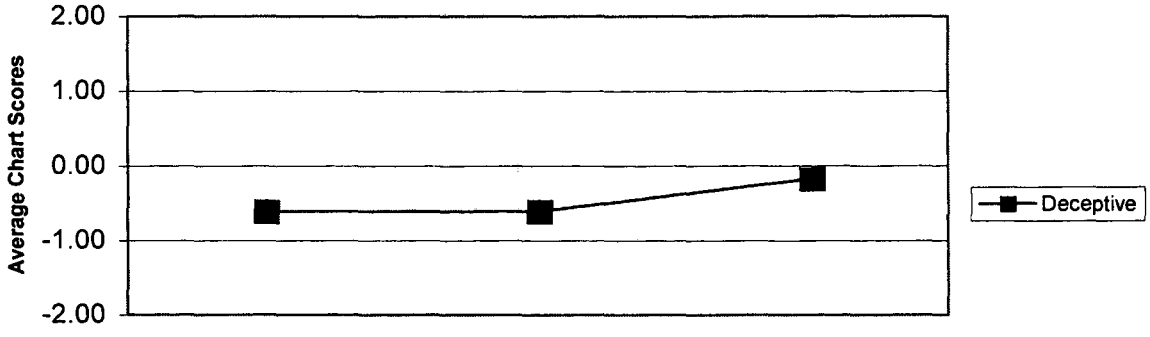
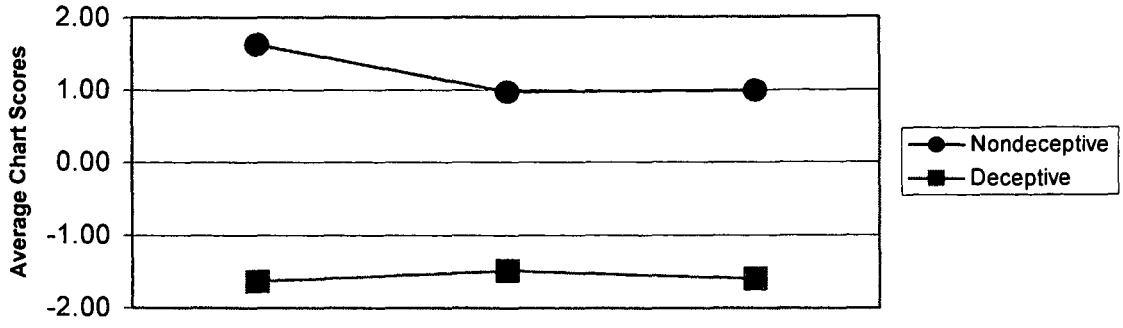


Figure 8. Sample 6: Average Respiration Scores for Deceptive and Nondeceptive Cases by Chart with the ZCT Format (from Krapohl & McManus, 1999)



Electrodermal

The TCMC dependability curve for the electrodermal channel suggests poor diagnosticity for that channel until late in the testing phase. Below are the profiles of electrodermal pattern produced from the averages of scores for charts 1, 2 and 3, with the profile suggested by the TCMC positioned first (Figures 9 – 15).

As noted with the respiration data signs of the TCMC dependability curve are not in evidence with the electrodermal channel. Average scores by chart vary up to one point, but no more, with common curve seen across data pools. There was no hint of the predicted TCMC effect in these data

Figure 9. Suggested profile of average electrodermal scores by channel for nondeceptive and deceptive cases for charts 1, 2 and 3 from Backster Total Chart Minutes Concept (1963a)

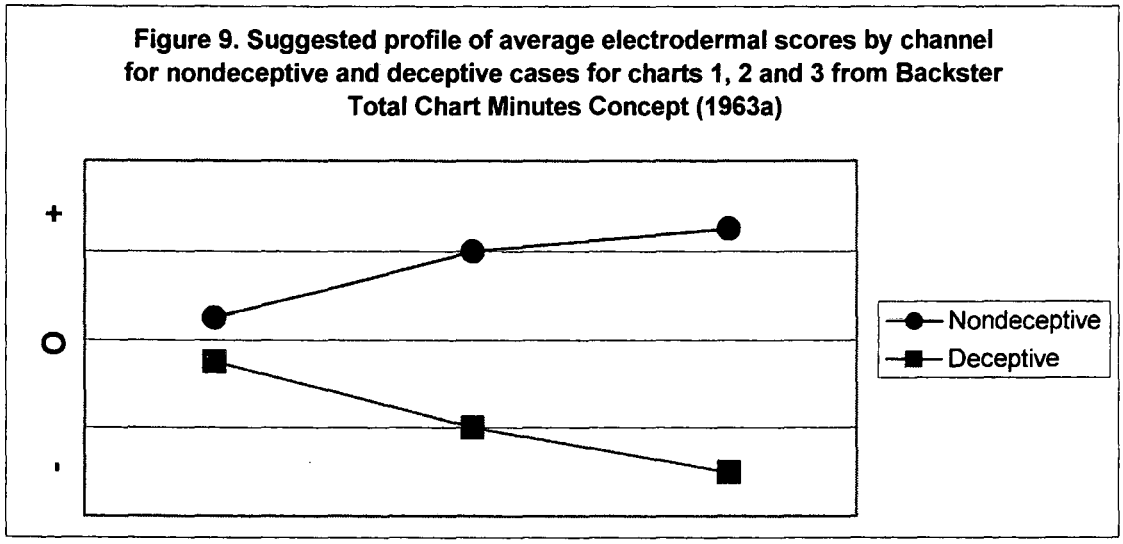


Figure 10. Sample 1: Average Electrodermal Scores for Deceptive and Nondeceptive Cases by Chart with the ZCT Format (from Capps & Ansley, 1992)

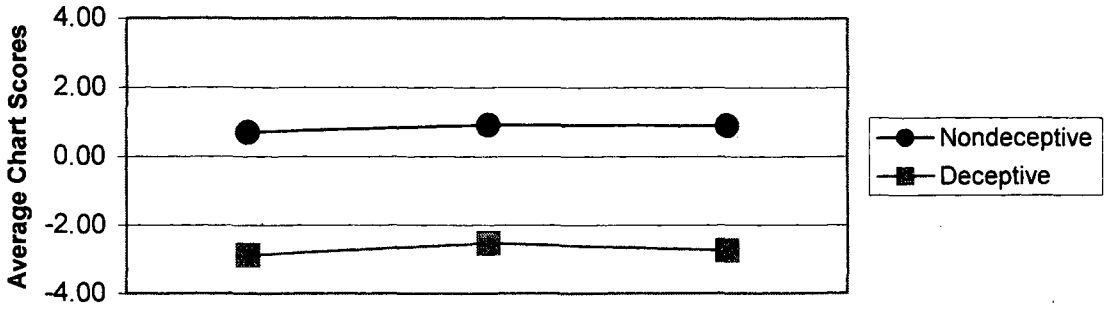


Figure 11. Sample 2: Average Electrodermal Scores for Deceptive and Nondeceptive Cases by Chart for the ZCT Format (from Blackwell, 1999)

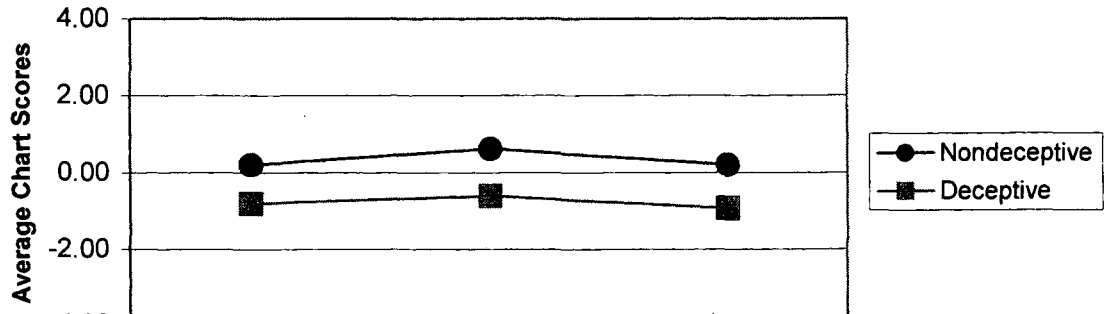


Figure 12. Sample 3: Average Electrodermal Scores for Deceptive and Nondeceptive Cases by Chart for the MGQT Format (from Blackwell, 1999)

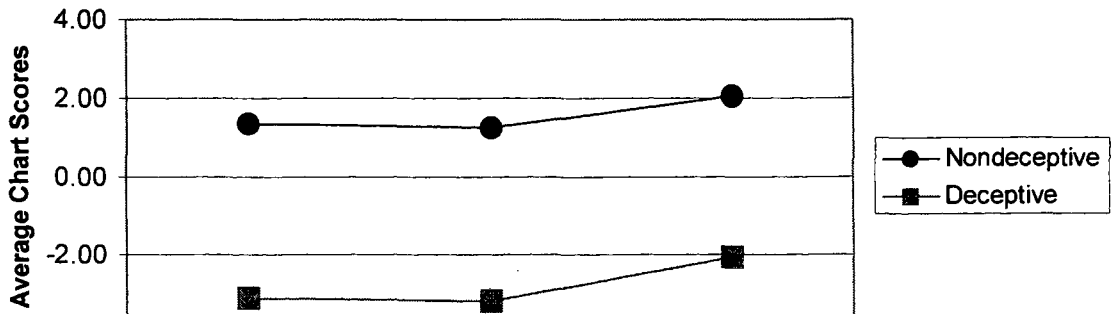


Figure 13. Sample 4: Average Absolute Electrodermal Scores for Deceptive and Nondeceptive Cases Combined, by Chart, for the ZCT Format (from Capps & Ansley, 1992)

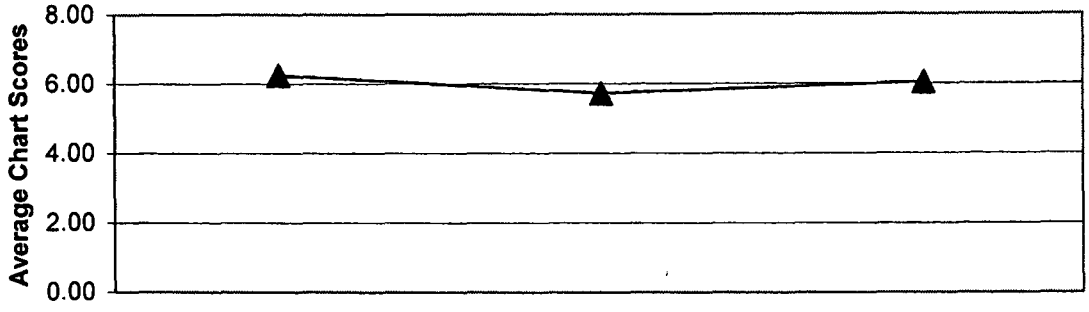


Figure 14. Sample 5: Average Electrodermal Scores for Deceptive Cases by Chart for the ZCT Format (Matte, unpublished data)

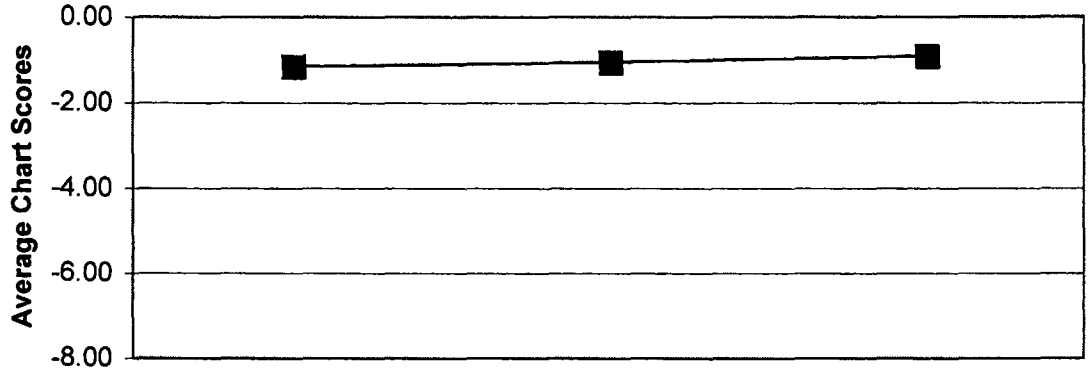
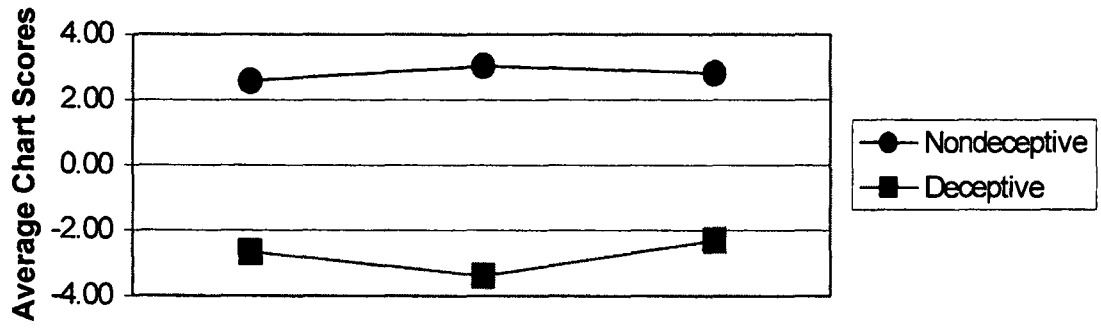


Figure 15. Sample 6: Average Electrodermal Scores for Deceptive and Nondeceptive Cases by Chart with the ZCT Format (from Krapohl & McManus, 1999)



Cardiovascular

The TCMC dependability curve shows that the cardiograph will provide its best diagnostic information in the 7- to 14-minute range. This period corresponds roughly between the end of the second chart, into the third chart. Therefore, we can expect that the third, and perhaps the second polygraph chart will provide scores more divergent from zero than does the first chart. Below is the profile of cardiograph pattern from the TCMC dependability curve, followed by those produced from the averages of scores for charts 1, 2 and 3 (Figures 16 – 22).

There is a poor correspondence between the suggested TCMC dependability curve profile and the field data. While inter-chart variability in scores was small, there were isolated changes between charts of one to two points for samples 1, 3, and 6. Most inter-chart variation was less than one point.

Figure 16. Suggested profile of average cardiograph scores by channel for nondeceptive and deceptive cases for charts 1, 2 and 3 from Backster Total Chart Minutes Concept (1963a)

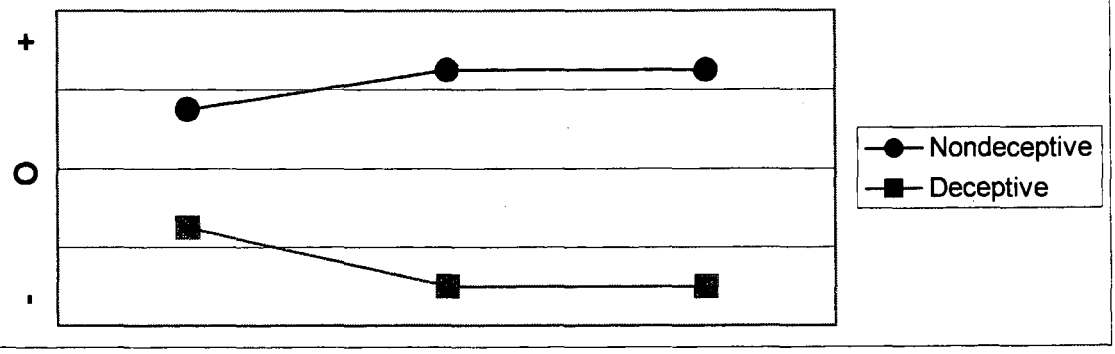


Figure 17. Sample 1: Average Cardiograph Scores for Deceptive and Nondeceptive Cases by Chart with the ZCT Format (from Capps & Ansley, 1992)

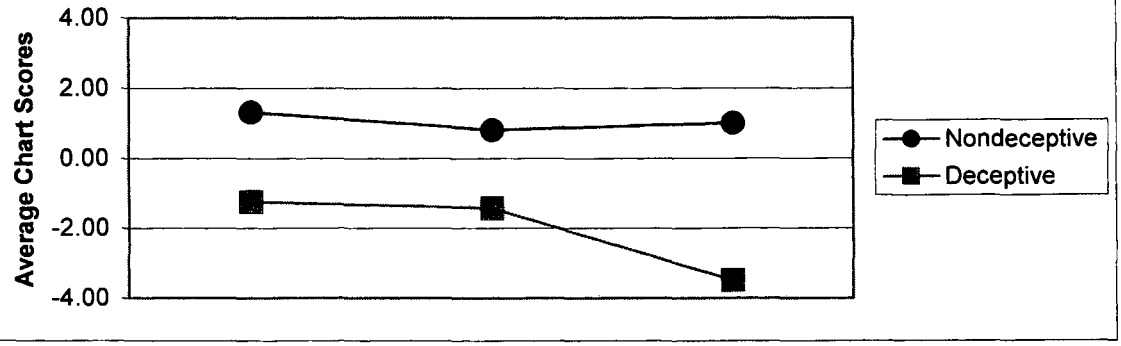


Figure 18. Sample 2: Average Cardiograph Scores for Deceptive and Nondeceptive Cases by Chart for the ZCT Format (from Blackwell, 1999)

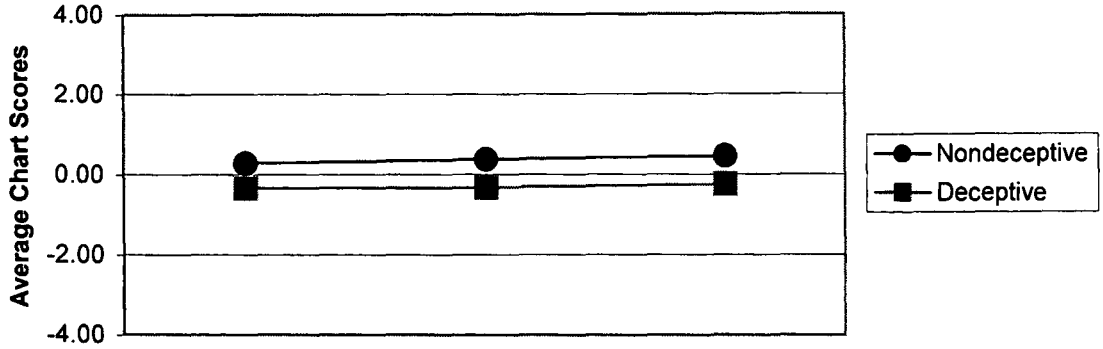


Figure 19. Sample 3: Average Cardiograph Scores for Deceptive and Nondeceptive Cases by Chart for the MGQT Format (from Blackwell, 1999)

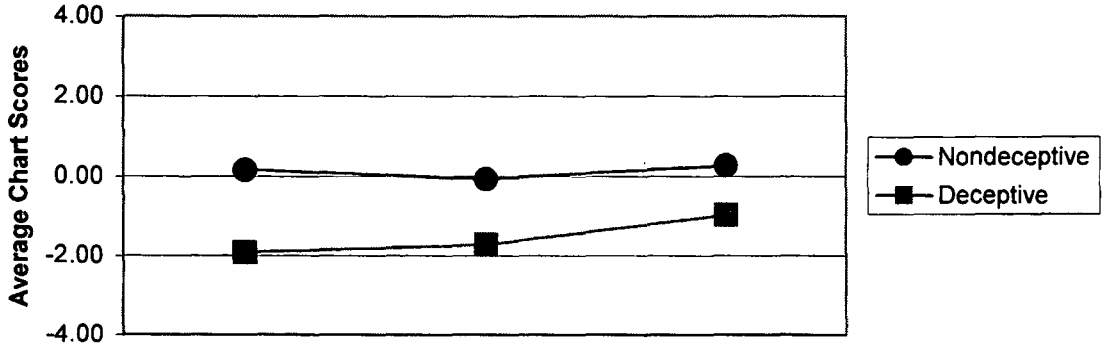
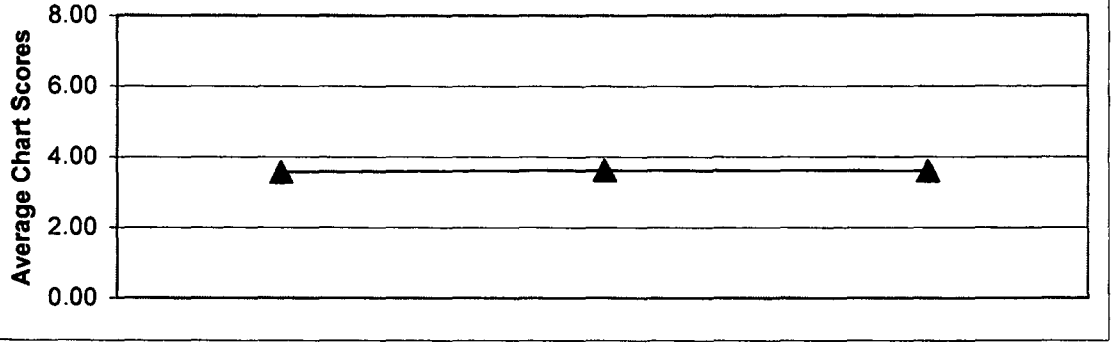
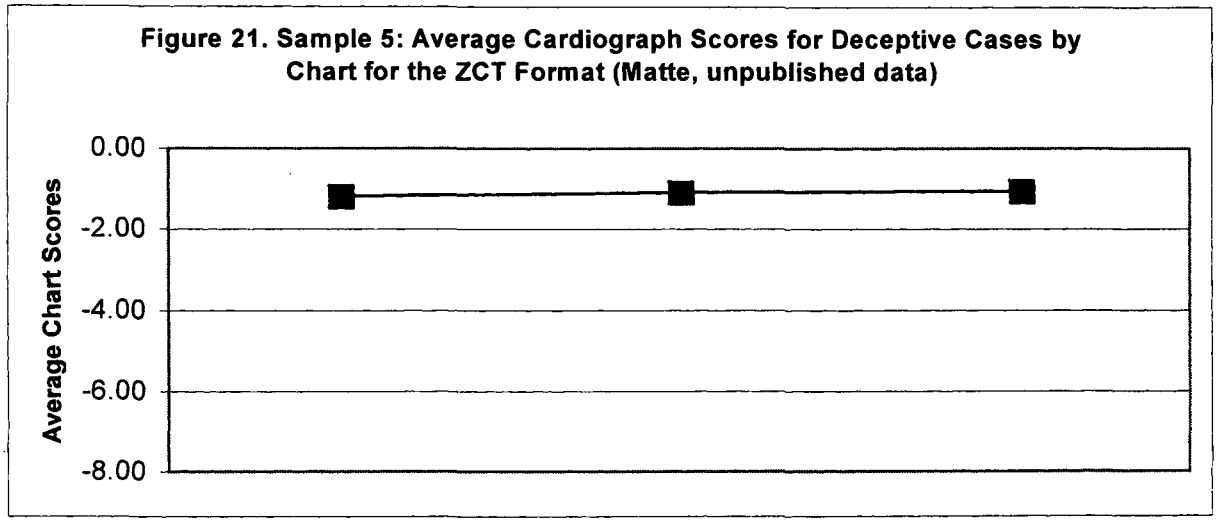
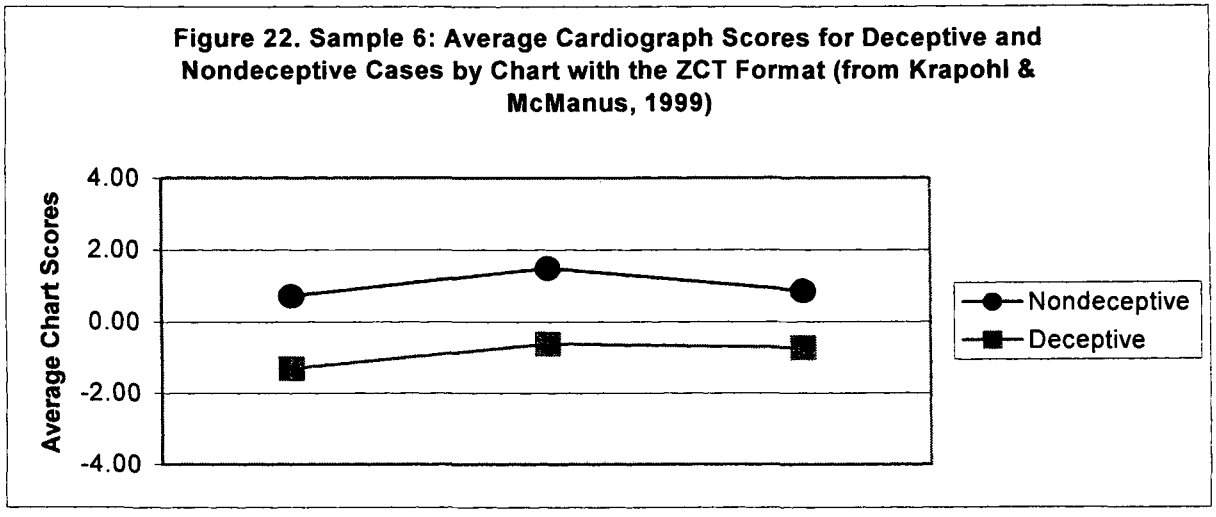


Figure 20. Sample 4: Average Absolute Cardiograph Scores for Deceptive and Nondeceptive Cases Combined, by Chart, for the ZCT Format (from Capps & Ansley, 1992)







Discussion

The striking lack of agreement between the various independent field data scoring profiles and the TCMC dependability curves is compelling evidence that the TCMC has little foundation in reality. Until supporting data are found, the TCMC should viewed with justifiable skepticism.

One possible criticism of the present project is that the data sets or scoring systems used in this study may have been different in some significant way from those on which the TCMC is based. This may be so, but is unknowable given the lack of information in the original concept paper. Backster did not specify whether the TCMC was technique-specific. However, Backster at that time was promoting his Zone Comparison Technique (Backster, 1960, 1962, 1963b), and one might reasonably expect that the TCMC would apply at least to that technique. In addition, he alludes to the relevant-irrelevant and peak of tension techniques in his concept paper. It would appear implicit from the article that the TCMC trend is a characteristic of all polygraph testing techniques.

As for the appropriateness of the scoring system used in the present paper to uncover the dependability curves, it would be

hard to envision any scoring system that, when used consistently across charts, would obscure a trend as robust as is asserted in the TCMC paper. All that can be stated from the present data is that contemporary 7-position scoring techniques failed to uncover evidence of the TCMC dependability curves with confirmed field cases. As such, the TCMC appears not to be useful to enhance scoring or testing methodologies, nor as a model of habituation during polygraph testing.

This effort was not without a brighter side, however. Two inter-esting trends were uncovered that are worthy of comment. The first deals with the question that prompted Backster to propose the TCMC originally: Which polygraph channel really *is* most diagnostic? One way to answer this question is to look at average channel scores per case, to see which channel showed the greatest separation of scores between deceptive and nondeceptive cases. In other words, if the average score for a given channel for the nondeceptive cases was not much different from the average score for that channel for the deceptive cases, one might conclude that channel provides little in the way of discriminative information for classifying a case as deceptive or nondeceptive. Greater separation between average nondeceptive scores and deceptive scores for a channel might signal that that channel was contributing more diagnostic information toward correct conclusions. Turning back to our six data samples again, three contained traditional 7-position scorings of field cases where deceptive and nondeceptive case data were broken out separately: Samples 1, 2, and 3. A post hoc analysis of those cases was conducted to address the question of channel diagnosticity.

There were 103 nondeceptive and 197 deceptive cases available in the three samples. Using weighted averages, it was determined that the distance between the average respiration score for the nondeceptive and for the deceptive was 3.81 per case. The corresponding differences for the electrodermal channel and cardiograph channel between the nondeceptive and deceptive cases were 9.00 and 5.52, respectively. Put into percentages, the electrodermal accounted for 49.1% of the total difference between deceptive case and nondeceptive case scores, with the cardiograph coming in at 30.1%, and respiration at 20.8%. These data are in concert with most research that shows that the electrodermal channel, on average, provides the most diagnostic information with both field and laboratory cases, and respiration the least. This trend has been confirmed with the relatively recent development of automated scoring algorithms, which, on average, weight the electrodermal channel in order to improve accuracy.

The second meaningful finding in the present data deals with the differences in the pattern of responding between deceptive and non-deceptive examinees. A cursory look at the graphs containing data from both deceptive and nondeceptive cases shows that, on average, nondeceptive examinees have total scores closer to zero than do deceptive examinees. Samples 1, 2, and 3 all show this trend. Sample 6, which contains deceptive and nondeceptive cases, has scoring rules that adjust for the nonsymmetrical response patterns, and therefore would not be expected to show the asymmetry in total scores that hand scorings do. The asymmetry in response patterns found here has been reported by other researchers previously (Franz, 1989; Krapohl, 1998; Raskin, Kircher, Honts, & Horowitz, 1988). The mounting evidence shows that scoring systems should have cutting scores that are not symmetrical around zero, since the psychophysiological phenomenon is asymmetrical. Both the Backster and Matte scoring systems have asymmetrical cutting scores in favor of the nondeceptive, a direction the data suggest they should be. The empirical foundation for the Backster and Matte thresholds is absent or incomplete, however, so it has not been established with any great confidence that their cutting scores are optimal. That said, no other major scoring system has even progressed to asymmetrical cutting scores. Though there exists a significant body of literature on cutting scores, there is little indication that what is known is reflected in what is done.

In summary, the present data sets do not support the Total Chart Minutes Concept with field cases conduct in the ZCT or MGQT formats. They provide additional evidence for

the need to adjust scoring methods or decision rules to adapt to the asymmetrical patterns of responding between deceptive and nondeceptive examinees. The data also provide further evidence that the electrodermal channel provides more diagnostic information than does either of the other two polygraph channels.

Acknowledgments

The writer is grateful to Dr. James Matte for data he supplied, and for the technical advice of Dr. Andrew Dollins and Stuart Senter. The views expressed in this report do not represent those of the Department of Defense Polygraph Institute, or the US Government.

References

Backster, C. (1960). Zone comparison technique. *Annual school research series of polygraph technique trends.* 8.

Backster, C. (1963a). Total Chart Minutes Concept. *Law and Order*, 11(10), 77-78.

Backster, C. (1963b). Polygraph professionalization through technique standardization. *Law and Order*, 11(4), 63-65.

Blackwell, N.J. (1999). PolyScore 3.3 and psycho-physiological detection of deception examiner rates of accuracy when scoring examinations from actual criminal investigations. *Polygraph*, 28(2), 149-175.

Capps, M.H., & Ansley, N. (1992). Numerical scoring of polygraph charts: What examiners really do. *Polygraph* 21(4), 264-320.

Franz, M.L. (1989). Technical report: Relative contributions of physiological recordings to detect deception. Contract Number MDA 904-88-M-6612.

Kircher, J.C., & Raskin, D.C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, 73(2), 291-302.

Krapohl, D.J. (1998). Short report: Proposed method for scoring electrodermal responses. *Polygraph*, 28(1), 82-84

Krapohl, D.J., & McManus, B. (1999). An objective method for manually scoring polygraph data. *Polygraph*, 29(3), 209-222.

Matte, J.A. (1996). *Forensic Psychophysiology Using the Polygraph: Scientific Truth Verification – Lie Detection.* J.A.M. Publications: Williamsville, New York.

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

Bibliography of Habituation Research

Donald J. Krapohl[1]

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

Baltissen, R., & Boucsein, W. (1986). Effects of a warning signal on reactions to aversive white noise stimulation: Does warning "short-circuit" habituation? *Psychophysiology*, 23(2), 224-231.

Barry, R. J. (1990). Scoring criteria for response latency and habituation in electrodermal research: A study in the context of the orienting research. *Psychophysiology*, 27(1), 94-100.

Ben-Shakhar, G. (1980). Habituation of the orienting response in complex sequences of stimuli. *Psychophysiology*, 17(6), 524-534.

Ben-Shakhar, G., Dymshitz, J., & Lieblich, I. (1982). Generalization of habituation of skin conductance responses to multidimensional sequences of stimuli. *Psychophysiology*, 19(2), 178-182.

Ben Shakhar, G., Gati, I., Ben Bassat, N., & Sniper, G. (2000). Orienting response reinstatement and dishabituation: Effects of substituting, adding, and deleting components of nonsignificant stimuli. *Psychophysiology*, 37(1), 102-110.

Ben-Shakhar, G., & Lieblich, I. (1982). The dichotomization theory for differential autonomic responsivity reconsidered. *Psychophysiology*, 19(3), 277.

Ben-Shakhar, G., & Lieblich, I. (1982). Similarity of auditory stimuli and generalization of skin conductance response habituation. *Physiological Psychology*, 10(3), 331-335.

Ben-Shakhar, G., Lieblich, I., & Kugelmass, S. (1975). Detection of information and GSR habituation: An attempt to derive detection efficiency from two habituation curves. *Psychophysiology*, 12(3), 283-288.

Churchill, M., Remington, B., & Siddle, D. A. T. (1987). The effects of context change on long-term habituation of the orienting response in humans. *The Quarterly Journal of Experimental Psychology*, 39B, 315-338.

Darr, R. F., Jr. (1973). Vasomotor behavior during semantic conditioning. *Polygraph*, 2(1), 65-72.

Dollins, A. B., Cestaro, V. L., & Pettit, D. J. (1998). Efficacy of repeated psychophysiological detection of deception testing. *Journal of Forensic Science*, 43(5), 1016-1023.

Ingram, E. M. (1994). Effects of Electrodermal Lability and Anxiety on the Electrodermal Detection of Deception With a Control Question Technique. (DoDPI94-R-0004). Fort McClellan, AL: Department of Defense Polygraph Institute.

Kircher, J. C., Raskin, D. C., & Honts, C. R. (1984). Electrodermal habituation in the detection of deception. *Psychophysiology*, 21(5), 585.

[1] Department of Defense Polygraph Institute, Fort Jackson, SC 29207

Kopp, M. S., Mihaly, K., Linka, E., & Bitter, I. (1987). Electrodermally differentiated subgroups of anxiety patients. I. Automatic and vigilance characteristics. *International Journal of Psychophysiology*, 5, 43-51.

Levenson, D. F., & Edelberg, R. (1985). Scoring criteria for response latency and habituation in electrodermal research: A critique. *Psychophysiology*, 22(4), 417-426.

Lobb, H., & Kaplun, J. (1970). Protection of GSR conditioning by dextroamphetamine. *Canadian Journal of Psychology*, 24(1), 15-26.

Malmierca, J. L. M., & Bernabé, J. R. Y. (1989). Condicionamiento instrumental de la actividad electrodérmica: Contingencia, conciencia y expectativa del castigo positivo [Instrumental conditioning of electrodermal activity: Contingency, awareness and expectancy of positive punishment]. *Revista Latinoamericana De Psicología*, 21(2), 219-242.

Michiro, K., Muranaka, T., & Miyata, Y. (1984). Effects of instructions on the skin conductance response. *Japanese Psychological Research*, 26(3), 159-167.

Miller, D. J., & Kotses, H. (1990). Habituation of phasic total respiratory resistance responses. *Psychological Reports*, 67(3), 1139-1145.

Nakayama, M., & Kizaki, H. (1990). Usefulness of the repeated presentation of questions on the psychophysiological detection of deception. *The Japanese Journal of Psychology*, 60(6), 390-393.

Porter, J. M. (1938). Adaptation of the galvanic skin response. *Journal of Experimental Psychology*, 23, 553-557.

Prystav, G. H. (1975). Autonomic responsivity to sensory stimulation in drug addicts. *Psychophysiology*, 12(2), 170-178.

Sagae, M. (1979). Effects of instruction on the physiological responses in the polygraph test. *Reports of the National Research Institute of Police Science*, 32, 22-25.

Siddle, D. A. T. (1985). Effects of stimulus omission and stimulus change on dishabituation of the skin conductance response. *Journal of Experimental Psychology*, 11(2), 206-216.

Suzuki, A., & Hikita, Y. (1964). An analysis of response on polygraph: A diminution of responses. *Reports of the National Research Institute of Police Science*, 17, 290-295.

Timm, H. W. (1984). Significant findings attributable to electrodermal habituation effects: Artifact or essence in detection of deception research. *Journal of Police Science and Administration*, 12(3), 267-276.

Watts, J. M. (1975). Anxiety and the habituation of skin conductance response. *Psychophysiology*, 12(5), 596-601.

Test of a Mock Theft Scenario for Use in the Psychophysiological Detection of Deception: IV

DoDPI Research Division Staff [1] [2]

Abstract

The study described in this report is a continuation of research to develop a participant manipulation to serve as a standard procedure for laboratory psychophysiological detection of deception (PDD) research. The manipulations used in this study were similar to one reported by Kircher (1983) of the University of Utah. In Experiments 1 and 2 two groups of 16 participants who were assigned to be either guilty or innocent of the mock theft of a ring were tested using the Zone Comparison Test, a PDD examination taught at the Department of Defense Polygraph Institute. Written and audio taped instructions were provided to all participants. All participants were promised $50 for participating in the study and an additional $25 if they were classified as nondeceptive following a PDD examination. Three human examiners evaluated each of the 32 sets of polygraph charts. For Experiment 1, the decisions made by examiners were correct 55% of the time, incorrect 23% of the time, and no opinion 22% of the time. Experiment 2 was identical to Experiment 1 except that participants completed two screening questionnaires prior to testing. Participants who did not complete the questionnaires satisfactorily were excluded from the study. For Experiment 2, the decisions made by examiners were correct 66% of the time, incorrect 10% of the time, and no opinion 24% of the time. It is concluded that the procedures used to manipulate participants in Experiments 1 and 2 did not meet the necessary requirements for a standard procedure, but that the screening procedure used in Experiment 2 did result in higher accuracy.

Key Words: mock crime scenarios, psychophysiological detection of deception, Zone Comparison Test

One research goal of the Department of Defense Polygraph Institute (DoDPI) is to determine the accuracy (i.e., validity) of psychophysiological detection of deception (PDD) procedures. For instance, is it more effective to use the directed lie or the probable lie as a comparison question? To answer such questions, investigators must manipulate one or more variables and determine which manipulations produce the most effective examination. In order to maintain scientific integrity within such a cumulative research program, the methodology of all studies should be as uniform as possible, varying only those procedures under investigation. One of the first steps toward obtaining such uniformity is to develop a participant manipulation procedure that produces reliable results.

One participant manipulation, the "mock" or simulated crime scenario, has been extensively investigated (for a review see Kircher, Horowitz, & Raskin, 1988). Despite criticism that mock crime scenarios and other laboratory research lack external validity (Furedy & Heslegrave, 1991; Lykken, 1981; Iacono, 1991; Office of Technology Assessment [OTA], 1983), mock crime scenarios are generally associated with statistically significant detection of deception, and they provide good experimental control over the experience of participants (Kircher et al., 1988). Not all mock crime scenarios are equally effective, however. In a review of 14 analogue studies, Kircher et al. found that mock crime scenarios involving an incentive for passing the polygraph examination were

[1] Department of Defense Polygraph Institute
[2] Correspondence addressed to Stuart Senter, Department of Defense Polygraph Institute, Fort Jackson, SC 29207

associated with higher detection of deception. They found a high correlation between incentives and accuracy, r = .73. Perhaps, responding in deceptive individuals during polygraph examinations is only elicited when there is a substantial risk associated with the detection of deception (Lykken, 1981; OTA, 1983). Although the potential loss of an incentive for passing the polygraph examination is a substantial risk, it is less substantial than the risk to one's freedom and reputation experienced in actual criminal investigations. The objective of the present study was to test the effectiveness of a specific mock theft scenario that incorporated a monetary incentive.

Experiment 1

The scenario used here is fundamentally the same as the scenario reported by Kircher (1983). This scenario was reported to be highly effective, with 87% of the participants correctly identified as guilty or innocent, and only 7% of the participants left unidentified (that is, inconclusive). We expected to find similar results in the present study. More formally, if the population accuracy rate for this scenario is 87% (which is the best estimate at the present time), then power analyses indicate that we should expect to obtain at least an 80% accuracy rate in roughly 90% of studies when using 32 participants per study (Glass & Hopkins, 1996, Eq. 13.9). Accordingly, the DoDPI set a goal of 80% accuracy for the present study, which employed 32 participants. If the present mock theft scenario meets the 80% goal, it will be considered as a possible standard scenario for use in the cumulative research program described above.

Method

Participants

Participants were 32 native English speaking civilians (12 males and 20 females), ranging in age from 19 to 44 years. Participants were recruited by a temporary employment service, and were instructed by the employment service to report to the DoDPI at a previously specified time. No attempt was made to select participants with specific demographics (i.e., gender, race, age, etc.) because previous research has shown little effect of demographics on PDD outcome (e.g., Reed, 1993). All participants reported themselves to be healthy, free from drugs and medication, and experimentally naive. No participant reported ever having taken a polygraph exam. Participants were paid $50 for participating in testing and promised and paid an additional $25 bonus for a nondeceptive PDD examination outcome. Half of the participants were assigned to the deceptive group and half were assigned to the nondeceptive group. Assignment to groups was predetermined using time of arrival as the only criterion. The procedures used in this project were reviewed and approved by the DoDPI Human Use Committee.

Examiners

Two experienced field examiners who were certified by the Department of Defense conducted the examinations. Three additional Department of Defense certified PDD examiners, who were unaware of the participants' group assignments and veracity, independently scored the examinations. The examination schedule allowed two hours to test each participant, with each examiner testing four participants per day on Tuesday, Wednesday, and Thursday, and two participants on Monday afternoon and Friday.

Apparatus

Equipment used in testing included: two portable cassette recorders which were used to play instructions during the participant manipulation; a simulated diamond ring which was "stolen" by deceptive participants; and two video cameras, televisions, and digital audiovisual mixers which were used to record PDD examinations. Computerized polygraph systems (Axciton Systems, Inc., Houston, TX, Version 7.0) were used to record, and subsequently print, electrodermal, respiratory, and cardiovascular activity. Participants were seated in an adjustable-arm chair (Lafayette, Lafayette, IN, Model 76871) during the polygraph examination.

Procedure

Each participant was instructed by the temporary employment agency to proceed to a specific room in Ft. McClellan building 3165, and to read the instructions found in an envelope taped to the door of the room. The instructions directed the participant to enter

the room, read and sign a volunteer agreement affidavit, and to listen to a tape recording of further instructions.

The tape recorded instructions for deceptive participants directed them to participate in the following scenario. They were to proceed to a specific office and ask the secretary where Mr. Mitchell could be found, knowing there was no Mr. Mitchell in the building. They were told to leave the secretary's office when told that no one named Mitchell worked in the building. They were instructed to wait out of the secretary's sight, until the secretary left the office. The secretary waited approximately three minutes before leaving the office. The participant was further directed to enter the secretary's office and take a "diamond" ring from an envelope in a metal cash box in the secretary's desk. The participant was instructed to destroy the envelope and to conceal the ring on their person. The tape recorded instructions also directed deceptive participants to accomplish their task and return to the room with the tape recorder within 15 minutes, prepare an excuse in case they were caught, and be careful not to leave fingerprints in the secretary's office.

Nondeceptive participants were informed, via tape recorded instructions, that a ring was being stolen by some other participants, but that they were innocent of the theft. They were directed to proceed to the clearly marked building lobby where they were to wait for 15 minutes before returning to the room with the tape recorder.

The tape recorded instructions informed all of the participants that they would be given a lie detector test by an expert polygraph examiner who did not know if they were guilty of the theft. They were also cautioned that they would be disqualified from receiving any payment if they revealed details of their activities. Finally, participants were told that they would receive a bonus only if the PDD examiner found them to be nondeceptive.

Examiners always met participants in the room where the participants heard their instructions. The examiner introduced himself to the participant as the person who would administer the polygraph, then escorted the participant to the examination room. The examiner reminded the participant that the $25 bonus was contingent on a truthful outcome on the test. The examiner then seated the participant in a Lafayette polygraph chair and began the pretest (DoDPI, 1994) by asking the participant four pretest interview questions. If participants confessed or incriminated themselves by revealing knowledge that only a guilty participant would know, their participation in the study was terminated. The examiner then obtained biographical information from the participant. Next, the examiner reviewed the test questions. If, during this review, the participant answered any of the comparison questions with a "yes", the question was reworded to elicit an answer of "no."

Sensors were attached to the participant in the following locations: electrodermal finger plates on the distal-medial phalanges of the first and third fingers of the (typically) nondominant hand, blood pressure cuff on the (typically) dominant arm above the brachial artery, and pneumographic chest assemblies across the pectoralis major but under the arm ("thoracic" sensor) and across the rectus abdominis immediately above the navel ("abdominal" sensor). Placement of the sensors was governed by visual cues only, and was therefore only approximate.

After placement of the sensors, an acquaintance test (DoDPI, 1999) was conducted. The acquaintance test consisted of requiring the participant to choose a number between 3 and 6 and then informing the examiner of the chosen number. The participant was then told to deny selecting any number during testing. The participant was then tested on the numbers 2-7. The results of the acquaintance test were presented to the participant as a demonstration of the validity of the lie detection technique.

A Zone Comparison Test (ZCT) immediately followed the acquaintance test. The ZCT is composed of 10 questions, with each question presented approximately 25 seconds after the onset of the previous question (DoDPI, 1992). After each test, the examiner asked the participant how he or she felt about the questions and whether there was any problem with any of them, focusing

specifically on the probable lie comparison questions. This procedure, according to Raskin, Barland, and Podlesny (1977), maintains or increases the salience of the comparison questions. After the fifth test, the sensors were removed. A research assistant escorted the participant from the examination room to a nearby small office. The participant was then debriefed and told the examination result.

Data Reduction

The polygraph charts were independently evaluated by three examiners using the 7-position scoring method taught at the DoDPI (DoDPI, 1992). The examiners were blind to participant group membership and veracity.

Results

Decision frequencies are provided in Table 1. The average percentage of correct, incorrect, and NO decisions was 55%, 23%, and 22%, respectively. Excluding NO decisions, the average percentage of correct decisions was 71%. Collapsing across participant veracity, Cochran's Q tests (Siegal & Castellan, 1988) indicated that there were no significant differences among evaluators in the proportion of correct, incorrect, and NO decisions.

Table 1
Frequency of Decisions for 16 Deceptive and 16 Nondeceptive Participants

Deceptive				Nondeceptive			
Evaluator	Correct	Error	NO	Correct	Error	NO	Total
1	10	3	3	8	4	4	32
2	11	1	4	4	6	6	32
3	13	1	2	7	7	2	32

Note. NO = no opinion.

Table 2 shows the pairwise proportion of agreement between each evaluator, in addition to the proportion of correct decisions for each evaluator. As Table 2 shows, the proportion of agreement within evaluators ranged from .69 to .75. The proportion of correct decisions ranged from .47 to .63.

Table 2
Pairwise Proportion of Agreement Between Evaluators

Evaluator	2	3	Accuracy
1	.69	.70	.56
2		.75	.47
3			.63

Discussion

The mock crime procedure used in this study did not meet the DoDPI goal of 80% correct (against 50% chance). The procedure also produced slightly greater than 20% inconclusive (or NO) decisions. Since the study was not designed to measure the effect of a monetary incentive, no assessment of this factor was made.

A possible source of the low accuracy achieved in Experiment 1 may have been a lack of comprehension on the part of participants either with respect to the mock crime scenario or in terms of the instructions provided by the polygraph examiners. Experiment 2 was conducted with a screening process in order to correct for this possibility.

Experiment 2

Experiment 2 served to replicate Experiment 1 procedurally, but with an added component. Experiment 2 sought to remedy the low accuracy achieved in Experiment 1 by requiring that participants complete two screening questionnaires prior to completing either the innocent or guilty scenario. These instruments were included for two reasons. First, they were included to ensure that participants were mentally competent. Second, they served to ensure that participants had a clear understanding of the features of the study in which they were participating. Only those participants who performed to criterion on these instruments were allowed to further participate in the study.

Method

Participants

Sixty participants were screened prior to the experiment. Twenty-one participants did not pass the screening instruments (described below). Of the remaining 39 participants, two were eliminated due to problems during polygraph testing, and five were not included because participant capacity (N=32) had been reached. The 32 participants who succeeded in passing the screening instruments and in completing the study were native English speaking civilians (17 males and 15 females), ranging in age from 19 to 44 years. With the exception of the screening process, participants were selected and compensated in the same way as those in Experiment 1.

Examiners

Two experienced field examiners who were certified by the Department of Defense conducted the examinations. Three additional Department of Defense certified PDD examiners, who were unaware of the participants' group assignments and veracity, independently scored the examinations. The examination schedule allowed two hours to test each participant, with each examiner testing four participants per day on Tuesday, Wednesday, and Thursday, and two participants on Monday afternoon and Friday.

Apparatus

The polygraph instrumentation was identical to that used in Experiment 1. Two screening questionnaires were added in Experiment 2. The first instrument was the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975. This instrument includes a set of simple factual questions, a probed memory recall task, some simple motoric instructions, and some simple reading and writing instructions. The Mini Mental State Examination is very brief, containing only a total of 22 items.

The second instrument was a brief reading comprehension task. The instrument included a written statement (1.5 pages, single-spaced) describing the features and purpose of a hypothetical study, in addition to a participant's role in the study. Included with the written statement were 14 multiple choice items to ensure that participants could read and understand the passage of text. The hypothetical study described was similar, though not identical to the actual study. This passage was included to insure that participants could understand the procedures of a study very similar to the one actually being used, without disclosing the details of the actual study.

Both the Mini-Mental State Examination and the reading comprehension task were selected, based on the expertise of the research staff, in order to screen out participants who would not, or could not, follow oral and written instructions. Of the

twenty-one participants who did not pass the screening instruments, eighteen failed only the reading comprehension task, and three failed both the reading comprehension task and the Mini-Mental State Examination.

Procedure

The procedure was identical to that used in Experiment 1, excepting the inclusion of the two screening instruments. Prior to being scheduled in the study, participants met a confederate at the employment agency where they completed the Mini-Mental State Examination and the reading comprehension test. Participants were informed that they were to complete the two questionnaires so that their memory and ability to concentrate could be assessed. In addition, participants were told that it was necessary for them to demonstrate that they could follow both written and oral instructions.

The Mini-Mental State Examination was administered orally, and participants responded orally. This was done to make sure that participants could follow oral instructions. Next, participants were required to complete the reading comprehension task in written format. Each instrument took approximately 15 minutes to complete, for a total of 30 minutes to complete both instruments. With the exception of two questions on the reading comprehension task having to do with subjective opinions, participants were to answer all items on both instruments correctly. Failure to do so resulted in disqualification from the study. Participants who met the criteria were then scheduled for the study.

Data Reduction

The polygraph charts were independently evaluated by three examiners using the 7-position scoring method taught at the DoDPI (DoDPI, 1992). The examiners were again blind to participants group membership and veracity.

Results

The frequencies of evaluation decisions for Experiment 2 are provided in Table 3. The average percentage of correct, incorrect, and NO decisions produced by human scorers was 66%, 10%, and 24%, respectively. Excluding NO decisions, the average percentage of correct decisions was 86%. Collapsing across participant veracity, Cochran's Q tests indicated that there were no significant differences among evaluators in the proportion of correct, incorrect, and NO decisions.

Table 4 shows the pairwise proportion of agreement between each evaluator, in addition to the proportion of correct decisions for each evaluator. As Table 4 shows, the proportion of agreement was comparable to that found in Experiment 1, ranging from .63 to .78.

Table 3
Frequency of Decisions for 16 Deceptive and 16 Nondeceptive Participants

Deceptive				Nondeceptive			
Evaluator	Correct	Error	NO	Correct	Error	NO	Total
1	14	1	1	7	2	7	32
2	11	0	5	9	2	5	32
3	13	1	2	9	4	3	32

Note. NO = no opinion

Table 4
Pairwise Proportion of Agreement Between Evaluators

Evaluator	2	3	Accuracy
1	.63	.69	.66
2		.78	.63
3			.69

Discussion

Compared to Experiment 1, Experiment 2 produced a non-significant (z = .834, p > .05), but notable increase in accuracy (55% to 66%). The boost in accuracy resulted from an increase in the number of correct calls with a corresponding decrease in the number of errors. However, the proportion of NO decisions remained constant across the two experiments. The average proportion of agreement was also larger for Experiment 2 relative to Experiment 1.

General Discussion

The results indicate that Experiments 1 and 2 did not meet the 80% accuracy criterion or the 20% or fewer NO decision proportion. Experiment 2 did show a substantial (though non-significant) increase in veracity decision accuracy relative to Experiment 1. This increase in accuracy could be attributable to the screening of participants using the two instruments described above. However, this conclusion is tentative in that Experiments 1 and 2 used different examiners, different human evaluators, and were conducted at different times (Experiment 2 was conducted following the completion of Experiment 1).

Kircher (1983) reported results with a single unaware evaluator of 87% correct, 6% incorrect, and 7% inconclusives with 100 participants using five tests when necessary. Overall, this discrepancy in evaluator accuracy suggests that the participant manipulation used in the mock crime scenario of the present study may have been carried out in a different fashion from that of Kircher or that the participant sub-populations may have been different in the two studies. There are other noted procedural differences that may also account for some of the differences between Kircher's results and those of the present study. One major difference was that Kircher solicited participants using a classified advertisement in a local newspaper. Kircher's participants had no direct contact with experimenters until they met the examiner. In contrast, the participants in this study were obtained via an employment agency. Other differences include the number of charts used,

rules to discern participant veracity with assigned scores, and the number of data channels used (Senter, Dollins, & Krapohl, 2000). Procedural differences may also exist with respect to the way in which the pretest interviews were conducted, and the way in which comparison questions were emphasized to participants. Future research should investigate these possibilities in order to determine the source of differences and to discover the optimal set of procedures for conducting polygraph examinations.

The accuracy in the current study may also be an example of the accuracy variability observed in other analog PDD studies. Honts and Quick (1995) reported accuracy rates ranging from a high of 88% to a low of 53% for four laboratory studies conducted since 1986. Given the degree of accuracy variability seen in many analog studies, accuracy could be a function of more than the scenario that participants enact. Other important factors contributing to accuracy may include participant characteristics, instrumentation, and examiner variables.

In conclusion, when considering the high accuracy achieved with this paradigm in previous studies, it is recommended that research using this standard procedure continue, perhaps in conjunction with a screening mechanism such as that used in Experiment 2. However, attention should be paid to the identification of factors that mediate deception detection accuracy, including those beyond the scenario script. The research should identify and include those components that will likely contribute to the degree of differential responding necessary for a good standard methodology. Finally, the methodology must be repeatable and transportable.

Acknowledgments

The design and implementation of this project are primarily the work of Eben M. Ingram, Ph.D., who is now with the Centers for Disease Control and N. Joan Blackwell, M.S. The data analysis and report are primarily the work of Stuart M. Senter, Ph.D., and Andrew B. Dollins, Ph.D. Thanks go to Department of Defense Polygraph Institute (DoDPI) staff Joan Harrison-Woodard, Jennifer Justice, Brenda Smith, and Randall Reynolds whose assistance was indispensable. Appreciation is also extended to DoDPI instructors Donnie Dutton, Bill Gary, Bill Gashler, Richard Giraud, Vanecca Green, Joe Phipps, and Johnny Rodgerson, as well as Craig Malmfeldt of the United States Army Criminal Investigation Division and Special Agent Robert Fritzsche of the United States Naval Criminal Investigative Service for assisting with the project. This research was funded by the Department of Defense Polygraph Institute, Fort McClellan, Alabama, as project DoDPI97-P-0004 and DoDPI97-P-0004A. The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Department of Defense Polygraph Institute. (1992). Zone comparison test. Fort McClellan, AL: Author.

Department of Defense Polygraph Institute. (1994). Pretest interview. Fort McClellan, AL: Author.

Department of Defense Polygraph Institute. (1999). Acquaintance Test. Fort Jackson, SC: Author.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189-198.

Furedy, J. J., & Heslegrave, R. J. (1991). The forensic use of the polygraph: A psychophysiological analysis of current trends and future prospects. In P. K. Ackles, J. R. Jennings & M. G. H. Coles (Series Eds.), Advances in Psychophysiology: Vol. 4. (4th ed., pp. 157-189). London: Jessica Kingsley Publishers.

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology (3rd ed.). Boston: Allyn & Bacon.

Honts, C. R., & Quick, B. D. (1995). The polygraph in 1995: Progress in science and the law. North Dakota Law Review, 71, 987-1020.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In P. K. Ackles, J. R. Jennings & M. G. H. Coles (Series Eds.), Advances in Psychophysiology: Vol.4. (4th ed., pp. 201-207). London: Jessica Kingsley Publishers.

Kircher, J. C. (1983). Computerized decision-making and patterns of activation in the detection of deception. Unpublished doctoral dissertation, University of Utah, Salt Lake City.

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

Lykken, D. T. (1981). A tremor in the blood: Uses and abuses of the lie detector. New York: McGraw-Hill.

Office of Technological Assessment. (1983). Scientific validity of polygraph testing: A research review-A technical memorandum (OTA-TM-H-15). Washington, DC: U. S. Congress, Office of Technology Assessment.

Raskin, D. C., Barland, G. H., & Podlesny, J. A. (1977). Validity and reliability of detection of deception. Psychophysiology, 6, 1-39.

Reed, S. D. (1993). Effect of demographic variables on psychophysiological detection of deception outcome accuracies. (Report No. DoDPI90-R-0003). Fort McClellan, AL: Department of Defense Polygraph Institute.

Senter, S. M., Dollins, A. B., & Krapohl, D. J. (2000). Comparison of Utah and DoDPI scoring accuracy: Equating veracity decision rule, chart rule, and number of data channels used. (Report No. DoDPI00-R-0001). Fort Jackson, SC: Department of Defense Polygraph Institute.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). New York: McGraw-Hill.

Review of Polygraph Screening Assessment Method

William J. Gaschler[1,2], James P. McGettigan[1], Paul M. Menges[1], and James F. Waller[1]

Polygraph screening examinations, used as assessment instruments in the area of counterintelligence and applicant screening in some intelligence and security organizations, have increased dramatically in the past 15 years. This increased usage has fueled increased debate and discussion concerning the impact of this assessment process on individual rights to privacy in the United States. Additionally, as arguments for and against routine polygraph screening have taken center stage, one argument against this form of assessment has taken the tact of questioning its validity and reliability in light of the abundance of polygraph countermeasure information now available to the layperson. Attempts by examinees to manipulate polygraph data in polygraph examinations by employing mental, physical, and drug induced countermeasures was once thought by examiners to be futile on the part of the examinee and something not worthy of serious concern by examiners. While research has demonstrated this attitude to be simplistic and incorrect (Honts, Hodes, & Raskin, 1985; Honts, Raskin, & Kircher, 1987; Honts, Raskin, & Kircher, 1994), research has also clearly demonstrated the continued validity and reliability of the polygraph assessment process in spite of spontaneous countermeasures employed by examinees (Honts, Amato, & Gordon, 2001). Polygraph screening in the counterintelligence and applicant testing arenas continues to be a valuable assessment tool in the hands of federally trained polygraph examiners.

The United States Government began using polygraph screening examinations in the national security arena in 1949. The Central Intelligence Agency (CIA) was the first agency to routinely polygraph new employees and require periodic exams of current employees. This type of assessment was deemed necessary as a security precaution to be used in areas where our most sensitive national secrets were held. With time, similar security requirements were put into place in other counterintelligence and security organizations, specifically, the National Security Agency and elements of the Department of Defense.

After a victory in World War II, American technology and economic strength, particularly in the field of nuclear science, gave the United States a significant military advantage over the Soviet Union. Subsequent international tensions gave rise to the Cold War between the two superpowers. The conflict between democracy and communism lasted for the next half century, until the fall of communism in the Soviet Union in 1991. Today the challenges to U.S. national security come from the People's Republic of China, as well as rogue nations and terrorist factions believed to threaten the well being of the remainder of the world.

In this environment, societal attitudes, morals, ethics, philosophies, and individual interests developed and changed. After surviving the war years from 1941-1945, what was called "the greatest generation" recognized the Soviet Union as a significant threat to world peace. Most U.S. citizens found it easy to understand and accept the security measures initiated for protecting government secrets. Although the number of persons subject to polygraph examinations as a security assessment instrument was relatively small, most understood that individual rights to privacy had to be balanced when it came to safeguarding the democracy of the United States. Few individuals questioned stringent security requirements and close scrutiny for the award of a security clearance. It was

[1] Department of Defense Polygraph Institute
[2] To whom correspondence should be addressed at Department of Defense Polygraph Institute, Fort Jackson, SC 29207

recognized that access to classified information was a privilege, and not a right acquired by residence or citizenship. During the last 50 years, it appears that concern for individual rights and freedoms has caused a shift in emphasis and acceptance by the general public. An attitude of questioning government and authority has gained ground. Individuals now question laws and procedures more freely and openly. What might once have appeared as an anarchist view, today might appear as normal questioning of policy. One might even suggest that such a shift is healthy, normal, and necessary in a free and democratic society. The current argument is over how, what, and when individual freedoms must be adjusted or impacted by measures necessary for the normal rule of law for the greater good.

Individuals subject to polygraph screening examinations for sensitive information access, or employment with some security organizations, have recently become more vocal and organized in their opposition to the polygraph screening. This has taken place is in spite of changes in regulations and hiring practices of federal agencies to become more tolerant in the acceptance of applicants who would have not been considered a decade earlier. At one time, many physical disabilities and or lifestyle preferences would have precluded many people from obtaining the productive positions they hold today.

Technological advancements in communication and the Internet now allow individuals to sit at home at their personal computers and research over 90,000 references regarding polygraph technology, examination techniques, and countermeasures. Individuals subject to polygraph examination by federal agencies also have access to the pro and con arguments aired on the Internet. The information highway has made it very easy for individuals who want to research this investigative tool for employment screening. Prior to the advent of the Internet, someone who wanted this information would have to spend weeks in libraries conducting this research. Today, it's available in only minutes with computer technology.

In addition to the vocal opposition of some federal applicants, the news media tend to highlight only the negative, sensational issues regarding polygraph screening examinations. However, in spite of concerns raised about the effects of false positive results upon individual careers and job applications, the results to date of federally mandated polygraph screening programs tend to invalidate these concerns. According to the Congressional Record in 2000, federal agencies performing counterintelligence screening examinations have conducted a total of 43, 648 since inception of the program in 1995. Of that total, 963 made admissions to relevant national security issues. Further, investigation revealed that of the 963 subjects making admissions to relevant issues, 121 subjects had adverse adjudicative action taken against them by their respective agencies. (Note: Adverse action is defined as administrative action and or criminal prosecution) If a screening program is to be effective and is conducted in accordance with governing standards, one must recognize that some of the examinees will be identified as having violated one or more of the relevant issues under scrutiny in the procedure. The fact that only 121 subjects out of a total of 43, 648 had their careers adversely impacted by polygraph screening is testimony that the fear of large numbers of the test population falling victim to high number of false positive results and the concomitant adverse action, is unwarranted.

As mentioned above, early use of polygraph as an assessment tool was related primarily to a small number of personnel involved with intelligence and security agencies, or in law enforcement use in specific issue examinations. While there may always have been some concern regarding reliability and validity of the polygraph process by the population subject to exams, the larger population pool today and availability of information regarding this assessment tool has resulted in a louder cry for scrutiny regarding the instrument's validity and reliability. Recently the issue has risen to the level of congressional scrutiny. Perceived or reported failures in investigations involving suspected spies such as the case involving a Department of Energy (DOE) employee suspected of espionage who had failed a polygraph

examination and the case of a senior employee of the Federal Bureau of Investigation (FBI) who was charged with espionage, keep the issue of counterintelligence polygraph screening exams in the forefront. The DOE employee had been investigated for years and had failed a polygraph examination related to espionage, but was eventually released after the investigation failed to provide sufficient solid evidence for prosecution (Report on the Investigation of Espionage Allegations Against Dr. Wen Ho Lee, 2000). The FBI employee was a senior counterintelligence investigator who had never been subject to any type of polygraph screening by the FBI.

While this issue is in the forefront in current press reports, a small group of scientists within a larger population subject to the counterintelligence polygraph screening process has raised the issue of validity and reliability in view of the abundance of polygraph countermeasure information available to the average examinee. The basic position they raise is that if countermeasures can be effective in defeating the polygraph assessment process, why maintain or further implement such polygraph screening. Others have raised similar questions related to validity and reliability. Some of these opponents of polygraph screening have created, organized, and maintained complex, professional web sites to address their issues and provide a forum for opposition. *Anti-polygraph.org*, *NoPolygraph.Com*, *Stoppolygraph.Com*, and *Polygraph.com*, among others, attempt to further their stated goal of abolishing all polygraph screening in the U.S. They all advocate that polygraph is not based on science and the only real way to insure a favorable outcome is to employ countermeasures during an examination.

Polygraph countermeasures can be defined as any measure employed by an examinee to defeat the examination. Some go so far as to define countermeasures as any effort by an examinee to influence the exam. Whether countermeasures should by definition only refer to guilty persons attempting to hide their guilt from an examiner or whether it should include the innocent person doing what he or she feels needs to be done to be called truthful is moot to this discussion. Anti-polygraph web sites advise all examinees to employ countermeasures. Some of the site founders feel that deceptive persons may choose to employ countermeasures in order to appear non-deceptive, truthful persons may choose to use them to protect themselves against a false positive outcome (Maschke & Scalabrini, 2000). In the past, those employing countermeasures were thought to be primarily guilty subjects in specific issue criminal examinations trying to get through the exam without being correctly identified. Today, based on the advice of anti-polygraph web sites and the huge growth in polygraph screening examinations, a dramatic increase is expected in innocent subjects employing countermeasures to ensure they pass their screening exams.

There are two basic types of countermeasures addressed on the Internet: behavioral and physical or manipulation or the data being recorded. Behavioral countermeasures include those things the subject can do to appear honest and truthful to the examiner. Physical countermeasures or chart manipulation are those countermeasures that will actually affect the physiological data recorded by the polygraph instrument (Maschke & Scalabrini, 2000).

In an attempt to deal with the concerns over validity and reliability of the polygraph as an assessment tool in determining truth or deception, the Department of Defense Polygraph Institute (DODPI) was tasked to conduct scientific research especially in the area of screening exams. Two studies concerning the accuracy of a screening test known as the Test for Espionage and Sabotage (TES) were conducted using mock crime scenarios. In the first experiment, 83.3% of the guilty subjects were identified and 98% of the innocent subjects were properly identified (Research Division Staff 1995a). The second study replicated the results of the first study (Research Division Staff, 1995b). These studies validated the process and demonstrated reliability through the replication evidenced in the second study. The accuracy achieved in the studies is acceptable to all but the most adamant critics of the process.

The concern over the reliability and validity of individual assessment tools is

not something unique to polygraph. None of the assessment tools used in individual assessment or the forensic sciences are 100% accurate. Good examples are the Minnesota Mutiphasic Personality Inventory-2 (MMPI-2) and MMPI-Adolescent. Both inventories are excellent assessment tools but they are vulnerable to faking because of the transparency of some of the items. In spite of special scales to identify faking, it is still believed to be easy for the client to *employ countermeasures* and slant answers to give a favorable or unfavorable impression (MMPI-A, 1999). This type of evidence would suggest that polygraph screening examinations are as valid and reliable as other currently used individual assessment tools. In fact, in some cases they appear to have a higher rate of accuracy. It is also clear that the susceptibility to countermeasures is not unique to polygraph examinations.

The validity and reliability of polygraph tests as affected by various medications have had minimal research inquiries (Iacono, 2001). There is no known single drug or combination of drugs, which is able to selectively influence response(s) to any single question upon a test. According to Andrew B. Dollins, PhD. (personal communication, June 11, 2001), of the Department of Defense Polygraph Institute, an examinee must demonstrate a capability to respond to external stimulus to effect an interpretable recording on at least one channel being recorded. All ingested drugs have at least one side effect because they are foreign to the organism. Hence, there is a plethora of potential drugs that may be consumed congruent to an examination. Not all in the pharmacopoeia will influence test recordings. For example, for someone who regularly drinks caffeinated liquids, one cup of coffee or tea will not cause anomalies. However, sudden consumption of hydrochlorothiazide, a blood pressure stabilizer, will experience lower blood pressure.

In order to bring into question the validity and reliability of polygraph results, a chemical substance would have to demonstrate differential effect, meaning it would have to suppress responses to some questions but not others (A. B. Dollins, personal communication, June 11, 2001). No known drug is capable of selecting only certain questions on which to exert an effect.

The most commonly prescribed drugs in America do not influence the results of a polygraph examination. Although selected drugs target a body system, they do not assert influence upon all recorded channels (respiration, galvanic skin behavior, and cardiograph protocol) simultaneously. Accuracy of an examination may be influenced by drug consumption, however when recorded data appears untoward, the examiner usually renders No Opinion, due to uncertainty rather than commit to truthfulness or deception.

Countermeasures concerns have forced the polygraph community to re-evaluate the position previously held by many examiners, who felt that countermeasures were of no concern and that anyone attempting to employ countermeasures during testing would be easily identified by the examiner. Research has demonstrated the possible impact of countermeasures on the reliability of test results. Honts et al.(2001) have demonstrated that the use of sophisticated countermeasures by trained subjects could result in high False Negatives rates when examiners employed traditional scoring methods. Modern examiners must expect sophisticated countermeasure efforts by examinees that do extensive research and study regarding countermeasures prior to their polygraph examination. They must also be alert for the less sophisticated efforts of those persons who simply obtain some basic instruction in how to defeat the polygraph prior to their test. Moreover, the modern examiner must employ anti-countermeasures to counter this threat and thereby maintain a high level of accuracy and reliability in the polygraph examination process.

Modern examiners must adapt to the countermeasure threat present today and improve methods and measures, if this assessment process is to retain high levels of accuracy, reliability, and validity. To this end, research has provided some insight by demonstrating that the employment of countermeasures by a guilty subject does little to diminish his or her capability to respond physiologically to relevant questions during testing. Honts et al. (2001) found in one lab study that regardless of the use of

spontaneous countermeasures by examinees, the polygraph examinations disclosed consistent, significant, and timely physiological responses to relevant questions by guilty subjects. Countermeasures are most successful when examiners rely solely or almost entirely upon one form of evaluation, a strict spot numerical analysis, versus using a more all-encompassing global view in conjunction with a conservative numerical analysis. Anecdotal evidence collected from actual field examinations conducted in 2000 and 2001 by several federal agencies confirms that examinees are being correctly identified as employing countermeasures in federally administered polygraph examinations. In many of the examinations, the subjects confessed their attempts at or use of countermeasures during testing. In all cases, the subjects were recognized as responding significantly to the relevant issues during testing. The type of test data analysis utilized in these cases is indicative of the more all-encompassing analysis that must become commonplace in polygraph examinations. Given the growing efforts by guilty and innocent alike to actively influence the outcome of their examinations, examiners must be constantly alert to the threat. They must also strive to become experts at their profession, including becoming expert at identifying countermeasures and employing anti-countermeasures. The accuracy and validity of the process is at stake.

References

Department of Defense Polygraph Institute Research Division Staff (1995a). Psychological Detection of Deception Accuracy Rates obtained using the Test for Espionage and Sabotage (TES), Fort McClellan, AL: Author.

DoDPI Research Division Staff (1995b). A Comparison of Psychological Detection of Deception Accuracy Rates Obtained using the Counterintelligence Scope Polygraph (CSP) and the TES Questions Formats. Fort McClelland, AL: Author.

Honts, C. R., Amato, S. L., & Gordon, A. K. (2001). Effects of spontaneous countermeasures used against the comparison question test. *Polygraph*, 30(1), 1-8.

Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Psychology*, 70, 177-187.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology*, 1, 241-247.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, 79, 252-259.

Iacono, W. G. (2001). Letter, Re: Response to follow-up questions regarding polygraph testing from Senators Leahy and Grassley. Retrieved on May 16, 2001, from the World Wide Web: http://antipolygraph.org/hearings/senate-judiciary-2001/iacono-letter.shtml

Machke, G. W. & Scalabrini, G. J. (2000). The lie behind the lie detector. AntiPolygraph.org. Retrieved on June 4, 2001, from the World Wide Web: http://www.antipolygraph.org/pubs.shtml

MMPI-A (1999). Minnesota Multiphasic Personality Inventory-Adolescent. National Computer Systems, Inc., Retrieved on 14 May 2001 from the World Wide Web: http://assessments.ncs.com/assessments/tests/mmpia.htm.

Report on the investigation of espionage allegations against Dr. Wen Ho Lee, 8 March 2000, retrieved on June 12, 2001, from the World Wide Web: http://www.senate.gov/~specter/rept01.pdf

A Computational Guide to Power Analysis of Fixed Effects in Balanced Analysis of Variance Designs

Andrew B. Dollins[1,2]

Abstract

This manuscript provides a step-by-step guide to statistical power calculation for the fixed effects of analysis of variance (ANOVA) designs with an equal number of observations in each cell. A brief history of ANOVA hypothesis testing theory is included to explain why power calculation is important and how the results can be used. The relationship between lambda (λ), the noncentrality parameter used to calculate power in the ANOVA, and Cohen's (1988) measure of effect size is provided. Algorithms are provided for power calculation and for conversion between λ, Cohen's measure of effect size, and phi--the parameter used in many tables of the noncentral F distribution. The appendices contain power calculation examples for the main and interaction effects of 2 x 3 x 3 between- and within- subjects designs.

Key Words: Computation guide, analysis of variance (ANOVA), statistical power, lambda (λ), alpha (α), beta (β), effect size, algorithm.

Scheffe' (1959, p. 3) roughly defines the analysis of variance (ANOVA) as "a statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important and to estimate the effects." Scheffe' (1959, p. 3) attributes the development of ANOVA techniques chiefly to R. A. Fisher (1918, 1935), who was attempting to address agricultural rather than psychological research.

In practice, the ANOVA is a set of procedures for calculating the probability that a particular set of observations could have occurred by chance (i.e., randomly). Thus, a hypothesis may be rejected, with some degree of confidence, that a similar set of observations would not occur by chance. The hypothesis tested is usually the null hypothesis that two or more means (dependent variables), observed during two or more experimental manipulations (independent variables) are equal. This hypothesis may only be rejected (i.e., the groups of values are not equal) based on the ANOVA of the observed values. It is important to note that failure to reject a hypothesis does not, according to Fisherian logic, indicate acceptance of the hypothesis (Fisher, 1966, p. 16). (Cohen [1990] argues that this is a flaw in Fisherian logic because the null hypothesis is always false in the real world - given a large enough sample size.) If the probability that the observed values could have occurred by chance is less than a preset probability level (i.e., referred to as the significance criterion or alpha [α]), the null hypothesis is rejected.

Neyman and Pearson (1928a, 1928b) proposed that the specification of an alternative hypothesis be added to the ANOVA. (This concept was, according to Cohen [1990, p. 1307], violently opposed by Fisher.) Inclusion of an alternative hypothesis, to be accepted if the null hypothesis was rejected, revolutionized the decision process associated with the ANOVA. Now the ANOVA could be used to both support and reject hypotheses. Including an alternative hypothesis, with an

[1] Department of Defense Polygraph Institute
[2] To whom correspondence should be addressed at Department of Defense Polygraph Institute, Fort Jackson, SC 29207

associated effect size, allows the calculation of the probability that the null hypothesis is not rejected when it is false, as well as the probability of rejecting the null hypothesis, given that the alternative hypothesis is true--referred to as beta (β) or Type II error. Calculation of β allows calculation of its compliment (i.e., 1 - β), power, the probability that the null hypothesis is correctly rejected. The probability that the null hypothesis will be rejected when it is true, alpha (α), is referred to as the Type I error rate.

The number of observations necessary to support a hypothesis can thus be calculated--given the desired α and β probabilities and the magnitude of the difference between the null and alternate hypotheses. The power of an ANOVA test can also be calculated--given the desired α level, the number of observations, and the magnitude of the difference between the null and alternate hypotheses. Power analysis is primarily used to determine the probability that a statistically significant difference will be obtained, given a specified difference among the observations, and a specified number of observations; or the probability that a statistically significant effect would have been obtained (where none is found) if one had existed. While it is possible to calculate and use the parameters necessary to support a hypothesis with a relatively high degree of confidence it is, apparently, rarely done. This is documented by the relatively low power (< .60) of the majority of studies, in numerous research fields, to detect small and medium effects (Bones, 1972; Brewer 1972; Brewer & Owne, 1972; Brown & Hale, 1992; Chase & Tucker, 1975; Chase & Chase, 1976; Chase & Barnum, 1976; Christensen & Christensen, 1977; Cohen, 1962; Crane, 1976; Daly & Hexamer, 1983; Fagley, 1985; Frieman, Chalmers, Smith, & Kuebler, 1978; Haase, 1974; Haase, Waechter, Solomon, 1982; Hall, 1982; Jones & Brewer, 1972; Julnes & Mohr, 1989; Kosciulek, 1993; Kosciulek & Szymanski, 1993; Kroll & Chase, 1975; Orme & Combs-Orme, 1986; Orme & Tolman, 1986; Ottenbacher, 1982; Penick & Brewer, 1972; Rossi, 1990; Rothpearl, Mohs, & Davis, 1981; Sawyer & Ball, 1981; Sedlmeier & Gigerenzer, 1989; Wolley, 1983; Wooley & Dawson, 1983). According to S. E. Edgell (personal communication, August 14, 1995) the main problem with low power is that the researcher wastes time by running studies that have little chance of finding the result desired.

Perhaps one of the reasons that the power of F ratios are not calculated or reported more frequently is the difficulty associated with power calculations. Calculating the power of F ratios in ANOVA designs can be difficult, particularly for designs with more than one factor and/or repeated factors, because the majority of the calculations must be completed by hand. The most complete text on the topic of power analysis is Jacob Cohen's Statistical Power Analysis for the Behavioral Sciences (1988)--which addresses power calculation for most commonly used parametric and non-parametric statistics. Unfortunately, Cohen's (1988) calculations for the ANOVA (pp. 273-406, 550-551) are appropriate for one-way ANOVA designs but underestimate the power and overestimate the sample sizes of higher level designs (Koele, 1982). (Note: Koele was referring to the calculations described in the 1977 edition of Cohen's book - which remain the same in the more current 1988 version.) As can be seen in Appendices B and C, this is true for between- subjects designs, but the reverse is true for within-subjects designs. Cohen (1988) does not describe power calculations for repeated measure ANOVA designs in any detail, but suggestions may be found elsewhere (Bavry, 1991, pp. 63-76; Davidson, 1972, p. 448; Koele, 1982; Kraemer & Thiemann, 1987, pp. 45-52; Lipsey, 1990, pp. 79-84; Winer, 1971, p. 516).

Cohen (1988) does, however, note several important observations concerning power analysis. Statistical significance levels have generally been set by convention to .05 or .01 (Cowles & Davis, 1982). No such convention exists for power levels, however, Cohen (1988, p. 56) suggests that the value of .80 be used when the investigator has no other basis for setting the desired power value. Cohen (1988, pp. 284-288, 355) further proposes that ANOVA effect sizes, for the behavioral sciences, be categorized into small (.10), medium (.25), and large (.40) for theoretical purposes. Cohen (1988, pp. 364-367) also notes that it is possible to calculate power for separate effects of a complex factorial design independently. This is somewhat analogous to the independent

calculation of the effects in a complex factorial design.

The following guide to calculating the power of fixed effects in balanced ANOVA design F tests is designed to summarize what can be a very confusing process. The works of Bavry (1991), Borenstein and Cohen (1988), Cohen (1988), Koele (1982), and Winer (1971) were relied upon most heavily during the development of this guide. It should be noted that the processes described are based primarily on statistical theory rather than empirical evidence. Monte Carlo studies of the statistical power of ANOVA designs have, however, been reported (Cole, Maxwell, Arvey, & Salas, 1994; Cornell, Young, Seaman, & Kirk, 1992; Keselman, Rogan, Mendoze, & Breen, 1980; Klockars & Hancock, 1992). The description below pertains only to power analysis of a complex fixed effect between- and within- subjects factorial ANOVA designs with an equal number of observations in each cell (Cohen's, 1988, case 2). Adjustments for an unequal number of observations in each cell are described by Cohen.

Power Calculation

To calculate the power of the F ratio of a complex fixed effect ANOVA design, it is necessary to know the: significance criterion of the F ratio for which power is calculated (α); degrees of freedom of the numerator of the F ratio for which power is calculated; degrees of freedom of the denominator of the F ratio for which power is calculated; and, the noncentrality parameter associated with the F ratio. As detailed below, the noncentrality parameter can be calculated using Cohen's effect size--f (1988). If predicting the power of a repeated measure design using data from a between-subjects design, it is also necessary to calculate the (assumed constant) correlation between pairs of observations on the same element and factor level, as detailed below.

According to Koele (1982), the power of a fixed effect ANOVA F test is the probability that (F > Fc given df1, df2, lambda [λ]). Koele defines λ as the noncentrality parameter; df1 and df2 as the numerator and denominator, respectively, degrees of freedom of the F ratio for which power is being calculated, and Fc as the critical F value (with df1 and df2 degrees of freedom) that the F ratio must exceed at a given significance level. It is distributed as a noncentral F distribution.

Significance Criterion / Fc - Critical F Value

Fc is the F ratio, associated with a given probability (α) level which the calculated F statistic must exceed to be significantly different from chance. For instance, an observed F ratio with df1 = 3 and df2 = 20 must exceed Fc = 3.10 to be statistically significant at an α level of 0.05 and must exceed Fc = 4.94 to be statistically significant at an α level of 0.01. This value can be calculated from the central F distribution given the α level and the numerator and denominator degrees of freedom. It can also be obtained from tables of the central F distribution given in most textbooks of statistical analyses (e.g. Winer, 1971; Keppel, 1991).

Numerator Degrees of Freedom - df1

These are the degrees of freedom associated with the numerator of the F ratio for which power is calculated.

Denominator Degrees of Freedom - df2

These are the degrees of freedom associated with the denominator of the F ratio for which power is calculated.

The Noncentrality Parameter - λ

The noncentrality parameter is equal to the F statistic numerator sum of squares, with each term replaced by its expectation, divided by the within-cells error variance (i.e., the mean squares error term; Kendall & Stuart, 1966, p. 5; Scheffe', 1959, p. 39). The noncentrality parameter, λ is thus equal to the calculated F ratio times its numerator degrees of freedom. For example, the λ associated with F(2, 8) = 63.389 would be 2 * 63.389 or 126.778, and the λ associated with F(4, 16) = 0.357 would be 4 * 0.357 or 1.427 (see Appendices A, B, and C for more examples).

The Noncentral F Distribution

Once Fc, df1, df2, and k are determined, power calculation is completed by use of the noncentral F distribution. Tables of this distribution are provided by Rotton and Schonemann (1978), Tiku (1967), and most textbooks on ANOVA. Table powers are usually indexed by df1, df2, and phi (ϕ)--rather than λ.

According to Winer, Brown, and Michels (1991, p. 408), λ can be converted to ϕ using the following algorithm.

ϕ = SQRT[λ / (number of effect levels)]

Laubscher (1960) describes a square root normal approximation of the noncentral F distribution (formula 6) which may be used to calculate the power of an F ratio using a hand calculator and tables of the central F and Z distributions. While both Cohen (1988, p. 550) and Laubscher (1960) describe a cube root normal approximation, Laubscher concluded that the square root approximation was slightly more accurate for the tested data set. Cohen (1988, p. 550) comments that Laubscher's square root normal approximation of noncentral F "gave excellent agreement with exact value determinations given in the literature...except when n and f are small," but does not define small. A somewhat simplified version of Cohen's adaptation of Laubscher's square root approximation of the probabilities given by the noncentral F distribution is:

$$X1 = (df1 + 2 * \lambda) / (df1 + \lambda)$$
$$X2 = (df1 * Fc) / df2$$

$$Z = \frac{SQRT[2 * (df1 + \lambda) - X1] - SQRT[(2 * df2 - 1) * X2]}{SQRT(X1 + X2)}$$

Power >= Probability of (Z)

Where:

- df1 = numerator degrees of freedom of the original F ratio.
- df2 = denominator degrees of freedom of the original F ratio.
- λ = the non-centrality parameter.
- Fc = the value of the critical F ratio given the original F ratio degrees of freedom and significance criterion.
- Z = A Z value, the probability of which may be determined using a table of proportions of area under the standard normal curve. This probability is the probability of a Type II error (i.e., β).

The following computer programs and associated manuals were used in the preparation of this manuscript: Statistical design analysis software (Bavry, 1996); Stat-Power statistical design analysis system (Bavry, 1991); and Statistical power analysis: A computer program (Borenstein & Cohen, 1988). A review of computer programs used to calculate power analyses may be found elsewhere (Goldstein, 1989).

Effect Size

Calculating Effect Size

Calculating the power of a completed F test is thus a relatively straightforward task given the significance criterion, F ratio degrees of freedom, and λ. As mentioned above, however, power analysis is primarily useful in predicting the number of observations needed to obtain a significant effect, if one exists, with a given power, or the probability that a statistically significant effect would have been obtained if one had existed. In both cases, the F ratio necessary to predict λ does not exist and must be estimated. The discerning reader will realize that it may be difficult to estimate λ on an a priori basis. Several investigators have proposed ANOVA-based measures of effect size to assist in λ estimation, as reviewed by Tatsuoka (1993). Probably the most intuitive is Cohen's f, which is defined as the standard deviation of the effect means divided by the (common) within- cell standard deviation (Cohen, 1988, pp. 274-275). While Cohen (1988, pp. 215-406) provides several examples of the standard deviation of the effect means calculations, a detailed explanation of the (common) within-cell standard deviation is not found. Hedges (1981), however, demonstrated that the square root of the F ratio's within-cell mean square error term provides the best

unbiased estimator of the within-cell standard deviation. Thus, the terminology of Cohen (1988) and Hedges (1981) are adapted as:

Effect size (f) = SDm / SDe
Where:

f = Cohen's ANOVA-based effect size (Cohen uses the letter f to indicate effect size - this should not be confused with the uppercase F which is used to denote the F ratio).

SDm = The standard deviation of the effect means.

Sde = The square root of the within-cell mean square error term.

The effect size numerator (SDm) is calculated using one of three techniques depending on the type of factor (main effect vs. interaction) and the number of levels. Calculation procedures for the effect size denominator (SDe) for a between- subjects ANOVA design differs from those for a within-subjects ANOVA design. These are detailed below and numerical examples are provided in Appendix D. Before proceeding with the examples, a short description of the notation used is necessary. The capital letter "M" is used to indicate the mean of a cell, lower case letters are used to indicate the factor, and arabic numbers are used to indicate the factor level. A period will be used to indicate that a particular factor has been averaged. Thus: "Ma.." indicates the means associated with factor A; "Ma1.." indicates the mean of factor A, level 1; "M.b." indicates the means associated with factor B; "M..c" indicates the means associated with factor C; "Mabc" indicates the cell means associated with the A x B x C interaction; and "M..." indicates the grand mean of all values in the data set. For within-subject designs, the notation for specific observations follows the same pattern where: "Ma1..s1" indicates the average of subject 1's scores over level 1 of factor A and "M..c4s3" indicates the average of subject 3's scores over level 4 of factor C.

The following examples are for an A (2 levels) x B (3 levels) x C (4 levels) design with 5 observations per cell. The SDm is calculated in the same manner for both the within- and between-subjects designs. The same SDe term is used to calculate the effect size of each factor in a between-subject design - in the same manner as a common mean square error term is used when calculating the F ratio for each test of a between subjects design. The SDe term is used to calculate the effect size of each factor in a within-subjects design varies, as does the mean square error term used when calculating the F ratios of a within-subjects design.

The SDe term for the A (2 levels) x B (3 levels) x C (4 levels) example with 5 independent observations in each cell is:

$$SDe = SQRT\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}\sum_{l=1}^{5}(Xaibjcksl - Maibjck.)^2}{2*3*4*(5-1)}\right]$$

Note: ^ = exponentiation, thus X^2 = X*X.

Or, more simply, the square root of the average cell variance:

Factor	SDe
A	SQRT[(VARa1b1c1 + VARa2b1c1 +...+ VARa2b3c4) / 24]
B	SQRT[(VARa1b1c1 + VARa2b1c1 +...+ VARa2b3c4) / 24]
.	SQRT[(VARa1b1c1 + VARa2b1c1 +...+ VARa2b3c4) / 24]
A x B x C	SQRT[(VARa1b1c1 + VARa2b1c1 +...+ VARa2b3c4) / 24]

Where: VAR is the variance.

The general SDe term for a within-subjects ANOVA is the square root of the within-cell mean square error term used in the F ratio for which the power is being calculated. A general example is given below and examples of specific calculations for the various effects may be found in Appendix D:

$$SDe = SQRT\left[\frac{\sum_{x=1}^{2}\left[\left(\sum_{y=1}^{5}(Max..sy - Max...)\right)^2\right]}{8 \text{ (i.e., the F ratio denominator df)}}\right]$$

(1) Effect size of a main effect with 2 levels is calculated using:

$$f = \left|\frac{0.5 * (\text{maximum Ma.. - minimum Ma..})}{SDe}\right|$$

Note: The standard deviation of two values is 0.5 * the difference between the two values.

(2) Effect size of a main effect with more than 2 levels is calculated using:

$$f = \frac{SQRT\{[\sum_{x=1}^{N}(M..cx - M...)^2] / N\}}{SDe}$$

(3) Interaction effect sizes are the square root of the summed squares of the contribution of each cell to the effect divided by the number of cells. The contribution of each cell's effect is calculated by removing the contributions of other factors to that cells effect (i.e., using the linear model). The process is similar to that used to calculate the sum of squares for an F ratio interaction. For example, the effect size for Cohen's (1988) example 8.6 (pp. 368-372) A(2 levels) x B(3 levels) interaction would be calculated as:

Xa1b1. = Ma1b1. - Ma1.. - M.b1. + M...
Xa1b2. = Ma1b2. - Ma1.. - M.b2. + M...
Xa1b3. = Ma1b3. - Ma1.. - M.b3. + M...
Xa2b1. = Ma2b1. - Ma2.. - M.b1. + M...
Xa2b2. = Ma2b2. - Ma2.. - M.b2. + M...
Xa2b3. = Ma2b3. - Ma2.. - M.b3. + M...

$$f = \frac{SQRT\{\sum_{x=1}^{2}[\sum_{y=1}^{3}(Xaxby \wedge 2)] / (2*3)\}}{SDe}$$

Note: Calculating the cell contributions can become quite complex. A good guide for the factors and signs may be found in Kirk (1968). The X???s used to calculate the SDm for Cohen's example 8.6 A x B x C effect would be:

Xabc = Mabc - Mab. - Ma.c - M.bc + Ma.. + M.b. + M..c - M...

Converting Cohen's Effect Size to λ

Cohen's (1988) ANOVA-based measure of effect size can be converted to k using the following algorithm.

λ = f^2 * (the total number of observations analyzed for the effect)

The total number of observations analyzed for an effect is the number of observations used in calculating the error term and will differ for within- and between-subjects ANOVA designs. For example, the number of observations for the effects of an A (2 levels) x B (3 levels) x C (4 levels) ANOVA with 5 observations per cell, analyzed as a within- or between-subjects design would be:

Effect	Total Number of Observations Within-subjects	Total Number of Observations Between-subjects
A	10	120
B	15	120
AxB	15	120
C	20	120
AxC	20	120
BxC	120	120
AxBxC	120	120

A Note Concerning Cohen's Description of ANOVA Power Calculation

The power tables for ANOVA designs provided by Cohen (1988, pp. 273-406) require specification of: a desired significance criterion; an effect size; the F ratio numerator degrees of freedom; and the sample size. Cohen (1988, p. 365) indicates that it is necessary to use an adjusted samples size to cope with the discrepancy in denominator (error) degrees of freedom between one-way and higher-way ANOVA designs. Cohen (1988, p. 365) describes the calculation of sample size (n') as follows:

$$\text{sample size} = n' = \frac{\text{denominator } df}{u + 1} + 1$$

Where:

u = the degrees of freedom associated with the numerator of the F ratio for which power is to be calculated.

denominator df = total number of observations in the analysis minus the total number of cells in the analysis.

An example calculation of n for each of the effects in a 2(A) x 3(B) x 4(C) ANOVA with 5 observations per cell (Cohen's example 8.6, p. 368-372) would be:

Total observations = 120 (i.e., 2 * 3 * 4 * 5)
Total number of cells = 24 (i.e., 2 * 3 * 4)
Denominator df = 120 - 24 = 96

Effect	Numerator df	n
A	1	49.0
B	2	33.0
C	3	25.0
AxB	2	33.0
AxC	3	25.0
BxC	6	14.7
AxBxC	6	14.7

This adjustment works well for a one-way ANOVA design. However, as noted by Koele (1982), and illustrated in Appendices B and C, using Cohen's technique to calculate the power of effects in higher-way ANOVA designs will result in an underestimation of the power of between-subjects design effects and overestimation of the power of within-subjects effects. It is thus suggested that Cohen's ANOVA-based effect size measure be converted to and/or from λ and noncentral F distribution probabilities be used to estimate power. This will ensure accurate results and is, in addition, less complicated computationally.

Constant Correlation

An assumption in repeated measures ANOVA is that there is a "constant" correlation between pairs of observations on the same subject under different conditions (Winer, 1971, p. 516). Winer (1971, p. 516) and others (Lipsey, 1990, pp. 79-84; Davidson, 1972, p. 448; Kraemer & Thiemann, 1987, pp. 45-52) suggest that SDe should be increased or decreased according to the constant correlation when attempting to estimate the SDe for a within- subjects ANOVA design using existing data from a study with a between-subjects ANOVA design (details below). A problem occurs when deciding how to estimate the constant correlation. When comparing only two observations, the product-moment correlation may be used as an estimate of the constant correlation. Dr. Bavry (personal communication) and others (Silver & Dunlap, 1987; Silver & Hollingsworth, 1989; Viana, 1980, 1993) suggest that the best estimate of the constant correlation is calculated by averaging the Fisher's Z transform (Fisher, 1921) of all of the within-subjects between-cell correlations, then converting that Fisher's Z transform average back to a correlation coefficient. An numerical example of constant correlation calculation for data presented in Appendix A is given in Appendix E. Fisher's Z transform and its inverse are as follows (Silver & Dunlap, 1987).

Fisher's Z transform is: $Z = 0.5 * \log_e [(1+r)/(1-r)]$

The inverse transform is: $r = (X - 1) / (X + 1)$

Where:

r = the correlation coefficient

$X = \exp_e (2 * Z)$

Note: The constant correlation correction is only necessary when attempting to estimate the SDe for a within-subjects ANOVA design using existing data from a study with a between-subjects ANOVA design.

According to Winer (1971, p. 516), the following correction should be used to adjust estimates of SDe obtained from between-subjects designs when calculating power analyses of F ratios involving repeated measures. The SDe of repeated measure interaction and main effects should be adjusted by multiplying SDe by $(1-r)$, where r is the constant correlation for that effect. The SDe of between groups effects which are composed of repeated measures on each member of a group should be adjusted by multiplying SDe by $(1 + W * r)$, where W is the tested effect degrees of freedom and r is the constant correlation for that effect.

Description of the Appendices

Appendix A contains the results of between-subjects and within-subjects ANOVA of data presented by Winer (1962, p. 324; 1971, p. 546). Appendices B and C contain the results of a power analysis of the data in Appendix A using the suggested noncentral F distribution and Cohen's tables, respectively. A comparison of the results obtained using the

two methods illustrates the tendency of Cohen's technique to overestimate between-subjects and underestimate within-subjects higher-way ANOVA effect powers. Appendix D contains a numerical example of the calculations necessary to obtain the data presented in Appendices B and C. Appendix E contains a numerical example of the use of Fisher's Z transform to calculate the average correlation of data in Appendix A. Appendix F contains algorithms for converting values among λ, Φ, and Cohen's effect size for ANOVA (f).

Acknowledgments

The author would like to express special thanks to James L. Bavry, Ph.D. for his many suggestions; Victor L. Cestaro, Ph.D. for his supportive comments throughout the process of deciphering and understanding the notation used by various authors; and Stephen E. Edgell, Ph.D. who, in addition to providing many useful suggestions, wrote and ran computer programs to simulate and verify the accuracy of the equations and procedures presented on pages 5 through 8 of this manuscript. This project was supported by funds from the Department of Defense Polygraph Institute as project DoDPI95-P- 0007. This manuscript is excerpted from a Federal Technical report of the same title. The full report is available from the Defense Technical Information Service #ADA300769. The views expressed in this manuscript are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Bavry, J. L. (1991). Stat-Power statistical design analysis system. Chicago, IL: Scientific Software, Inc.

Bavry, J. L. (1996). Statistical design analysis software Portland, OR: QEI Systems.

Bones, J. (1972). Statistical power analysis and geography. Professional Geographer, 24, 229-232.

Borenstein, M., & Cohen, J. (1988). Statistical power analysis: A computer program. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brewer, J. K. (1972). On the power of statistical tests in the American educational research journal. American Educatonal Research Journal, 9, 391-401.

Brewer, J. K., & Owne, P. W. (1972). A note on the power of statistical tests in the journal of educational measurement. Journal of Educational Measurement, 10, 71-74.

Brown, J., & Hale, M. S. (1992). The power of statistical studies in consultation-liaison psychiatry. Psychosomatics, 33, 437-443.

Chase, L. J., & Tucker, R. K. (1975). A power-analytic examination of contemporary communication research. Speech Monographs, 42, 29-41.

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. Journal of Applied Psychology, 61, 234-237.

Chase, L. J., & Barnum, S. J. (1976). An assessment of quantitative research in mass communications. Journalism Quarterly, 53, 308-311.

Christensen, J. E., & Christensen, C. E. (1977). Statistical power analysis of health, physical

education, and recreation research. Research Quarterly, 48, 204-208.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among dependent variables. Psychological Bulletin, 115, 465-474.

Cornell, J. E., Young, D. M., Seaman, S. L., & Kirk, R. E. (1992). Power comparisons of eight tests for sphericity in repeated measures designs. Journal of Educational Statistics, 17, 233-249.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 37, 553- 558.

Crane, J. A. (1976). The power of social intervention experiments to discriminate differences between experimental and control groups. Social Service Review, 50, 224-242.

Daly, J. A., & Hexamer, A. (1983). Statistical power in research in English education. Research in the Teaching of English, 17, 157-164.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. Psychological Bulletin, 77, 446-452.

Fagley, N. S. (1985). Applied statistical power analysis and the interpretation of non-significant results by research consumers. Journal of Counseling Psychology, 32, 391-396.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society, Edinburgh, 52, 399-433.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron, 1, 1-32.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1966). The design of experiments (8th ed.). New York: Hafner Publishing.

Freiman, J. A., Chalmers, T. C., Smith, H. Jr., & Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative trials". The New England Journal of Medicine, 299, 690-694.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. The American Statistician, 43, 253-260.

Haase, R. F. (1974). Power analysis of research in counselor education. Counselor Education

and Supervision, 14, 124-132.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. Journal of Counseling Psychology, 29, 58-65.

Hall, J. C. (1982). The other side of statistical significance: A review of type II errors in the Australian medical literature. Australia and New Zealand Journal of Medicine, 12, 7-9.

Hedges, L. B. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Jones, B. J., & Brewer, J. K. (1972). An analysis of the power of statistical tests reported in the Research Quarterly. Research Quarterly, 43, 23-30.

Julnes, G. Mohr, L. B. (1989). Analysis of no-difference findings in evaluation research. Evaluation Review, 13(6), 628-655.

Kendall, M. G., & Stuart, A. (1966). The advanced theory of statistics (Vol. 3). London: Griffin.

Keppel, G. (1991). Design and analysis, a researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keselman, H. J., Rogan, J. C., Mendoza, J, L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures F tests. Psychological Bulletin, 87, 479-481.

Kirk, R. E. (1968). Experimental Design: Procedures for the behavioral sciences. Belmont, CA: Brooks/Cole Publishing Company.

Klockars, A. J., & Hancock, G. R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. Psychological Bulletin, 111, 505-510.

Koele, P. (1982). Calculating power in analysis of variance. Psychological Bulletin, 92, 513-516.

Kosciulek, J. F. (1993). The statistical power of vocational evaluation research. Vocational Evaluation and Work Adjustment Bulletin, 26, 142-145.

Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. Rehabilitation Counseling Bulletin, 36, 212-219.

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. London: Sage.

Kroll, R. M., & Chase, L. J. (1975). Communication disorders: A power analytic assessment of recent research. Journal of Communication Disorders, 8, 237-247.

Laubscher, N. F. (1960). Normalizing the noncentral t and F distributions. Annals of Mathematical Statistics, 31, 1105-1112.

Lipsey, M. W. (1990). Design sensitivity. Newbury Park, CA: Sage.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference (Part I). Biometrika, 20A, 175-240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference (Part II). Biometrika, 20A, 263-294.

Orme, J. G., & Tolman, R. M. (1986). The statistical power of a decade of social work education research. Social Service Review, 60, 620-632.

Orme, J. G., & Combs-Orme, T. D. (Fall / 1986). Statistical power and type II errors in social work research. Social Work Research & Abstracts, 22, 3-10.

Ottenbacher, K. (1982). Statistical power and research in occupational therapy. Occupational Therapy Journal of Research, 2, 13-25.

Penick, J. E., & Brewer, J. K. (1972). The power of statistical tests in science teaching research. Journal of Research in Science Teaching, 9, 377-381.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Rothpearl, A. B., Mohs, R. C., & Davis, K. L. (1981). Statistical power in biological psychiatry. Psychiatry Research, 5, 257-266.

Rotton, J., & Schonemann, P. H. (1978). Power tables for analysis of variance. Educational and Psychological Measurement, 38, 213-229.

Sawyer, A. G., & Ball, A. D. (1981). Statistical power and effect size in marketing research. Journal of Marketing Research, 18, 275-290.

Scheffe', H. (1959). The analysis of variance. New York: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's Z transform be used? Journal of Applied Psychology, 72, 146-148.

Silver, N. C., & Hollingsworth, S. C. (1989) A Fortran 77 program for averaging correlation coefficients. Behavior Research Methods, Instruments, and Computers, 21, 647-650.

Tatsuoka, M. (1993). Effect size. In G. Keren & C. Lewis (Eds.), The handbook for data analysis in the behavioral sciences: Methodological issues (pp. 461-479). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tiku, M. L. (1967). Tables of the power of the F test. Journal of the American Statistical Association, 62, 525-539.

Viana, M. A. G. (1980). Statistical methods for summarizing independent correlational results. Journal of Educational Statistics, 5, 83-104.

Viana, M. A. G. (1993). On a criterion for combining correlational data. Journal of Educational Statistics, 18, 261-270.

Winer, B. J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design (3rd ed.). New York: McGraw-Hill.

Woolley, T. W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 58, 710-715.

Woolley, T. W., & Dawson, G. O. (1983). A follow-up power analysis of the tests used in the journal of research in science teaching. Journal of Research in Science Teaching, 20, 673-681.

Appendix A

Example Data and Analyses of Variance

The Raw Data (Winer, 1962, p. 324, 1971, p. 546):

		Period	P1			P2			P3		
Subject	Noise	Dial	D1	D2	D3	D1	D2	D3	D1	D2	D3
1	N1		45	53	60	40	52	57	28	37	46
2	N1		35	41	50	30	37	47	25	32	41
3	N1		60	65	75	58	54	70	40	47	50
4	N2		50	48	61	25	34	51	16	23	35
5	N2		42	45	55	30	37	43	22	27	37
6	N2		56	60	77	40	39	57	31	29	46

Results of the analysis of Winer's data (1962, p. 324; 1971, p. 546) as a 2(NOISE-between) x 3(PERIOD-within) x 3(DIAL-within) ANOVA

BETWEEN SUBJECTS

SOURCE	SS	DF	MS	F	P
NOISE	468.167	1	468.167	0.752	0.435
ERROR	2491.111	4	622.778		

WITHIN SUBJECTS

SOURCE	SS	DF	MS	F	P	G-G	H-F
PERIOD	3722.333	2	1861.167	63.389	0.000	0.000	0.00
NOISE*PERIOD	333.000	2	166.500	5.671	0.029	0.057	0.02
ERROR	234.889	8	29.361				
GREENHOUSE-GEISSER EPSILON:	0.6476		HUYNH-FELDT EPSILON:			1.0000	
DIAL	2370.333	2	1185.167	89.823	0.000	0.000	0.00
NOISE*DIAL	50.333	2	25.167	1.907	0.210	0.215	0.21
ERROR	105.556	8	13.194				
GREENHOUSE-GEISSER EPSILON:	0.9171		HUYNH-FELDT EPSILON:			1.0000	
PERIOD*DIAL	10.667	4	2.667	0.336	0.850	0.729	0.85
NOISE*PERIOD*DIAL	11.333	4	2.833	0.357	0.836	0.716	0.83
ERROR	127.111	16	7.944				
GREENHOUSE-GEISSER EPSILON:	0.5134		HUYNH-FELDT EPSILON:			1.0000	

Results of the analysis of Winer's data (1962, p. 324; 1971, p. 546) as a 2(NOISE-between) x 3(PERIOD-between) x 3(DIAL-between) ANOVA

SOURCE	SS	DF	MS	F	P
NOISE	468.167	1	468.167	5.696	0.022
PERIOD	3722.333	2	1861.167	22.646	0.000
DIAL	2370.333	2	1185.167	14.421	0.000
NOISE*PERIOD	333.000	2	166.500	2.026	0.147
NOISE*DIAL	50.333	2	25.167	0.306	0.738
PERIOD*DIAL	10.667	4	2.667	0.032	0.998
NOISE*PERIOD*DIAL	11.333	4	2.833	0.034	0.998
ERROR	2958.667	36	82.185		

Appendix B

Power Calculations using the noncentral F distribution

POWER of Winer's (1962, p. 324; 1971, p. 546) 2(NOISE-between) x 3(PERIOD-within) x 3(DIAL-within) ANOVA example using Bavry's Non-central Cumulative F Probability calculation.

Factor	df		λ	Non-central Cumulative Probability (β)	Power (1 - β)	
	Numerator	Denominator				
NOISE	1	4	0.752	0.896	0.104	+
PERIOD	2	8	126.778	0.000	0.999	
NOISE*PERIOD	2	8	11.342	0.303	0.697	
DIAL	2	8	179.652	0.000	0.999	
NOISE*DIAL	2	8	3.815	0.714	0.286	+
PERIOD*DIAL	4	16	1.343	0.893	0.107	+
NOISE*PERIOD*DIAL	4	16	1.427	0.889	0.111	+

POWER of Winer's (1962, p. 324; 1971, p. 546) 2(NOISE-between) x 3(PERIOD-between) x 3(DIAL-between) ANOVA data using Bavry's Non-central Cumulative F Probability calculation.

Factor	df		λ	Non-central Cumulative Probability (β)	Power (1 - β)	
	Numerator	Denominator				
NOISE	1	36	5.697	0.358	0.642	
PERIOD	2	36	45.292	0.000	1.000	
DIAL	2	36	28.841	0.002	0.998	
NOISE*PERIOD	2	36	4.052	0.610	0.390	+
NOISE*DIAL	2	36	0.612	0.905	0.095	+
PERIOD*DIAL	4	36	0.130	0.944	0.056	+
NOISE*PERIOD*DIAL	4	36	0.138	0.944	0.056	+

+ These power values are given to illustrate the use of the cited formulae. They are not indicative of the power of the original F ratio because the original F ratio did not reach significance at the 0.05 level.

Appendix C

Power Calculations using Cohen's tables and technique

POWER of Winer's (1962, p. 324; 1971, p. 546) 2(NOISE-between) x 3(PERIOD-within) x 3(DIAL-within) ANOVA example using Cohen's method.

	u	n/group	n'	SDm	SDe	f	POWER
NOISE	1.000	3.000	3.000	2.944	8.318	0.354	0.111 +
PERIOD	2.000	6.000	6.000	8.302	3.129	2.653	0.999
NOISE*PERIOD	2.000	6.000	6.000	2.483	3.129	0.794	0.705 +
DIAL	2.000	6.000	6.000	6.625	2.097	3.159	0.999
NOISE*DIAL	2.000	6.000	6.000	0.965	2.097	0.460	0.337 +
PERIOD*DIAL	4.000	9.800	10.800	0.444	2.818	0.157	0.117 +
NOISE*PERIOD*DIAL	4.000	9.800	10.800	0.458	2.818	0.163	0.123 +

POWER of Winer's (1962, p. 324; 1971, p. 546) 2(NOISE-between) x 3(PERIOD-between) x 3(DIAL-between) ANOVA data using Cohen's method.

	u	n/group	n'	SDm	SDe	f	POWER
NOISE	1.000	27.000	19.000	2.944	9.065	0.324	0.573
PERIOD	2.000	18.000	13.000	8.302	9.065	0.916	0.999
DIAL	2.000	18.000	13.000	6.625	9.065	0.731	0.980
NOISE*PERIOD	2.000	9.000	13.000	2.483	9.065	0.274	0.293 +
NOISE*DIAL	2.000	9.000	13.000	0.965	9.065	0.106	0.078 +
PERIOD*DIAL	4.000	6.000	8.200	0.444	9.065	0.049	0.051 +
NOISE*PERIOD*DIAL	4.000	3.000	8.200	0.458	9.065	0.051	0.053 +

+ These power values are given to illustrate the use of the cited formulae. They are not indicative of the power of the original F ratio because the original F ratio did not reach significance at the 0.05 level.

Appendix D

Numerical Examples of Power Calculation

Calculation of SDe for the data from Winer's (1962, p. 324; 1971, p. 546) 2 x 3 x 3 example analyzed as a 2(NOISE-between) x 3(PERIOD-between) x 3(DIAL-between) design.

N#P#D#S					
1 1 1 .	Mn=	46.6667	Var=	158.3333
1 1 2 .	Mn=	53.0000	Var=	144.0000
1 1 3 .	Mn=	61.6667	Var=	158.3333
1 2 1 .	Mn=	42.6667	Var=	201.3333
1 2 2 .	Mn=	47.6667	Var=	86.3333
1 2 3 .	Mn=	58.0000	Var=	133.0000
1 3 1 .	Mn=	31.0000	Var=	63.0000
1 3 2 .	Mn=	38.6667	Var=	58.3333
1 3 3 .	Mn=	45.6667	Var=	20.3333
2 1 1 .	Mn=	49.3333	Var=	49.3333
2 1 2 .	Mn=	51.0000	Var=	63.0000
2 1 3 .	Mn=	64.3333	Var=	129.3333
2 2 1 .	Mn=	31.6667	Var=	58.3333
2 2 2 .	Mn=	36.6667	Var=	6.3333
2 2 3 .	Mn=	50.3333	Var=	49.3333
2 3 1 .	Mn=	23.0000	Var=	57.0000
2 3 2 .	Mn=	26.3333	Var=	9.3333
2 3 3 .	Mn=	39.3333	Var=	34.3333
Grand Average =		44.2777	Var=	82.1852

SDe = SQRT(82.1852) = 9.0656

Power Calculation for NOISE Effect:

Calculation of SDm for the NOISE effect:

N#P#D#S#	Xi	=Mn...	-M....
1 . . .	2.944	=47.222	44.278
2 . . .	-2.944	=41.333	44.278

Xi		Xi^2	Sum of Squares (Xi^2) * (Observations/average)
2.944		8.670	234.083
-2.944		8.670	234.083
	SUM =	17.340	SUM = 468.167

$$SDm = \text{Sqrt}\left(\sum_{i=1}^{2} (Xi^2) / I\right) = SQRT(17.340 / 2) = 2.944$$

Calculation of SDe for the NOISE effect:

N#P#D#S#	Yj	=Mn..s	-Mn...
1 . . 1	-0.778	46.444	47.222
1 . . 2	-9.667	37.556	47.222
1 . . 3	10.444	57.667	47.222
2 . . 4	-3.222	38.111	41.333
2 . . 5	-3.778	37.556	41.333
2 . . 6	7.000	48.333	41.333

Yj	Yj^2	Sum of Squares (Yj^2) * (Observations/average)
-0.778	0.605	5.444
-9.667	93.444	841.000
10.444	109.086	981.778
-3.222	10.383	93.444
-3.778	14.272	128.444
7.000	49.000	441.000
SUM =	276.790	SUM = 2491.111

$$\underline{SDe} = \text{Sqrt}\left(\sum_{j=1}^{6} (Yj^2) \text{ / denominator } \underline{df}\right) = \text{SQRT}(276.790 / 4) = 8.318$$

COHEN	BAVRY
Effect Size = $\underline{f}$ = SDm / SDe	λ = SSm / MSe
Effect Size = $\underline{f}$ = 2.944 / 8.318	λ = 468.167 / (2491.111/4)
Effect Size = 0.354	λ = 0.752
$\underline{u}$ = 1.0	DFm = 1.0
$\underline{n}'$= (4 / (1 + 1)) + 1 = 3	DFe = 4.0
Effect power = 0.111	Effect power = 0.104

Note: Cohen's Denominator df = (6/2 - 1) * 2 = 4

Power Calculation for the PERIOD and NOISE*PERIOD Effects:

Calculation of SDm for the PERIOD effect:

N#P#D#S#	Xi	=M.p..	-M....
. 1 . .	10.055	54.333	44.278
. 2 . .	0.222	44.500	44.278
. 3 . .	-10.278	34.000	44.278

Xi		Xi^2	Sum of Squares (Xi^2) * (Observations/average)
10.055		101.103	1819.854
0.222		0.049	.882
-10.278		105.637	1901.466
	SUM =	206.789	SUM = 2722.202

$$\underline{SDm} = \text{Sqrt}\left(\sum_{i=1}^{3} (Xi^2) / I\right) = SQRT(206.789 / 3) = 8.302$$

Calculation of SDm for the NOISE*PERIOD effect:

N#P#D#S#	Xi	=Mnp..	-M.p..	-Mn...	+M....
1 1 . .	-3.4995	53.778	54.333	47.222	44.277
1 2 . .	2.0001	49.444	44.500	47.222	44.277
1 3 . .	1.5001	38.444	34.000	47.222	44.277
2 1 . .	3.5006	54.889	54.333	41.333	44.277
2 2 . .	-1.9997	39.556	44.500	41.333	44.277
2 3 . .	-1.4997	29.556	34.000	41.333	44.277

Xi	Xi^2	Sum of Squares (Xi^2) * (Observations/average)
-3.499	12.246	110.219
2.000	4.000	36.004
1.500	2.250	20.253
3.500	12.254	110.288
-1.999	3.999	35.989
-1.499	2.249	20.242
	SUM = 36.999	SUM = 332.9937

$$\underline{SDm} = \text{Sqrt}\left(\sum_{i=1}^{6} (Xi^2) / I\right) = SQRT(36.999 / 6) = 2.483$$

Calculation of SDe for the PERIOD and NOISE*PERIOD Effects:

N#	P#	D#.S#	Yj	=Mnp.s	-Mnp..	-Mn..s	+Mn...
1	1	. 1	-0.333	52.667	53.778	46.445	47.222
1	1	. 2	-2.111	42.000	53.778	37.556	47.222
1	1	. 3	2.444	66.667	53.778	57.667	47.222
2	1	. 4	1.333	53.000	54.889	38.111	41.333
2	1	. 5	-3.778	47.333	54.889	37.556	41.333
2	1	. 6	2.445	64.333	54.889	48.333	41.333
1	2	. 1	1.000	49.667	49.445	46.445	47.222
1	2	. 2	-1.778	38.000	49.445	37.556	47.222
1	2	. 3	0.778	60.667	49.445	57.667	47.222
2	2	. 4	0.333	36.667	39.556	38.111	41.333
2	2	. 5	0.889	36.667	39.556	37.556	41.333
2	2	. 6	-1.222	45.333	39.556	48.333	41.333
1	3	. 1	-0.667	37.000	38.445	46.445	47.222
1	3	. 2	3.889	32.667	38.445	37.556	47.222
1	3	. 3	-3.222	45.667	38.445	57.667	47.222
2	3	. 4	-1.667	24.667	29.556	38.111	41.333
2	3	. 5	2.889	28.667	29.556	37.556	41.333
2	3	. 6	-1.222	35.333	29.556	48.333	41.333

		Sum of Squares
Yj	Yj^2	(Yj^2) * (Observations/average)
-0.333	0.111	0.333
-2.111	4.457	13.372
2.444	5.975	17.926
1.333	1.778	5.333
-3.778	14.273	42.820
2.445	5.976	17.929
1.000	1.000	3.001
-1.778	3.161	9.483
0.778	0.605	1.815
0.333	0.111	0.333
0.889	0.790	2.371
-1.222	1.494	4.482
-0.667	0.445	1.334
3.889	15.125	45.375
-3.222	10.383	31.148
-1.667	2.778	8.333
2.889	8.346	25.039
-1.222	1.494	4.482
	SUM = 78.303	SUM = 234.910

$$SDe = \sqrt{\sum_{j=1}^{18} (Yj^2) / \text{denominator df}} = SQRT(78.303 / 8) = 3.129$$

Power Calculation for the PERIOD Effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 8.302 / 3.129	k = 3722.333 / (234.889/8)
Effect Size = 2.653	k = 126.777
u = 2.0	DFm = 2.0
n'= (15 / (2 + 1)) + 1 = 6.00	DFe = 8.0
Effect power = 1.000 (0.9999)	Effect power = 1.000 (0.9999)

Power Calculation for the NOISE*PERIOD Effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 2.483 / 3.129	k = 333.000 / (234.889/8)
Effect Size = 0.794	k = 11.342
u = 2.0	DFm = 2.0
n'= (15 / (2 + 1)) + 1 = 6.00	DFe = 8.0
Effect power = 0.7055	Effect power = 0.6973

Note: Denominator df = (18/3 - 1) * 3 = 15

Power Calculation for the DIAL and NOISE*DIAL Effects:

Calculation of SDm for the DIAL effect:

N#P#D#S#	Xi	=M..d.	-M....
. . 1 .	-6.889	37.389	44.278
. . 2 .	-2.056	42.222	44.278
. . 3 .	8.944	53.222	44.278

	Sum of Squares	
Xi	Xi^2	(Xi^2) * (Observations/average)
-6.889	47.458	854.250
-2.056	4.227	76.088
8.944	79.995	1439.912
SUM = 131.681	SUM = 2370.251	

$$SDm = \sqrt{\sum_{i=1}^{3} (Xi^2) / I} = SQRT(131.681 / 3) = 6.625$$

Calculation of SDm for the NOISE*DIAL effect:

N#P#D#S#	Xi	=Mn.d.	-M..d.	-Mn...	+M....
1 . 1 .	-0.221	40.111	37.389	47.222	44.278
1 . 2 .	1.278	46.444	42.222	47.222	44.278
1 . 3 .	-1.055	55.111	53.222	47.222	44.278
2 . 1 .	0.222	34.667	37.389	41.333	44.278
2 . 2 .	-1.277	38.000	42.222	41.333	44.278
2 . 3 .	1.056	51.333	53.222	41.333	44.278

	Sum of Squares	
Xi	Xi^2	(Xi^2) * (Observations/average)
-0.221	0.049	0.439
1.278	1.634	14.702
-1.055	1.113	10.021
0.222	0.050	0.446
-1.277	1.632	14.684
1.056	1.115	10.036
	SUM = 5.592	SUM = 50.326

$$SDm = \sqrt{\sum_{i=1}^{6} (Xi^2) / I} = SQRT(5.592 / 6) = 0.965$$

Calculation of SDe for the DIAL and NOISE*DIAL Effects:

N#	P#	D#	S#	Yj	=Mn.ds	-Mn.d.	-Mn..s	+Mn...
1	.	1	1	-1.667	37.667	40.111	46.444	47.222
1	.	1	2	-0.444	30.000	40.111	37.556	47.222
1	.	1	3	2.111	52.667	40.111	57.667	47.222
2	.	1	4	-1.111	30.333	34.667	38.111	41.333
2	.	1	5	0.444	31.333	34.667	37.556	41.333
2	.	1	6	0.667	42.333	34.667	48.333	41.333
1	.	2	1	1.667	47.333	46.444	46.444	47.222
1	.	2	2	-0.111	36.667	46.444	37.556	47.222
1	.	2	3	-1.556	55.333	46.444	57.667	47.222
2	.	2	4	0.222	35.000	38.000	38.111	41.333
2	.	2	5	2.111	36.333	38.000	37.556	41.333
2	.	2	6	-2.333	42.667	38.000	48.333	41.333
1	.	3	1	0.000	54.333	55.111	46.444	47.222
1	.	3	2	0.556	46.000	55.111	37.556	47.222
1	.	3	3	-0.556	65.000	55.111	57.667	47.222
2	.	3	4	0.889	49.000	51.333	38.111	41.333
2	.	3	5	-2.556	45.000	51.333	37.556	41.333
2	.	3	6	1.667	60.000	51.333	48.333	41.333

Yj	Yj^2	Sum of Squares (Yj^2) * (Observations/average)
-1.667	2.778	8.333
-0.444	0.198	0.593
2.111	4.457	13.370
-1.111	1.235	3.704
0.444	0.198	0.593
0.667	0.444	1.333
1.667	2.778	8.333
-0.111	0.012	0.037
-1.556	2.420	7.259
0.222	0.049	0.148
2.111	4.457	13.370
-2.333	5.444	16.333
0.000	0.000	0.000
0.556	0.309	0.926
-0.556	0.309	0.926
0.889	0.790	2.370
-2.556	6.531	19.593
1.667	2.778	8.333
	SUM = 35.185	SUM = 105.556

$$SDe = \sqrt{\sum_{j=1}^{18} (Yj^2) / \text{denominator df}} = SQRT(35.185 / 8) = 2.097$$

Power Calculation for the DIAL Effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 6.625 / 2.097	k = 2370.333 / (105.556/8)
Effect Size = 3.159	k = 179.652
u = 2.0	DFm = 2.0
n'= (15 / (2 + 1)) + 1 = 6.00	DFe = 8.0
Effect power = 1.000 (0.9999)	Effect power = 1.000 (0.9999)

Power Calculation for the NOISE*DIAL Effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 0.965 / 2.097	k = 50.333 / (105.556/8)
Effect Size = 0.460	k = 3.815
u = 2.0	DFm = 2.0
n'= (15 / (2 + 1)) + 1 = 6.00	DFe = 8.0
Effect power = 0.337	Effect power = 0.286

Note : Denominator df = (18/3 - 1) * 3 = 15

Power Calculation for the PERIOD*DIAL and NOISE*PERIOD*DIAL Effects:

Calculation of SDm for the PERIOD*DIAL effect:

N#P#D#S#	Xi	=M.pd..	-M..d..	-M.p...	+M....
. 1 1 .	0.556	48.000	37.389	54.333	44.278
. 1 2 .	-0.277	52.000	42.222	54.333	44.278
. 1 3 .	-0.277	63.000	53.222	54.333	44.278
. 2 1 .	-0.444	37.167	37.389	44.500	44.278
. 2 2 .	-0.277	42.167	42.222	44.500	44.278
. 2 3 .	0.723	54.167	53.222	44.500	44.278
. 3 1 .	-0.111	27.000	37.389	34.000	44.278
. 3 2 .	0.556	32.500	42.222	34.000	44.278
. 3 3 .	-0.444	42.500	53.222	34.000	44.278

Xi	Xi^2	Sum of Squares (Xi^2) * (Observations/average)
0.556	0.309	1.855
-0.277	0.077	0.460
-0.277	0.077	0.460
-0.444	0.197	1.183
-0.277	0.077	0.460
0.723	0.523	3.136
-0.111	0.012	0.074
0.556	0.309	1.855
-0.444	0.197	1.183
	SUM = 1.778	SUM = 10.667

$$SDm = \sqrt{\sum_{i=1}^{9} (Xi^2) / I} = SQRT(1.778 / 9) = 0.444$$

Calculation of SDm for the NOISE*PERIOD*DIAL effect:

N#	P#	D#	S#	Xi	= Mnpd	-Mnp..	-Ma.d.	-M.pd.	+Mn...	+M.p..	+M..d.	-M....
1	1	1	.	-0.555	46.667	53.778	40.111	48.000	47.222	54.333	37.389	44.277
1	1	2	.	0.278	53.000	53.778	46.444	52.000	47.222	54.333	42.222	44.277
1	1	3	.	0.278	61.667	53.778	55.111	63.000	47.222	54.333	53.222	44.277
1	2	1	.	0.779	42.667	49.444	40.111	37.167	47.222	44.500	37.389	44.277
1	2	2	.	-0.721	47.667	49.444	46.444	42.167	47.222	44.500	42.222	44.277
1	2	3	.	-0.055	58.000	49.444	55.111	54.167	47.222	44.500	53.222	44.277
1	3	1	.	-0.221	31.000	38.444	40.111	27.000	47.222	34.000	37.389	44.277
1	3	2	.	0.446	38.667	38.444	46.444	32.500	47.222	34.000	42.222	44.277
1	3	3	.	-0.221	45.667	38.444	55.111	42.500	47.222	34.000	53.222	44.277
2	1	1	.	0.555	49.333	54.889	34.667	48.000	41.333	54.333	37.389	44.277
2	1	2	.	-0.278	51.000	54.889	38.000	52.000	41.333	54.333	42.222	44.277
2	1	3	.	-0.278	64.333	54.889	51.333	63.000	41.333	54.333	53.222	44.277
2	2	1	.	-0.778	31.667	39.556	34.667	37.167	41.333	44.500	37.389	44.277
2	2	2	.	0.722	36.667	39.556	38.000	42.167	41.333	44.500	42.222	44.277
2	2	3	.	0.055	50.333	39.556	51.333	54.167	41.333	44.500	53.222	44.277
2	3	1	.	0.222	23.000	29.556	34.667	27.000	41.333	34.000	37.389	44.277
2	3	2	.	-0.445	26.333	29.556	38.000	32.500	41.333	34.000	42.222	44.277
2	3	3	.	0.222	39.333	29.556	51.333	42.500	41.333	34.000	53.222	44.277

Xj	Xj^2	Sum of Squares (Xj^2) * (Observations/average)
-0.555	0.308	0.924
0.278	0.077	0.232
0.278	0.077	0.232
0.779	0.607	1.821
-0.721	0.520	1.560
-0.055	0.003	0.009
-0.221	0.049	0.147
0.446	0.199	0.597
-0.221	0.049	0.147
0.555	0.308	0.924
-0.278	0.077	0.232
-0.278	0.077	0.232
-0.778	0.605	1.816
0.722	0.521	1.564
0.055	0.003	0.009
0.222	0.049	0.148
-0.445	0.198	0.594
0.222	0.049	0.148
	SUM = 3.778	SUM = 11.333

$$\underline{SDm} = \sqrt{\sum_{i=1}^{18} (Xi^2) / I} = SQRT(3.778 / 18) = 0.458$$

Calculation of SDe for the PERIOD*DIAL and NOISE*PERIOD*DIAL Effects:

N#P#D#S#		Yj^2	Yi = Mnpds	-Mnpd.	-Mnp.s	-Mn.ds	+Mnp..	+Mn.d.	+Mn..s	Mn...
1 1 1 1	1.234	1.111	45.000	46.666	52.667	37.667	53.778	40.111	46.444	47.222
1 1 1 2	0.310	0.557	35.000	46.666	42.000	30.000	53.778	40.111	37.556	47.222
1 1 1 3	2.776	-1.666	60.000	46.666	66.667	52.667	53.778	40.111	57.667	47.222
2 1 1 4	13.454	3.668	50.000	49.333	53.000	30.333	54.889	34.667	38.111	41.333
2 1 1 5	0.048	-0.220	42.000	49.333	47.333	31.333	54.889	34.667	37.556	41.333
2 1 1 6	11.854	-3.443	56.000	49.333	64.333	42.333	54.889	34.667	48.333	41.333
1 1 2 1	0.309	-0.556	53.000	53.000	52.667	47.333	53.778	46.444	46.444	47.222
1 1 2 2	0.012	-0.111	41.000	53.000	42.000	36.667	53.778	46.444	37.556	47.222
1 1 2 3	0.445	0.667	65.000	53.000	66.667	55.333	53.778	46.444	57.667	47.222
2 1 2 4	1.777	-1.333	48.000	51.000	53.000	35.000	54.889	38.000	38.111	41.333
2 1 2 5	0.307	-0.554	45.000	51.000	47.333	36.333	54.889	38.000	37.556	41.333
2 1 2 6	3.568	1.889	60.000	51.000	64.333	42.667	54.889	38.000	48.333	41.333
1 1 3 1	0.309	-0.556	60.000	61.667	52.667	54.333	53.778	55.111	46.444	47.222
1 1 3 2	0.197	-0.444	50.000	61.667	42.000	46.000	53.778	55.111	37.556	47.222
1 1 3 3	1.000	1.000	75.000	61.667	66.667	65.000	53.778	55.111	57.667	47.222
2 1 3 4	5.443	-2.333	61.000	64.333	53.000	49.000	54.889	51.333	38.111	41.333
2 1 3 5	0.607	0.779	55.000	64.333	47.333	45.000	54.889	51.333	37.556	41.333
2 1 3 6	2.421	1.556	77.000	64.333	64.333	60.000	54.889	51.333	48.333	41.333
1 2 1 1	1.498	-1.224	40.000	42.667	49.667	37.667	49.444	40.111	46.444	47.222
1 2 1 2	0.605	-0.778	30.000	42.667	38.000	30.000	49.444	40.111	37.556	47.222
1 2 1 3	3.996	1.999	58.000	42.667	60.667	52.667	49.444	40.111	57.667	47.222
2 2 1 4	7.113	-2.667	25.000	31.667	36.667	30.333	39.555	34.667	38.111	41.333
2 2 1 5	0.605	0.778	30.000	31.667	36.667	31.333	39.555	34.667	37.556	41.333
2 2 1 6	3.568	1.889	40.000	31.667	45.333	42.333	39.555	34.667	48.333	41.333
1 2 2 1	5.968	2.443	52.000	47.667	49.667	47.333	49.444	46.444	46.444	47.222
1 2 2 2	0.789	0.888	37.000	47.667	38.000	36.667	49.444	46.444	37.556	47.222
1 2 2 3	11.116	-3.334	54.000	47.667	60.667	55.333	49.444	46.444	57.667	47.222
2 2 2 4	0.000	-0.001	34.000	36.667	36.667	35.000	39.555	38.000	38.111	41.333
2 2 2 5	1.234	1.111	37.000	36.667	36.667	36.333	39.555	38.000	37.556	41.333
2 2 2 6	1.237	-1.112	39.000	36.667	45.333	42.667	39.555	38.000	48.333	41.333
1 2 3 1	1.496	-1.223	57.000	58.000	49.667	54.333	49.444	55.111	46.444	47.222
1 2 3 2	0.012	-0.111	47.000	58.000	38.000	46.000	49.444	55.111	37.556	47.222
1 2 3 3	1.777	1.333	70.000	58.000	60.667	65.000	49.444	55.111	57.667	47.222
2 2 3 4	7.108	2.666	51.000	50.333	36.667	49.000	39.555	51.333	38.111	41.333
2 2 3 5	3.568	-1.889	43.000	50.333	36.667	45.000	39.555	51.333	37.556	41.333
2 2 3 6	0.605	-0.778	57.000	50.333	45.333	60.000	39.555	51.333	48.333	41.333
1 3 1 1	0.012	0.110	28.000	31.000	37.000	37.667	38.444	40.111	46.444	47.222
1 3 1 2	0.049	0.222	25.000	31.000	32.667	30.000	38.444	40.111	37.556	47.222
1 3 1 3	0.112	-0.334	40.000	31.000	45.667	52.667	38.444	40.111	57.667	47.222
2 3 1 4	1.000	-1.000	16.000	23.000	24.667	30.333	29.555	34.667	38.111	41.333
2 3 1 5	0.308	-0.555	22.000	23.000	28.667	31.333	29.555	34.667	37.556	41.333
2 3 1 6	2.421	1.556	31.000	23.000	35.333	42.333	29.555	34.667	48.333	41.333
1 3 2 1	3.572	-1.890	37.000	38.667	37.000	47.333	38.444	46.444	46.444	47.222
1 3 2 2	0.607	-0.779	32.000	38.667	32.667	36.667	38.444	46.444	37.556	47.222
1 3 2 3	7.108	2.666	47.000	38.667	45.667	55.333	38.444	46.444	57.667	47.222
2 3 2 4	1.777	1.333	23.000	26.333	24.667	35.000	29.555	38.000	38.111	41.333
2 3 2 5	0.308	-0.555	27.000	26.333	28.667	36.333	29.555	38.000	37.556	41.333
2 3 2 6	0.605	-0.778	29.000	26.333	35.333	42.667	29.555	38.000	48.333	41.333
1 3 3 1	3.158	1.777	46.000	45.667	37.000	54.333	38.444	55.111	46.444	47.222
1 3 3 2	0.308	0.555	41.000	45.667	32.667	46.000	38.444	55.111	37.556	47.222
1 3 3 3	5.448	-2.334	50.000	45.667	45.667	65.000	38.444	55.111	57.667	47.222
2 3 3 4	0.112	-0.334	35.000	39.333	24.667	49.000	29.555	51.333	38.111	41.333
2 3 3 5	1.234	1.111	37.000	39.333	28.667	45.000	29.555	51.333	37.556	41.333
2 3 3 6	0.605	-0.778	46.000	39.333	35.333	60.000	29.555	51.333	48.333	41.333
	127.111	-0.006								

$$SDe = \text{Sqrt}\left(\sum_{j=1}^{18} (Yj^2) / \text{denominator df}\right) = \text{SQRT}(127.111 / 16) = 2.818$$

Power Calculation for the PERIOD*DIAL effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 0.444 / 2.818	k = 10.667 / (127.111/16)
Effect Size = 0.157	k = 1.343
u = 4.0	DFm = 4.0
n' = (49 / (4 + 1)) + 1 = 10.8	DFe = 16.0
Effect power = 0.117	Effect power = 0.1068

Power Calculation for the NOISE*PERIOD*DIAL Effect:

COHEN	BAVRY
Effect Size = f = SDm / SDe	k = SSm / MSe
Effect Size = f = 0.458 / 2.818	k = 11.333 / (127.111/16)
Effect Size = 0.163	k = 1.427
u = 4.0	DFm = 2.0
n' = (49 / (4 + 1)) + 1 = 10.8	DFe = 16.0
Effect power = 0.123	Effect power = 0.110

Note: Denominator df = (54/5 - 1) * 5 = 49

Appendix E

Calculation of the Average Correlation of Data from Appendix A

Correlation Matrices:

	P1			P2			P3		
	D1	D2	D3	D1	D2	D3	D1	D2	D3
P1 D1		.9315	.9567	.6881	.4275	.8151	.5789	.3888	.5396
P1 D2			.9514	.8737	.6726	.9195	.8066	.6321	.8021
P1 D3				.7116	.4163	.7959	.6693	.3922	.6568
P2 D1					.8655	.9157	.9665	.9183	.9212
P2 D2						.8019	.7890	.9188	.8402
P2 D3							.8332	.8072	.8464
P3 D1								.8928	.9554
P3 D2									.8642
P3 D3									

Fisher's Z Transform of Correlation Matrices*:

	P1			P2			P3		
	D1	D2	D3	D1	D2	D3	D1	D2	D3
P1 D1		1.669	1.905	0.844	0.457	1.142	0.661	0.410	0.604
P1 D2			1.847	1.348	0.815	1.586	1.117	0.745	1.104
P1 D3				0.890	0.443	1.087	0.810	0.414	0.787
P2 D1					1.315	1.561	2.036	1.578	1.597
P2 D2						1.104	1.069	1.581	1.222
P2 D3							1.198	1.119	1.243
P3 D1								1.435	1.890
P3 D2									1.310
P3 D3									

* Values were rounded to three significant digits.

Average Fisher's $\underline{Z} = 1.1652$

$x = \exp_e (2 * 1.1652) = 10.28$

Average Correlation = (10.28 - 1) / (10.28 + 1) = .8226

Appendix F

Useful Conversion Algorithms

u = SQRT(k / (number of effect levels))
(Winer et al., 1991, p. 408)

u = effect size * SQRT(the total number of observations on which the effect estimate is based)
(Winer et al., 1991, p. 409)

k = effect size^2 * (the total number of observations on which the effect estimate is based)
(adapted from Cohen, 1988, p. 550)

k = F * df1 Where: df1 = the numerator degrees of freedom
(J. L. Bavry, personal communication, September 14, 1995)

k = (Sum of Squares Between) / (Mean Square Error)
(J. L. Bavry, personal communication, September 14, 1995)

The ASTM Exclusionary Standard and the APA "Litigation Certificate" Program

Jonathan Marin[1]

What is the Exclusionary Standard?

The Exclusionary Standard is an implementation of ideas proposed in the article (1) "He Said / She Said" published in Polygraph, Volume 29 Number 4 (2000) p 299. (An extended version appears on the web at URL: http://users.rcn.com/jonmarin//Polygraph1.htm .]

The concept applies established statistical principles to "He Said/She Said" situations where one of a pair of opposed witnesses is almost certainly lying. In such situations, Psychophysiological Detection of Deception (PDD) results from both witnesses can be evaluated together and reliably used to *exclude untrustworthy testimony.* The underlying statistical concept is simple enough:

- If you roll one die, the chance of getting a 6 is 16.66%.
- Roll two, and the chance of getting two 6s is only 2.77%.

When applied to *paired* PDD results, the statistical gain is clear. Suppose that the probability that either result alone will be wrong -- false positive *or* false negative -- is comparable to getting a 6 on a single roll of a die, or 16.66%. If participation is limited to examiners using standardized techniques, who irrespective of other credentials have demonstrated an accuracy rate of at least 85% in a controlled protocol, then error rates lower than 2.77% can be confidently predicted. Because the standard utilizes paired results *to exclude untrustworthy testimony* rather than to admit the results themselves into evidence, longstanding precedent against the admissibility of polygraph results need not change.

What are the Social Implications?

Application of the proposed standard will sharply reduce the incidence of perjured testimony in both the criminal and civil justice systems, and therefore offers important benefit to society as a whole. Because it opens a whole new domain for the practice of PDD, its implementation will be especially beneficial to APA members, the APA, and the PDD profession as a whole.

Applied to criminal cases, paired testing will reduce the incidence of wrongful convictions due to perjured testimony of:

- Informants testifying with an expectation of leniency
- Witnesses with an undisclosed interest in the outcome
- Police officers testifying to the voluntariness of confessions and the circumstances surrounding searches and seizures.

Applied to civil cases, it will:

- Reduce the number of groundless lawsuits initiated.
- Reduce the incidence of meritorious suits stymied by specious, perjury-dependent defenses.
- Reduce the load on the courts, thereby speeding justice for meritorious litigants.
- Reduce the incidence of tried cases that are decided incorrectly due to perjured testimony.
- Increase courts' willingness to penalize frivolous litigants and their attorneys.

Litigants and their attorneys will understand that they have little hope of winning if their opponents' key witnesses will be allowed to testify, unopposed, about the important facts in the case. The high costs of litigation provide a strong incentive against sustaining a case in the face of

[1] To whom correspondence should be addressed at P.O. Box 840, Brooklyn, NY 11202

those odds. Litigants who nevertheless persevere will risk being found frivolous by the court, and burdened with their opponents' legal fees as well as their own. In their own best interest, rational plaintiffs' attorneys will advise their clients to abandon their case, and rational defendants' attorneys will advise them to offer a quick and equitable settlement.

Paired testing need not be applied to all testimony in dispute. At the least, however, it should be applied to witnesses in both criminal and civil litigation, where:

- The facts in dispute make it likely that the case will hinge on whom the jury believes.
- The nature of the transaction makes it unlikely that either party could be honestly mistaken.

Exclusion based upon paired test results, even under the tight constraints in the proposed Standard, does not provide the absolute certainty that DNA often can. It applies, however, to a much wider range of cases than DNA, whose applicability is essentially limited to paternity cases and crimes involving intimate violence. It will permit civil cases to be resolved more quickly and fairly. It will deter many constitutional violations and other forms of police and prosecutorial misconduct that lead to wrongful convictions, and it will help innocent persons who are nevertheless erroneously accused or wrongfully convicted. For these reasons, it will foster favorable public recognition for PDD, comparable to that now enjoyed by DNA science.

Why is Standardization Desirable?

The Exclusionary Standard is at present under consideration by the American Society for Testing and Materials (ASTM) for designation as a standard. The ASTM develops and promulgates standards for product specifications, quality control, test regimens and performance criteria, terminology definitions, and professional practices. ASTM standards play a preeminent role in product design and government regulations for products ranging from Aerospace and Aircraft to Vehicle-pavement systems, and for services from Emergency Medical Services to Waste Management. The standards are formulated by some 170 committees, each composed of persons having appropriate expertise. [2]

The province of the Committee on Forensic Sciences includes forensic engineering, criminalistics, questioned documents, fire debris analysis, drug-testing analysis, and collection and preservation of physical evidence. The Committee on Forensic Psychophysiology is responsible for PDD standards in research, polygraph instrumentation, quality control, examiner education and training, and ethics. [3]

ASTM designation of the Exclusionary Standard will be valuable for several reasons. Standardization will preclude protracted negotiation on a case-by-case basis that would add cost and offer scope for obstruction and delay. Trial courts will not risk being burdened with motions contesting procedural details, nor be required to make rulings outside their expertise. Appellate courts will have an unambiguous clear base from which to gauge departures and irregularities.

Requests for standardized paired testing will be simple, straightforward, and well defined in their meaning. Parties will not stand to gain from requesting paired testing as a spurious show of good faith, with the intention of impeding the process later. Standardization will therefore allow strong inferences regarding the merits of parties' cases to be drawn from both requests and refusals.

Standardization will allow experience data from multiple jurisdictions to be combined in order to widen acceptance, and evaluate and accommodate refinements. An ASTM standard will also provide a model for legislatures to use when incorporating paired testing into rules of civil and criminal procedure.

The standard will simultaneously make PDD both more flexible and more rigorous. Demonstrable advances in instrumentation and technique can be rapidly deployed but will have to meet stringent empirical criteria. Whether for particular phases and components or for whole examination protocols, there is no constraint other than that a proposed advance be sufficiently

plausible to justify the cost of the necessary trials. The standard will thus serve as a proving ground for sound pre-test procedures, instrumentation, and scoring systems, etc., which may then become standards in their own right.

Trials of individual phases and components obviously carry a lower risk of failure than do whole new protocols. In the context of the Exclusionary Standard, their standardization will increase diversity of PDD at the detail level, complicate the preparation of defensive measures by subjects who intend to be deceptive, and help make the exclusionary process increasingly robust against the possible evolution of effective countermeasures. In a broader context, their standardization will:

- Constitute a framework of knowledge within which the effectiveness of new techniques can be precisely assessed.
- Guarantee "apples-to-apples" meaningfulness of comparative field data
- Move PDD toward the Science end of the Science-Art continuum.
- Provide sound defenses against charges that PDD is "junk science"
- Clarify performance criteria for those aspects which remain an art
- Facilitate quality control at all levels

What are the Benefits to APA Members?

Enhanced Professional Prestige

Few aspects of a society are more important than that its legal system not harm citizens unjustly and that it deserve their confidence that they can themselves seek justice with a reasonable expectation of achieving it. The high status accorded to the legal profession and its practitioners reflects this. Dislike and distrust for lawyers is deep and widespread, however, and reflects serious legitimate grievances with the legal system. People abhor being cast as extras in their own movie, helplessly reliant on the advice of counsel whose industry, competence, and honesty they are unequipped to judge. They condemn the cynicism and moral anaesthesia that is an occupational hazard of professionals who spend much of their careers espousing the causes of clients they know to be in the wrong (4, 5, 6). They question a system where outcomes often have less to do with the merits of the case than with the ability to inflict and withstand stress and expense. Most of all, perhaps, people resent that they are unable to access justice directly but may approach it only through the intermediation of lawyers, who thereby function as an obligatory priesthood standing between them and it.

The Exclusionary Standard offers relief from these grievances. It promises honest litigants a fast and fair outcome dependant on the merits of cases rather than on the machinations of attorneys. People will appreciate that it reduces their exposure to malicious lawsuits and their strain and expense if one is brought against them. They will perceive that the Standard protects them from police excess and unjust prosecution. Their recognition that the protection they enjoy is not mediated by lawyers, but by the PDD profession, will enable the profession and its practitioners to claim a prestige commensurate with its new importance.

Higher Incomes and Expanded Opportunities

Laboratory studies have repeatedly demonstrated that that we humans aren't good lie detectors, e.g. [7][8]. PDD can trace its origin to that fact. That knowledge gives dishonest litigants hope that they may prevail, and reason to pursue their case, and together with the high costs of litigation often causes honest litigants to acquiesce to less-than-optimal settlements.

The Exclusionary Standard will enable honest litigants to improve their prospects of a favorable outcome. The implications are enormous. The number of lawsuits initiated or maintained because one of the parties is lying is so large that, despite the reduction in number of lawsuits initiated, and despite the cases that will be abandoned when PDD is demanded, it will open a vast new market for PDD services.

All practitioners stand to benefit. Practitioners holding Litigation Certificates will benefit directly. Litigation is expensive. The standard is litigants' avenue of escape from having to pay attorney fees of $200 per hour or more while having to endure an open-ended gauntlet of depositions, interrogatories, and

motions drafted and answered. Litigants will readily cost-justify, and gladly pay, high rates for the one-time services of Litigation Certificate holders.

The demand for the present range of PDD services will not be affected by the standard, but the pool of examiners available to address it will shrink as examiners become certified and concentrate on litigation work. The law of supply and demand ensures that the remaining uncertified practitioners will be able to book more examinations at higher rates than at present.

The examiner shortage will also cause significant increases in the number and attractiveness of positions for new practitioners, and in the number of candidates wanting to fill them. Existing training programs will expand, and new ones will be established. The need to man them will create additional opportunities for experienced examiners.

The private-sector opportunities spawned by the Standard will be attractive to many of the examiners who now work for federal, state, and local government. Many examiners will become certified and leave government service for private-sector opportunities, facing the agencies with an urgent personnel retention problem. Agency management will realize, perhaps after a brief period of trial and error, that they can stem the exodus only by offering better pay, reduced workload, and improved conditions. The Standard thus promises benefit to all public-sector examiners, whether they decide to leave or stay on.

What Will the Certification Process Involve?

The validity of the standard depends ultimately on a high underlying rate of accuracy. The credibility of any implementation rests on solid evidence that certified examiners are capable of achieving such a rate not only as a class, but individually. The certification protocol will be modeled closely on the certification protocol currently being used.

In broad outline, that means that candidates will:

- Demonstrate a clean record and meet minimum formal education requirements
- Correctly score charts from an archive of examinations where ground truth is known
- Administer examinations in accordance with a procedure which DOD has put through trials in a well-defined laboratory setting producing accuracy rates of at least 85%
- Receive approval of their correct administration of the procedure from examiners who following that procedure in a laboratory setting have personally achieved accuracy of at least 90%

Details of PDD protocols and novel countermeasures are regularly disseminated in print and on the Internet. To ensure continued validity of any approved protocol, it will be necessary periodically for some successful candidates to participate in full-scale laboratory studies and to have their results cross-validated.

Before being approved for certification purposes, new protocols and equipment will have to pass empirical trials.

Taken as a whole, this certification process will

- Protect the integrity of the standard
- Help APA distinguish itself and its members from examiners of uncertain capability who inhabit the fringes of the profession
- Enhance respect for PDD professionals
- Identify and distinguish effective and ineffective procedures
- Critically weaken the position of those who oppose PDD on the ground it lacks a proper scientific basis. It will marginalize studies that have shown low accuracy and establish repeatable high rates of accuracy as an observed phenomenon. It is the business of scientists to explain observed phenomena, not deny them.
- Establish a baseline from which to evaluate new hardware and methodology
- Control liability exposure of Litigation Certificate holders and of the APA

What are the Benefits to the APA?

Drives Revenue Growth

Because certified examiners will be able to book more examinations and command a higher rate per examination, the Litigation Certificate promises to be an extremely valuable asset. The APA will be able to charge candidates the full costs of their certification, plus a modest premium to defray the costs of interfacing with legislatures, education of the legal community and the public, and related programs. Certificate holders and their firms will have no difficulty cost-justifying a fee to the APA for the annual renewal of their certificates.

A substantial increase in APA membership can be expected, reflecting the enhanced prestige of the APA as well as the desire of current non-members to become eligible for certification.

The standard promises to integrate PDD into the legal system. Its new and vitally important role will make the APA an attractive recipient for a wide range of government and foundation grants:

- To subsidize establishment of an initial population of Certificate holders
- To cross-validate the certification protocol to lab studies
- To validate new technology and equipment
- To validate new approaches to question formation, scoring, etc.
- To assure no degradation due to familiarity or countermeasures
- To educate the public, the legal profession, and the courts
- To characterize and quantify the effects, through time:
- On court calendars
- On disposition of criminal cases
- On police and prosecutorial misconduct
- On lawsuits filed
- On intervals between filing and settlement

Helps APA Enhance Prestige, Form Alliances and Promote its Agenda

The various education and public information programs will serve, incidentally but effectively, as powerful public relations engines for the APA. Major improvements of the legal system don't happen often. When they do, they're real news. At first the promise of improvement, and later the results in practice, will attract favorable media coverage, e.g.:

- "Manna for the Honest Litigant"
- "Unclogging the Courts"
- "Faster and Fairer"
- "Righted Wrongs

The media will have a reason to take an interest in experiments that demonstrate how poor we humans are at detecting deception. In high profile cases such as the recent O.J. Simpson "road rage" trial and the Rabbi Neulander murder trial, where witnesses tell diametrically opposite stories, PDD will be central to the case, and to the story.

Publicity and prestige will enhance the influence of the APA in its efforts to advance licensing and other items on its public policy agenda.

The APA will also be better positioned to attract a variety of influential allies:

- Bar Associations interested in promoting the credibility and integrity of the legal profession, an interest that will outweigh the parochial interests of trial lawyers, who are in the minority
- Courts interested in correct outcomes and shortened calendars
- Legal Aid and others vulnerable to "strategic" abuse of motions and the discovery process
- Public Defenders interested in focusing resources on the actually innocent
- Insurance industry interested in reducing groundless tort claims, legal malpractice claims, and litigation costs
- Corporations that design, manufacture or sell consumer products, interested in reducing exposure to inflated and fraudulent product liability suits
- Consumer advocacy groups interested in streamlining the relief process for true victims of actually defective products

Implementation of the Standard will place PDD squarely at the center of profound and highly visible improvements in the legal system, yielding many important benefits to the APA and the profession. It is in the interests of all PDD professionals to actively support ASTM acceptance of the standard,

and aggressive pursuit by the APA of the opportunities it opens.

Notes

[1] "He Said / She Said", Polygraph, Volume 29 Number 4 (2000) p 299.

[2] American Society for Testing and Materials, web site at URL: http://www.astm.org/

[3] A complete alphabetical listing of the ASTM technical committees with links to all of them is on the web at URL: http://www.astm.org/cgi-bin/SoftCart.exe/COMMIT/numcomm.html?L+mystore+azpk6464+1004357034 The links to the committees on Forensic Science and Forensic Psychophysiology are E30 and E52 respectively.

[4] James R. Elkins: The Moral Labyrinth of Zealous Advocacy, 21 Cap. U. L. Rev. 735 (1992)

[5] Charles W. Joiner: Our System of Justice and the Trial Advocate, 24 U. San Francisco L. Rev. 1, 15-19 (1989)(At the time the article was written, the author was Senior United States District Judge for the Eastern District of Michigan)

[6] Lawrence K. Hellman: The Effects of Law Office Work on the Formation of Law Students' Professional Values: Observation, Explanation, Optimization, 4 Geo. J. Leg. Ethics 537 (1991) Hellman refers to the problem as "moral malaise".

Note: The problems analyzed by [4],[5], and [6] are specific to the legal profession, and are distinct from the problem of incompetent and unscrupulous practitioners that afflicts every occupation to some extent. Canons of ethics require the litigation attorney to be the zealous advocate of his client's cause. He transgresses the bounds of his role if he presumes to restrain his advocacy in the interests of "justice" based on his personal judgement of his client's cause. Justice is presumed to emerge from the adversary system when all participants perform their defined roles well. Judgment is the province of the jury or the court. The conflict between normal values and the mandated indifference to them leads to the deadening of moral awareness that I call "moral anaesthesia".

[7] Vrij, Aldert: The impact of information and setting on detection of deception by police detectives. Journal-of-Nonverbal-Behavior; Vol. 18(2) 117-136 (Sum 1994)
360 police officers evaluated videotaped interviews for deception. They were separated into 12 groups varying by setting and information. Accuracy was low – the best being about 60%.

[8] DePaulo, Bella M.: Spotting lies: Can humans learn to do better?
Current-Directions-in-Psychological-Science; Vol. 3(3) (Jun 1994)
Found no significant difference at detecting deception between college students and trained law enforcement officers.