Polygraph

VOLUME 32 2003 NUMBER 2

Contents

A Comparative Analysis of Polygraph with other Screening and Diagnostic Tools 57
Philip E. Crewson

Scoring Cutoffs – Picking the Best 86
James R. Wygant

Detection of Deception 97
Jennifer M. C. Vendémia

Objective Assessment of Comparison Question Polygraphy 107
Vance V. MacLaren & Donald J. Krapohl

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

A Comparative Analysis of Polygraph with other Screening and Diagnostic Tools

Philip E. Crewson[1]

Abstract

The purpose of this study was to conduct a limited review of literature published between January 1986 and May 2001 concerning the accuracy and reliability of screening and diagnostic tests in polygraph, medicine, and psychology. Out of the 5,189 hits produced by the literature search, 1,158 articles and abstracts were reviewed, 145 were found to be useful resulting in data on 198 studies. For field screening assessments, the sensitivity of polygraph, medical, and psychological tools was .59, .79, and .74 respectively. Specificity of polygraph, medical, and psychological screening was .90, .94, and .78. For field diagnostic assessments, the sensitivity of polygraph, medical, and psychological tools was .92, .83, and .72. Specificity of polygraph, medical, and psychological diagnostic testing was .83, .88, and .67 respectively. Agreement was measured with kappa. Among readers in polygraph, medicine, and psychology kappa was .77, .56, and .79 respectively. Reports in the literature of polygraph's accuracy and reliability (agreement) on specific issues appear to be consistent with published studies on medical and psychological assessment tools. However, there is an enormous range of accuracy and agreement not only within polygraph but also medicine and psychology. Although there were very few polygraph screening studies, accuracy reports were lower than those in medicine and psychology.

Introduction

The purpose of this study was to conduct a limited review of the literature concerning the accuracy and reliability of screening and diagnostic tests in polygraph, medicine, and psychology. Measures in common use today for evaluating assessment tools assume perfection is the benchmark of a tool's efficacy. This inevitably causes disappointment in the performance of assessment tools, because they rarely produce 100% accuracy or reliability unless significant tradeoffs are made. The premise of this study is that something less than perfection is the common outcome of assessment tool studies. What follows is an effort to put the reported accuracy and reliability of polygraph in context with studies from the medical and psychological literature. It is important to recognize that comparing assessment tools across different disciplines and technologies will not clarify whether or not polygraph is an accurate or reliable means for detecting truth and deception. It does, however, place the less than perfect performance of polygraph along side other commonly used diagnostic and screening tools.

The literature review focuses on the validity (accuracy) and reliability (agreement) of polygraph as it compares to other assessment tools outside the framework of the detection of deception. The primary focus is on common medical (diagnostic radiology) assessment tools such as ultrasound (US), x-rays, computed tomography (CT), and magnetic resonance imaging (MRI) along with psychological assessment tools such as the Minnesota Multiphasic Personality Inventory (MMPI) and the Diagnostic and Statistical Manual of Mental Disorders (DSM-III and IV). Polygraph's approach involves a human reader using technology to measure and interpret physiological conditions and responses to make a diagnosis.

[1] This manuscript was completed in 2001 as a final report to the Department of Defense Polygraph Institute.

perfect agreement. A value of 0.0 represents no agreement. Kappa can also range to -1.0 (perfect disagreement), but there are no negative kappas reported in this study. A list of common terms there are not used is also provided in Table 2.

Upon review, a study was categorized as either analog (laboratory) or field-based (actual cases) and whether they were measuring an assessment tool in a screening or diagnostic application. Screening applications involve the use of an assessment tool on a general population in which there is no specific evidence of disease. As an example, screening mammography is routinely used on asymptomatic women in the hope of finding disease at an early stage. Diagnostic correlates with the polygraph specific issue test and is reserved for studies where there is prior evidence a condition exists, such as when a test is ordered after a clinical examination of a patient suggests an abnormality. As an example, diagnostic mammography is used on symptomatic women; those who have discovered a lump or other abnormality in the breast.

Abstracts were reviewed for evidence of comparable measures of accuracy and agreement. Exploratory studies, newsletters, commentaries, non-established scales, duplications, and studies that were unlikely to produce appropriate statistics were avoided. Agreement studies that didn't compare interpretations between two or more human raters were not used. If accuracy or agreement were reported separately by various control groups (sex, race, age) an effort was made to calculate an average. Scales that did not have an established cutoff for disease were not used. Only studies investigating a procedure in common use were used. This was determined by words and phrases in the text such as "preliminary," "could be used," "potential for." When accuracy was presented for both a newly proposed versus old established technique, only the data for the established technique were used. Studies involving the use of two or more procedures to form a decision and studies designed to stage the progression of known disease were also excluded. Medical studies involving invasive scopes were not used; nor did this review include any medical diagnostic tests outside radiology, such as pathology or cardiology. When accuracy was reported at several cutoffs, the first diseases cutoff was used if there was no other indication of recommended practice. Contrary to the review conducted by the Office of Technology Assessment (OTA) in 1983, inconclusive results were not used in the accuracy estimates. Although inconclusives were rarely mentioned in the medical and psychological literature, when they were mentioned, they were explicitly excluded from the accuracy estimates. Inconclusive interpretations were used for agreement statistics when the data were available.

Data collected on accuracy, agreement, number of subjects, and number of studies were entered into a spreadsheet. These data were double-verified for accuracy. The spreadsheet was used to sort studies and quantify summary measures. A mean, median, minimum, and maximum value were calculated to summarize the overall results of the screening and diagnostic studies found in this inquiry. This approach is similar to prior reviews. No statistical analysis was conducted nor is it recommended.

Weaknesses

Before continuing, the results of this review should be put into context by clearly noting several weaknesses in both study design and application. This report contains a fifteen-year snapshot of three literature domains, not definitive estimates of diagnostic test performance. Therefore the greatest concern is overgeneralization of the results beyond their simple intent to frame the science of accuracy and reliability. In addition, this review should be viewed with the following caveats in mind

1. This is not a systematic review of the literature in polygraph, medicine, or psychology. Specific rules were followed to collect examples of the relevant body of literature, but there was also a very utilitarian perspective taken in obtaining a sampling of accuracy and agreement reports on commonly known assessment tools.

2. The summary statistics reported for polygraph, medicine, and psychology should not be interpreted as generalizable to all assessment tools or applications within these professions. The summary statistics are simply a method of conveying the central tendency and variation of accuracy and reliability estimates reported in the literature. They are not statements of accuracy for a particular procedure or profession. There is much more that would need to be done to develop that level of precision. As an example, a systematic and replicable review should include special analytical techniques such as meta-analysis, study quality scoring, exclusion of low quality studies, and exhaustive disease-technology specific literature searches. This would be an enormous undertaking that far exceeds the objective of this study.

3. Similar to one of the weaknesses mentioned above, this review did not make any effort to determine the quality of the research that produced the statistics reported in the tables that follow.

4. It should be noted that the tools used for polygraph, medicine, and psychology are not directly comparable in either their technology, application, or patient populations.

Results

The results of this literature review are separated into several sections. After reviewing the results of the search effort, the overall results for screening and diagnostic accuracy will be presented. This will be followed by a rank-ordered comparison of accuracy as it relates to common medical and psychological diseases (e.g. appendicitis, depression). A rank-ordered comparison will also be provided by assessment technique (MRI, MMPI, etc.). The results section will conclude with a review of reader agreement.

Literature Search

A search for polygraph studies was conducted on April 1, 2001 through the index provided by National Polygraph Consultants (www.nationalpolygraphconsultants.com). Out of 152 articles found, 42 were reviewed, data from 16 articles were used representing 51 separate studies (see Table 3). A search for medical studies was conducted on April 30, 2001 through the PubMed index (www.ncbi.nlm.nih.gov/PubMed). The search looked for keywords in both the title and abstract and covered the time frame 1/1/1986 through 4/30/2001. Because there were tens of thousands of hits in PubMed, the search was refined to focus on any keywords in the title or abstract that contained both a common imaging modality (plain film, mammography, ultrasound, CT, MRI) and kappa, sensitivity, specificity, or receiver operating characteristic curve (ROC). Abstracts and/or articles from 933 articles were reviewed. Data from 90 of these articles were used representing 90 separate studies. A search for psychological literature was conducted via PsycInfo Direct (http://www.psycinfo.com) on April 29, 2001 for keywords in the abstract. The search covered 1/1/1985 to 4/29/2001. Out of 3,975 articles found, 183 were reviewed; data from 39 articles were used representing 57 separate studies.

Although analog studies were very common in the polygraph literature, none were found in the psychology literature and only two were found in the medical literature. As a result, most comparisons mentioned in the report focus on describing field polygraph accuracy (bolded in tables); however, the tables also include results for analog polygraph along with analog and field studies averaged into one combined accuracy measure. Some articles reported more than one accuracy or agreement estimate. As a result, the count of studies provided in some of the tables may sum to more than what is reported in Table 3.

Table 3
Search Results.

Field	Hits	Reviewed	Articles Used	**Studies Reported**
Polygraph	152	42	16	**51**
Medical	1,065	933	90	**90**
Psychological	3,975	183	39	**57**
Total	5,189	1,158	145	**198**

A complete listing of the data used in this report is contained in the Appendix.

Accuracy of Screening Techniques

Table 4 shows the accuracy reported by screening studies. Five polygraph screening studies were found. Based on three analog studies, the mean sensitivity of polygraph screening (.76) is greater than that reported in the two field polygraph screening studies (.59). Specificity in analog polygraph screening studies (.82) is less than field screening studies (.90). For ten medical screening studies, both the mean sensitivity (.79) and specificity (.94) are greater than polygraph. Psychology screening (36 studies) reports have greater sensitivity (.74) than polygraph, but lower specificity (.78).

Overall, the mean reported combined accuracy of screening polygraph (.74) is similar to screening psychology studies (.76), but lower than the mean combined screening accuracy for medical (.86) studies. The range between the minimum and maximum combined accuracy estimates from the literature is very different for polygraph (.69 to .80), medicine (.76 to .99), and psychology (.42 to .98). On average, polygraph screening studies use about half (467) as many subjects as psychology (996) studies and far less than reported for medical screening studies (56,581).

Accuracy of Diagnostic Techniques

There are 37 field polygraph, 94 medical, and 51 psychology diagnostic studies reported in Table 5. Mean sensitivity and specificity reported in field polygraph diagnostic studies are greater than those based on analog diagnostic studies. The mean sensitivity reported for medicine (.83) and psychology (.72) are lower than field polygraph (.92) studies. Although polygraph field studies have a mean specificity (.83) that is greater than psychology studies (.67), polygraph's specificity is similar but lower than that reported in the medical studies (.88). Overall, the mean combined diagnostic accuracy of polygraph (.88) and medical (.86) studies are very similar. The range between the minimum and maximum combined accuracy estimates from the literature are very similar for polygraph (.64 to 1.0), medicine (.60 to 1.0), and psychology (.50 to .93). On average, polygraph diagnostic studies use about half (108) as many subjects as medicine (284) and psychology (218).

Accuracy by Target Condition

Table 6 reports screening and diagnostic accuracy by target condition and assessment technique used. The list is ordered from highest to lowest mean combined accuracy. Diagnosing acute appendicitis with computed tomography (CT) has the greatest combined accuracy (.96). Based on five studies, CT has a sensitivity of .95 and specificity of .98 in the diagnosis of acute appendicitis. Diagnosing depression with the MMPI has the lowest mean combined accuracy (.67) reported in Table 6. Based on 37 studies, diagnostic field polygraph studies have an average combined accuracy of .88.

Table 4
Accuracy of Screening Techniques in Polygraph, Medicine, and Psychology.

	Polygraph				
	Analog	Field	Combined*	Medicine	Psychology
Sensitivity (TPR)					
Mean	0.76	**0.59**	0.67	**0.79**	**0.74**
Median	0.67	0.59	0.63	0.78	0.79
Minimum	0.61	0.45	0.53	0.51	0.11
Maximum	1.00	0.73	0.86	0.97	1.00
Studies	3	2	5	10	36
Specificity (TNR)					
Mean	0.82	**0.90**	0.86	**0.94**	**0.78**
Median	0.83	0.90	0.87	0.93	0.85
Minimum	0.63	0.87	0.75	0.87	0.00
Maximum	1.00	0.93	0.97	1.00	1.00
Studies	3	2	5	10	36
Combined Accuracy					
Mean	0.79	**0.74**	0.77	**0.86**	**0.76**
Median	0.72	0.74	0.73	0.85	0.78
Minimum	0.65	0.69	0.67	0.76	0.42
Maximum	1.00	0.80	0.90	0.99	0.98
Studies	3	2	5	10	36
Number of Subjects					
Mean	50	**467**	258	**56,581**	**996**
Median	40	467	253	19,758	307
Minimum	40	200	120	79	55
Maximum	71	733	402	202,070	16,235
Studies	3	2	5	10	36

*Note**=(analog + field)/2

This is similar to using ultrasound to diagnose carotid artery disease (.91), acute appendicitis (.91), and breast cancer (.90). It is also similar to using MRI (.86) and plain film (.86) to diagnose breast cancer. The combined accuracy of screening polygraph is one of the lowest reported in Table 6.

Accuracy by Evaluation Tool

Table 7 reports accuracy by type of evaluation tool. Similar to Table 6, the list is ordered from highest to lowest mean combined accuracy. Based on 37 field studies, diagnostic (specific issue) polygraph has the highest combined accuracy (.88). Overall, however, the combined diagnostic accuracy reported in field polygraph studies is very similar to those reported in MRI (.87), CT (.86), and ultrasound (.86) diagnostic studies. The MMPI, either screening (.61) or diagnostic (.67), has the lowest average combined accuracy.

Inter-Rater Agreement

Agreement among raters is measured as either the percent of cases in which two raters agree on an interpretation or the proportion of agreement beyond that expected by chance, which is represented by the kappa coefficient.

Table 5
Accuracy of Diagnostic Techniques in Polygraph, Medicine, and Psychology.

	Polygraph				
	Analog	Field	Combined	Medicine	Psychology
Sensitivity (TPR)					
Mean	0.89	**0.92**	0.91	**0.83**	**0.72**
Median	0.92	0.95	0.94	0.85	0.71
Minimum	0.63	0.71	0.67	0.25	0.37
Maximum	1.00	1.00	1.00	1.00	0.96
Studies	18	37	55	94	51
Specificity (TNR)					
Mean	0.78	**0.83**	0.81	**0.88**	**0.67**
Median	0.79	0.90	0.85	0.93	0.65
Minimum	0.49	0.43	0.46	0.44	0.41
Maximum	0.97	1.00	0.99	1.00	0.95
Studies	18	37	55	94	51
Combined Accuracy					
Mean	0.84	**0.88**	0.86	**0.86**	**0.70**
Median	0.85	0.90	0.87	0.88	0.69
Minimum	0.60	0.64	0.62	0.60	0.50
Maximum	0.98	1.00	0.99	1.00	0.93
Studies	18	37	55	94	51
Number of Subjects					
Mean	72	**108**	90	**284**	**218**
Median	55	64	60	124	84
Minimum	15	16	16	23	29
Maximum	192	959	576	4,811	1,079
Studies	18	37	55	89	51

Although these are very common measures of agreement, neither of these measures were reported often in the literature reviewed for this study. As an example, there was only one study gathered in the search of psychology literature that reported between rater percent agreement. All agreement studies reported in Table 8 are based on field studies. There were only three screening studies found that reported agreement data and these were all polygraph. There were no analog studies found. The eight polygraph studies reporting percent agreement averaged 91% among polygraph examiners compared to 81% for physicians (based on five studies). Kappa coefficients were found in all three disciplines. Based on six studies in the psychology literature, the mean kappa among psychologists is .79. This is similar to polygraph examiners (.77), but greater than reports on physicians (.56). It is important to note that kappa is a chance corrected measure.

This means that the kappa coefficient depends on both agreement and the distribution of cases used in a particular study. Two studies with identical percent agreements can have dramatically different kappas if the distribution of subject diagnoses vary (proportion of subjects with and without disease). As a result, it is very difficult to compare kappa from one study to the next either within the same discipline or between two disciplines.

Table 6
Rank Ordered "Combined Accuracy" on Common Medical and Psychological Diseases.

		Average Accuracy			
Target Condition	Technique	Sensitivity (TPR)	Specificity (TNR)	**Combined Accuracy**	Number of Studies
Acute Appendicitis	CT	0.95	0.98	**0.96**	5
Brain Tumor	MRI	0.93	0.98	**0.95**	2
Carotid Artery Disease	US	0.89	0.93	**0.91**	14
Acute Appendicitis	US	0.84	0.97	**0.91**	2
Breast Cancer	US	0.92	0.87	**0.90**	3
Deception	**Polygraph**	**0.92**	**0.83**	**0.88**	**37**
Breast Cancer	MRI	0.98	0.74	**0.86**	3
Breast Cancer (screen)	Plain Film	0.79	0.92	**0.86**	4
Multiple Sclerosis	MRI	0.73	0.93	**0.83**	2
Breast Cancer	Plain Film	0.78	0.83	**0.80**	7
Alcohol Abuse (screen)	MAST*	0.80	0.78	**0.79**	4
Deception (screen)	**Polygraph**	**0.59**	**0.90**	**0.74**	**2**
Personality Disorders	DSM-IV**	0.84	0.60	**0.72**	3
Depression	MMPI	0.68	0.65	**0.67**	25

*Also included a study using MMPI
**Also included studies using ICD-10 and a Personality Index

Table 7
Rank Ordered "Combined Accuracy" of Diagnostic and Screening Tools.

	Average Accuracy			
Evaluation Tool	Sensitivity (TPR)	Specificity (TNR)	**Combined Accuracy**	Number of Studies
Polygraph	**0.92**	**0.83**	**0.88**	**37**
MRI	0.86	0.88	**0.87**	17
CT	0.83	0.89	**0.86**	19
US	0.84	0.87	**0.86**	38
Plain Film	0.77	0.85	**0.81**	12
MAST (screening)	0.64	0.92	**0.78**	3
Polygraph (screening)	**0.59**	**0.90**	**0.74**	**2**
DSM-IV	0.72	0.68	**0.70**	1
MMPI	0.68	0.65	**0.67**	17
MMPI (screening)	0.70	0.53	**0.61**	5

Table 8
Inter-rater Agreement on Diagnostic Cases among Polygraph Examiners, Physicians, and Psychologists.

		Polygraph Examiners	Physicians	Psychologists
Percent Agree				
	Mean	**91%**	**81%**	**88%**
	Median	91%	80%	88%
	Minimum	77%	77%	88%
	Maximum	100%	85%	88%
	Studies	8	5	1
Kappa (bi-rater)				
	Mean	**0.77**	**0.56**	**0.79**
	Median	0.80	0.60	0.79
	Minimum	0.53	0.34	0.64
	Maximum	1.00	0.72	0.91
	Studies	9	13	6
Number of Subjects				
	Mean	**102**	**150**	**174**
	Median	69	138	113
	Minimum	21	41	76
	Maximum	402	308	331
	Studies	9	14	6

Conclusion

The purpose of this study was to conduct a limited review and analysis of the literature concerning the accuracy and reliability of screening and diagnostic tests in polygraph, medicine, and psychology. Out of the 5,189 hits produced by the literature search, 1,158 articles and abstracts were reviewed, 145 were found to be useful resulting in data on 198 studies. The results of this review have shown there is an enormous range in reports of accuracy and agreement not only in polygraph but also medicine (limited to diagnostic radiology) and psychology. Overall, polygraph research on specific issue tests reports accuracy results similar to medicine. In contrast, polygraph screening studies report lower accuracy than medical studies but are similar to what is reported in the psychology literature.

To put these results into perspective, its worth reviewing several methodological issues raised almost two decades ago in the Office of Technology Assessment's (OTA) report on the "Scientific Validity of Polygraph Testing." These issues, taken directly from the Conclusion of the OTA report, are as follows:

- Accuracy is affected by factors such as reader training, experience, personal bias, and examinee characteristics
- Cases and readers are often selectively chosen rather than randomly
- Criteria for ground truth are inadequate in some studies
- There is wide variability in results from multiple studies

This review found that these same methodological deficits are very evident in the medical and psychological literature.

Polemics on polygraph often correctly identify these issues, but either overstate or fail to mention that these same problems afflict much of the research in medicine and psychology. [3]

To put the results of this review in context, Table 9 contrasts the average accuracy found in field diagnostic studies with those reported in the OTA study in 1983. Since the OTA study included inconclusive results in accuracy estimates (and this study does not), it is not surprising that this current literature review reports a somewhat greater level of accuracy. As can be seen, however, the level of polygraph accuracy found by OTA is in the same "ballpark" as that reported in medicine and psychology.

Table 9
OTA Findings Compared to Study Results (Field Diagnostic Cases.)

	Average Accuracy			
Aggregate Measures	Sensitivity (TPR)	Specificity (TNR)	**Combined Accuracy**	Number of Studies
OTA Findings (w/Incl.)	.86	.76	**.81**	10
Current Polygraph Findings	**.92**	**.83**	**.88**	**37**
Medicine	.83	.88	**.86**	94
Psychology	.72	.67	**.70**	51

The findings presented in Table 9 are consistent with OTA's conclusion that research into specific issue polygraph testing has shown the technique has some validity. It does not, however, answer the question of polygraph validity in screening tests. Although not very extensive, this review reports three analog and two field screening polygraph studies. Table 10 puts these alongside what was found for medical and psychological screening. As can be seen, the mean combined accuracy of medical screening (.86) is greater than the mean reported for analog or field polygraph screening studies. Both field (.74) and analog (.79) polygraph studies are similar to psychology screening studies (.76), but the reported sensitivity of field (.59) polygraph screening is noticeably lower than the mean found for medical and psychological screening.

Table 10
Rank ordered findings for screening studies.

	Average Accuracy			
Aggregate Measures	Sensitivity (TPR)	Specificity (TNR)	**Combined Accuracy**	Number of Studies
Medicine (Field)	.79	.94	**.86**	10
Current Polygraph Findings (Analog)	**.76**	**.82**	**.79**	**3**
Psychology (Field)	.74	.78	**.76**	36
Current Polygraph Findings (Field)	**.59**	**.90**	**.74**	**2**

[3] For a recent example, see Aftergood, Steven. "Polygraph testing and the DOE National Laboratories." SCIENCE Online. Vol 290 (5493), Nov 3, 2000: 939-940.

Executive Summary
A Comparative Analysis of Polygraph with other Screening and Diagnostic Tools

Summary

There has been much debate over the past 30 years about polygraph and its accuracy, reliability, utility, and lack of theoretical foundation. It should be recognized from this literature review, however, that many of these same issues could be raised about medical and psychological diagnostic tools. Based on the results of this review, it is unlikely polygraph research will be able to reach a level of accuracy and reliability to satisfy its opponents. It suffers from the same flaws of many other diagnostic tools: it will not be 100% accurate, nor will its application from one subject to the next or by one examiner to another be invariant.

The accuracy of humans assessing humans is unlikely to be 100%. As has been shown in this brief survey of the medical and psychological literature, there is wide variation in the accuracy of diagnostic tools from one application to the next. In fact, there is often wide variation between studies focused on one diagnostic tool, such as has been seen in past polygraph reviews. Since perfection remains elusive, some professions have learned to accept and manage this uncertainty. As an example, training, standardization, and an ongoing review of procedures are used to establish a baseline of acceptable practice and alternative mechanisms are developed and employed to help clarify equivocal test results.

Recommendations

One of the goals of this study is to compare polygraph research methodology to that used in medicine (diagnostic radiology) and psychology. The author does not claim to be an expert in all these methodologies, so what follows are general impressions from the literature review and personal experience.

1. *Room for the "Medical" Perspective?* Most research found in the medical profession uses very specific terminology (sensitivity, specificity, odds ratios) and other methodological approaches (receiver operating characteristic curves) for assessing diagnostic tests. Greater use of similar terminologies and methodological approaches in polygraph may make the results of polygraph research more meaningful to those outside the polygraph and psychology community. There may also be a reservoir of methodologies and knowledge in the health technology assessment field that could be applied to polygraph research.

2. *Greater Focus on Accuracy and Reliability?* The polygraph literature reviewed for this study occasionally delved more into providing a sophisticated analysis of process and metrics instead of clarifying how the outcomes of these factors affect accuracy and reliability. In several cases, basic measures of accuracy or reliability were not immediately apparent (sensitivity, specificity, and kappa had to be hand calculated from frequencies in tables).

3. *Analog Generalizability?* Although analog studies are practically nonexistent in the medical and psychological literature, they appear to serve an important role in polygraph research and are very valuable in assessing the internal validity of a test. It is difficult, however, to see how analog studies are generalizable to an applied (clinical) setting. As with medical and psychological tools, its unlikely polygraph will be able to demonstrate efficacy with a heavy reliance on laboratory studies.

4. *Screening Studies?* Compared to the number of specific issue polygraph studies, there were relatively few screening studies available for review. Although developing a suitable gold standard (ground truth) for an evaluation of field polygraph screening is a very difficult problem to surmount, similar issues have been faced by medicine and psychology. Based on the medical and psychological literature, there appears to be a considerable range in the quality of gold standards. The point suggested from the literature is that lack of a "pure" gold standard has not stopped screening research in psychology and medicine.

Too Many Inconclusives? Although inconclusive test results were a common

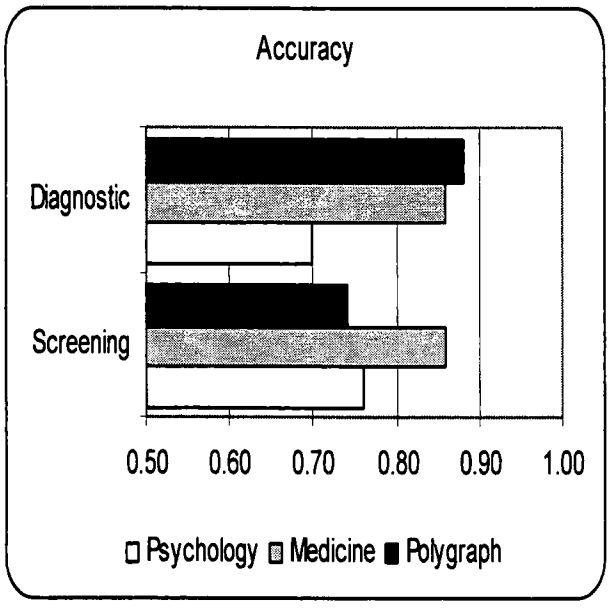
element in polygraph research, they were seldom mentioned in the medical and psychological literature. Any accuracy or reliability study which selects out only obvious interpretations will inflate accuracy estimates and threatens both the legitimacy of the research and the assessment technique.

Method

A limited review of literature published between January 1986 and May 2001 was conducted to evaluate studies reporting the accuracy and reliability of screening and diagnostic tests in polygraph, medicine, and psychology. Data for 198 studies were collected from 145 articles. Accuracy estimates are the combined average of sensitivity and specificity across all studies found within a particular category (1.00 = 100% accuracy).

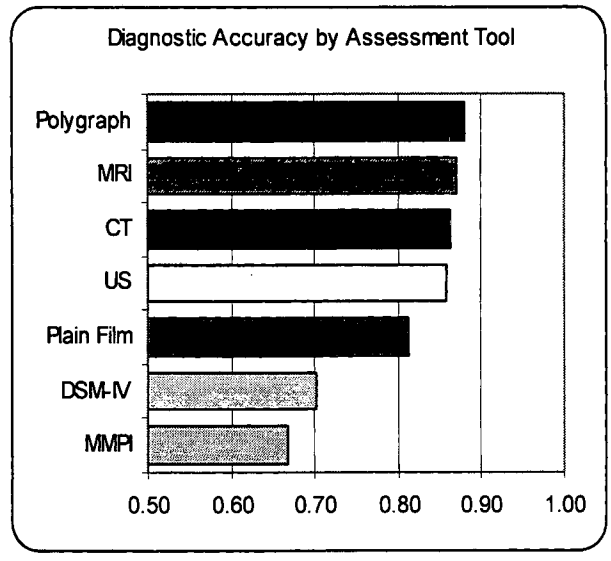
Diagnostic and Screening Accuracy

For field diagnostic assessments, the accuracy of polygraph, medical, and psychological tools was .88, .86, and .70 respectively. For field screening assessments, the accuracy of polygraph, medical, and psychological tools was .74, .86, and .76 respectively.



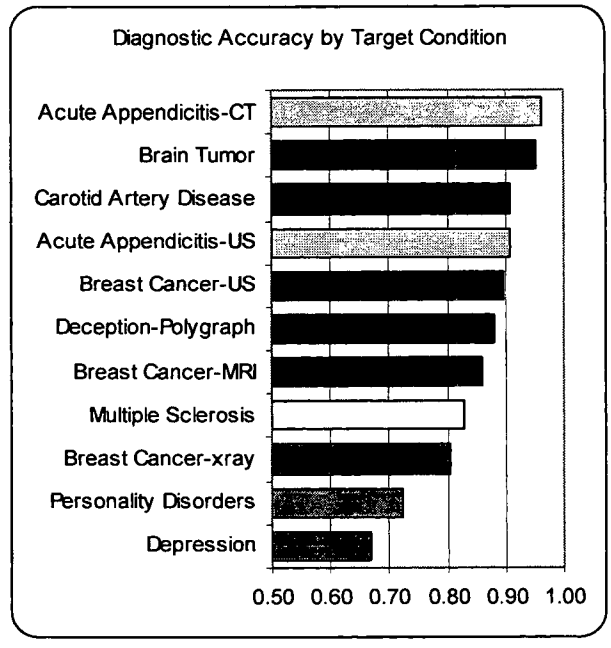
Accuracy of Various Diagnostic Tools

The average accuracy reported for 37 diagnostic polygraph studies (specific issue) was similar to MRI (17 studies), CT (19 studies), and ultrasound (38 studies). MMPI had the lowest reported accuracy (17 studies).



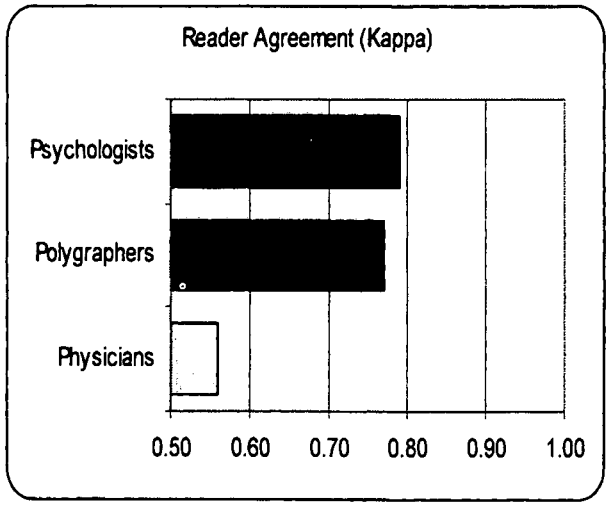
Accuracy by Target Condition

The average diagnostic accuracy for detecting deception with polygraph was similar to diagnosing breast cancer with MRI or ultrasound (US).



Agreement (kappa)

Averaging a standard measure of agreement across the reviewed literature suggests polygraph and psychology studies report similar levels of agreement. A kappa value of 1.0 represents 100% agreement beyond what would be expected by chance.



Conclusion

The level of accuracy and agreement reported in the polygraph literature is consistent with the medical and psychological literature.

Appendix

Polygraph Studies

		Field Agreement Studies					
		Screening			Diagnostic		
		Agree %	Kappa	Cases	Agree %	Kappa	Cases
	Mean	0.98		53	0.91	0.77	102
	Median	0.99		60	0.91	0.80	69
	Minimum	0.95		40	0.77	0.53	21
	Maximum	0.99		60	1.00	1.00	402
	Count	3	0	3	8	9	9
First Author	Year						
Edel & Moore	1984	0.95		40			
Yankee-Experienced examiners-with incl	1985	0.99		60			
Yankee-Inexperienced examiners-with incl	1985	0.99		60			
Elaad	1985				0.77	0.53	60
Elaad	1985				0.83	0.67	60
Patrick	1987				0.86	0.60	69
Honts	1988				0.90	0.81	21
Raskin	1988				0.91	0.80	70
Franz	1989				0.99	0.98	81
Matte	1989				1.00	1.00	114
Arellano	1990				1.00	1.00	40
Patrick & Iacono	1991					0.53	402

Medical Studies

			Field Accuracy Studies							
			Screening				Diagnostic			
			Se	Sp	Total %	Cases	Se	Sp	Total %	Cases
		Mean	0.79	0.94	0.86	56581	0.83	0.88	0.86	284
		Median	0.78	0.93	0.85	19758	0.85	0.93	0.88	124
		Minimum	0.51	0.87	0.76	79	0.25	0.44	0.60	23
		Maximum	0.97	1.00	0.99	202070	1.00	1.00	1.00	4811
		Count	10	10	10	10	94	94	94	89
First Author	Tech	Year								
Baines-breast cancer	Mammography	1988	0.75	0.94	0.85	44718				
Unk author-breast cancer	Mammography	1992	0.91	0.96	0.94	72706				
Burhenne-breast cancer	Mammography	1995	0.84	0.93	0.88	201937				
Beam-breast cancer	Mammography	1996	0.79	0.90	0.85	79				
Heinzen-breast cancer	Mammography	2000	0.78	0.92	0.85	202070				
Mettlin-prostate cancer	US	1991	0.77	0.89	0.83	2425				
Levi-congenital anomalies	US	1995	0.51	1.00	0.76	25046				
Strandell-endometrial pathology	US	1999	0.73	0.87	0.80	103				
Lennon-neural tube and ventral wall defects	US	1999	0.97	1.00	0.99	2257				
van Nagell-ovarian cancer	US	2000	0.81	0.99	0.90	14469				
Stark-Hepatic metastases	CT	1987					0.80	0.94	0.87	135
Shackford-minor head injuries	CT	1992					1.00	0.51	0.76	2166
Pasanen-unjaundiced cholestasis	CT	1994					0.53	0.86	0.70	33
van Gils-paraganglioma of the head/neck	CT	1994					0.73	0.94	0.84	60
Budoff-coronary artery disease	CT	1996					0.95	0.44	0.70	710
Rao- appendicitis	CT	1997					0.98	0.98	0.98	100
Mushlin-multiple sclerosis	CT	1997					0.25	0.95	0.60	303
Mushlin-brain tumor	CT	1997					0.93	1.00	0.97	303
Mushlin-cerebrovascular disease	CT	1997					0.88	0.95	0.92	303
D'Ippolito- appendicitis	CT	1998					0.91	1.00	0.96	52
Miller-acute flank pain	CT	1998					0.96	1.00	0.98	106
Vieweg- acute flank pain	CT	1998					0.98	0.98	0.98	105
Keberle- throat tumors	CT	1999					0.88	1.00	0.94	99
Lane- appendicitis	CT	1999					0.96	0.99	0.98	300
Garcia - appendicitis	CT	1999					0.94	0.94	0.94	139
Valk-colorectal cancer	CT	1999					0.69	0.96	0.83	115
Kurtz-ovarian cancer	CT	1999					0.92	0.89	0.91	213
Walker- appendicitis	CT	2000					0.94	1.00	0.97	65
Joseph-open-globe injuries	CT	2000					0.75	0.93	0.84	200
von Kummer-stroke damage	CT	2001					0.64	0.85	0.75	786
Stafford-blunt abdominal trauma	CT no contrast	1999					0.89	0.57	0.73	195
Stafford-blunt abdominal trauma	CT with contrast	1999					0.84	0.94	0.89	199
Martelli-breast cancer	Mammography	1990					0.73	0.80	0.77	1708
Elmore-breast cancer	Mammography	1994					0.70	0.94	0.82	150
Cwikla-breast cancer	Mammography	1998					0.70	0.57	0.64	70
Fenlon-breast cancer	Mammography	1998					0.81	0.82	0.82	44
Drew-breast cancer	Mammography	1999					0.88	0.89	0.88	285
Zonderland-breast cancer	Mammography	1999					0.83	0.97	0.90	4811
Moss-breast cancer	Mammography	1999					0.79	0.83	0.81	559
Stark-Hepatic metastases	MR	1987					0.82	0.99	0.91	135
Barronian-imaging of the knee	MR	1989					0.67	0.86	0.77	23
Glashow-anterior cruciate and meniscal lesions	MR	1989					0.83	0.84	0.84	47
Mooney-multiple sclerosis	MR	1990					0.88	0.94	0.91	
Mooney-brain infarct	MR	1990					0.88	1.00	0.94	
Mooney-brain tumor	MR	1990					0.93	0.95	0.94	
Mooney-other brain disease	MR	1990					0.91	0.92	0.92	
Young-carotid artery stenosis	MR	1994					0.89	0.82	0.86	70
Levine-osteomyelitis	MR	1994					0.77	1.00	0.89	26
Ascher-Endometriosis	MR	1995					0.76	0.60	0.68	31
Mussurakis-breast cancer	MR	1996					0.99	0.56	0.78	57
Mushlin-multiple sclerosis	MR	1997					0.58	0.91	0.75	303
Mushlin-brain tumor	MR	1997					0.93	1.00	0.97	303
Mushlin-cerebrovascular disease	MR	1997					1.00	1.00	1.00	303
Regan-acute cholecystitis	MR	1998					0.91	0.79	0.85	72
Kurtz-ovarian cancer	MR	1999					0.98	0.88	0.93	280
Drew-breast lesions	MR	1999					0.99	0.91	0.95	285
Blanchard-rotator cuff tears	MR	1999					0.79	0.81	0.80	38
Razumovsky-acute cerebral ischemia	MR	1999					0.84	1.00	0.92	30
Adamek-pancreatic cancer	MR	2000					0.84	0.97	0.91	124
Schroter-Creutzfeldt-Jakob disease	MR	2000					0.67	0.93	0.80	220
Imbriaco- breast masses	MR	2001					0.96	0.75	0.86	49
Scott-orthopedic fractures	PLAIN FILM	1993					0.79	0.83	0.81	60

Psychology Studies

		Field Accuracy Studies							
		Screening				Diagnostic			
		Se	Sp	Total %	Cases	Se	Sp	Total %	Cases
	Mean	0.74	0.78	0.76	996	0.72	0.67	0.70	218
	Median	0.79	0.85	0.78	307	0.71	0.65	0.69	84
	Minimum	0.11	0.00	0.42	55	0.37	0.41	0.50	29
	Maximum	1.00	1.00	0.98	16235	0.96	0.95	0.93	1079
	Count	36	36	36	36	51	51	51	51
First Author	Year								
Bradley-alcohol screening	1998	0.80	0.86	0.83	771				
Bradley-alcohol screening	1998	0.78	0.89	0.83	771				
Brooks-neuropsychological screening	1990	0.80	1.00	0.90	175				
Glascoe-developmental screening	1993	0.72	0.76	0.74	89				
Steer-major depression	1999	0.97	0.99	0.98	120				
Bradley-alcohol screening	1998	0.52	0.85	0.69	771				
Bradley-alcohol screening	1998	0.35	0.98	0.66	771				
Dent-memory problems in multiple sclerosis	2000	0.93	0.48	0.71	61				
Bradley-alcohol screening	1998	0.91	0.77	0.84	771				
Bradley-alcohol screening	1998	0.75	0.89	0.82	771				
Parikh-post-stroke depression	1988	0.86	0.90	0.88	80				
Baird-autism at 18 months of age	2000	0.38	0.98	0.68	16235				
Chen-attention-deficit hyperactivity	1994	0.23	1.00	0.62	122				
Scheinberg-eating disorders	1993	0.93	0.41	0.67	1112				
Scheinberg-eating disorders	1993	1.00	0.38	0.69	1112				
Gureje	1990	0.68	0.70	0.69	787				
Pomeroy-depression	2001	0.91	0.65	0.78	87				
Razavi-adjustment and major depressive disorders	1990	0.70	0.75	0.73	210				
Inwald-performance of government security personnel	1991	0.60	0.76	0.68	307				
Johnson-pathological gamblers	1998	1.00	0.85	0.93	423				
Benussi-alcoholism	1982	1.00	0.94	0.97	104				
Yersin-alcoholism	1989	0.70	0.92	0.81	268				
Erford-Math Essential skills screen	1998	0.98	0.88	0.93	100				
Uhlmann-dementia	1991	0.81	0.97	0.89	209				
Inwald-performance of government security personnel	1991	0.45	0.73	0.59	307				
Colligan-alcoholism	1988	0.74	0.84	0.79	2144				
Colligan-alcoholism	1988	1.00	0.00	0.50	2144				
Colligan-alcoholism	1988	0.62	0.34	0.48	2144				
Hirschfeld-bipolar spectrum disorder	2000	0.73	0.90	0.82	198				
Sherman-Pediatric Language Acquisition	1999	0.11	0.73	0.42	84				
Hiatt-job performance problems	1988	0.68	0.73	0.70	55				
de las Cuevas-Severity of Dependence Scale (SDS)	2000	0.98	0.94	0.96	100				
Birtchnell-depressive disorders	1989	0.94	0.86	0.90	133				
Bradley-alcohol screening	1998	0.48	0.86	0.67	771				
Bradley-alcohol screening	1998	0.91	0.72	0.82	771				
Bradley-alcohol screening	1998	0.82	0.85	0.83	771				
Berument-Autism	1999					0.85	0.75	0.80	200
Kogan-Geriatric Depression Scale	1994					0.64	0.73	0.69	59
Laprise- Geriatric Depression	1998					0.96	0.46	0.71	66
Blais-FRANTIC AVOIDANCE	1999					0.63	0.95	0.79	76
Blais-UNSTABLE RELATIONSHIPS	1999					0.94	0.91	0.93	76
Blais-IDENTITY DISTURBANCE	1999					0.73	0.84	0.79	76
Blais-IMPULSIVITY	1999					0.55	0.70	0.63	76
Blais-SUICIDAL	1999					0.96	0.41	0.69	76
Blais-AFFECTIVE INSTABILITY	1999					0.91	0.42	0.67	76
Blais-CHRONIC EMPTINESS	1999					0.52	0.79	0.66	76
Blais-POORLY CONTROLLED ANGER	1999					0.73	0.53	0.63	76
Blais-STRESS-RELATED PARANOIA	1999					0.51	0.60	0.56	76
Kogan-Geriatric Depression Scale	1994					0.79	0.69	0.74	59
Laprise-Geriatric Depression	1998					0.89	0.56	0.73	66
Merson-personality disorders	1994					0.95	0.50	0.73	29
Ivnick-MAYO VERBAL COMPREHENSION FACTOR SC	2000					0.55	0.85	0.70	1079
Chaffee-expressive and receptive language scales	1990					0.88	0.45	0.67	152
Ivnick- ATTENTION-CONCENTRATION SCORE	2000					0.71	0.70	0.71	1079
Ivnick-LEARNING FACTOR SCORE	2000					0.77	0.84	0.81	1079
Ivnick-PERCEPTUAL ORGANIZATION SCORE	2000					0.70	0.83	0.77	1079
Ivnick- RETENTION SCORE	2000					0.88	0.80	0.84	1079
BOONE 1994-depression	1994					0.61	0.62	0.62	62
WETZLER 1998-depression	1998					0.64	0.65	0.65	113
BEN-PORATH 1991-depression	1991					0.66	0.64	0.65	73
BEN-PORATH 1991-depression	1991					0.63	0.61	0.62	87
MUNLEY 1997-depression	1997					0.71	0.71	0.71	84
GREENBLATT 1999-depression	1999					0.54	0.56	0.55	75

Psychology Studies

		Field Agreement Studies					
		Screening			Diagnostic		
		Agree %	Kappa	Cases	Agree %	Kappa	Cases
	Mean		0.61	510	0.88	0.79	174
	Median		0.61	510	0.88	0.79	113
	Minimum		0.61	510	0.88	0.64	76
	Maximum		0.61	510	0.88	0.91	331
	Count	0	1	1	1	6	6
First Author	Year						
Lavigne- DSM-III–R with preschool children	1994		0.61	510			
Klin-autism	2000				0.88	0.71	131
DSM-III Phase Two Field Trials	1980					0.72	331
DSM-III Phase Two Field Trials	1980					0.64	331
Blais-NINE SCALE PERSONALITY DISORDER	1999					0.85	76
Hogervors-Alzheimer's disease	2000					0.90	82
Hilsenroth-Schizophrenia	1998					0.91	95

Articles Used-Polygraph

Aftergood, Steven. "Polygraph testing and the DOE National Laboratories." SCIENCE Online. Vol 290 (5493), Nov 3, 2000: 939-940.

Ansley, Norman. 1989. Accuracy and utility of RI screening by student examiners at DODPI. Polygraph and Personnel Security Research. Office of Security. National Security Agency. Fort George G. Meade, MD.

Ansley, Norman. 1990. "The validity and reliability of polygraph decisions in real cases. Polygraph 19(3): 169-181. (Table 2, pg 174)

Barland, Gordon H. Charles R. Honts, and Steven D. Barger. 1990. The detection of deception for multiple issues. DoDPI Project No. DoDPI89-P-0005.

Blackwell, Joan N. 1996. Polyscore: A comparison of accuracy. DoDPI Report no: DoDPI94-P-0006.

Brownlie C, Johnson GJ, and B. Knill. 1997. "Validation study of the relevant/irrelevant screening format" Unpublished report from government project. Provided by DoDPI.

Correa, Eileen Israel and Henry E. Adams. 1981. The validity of the preemployment polygraph examination and the effects of motivation. Polygraph. 143-155.

Edel, Eugene and Lane A. Moore Jr.1984. "Analysis of Agreement in polygraph charts. Polygraph 193-197. (Table II).

Honts CR and S. Amato. 1999. "The automated polygraph examination." Contract No. 110224-1998-MO.

Honts, Charles R and Lawrence N. Driscoll. 1987."An evaluation of the reliability and validity of rank order and standard numerical scoring of polygraph charts. Polygraph 241-255.

Honts, Charles R and Lawrence N. Driscoll. 1988. "A field validity study of the rank order scoring system (ROSS) in multiple issue control question tests." Polygraph 1-13

Jayne, Brian C. 1989. A comparison between the predictive value of two common preemployment screening procedures. The Investigator 5(3)

Krapohl, Donald J, Donnie W. Dutton, and Andrew H. Ryan. 2001. The Rank Order Scoring System: Replication and extension with field data. Forthcoming Polygraph 2001.

Krapohl, Donald J, Kendall W. Shull, and Andrew A. Ryan. 2000. Does the confession criterion in case selection inflate polygraph validity estimates? Paper submitted to the Forensic Science Communications, November 8, 2000. Polygraph Institute, Ft. Jackson SC.

Lykken, D.T. 1983. A tremor in the blood: Uses and abuses of the lie detector. New York: McGraw-Hill.

McCauley, Clark and Robert F. Forman. 1988. "A review of the Office of Technology Assessment report on polygraph validity." Basic and Applied Social Psychology 9(2): 73-84.

Patrick, C.J. and W.G. Iacono. 1991. "Validity of the control question polygraph test: The problem of sampling bias." Journal of Applied Psychology 76(2): 229-238.)

Scientific validity of polygraph testing: A research review and evaluation. Office of Technology Assessment. U.S. Congress: Report No. OTATMH15, November 1983.

Yankee, William J, James M. Powell, and Ross Newland. 1985. An investigation of the accuracy and consistency of polygraph chart interpretation by inexperienced and experienced examiners. Polygraph 14(2):108-117.

Articles Used-Medical

Adamek HE, Albert J, Breer H, Weitz M, Schilling D, Riemann JF. Pancreatic cancer detection with magnetic resonance cholangiopancreatography and endoscopic retrograde cholangiopancreatography: a prospective controlled study. Lancet 2000 Jul 15;356(9225):190-3

Ascher SM, Agrawal R, Bis KG, Brown ED, Maximovich A, Markham SM, Patt RH, Semelka RC. Endometriosis: appearance and detection with conventional and contrast-enhanced fat-suppressed spin-echo techniques. J Magn Reson Imaging 1995 May-Jun;5(3):251-7

Author unk; Specificity of screening in United Kingdom trial of early detection of breast cancer. BMJ 1992 Feb 8;304(6823):346-9

Ayers M, Prince M, Ahmadi S, Baran DT. Reconciling quantitative ultrasound of the calcaneus with X-ray-based measurements of the central skeleton. J Bone Miner Res 2000 Sep;15(9):1850-5

Baines CJ, McFarlane DV, Miller AB. Sensitivity and specificity of first screen mammography in 15 NBSS centres. Can Assoc Radiol J 1988 Dec;39(4):273-6

Baker JA, Kornguth PH, Lo JY, Floyd CE. 1996. Artificial neural network: Improving the quality of breast biopsy recommendations. Radiology 198:131-135.

Barronian AD, Zoltan JD, Bucon KA. Magnetic resonance imaging of the knee: correlation with arthroscopy. Arthroscopy 1989;5(3):187-91

Beam CA, Layde PM, Sullivan DC. 1996. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 156: 209-213

Belsky M, Gaitini D, Goldsher D, Hoffman A, Daitzchman M. Color-coded duplex ultrasound compared to CT angiography for detection and quantification of carotid artery stenosis. Eur J Ultrasound 2000 Sep;12(1):49-60

Blanchard TK, Bearcroft PW, Constant CR, Griffin DR, Dixon AK. Diagnostic and therapeutic impact of MRI and arthrography in the investigation of full-thickness rotator cuff tears. Eur Radiol 1999;9(4):638-42

Bozkurt T, Richter F, Lux G. Ultrasonography as a primary diagnostic tool in patients with inflammatory disease and tumors of the small intestine and large bowel. J Clin Ultrasound 1994 Feb;22(2):85-91

Braccini G, Lamacchia M, Boraschi P, Bertellotti L, Marrucci A, Goletti O, Perri G. Ultrasound versus plain film in the detection of pneumoperitoneum. Abdom Imaging 1996 Sep-Oct;21(5):404-12

Brant-Zawadzki MN, Jensen MC, Obuchowski N, Ross JS, Modic MT. Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. A comparison of two nomenclatures. Spine 1995 Jun 1;20(11):1257-63

Britton PD, Coulden RA. The use of duplex Doppler ultrasound in the diagnosis of breast cancer. Clin Radiol 1990 Dec;42(6):399-401

Budoff MJ, Georgiou D, Brody A, Agatston AS, Kennedy J, Wolfkiel C, Stanford W, Shields P, Lewis RJ, Janowitz WR, Rich S, Brundage BH. Ultrafast computed tomography as a diagnostic modality in the detection of coronary artery disease: a multicenter study. Circulation 1996 Mar 1;93(5):898-904

Burhenne LJ, Burhenne HJ, Kan L. 1995. Quality-oriented mass mammography screening. Radiology 194:185-188.

Cicinelli E, Romano F, Anastasio PS, Blasi N, Parisi C, Galantino P. Transabdominal sonohysterography, transvaginal sonography, and hysteroscopy in the evaluation of submucous myomas. Obstet Gynecol 1995 Jan;85(1):42-7

Currie IC, Jones AJ, Wakeley CJ, Tennant WG, Wilson YG, Baird RN, Lamont PM. Non-invasive aortoiliac assessment. Eur J Vasc Endovasc Surg 1995 Jan;9(1):24-8

Cwikla JB, Buscombe JR, Kelleher SM, Parbhoo SP, Thakrar DS, Hinton J, Deery AR, Crow J, Hilson AJ. Comparison of accuracy of scintimammography and X-ray mammography in the diagnosis of primary breast cancer in patients selected for surgical biopsy. Clin Radiol 1998 Apr;53(4):274-80

DePriest PD, Varner E, Powell J, Fried A, Puls L, Higgins R, Shenson D, Kryscio R, Hunter JE, Andrews SJ, et al.The efficacy of a sonographic morphology index in identifying ovarian cancer: a multi-institutional investigation. Gynecol Oncol 1994 Nov;55(2):174-8

D'Ippolito G, de Mello GG, Szejnfeld J. The value of unenhanced CT in the diagnosis of acute appendicitis. Rev Paul Med 1998 Nov-Dec;116(6):1838-45

Drew PJ, Turnbull LW, Chatterjee S, Read J, Carleton PJ, Fox JN, Monson JR, Kerin MJ. Prospective comparison of standard triple assessment and dynamic magnetic resonance imaging of the breast for the evaluation of symptomatic breast lesions. Ann Surg 1999 Nov;230(5):680-5

Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. 1994. Variability in Radiologist's Interpretations of Mammograms. NEJM. 331(22): 1493-1499.

Fenlon HM, Phelan N, Tierney S, Gorey T, Ennis JT. Tc-99m tetrofosmin scintigraphy as an adjunct to plain-film mammography in palpable breast lesions. Clin Radiol 1998 Jan;53(1):17-24

Freed KS, Paulson EK, Frederick MG, Preminger GM, Shusterman DJ, Keogan MT, Vieweg J, Smith RH, Nelson RC, Delong DM, Leder RA. Interobserver variability in the interpretation of unenhanced helical CT for the diagnosis of ureteral stone disease. J Comput Assist Tomogr 1998 Sep-Oct;22(5):732-7

Garcia Pena BM, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, Taylor GA. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. JAMA 1999 Sep 15;282(11):1041-6

Gebel M, Caselitz M, Bowen-Davies PE, Weber S.A multicenter, prospective, open label, randomized, controlled phase IIIb study of SH U 508 a (Levovist) for Doppler signal enhancement in the portal vascular system. Ultraschall Med 1998 Aug;19(4):148-56

Gentile AT, Moneta GL, Lee RW, Masser PA, Taylor LM Jr, Porter JM. Usefulness of fasting and postprandial duplex ultrasound examinations for predicting high-grade superior mesenteric artery stenosis. Am J Surg 1995 May;169(5):476-9

Glashow JL, Katz R, Schneider M, Scott WN. Double-blind assessment of the value of magnetic resonance imaging in the diagnosis of anterior cruciate and meniscal lesions. J Bone Joint Surg Am 1989 Jan;71(1):113-9

Grab D, Flock F, Stohr I, Nussle K, Rieber A, Fenchel S, Brambs HJ, Reske SN, Kreienberg R. Classification of asymptomatic adnexal masses by ultrasound, magnetic resonance imaging, and positron emission tomography. Gynecol Oncol 2000 Jun;77(3):454-9

Grotta JC, Chiu D, Lu M, Patel S, Levine SR, Tilley BC, Brott TG, Haley EC Jr, Lyden PD, Kothari R, Frankel M, Lewandowski CA, Libman R, Kwiatkowski T, Broderick JP, Marler JR, Corrigan J, Huff S, Mitsias P, Talati S, Tanne D. Agreement and variability in the interpretation of early CT changes in stroke patients qualifying for intravenous rtPA therapy. Stroke 1999 Aug;30(8):1528-33

Hedges, Larry V.; Hasselblad, Vic. Meta-analysis of screening and diagnostic tests. Psychological Bulletin. 1995 Jan Vol 117(1) 167-178

Heinzen MT, Yankaskas BC, Kwok RK. Comparison of woman-specific versus breast-specific data for reporting screening mammography performance. Acad Radiol 2000 Apr;7(4):232-6

Hill R, Conron R, Greissinger P, Heller M. Ultrasound for the detection of foreign bodies in human tissue. Ann Emerg Med 1997 Mar;29(3):353-6

Ihlberg L, Alback A, Roth WD, Edgren J, Lepantalo M. Interobserver agreement in duplex scanning for vein grafts. Eur J Vasc Endovasc Surg 2000 May;19(5):504-8

Imbriaco M, Del Vecchio S, Riccardi A, Pace L, Di Salle F, Di Gennaro F, Salvatore M, Sodano A. Scintimammography with 99mTc-MIBI versus dynamic MRI for non-invasive characterization of breast masses. Eur J Nucl Med 2001 Jan;28(1):56-63

Johansson M, Jensen G, Aurell M, Friberg P, Herlitz H, Klingenstierna H, Volkmann R. Evaluation of duplex ultrasound and captopril renography for detection of renovascular hypertension. Kidney Int 2000 Aug;58(2):774-82

Johnson MB, Wilkinson ID, Wattam J, Venables GS, Griffiths PD. Comparison of Doppler ultrasound, magnetic resonance angiographic techniques and catheter angiography in evaluation of carotid stenosis. Clin Radiol 2000 Dec;55(12):912-20

Joseph DP, Pieramici DJ, Beauchamp NJ Jr. Computed tomography in the diagnosis and prognosis of open-globe injuries. Ophthalmology 2000 Oct;107(10):1899-906

Keberle M, Kenn W, Tschammler A, Wittenberg G, Hilgarth M, Hoppe F, Hahn D. Current value of double-contrast pharyngography and of computed tomography for the detection and for staging of hypopharyngeal, oropharyngeal and supraglottic tumors. Eur Radiol 1999;9(9):1843-50

Kurtz AB, Tsimikas JV, Tempany CM, Hamper UM, Arger PH, Bree RL, Wechsler RJ, Francis IR, Kuhlman JE, Siegelman ES, Mitchell DG, Silverman SG, Brown DL, Sheth S, Coleman BG, Ellis JH, Kurman RJ, Caudry DJ, McNeil BJ.Diagnosis and staging of ovarian cancer: comparative values of Doppler and conventional US, CT, and MR imaging correlated with surgery and histopathologic analysis--report of the Radiology Diagnostic Oncology Group. : Radiology 1999 Jul;212(1):19-27

Lane MJ, Liu DM, Huynh MD, Jeffrey RB Jr, Mindelzun RE, Katz DS. Suspected acute appendicitis: nonenhanced helical CT in 300 consecutive patients. Radiology 1999 Nov;213(2):341-6

Lennon CA, Gray DL. Sensitivity and specificity of ultrasound for the detection of neural tube and ventral wall defects in a high-risk population. Obstet Gynecol 1999 Oct;94(4):562-6

Levi S, Schaaps JP, De Havay P, Coulon R, Defoort P. End-result of routine ultrasound screening for congenital anomalies: the Belgian Multicentric Study 1984-92. Ultrasound Obstet Gynecol 1995 Jun;5(6):366-71

Levine SE, Neagle CE, Esterhai JL, Wright DG, Dalinka MK. Magnetic resonance imaging for the diagnosis of osteomyelitis in the diabetic patient with a foot ulcer. Foot Ankle Int 1994 Mar;15(3):151-6

Lindsell DR. Ultrasound imaging of pancreas and biliary tract. Lancet 1990 Feb 17;335(8686):390-3

Ma OJ, Mateer JR. Trauma ultrasound examination versus chest radiography in the detection of hemothorax. Ann Emerg Med 1997 Mar;29(3):312-5

Maggino T, Gadducci A, D'Addario V, Pecorelli S, Lissoni A, Stella M, Romagnolo C, Federghini M, Zucca S, Trio D, et al. Prospective multicenter study on CA 125 in postmenopausal pelvic masses. Gynecol Oncol 1994 Aug;54(2):117-23

Martelli G, Pilotti S, Coopmans de Yoldi G, Viganotti G, Fariselli G, Lepera P, Moglia D. Diagnostic efficacy of physical examination, mammography, fine needle aspiration cytology (triple-test) in solid breast lumps: an analysis of 1708 consecutive cases. Tumori 1990 Oct 31;76(5):476-9

Masdeu JC, Van Heertum RL, Kleiman A, Anselmi G, Kissane K, Horng J, Yudd A, Luck D, Grundman M. Early single-photon emission computed tomography in mild head trauma. A controlled study. J Neuroimaging 1994 Oct;4(4):177-81

Mettlin C, Lee F, Drago J, Murphy GP. The American Cancer Society National Prostate Cancer Detection Project. Findings on the detection of early prostate cancer in 2425 men. Cancer 1991 Jun 15;67(12):2949-58

Miller OF, Rineer SK, Reichard SR, Buckley RG, Donovan MS, Graham IR, Goff WB, Kane CJ. Prospective comparison of unenhanced spiral computed tomography and intravenous urogram in the evaluation of acute flank pain. : Urology 1998 Dec;52(6):982-7

Mooney C, Mushlin AI, Phelps CE. 1990. Targeting assessments of magnetic resonance imaging in suspected multiple sclerosis. Med Decis Making 10:77-94.

Moss HA, Britton PD, Flower CD, Freeman AH, Lomas DJ, Warren RM. How reliable is modern breast imaging in differentiating benign from malignant breast lesions in the symptomatic population? Clin Radiol 1999 Oct;54(10):676-82

Mushlin AI, Mooney C, Holloway RG, Detsky AS, Mattson DH, Phelps CE. 1997. The cost-effectiveness of magnetic resonance imaging for patients with equivocal neurological symptons. International Journal of Technology Assessment in Health Care 13(1):21-34.

Mussurakis S, Buckley DL, Coady AM, Turnbull LW, Horsman A.Observer variability in the interpretation of contrast enhanced MRI of the breast. Br J Radiol 1996 Nov;69(827):1009-16

Nozoe T, Matsumata T, Sugimachi K. Usefulness of preoperative transvaginal ultrasonography for women with advanced gastric carcinoma. Am J Gastroenterol 1999 Sep;94(9):2509-12

Orlinsky M, Knittel P, Feit T, Chan L, Mandavia D. The comparative accuracy of radiolucent foreign body detection using ultrasonography. Am J Emerg Med 2000 Jul;18(4):401-3

Pasanen PA, Partanen K, Pikkarainen P, Alhava E, Pirinen A, Janatuinen E. A prospective study on the value of ultrasound, computed tomography and endoscopic retrograde cholangiopancreatography in the diagnosis of unjaundiced cholestasis. In Vivo 1994 Mar-Apr;8(2):227-30

Perre CI, Koot VC, de Hooge P, Leguit P. The value of ultrasound in the evaluation of palpable breast tumours: a prospective study of 400 cases. Eur J Surg Oncol 1994 Dec;20(6):637-40

Rao PM, Rhea JT, Novelline RA, Mostafavi AA, Lawrason JN, McCabe CJ. Helical CT combined with contrast material administered only through the colon for imaging of suspected appendicitis. AJR Am J Roentgenol 1997 Nov;169(5):1275-80

Razumovsky AY, Gillard JH, Bryan RN, Hanley DF, Oppenheimer SM. TCD, MRA and MRI in acute cerebral ischemia. Acta Neurol Scand 1999 Jan;99(1):65-76

Regan F, Schaefer DC, Smith DP, Petronis JD, Bohlman ME, Magnuson TH. The diagnostic utility of HASTE MRI in the evaluation of acute cholecystitis. Half-Fourier acquisition single-shot turbo SE. J Comput Assist Tomogr 1998 Jul-Aug;22(4):638-42

Robertson PL, Berlangieri SU, Goergen SK, Waugh JR, Kalff V, Stevens SN, Hicks RJ, Fabiny RP, Ugoni A, Kelly MJ. Comparison of ultrasound and blood pool scintigraphy in the diagnosis of lower limb deep venous thrombosis. Clin Radiol 1994 Jun;49(6):382-90

Rosen CL, Brown DF, Sagarin MJ, Chang Y, McCabe CJ, Wolfe RE. Ultrasonography by emergency physicians in patients with suspected ureteral colic. J Emerg Med 1998 Nov-Dec;16(6):865-70

Rozycki GS, Ballard RB, Feliciano DV, Schmidt JA, Pennington SD. Surgeon-performed ultrasound for the assessment of truncal injuries: lessons learned from 1540 patients. Ann Surg 1998 Oct;228(4):557-67

Saidi MH, Sadler RK, Theis VD, Akright BD, Farhart SA, Villanueva GR. Comparison of sonography, sonohysterography, and hysteroscopy for evaluation of abnormal uterine bleeding. J Ultrasound Med 1997 Sep;16(9):587-91

Schroter A, Zerr I, Henkel K, Tschampa HJ, Finkenstaedt M, Poser S. Magnetic resonance imaging in the clinical diagnosis of Creutzfeldt-Jakob disease. Arch Neurol 2000 Dec;57(12):1751-7

Schwerk WB, Wichtrup B, Rothmund M, Ruschoff J. Ultrasonography in the diagnosis of acute appendicitis: a prospective study. Gastroenterology 1989 Sep;97(3):630-9

Scott WW Jr, Rosenbaum JE, Ackerman SJ, Reichle RL, Magid D, Weller JC, Gitlin JN. Subtle orthopedic fractures: teleradiology workstation versus film interpretation. Radiology 1993 Jun;187(3):811-5

Shackford SR, Wald SL, Ross SE, Cogbill TH, Hoyt DB, Morris JA, Mucha PA, Pachter HL, Sugerman HJ, O'Malley K, et al. The clinical utility of computed tomographic scanning and neurologic examination in the management of patients with minor head injuries. J Trauma 1992 Sep;33(3):385-94

Simonovsky V. The diagnosis of cirrhosis by high resolution ultrasound of the liver surface. Br J Radiol 1999 Jan;72(853):29-34

Srinivasan J, Mayberg MR, Weiss DG, Eskridge J. Duplex accuracy compared with angiography in the Veterans Affairs Cooperative Studies Trial for Symptomatic Carotid Stenosis. Neurosurgery 1995 Apr;36(4):648-53

Stafford RE, McGonigal MD, Weigelt JA, Johnson TJ. Oral contrast solution and computed tomography for blunt abdominal trauma: a randomized study. Arch Surg 1999 Jun;134(6):622-6

Stark DD, Wittenberg J, Butch RJ, Ferrucci JT Jr. Hepatic metastases: randomized, controlled comparison of detection with MR imaging and CT. Radiology 1987 Nov;165(2):399-406

Strandell A, Bourne T, Bergh C, Granberg S, Asztely M, Thorburn J. The assessment of endometrial pathology and tubal patency: a comparison between the use of ultrasonography and X-ray hysterosalpingography for the investigation of infertility patients. Ultrasound Obstet Gynecol 1999 Sep;14(3):200-4

Thomas B, Falcone RE, Vasquez D, Santanello S, Townsend M, Hockenberry S, Innes J, Wanamaker S. Ultrasound evaluation of blunt abdominal trauma: program implementation, initial experience, and learning curve. J Trauma 1997 Mar;42(3):384-8; discussion 388-90

Thourani VH, Pettitt BJ, Schmidt JA, Cooper WA, Rozycki GS. Validation of surgeon-performed emergency abdominal ultrasonography in pediatric trauma patients. J Pediatr Surg 1998 Feb;33(2):322-8

Tsuda H, Kawabata M, Kawabata K, Yamamoto K, Hidaka A, Umesaki N. Comparison between transabdominal and transvaginal ultrasonography for identifying endometrial malignancies. Gynecol Obstet Invest 1995;40(4):271-3

Valk PE, Abella-Columna E, Haseman MK, Pounds TR, Tesar RD, Myers RW, Greiss HB, Hofer GA. Whole-body PET imaging with [18F]fluorodeoxyglucose in management of recurrent colorectal cancer. Arch Surg 1999 May;134(5):503-11

van Gils AP, van den Berg R, Falke TH, Bloem JL, Prins HJ, Dillon EH, van der Mey AG, Pauwels EK. MR diagnosis of paraganglioma of the head and neck: value of contrast enhancement. AJR Am J Roentgenol 1994 Jan;162(1):147-53

van Nagell JR Jr, DePriest PD, Reedy MB, Gallion HH, Ueland FR, Pavlik EJ, Kryscio RJ. The efficacy of transvaginal sonographic screening in asymptomatic women at risk for ovarian cancer. Gynecol Oncol 2000 Jun;77(3):350-6

Vieweg J, Teh C, Freed K, Leder RA, Smith RH, Nelson RH, Preminger GM. Unenhanced helical computerized tomography for the evaluation of patients with acute flank pain. J Urol 1998 Sep;160(3 Pt 1):679-84

von Kummer R, Bourquain H, Bastianello S, Bozzao L, Manelfe C, Meier D, Hacke W. Early prediction of irreversible brain damage after ischemic stroke at CT Radiology 2001 Apr;219(1):95-100

Walker S, Haun W, Clark J, McMillin K, Zeren F, Gilliland T. The value of limited computed tomography with rectal contrast in the diagnosis of acute appendicitis. Am J Surg 2000 Dec;180(6):450-4; discussion 454-5

Weishaupt D, Zanetti M, Boos N, Hodler J. MR imaging and CT in osteoarthritis of the lumbar facet joints. Skeletal Radiol 1999 Apr;28(4):215-9

Wong TW, Lau CC, Yeung A, Lo L, Tai CM. Efficacy of transabdominal ultrasound examination in the diagnosis of early pregnancy complications in an emergency department. J Accid Emerg Med 1998 May;15(3):155-8

Young GR, Humphrey PR, Shaw MD, Nixon TE, Smith ET. Comparison of magnetic resonance angiography, duplex ultrasound, and digital subtraction angiography in assessment of extracranial internal carotid artery stenosis. : J Neurol Neurosurg Psychiatry 1994 Dec;57(12):1466-78

Zielke A, Hasse C, Sitter H, Rothmund M. Influence of ultrasound on clinical decision making in acute appendicitis: a prospective study. Eur J Surg 1998 Mar;164(3):201-9

Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, van Voorthuisen AE. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. Radiology 1999 Nov;213(2):413-22

Zuberi SM, Matta N, Nawaz S, Stephenson JB, McWilliam RC, Hollman A. Muscle ultrasound in the assessment of suspected neuromuscular disease in childhood. Neuromuscul Disord 1999 Jun;9(4):203-7

Articles Used-Psychological

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition. Washington, D.C., APA, 1980.

Baird, Gillian; Charman, Tony; Baron-Cohen, Simon; Cox, Anthony; Swettenham, John; Wheelwright, Sally; Drew, Auriol. A screening instrument for autism at 18 months of age: A 6-year follow-up study. Journal of the American Academy of Child & Adolescent Psychiatry. 2000 Jun Vol 39(6) 694-702

Berument, Sibel Kazak; Rutter, Michael; Lord, Catherine; Pickles, Andrew; Bailey, Anthony. Autism screening questionnaire: Diagnostic validity. British Journal of Psychiatry. 1999 Nov Vol 175 444-451

Birtchnell, John; Evans, Chris; Deahl, Martin; Masters, Nigel. The Depression Screening Instrument (DSI): A device for the detection of depressive disorders in general practice. Journal of Affective Disorders. 1989 Mar-Jun Vol 16(2-3) 269-281

Blais, Mark A.; Hilsenroth, Mark J.; Fowler, J. Christopher. Diagnostic efficiency and hierarchical functioning of the DSM-IV borderline personality disorder criteria. Journal of Nervous & Mental Disease. 1999 Mar Vol 187(3) 167-173

Bradley KA, Boyd-Wickizer J, Powell SH, Burman ML. 1998. Alcohol screening questionnaires in women: A critical review. JAMA 280(2):166-171

Brooks, David A.; Williams, Ronald N.; Dean, Raymond S.; Wood, Tina M.; et al. The predictive validity of a neuropsychological screening measure. International Journal of Neuroscience. 1990 Mar Vol 51(1-2) 83-88

Chaffee, C. Anne; Cunningham, Charles E.; Secord-Gilbert, Margaret; Elbard, Heather; et al Screening effectiveness of the Minnesota Child Development Inventory expressive and receptive language scales: Sensitivity, specificity, and predictive value. Psychological Assessment. 1990 Mar Vol 2(1) 80-85

Chen, Wei J.; Faraone, Stephen V.; Biederman, Joseph; Tsuang, Ming T. Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. Journal of Consulting & Clinical Psychology. 1994 Oct Vol 62(5) 1017-1025

Colligan, Robert C.; Davis, Leo J.; Morse, Robert M.; Offord, Kenneth P. Screening medical patients for alcoholism with the MMPI: A comparison of seven scales. Journal of Clinical Psychology. 1988 Jul Vol 44(4) 582-592

de las Cuevas, Carlos; Sanz, Emilio J.; de la Fuente, Juan A.; Padilla, Jonathan; Berenguer, Juan C. The Severity of Dependence Scale (SDS) as screening test for benzodiazepine dependence: SDS validation study. Addiction. 2000 Feb Vol 95(2) 245-250

Dent, Alexandra; Lincoln, Nadina B. Screening for memory problems in multiple sclerosis. British Journal of Clinical Psychology. 2000 Sep Vol 39(3) 311-315

Erford, Bradley T.; Bagley, Donna L.; Hopper, James A.; Lee, Ramona M.; Panagopulos, Kathleen A.; Preller, Denise B. Reliability and validity of the Math Essential Skill Screener-Elementary Version (MESS-E). Psychology in the Schools. 1998 Apr Vol 35(2) 127-135

Foa, Edna B.; Cashman, Laurie; Jaycox, Lisa; Perry, Kevin. The validation of a self-report measure of posttraumatic stress disorder: The Posttraumatic Diagnostic Scale. Psychological Assessment. 1997 Dec Vol 9(4) 445-451

Glascoe, Frances Page; Byrne, Karen E. The accuracy of three developmental screening tests. Journal of Early Intervention. 1993 Fal Vol 17(4) 368-379

Gross, Kristin; Keyes, Megan D.; Greene, Roger L. Assessing depression with the MMPI and MMPI-2. Journal of Personality Assessment. 2000 Dec Vol 75(3) 464-477

Gureje, O.; Obikoya, B. The GHQ-12 as a screening tool in a primary care setting. Social Psychiatry & Psychiatric Epidemiology. 1990 Sep Vol 25(5) 276-280

Hiatt, Deirdre; Hargrave, George E. Predicting job performance problems with psychological screening. Journal of Police Science & Administration. 1988 Jun Vol 16(2) 122-135

Hilsenroth, Mark J.; Fowler, J. Christopher; Padawer, Justin R. The Rorschach Schizophrenia Index (SCZI): An examination of reliability, validity, and diagnostic efficiency. Journal of Personality Assessment. 1998 Jun Vol 70(3) 514-534

Hirschfeld, Robert M. A.; Williams, Janet B. W.; Spitzer, Robert L.; Calabrese, Joseph R.; Flynn, Laurie; Keck, Paul E., Jr.; Lewis, Lydia; McElroy, Susan L.; Post, Robert M.; Rapport, Daniel J.; Russell, James M.; Sachs, Gary S.; Zajecka, John. Development and validation of a

screening instrument for bipolar spectrum disorder: The Mood Disorder Questionnaire. American Journal of Psychiatry. 2000 Nov Vol 157(11) 1873-1875

Hogervorst, Eef; Barnetson, L.; Jobst, K. A.; Nagy, Zs.; Combrinck, M.; Smith, A. D. Diagnosing dementia: Interrater reliability assessment and accuracy of the NINCDS/ADRDA criteria versus CERAD histopathological criteria for Alzheimer's disease. Dementia & Geriatric Cognitive Disorders. 2000 Mar-Apr Vol 11(2) 107-113

Hyler, Steven E.; Skodol, Andrew E.; Kellman, H. David; Oldham, John M.; et al. Validity of the Personality Diagnostic Questionnaire--Revised: Comparison with two structured interviews. American Journal of Psychiatry. 1990 Aug Vol 147(8) 1043-1048

Inwald, Robin E.; Brockwell, Albert L. Predicting the performance of government security personnel with the IPI and MMPI. Journal of Personality Assessment. 1991 Jun Vol 56(3) 522-535

Ivnik, Robert J.; Smith, Glenn E.; Petersen, Ronald C.; Boeve, Bradley F.; Kokmen, Emre; Tangalos, Eric G. Diagnostic accuracy of four approaches to interpreting neuropsychological test data. Neuropsychology. 2000 Apr Vol 14(2) 163-177

Johnson, Edward E.; Hamer, Robert M.; Nora, Rena M. The Lie/Bet questionnaire for screening pathological gamblers: A follow-up study. Psychological Reports. 1998 Dec Vol 83(3, Pt 2) 1219-1224

Klin, Ami; Lang, Jason; Cicchetti, Domenic V.; Volkmar, Fred R. Brief report: Interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: Results of the DSM-IV Autism Field Trial. Journal of Autism & Developmental Disorders. 2000 Apr Vol 30(2) 163-167

Kogan, Evan S.; Kabacoff, Robert I.; Hersen, Michel; Van Hasselt, Vincent B. Clinical cutoffs for the Beck Depression Inventory and the Geriatric Depression Scale with older adult psychiatric outpatients. Journal of Psychopathology & Behavioral Assessment. 1994 Sep Vol 16(3) 233-242

Laprise, Rejeanne; Vezina, Jean. Diagnostic performance of the Geriatric Depression Scale and the Beck Depression Inventory with nursing home residents. Canadian Journal on Aging. 1998 Win Vol 17(4) 401-413

Lavigne, John V.; Arend, Richard; Rosenbaum, Diane; Sinacore, James; et al. Interrater reliability of the DSM-III--R with preschool children. Journal of Abnormal Child Psychology. 1994 Dec Vol 22(6) 679-690

Maki, Naruhiko; Ikeda, Manabu; Hokoishi, Kazuhiko; Nebu, Akihiko; Komori, Kenjiro; Shigenobu, Kazue; Fukuhara, Ryuji; Hirono, Nobutsugu; Nakata, Hideki; Tanabe, Hirotaka. Validity of the Short-Memory Questionnaire in vascular dementia. International Journal of Geriatric Psychiatry. 2000 Dec Vol 15(12) 1143-1146

Merson, S.; Tyrer, P.; Duke, Peter J.; Henderson, F. Interrater reliability of ICD-10 guidelines for the diagnosis of personality disorders. Journal of Personality Disorders. 1994 Sum Vol 8(2) 89-95

Parikh, Rajesh M.; Eden, Dianne T.; Price, Thomas R.; Robinson, Robert G. The sensitivity and specificity of the Center for Epidemiologic Studies Depression Scale in screening for post-stroke depression. International Journal of Psychiatry in Medicine. 1988 Vol 18(2) 169-181

Pomeroy, Ian M.; Clark, Christopher R.; Philp, Ian. The effectiveness of very short scales for depression screening in elderly medical patients. International Journal of Geriatric Psychiatry. 2001 Mar Vol 16(3) 321-326

Razavi, Darius; Delvaux, Nicole; Farvacques, Christine; Robaye, Edmond. Screening for adjustment disorders and major depressive disorders in cancer in-patients. British Journal of Psychiatry. 1990 Jan Vol 156 79-83

Scheinberg, Zvi; Koslowsky, Meni; Bleich, Avi; Mark, Mordechai; et al. Sensitivity, specificity, and positive predictive value as measures of prediction accuracy: The case of the EAT-26. Educational & Psychological Measurement. 1993 Fal Vol 53(3) 831-839

Sherman, Tracy; Shulman, Brian B. Specificity and sensitivity ratios of the Pediatric Language Acquisition Screening Tool for Early Referral-Revised. Infant-Toddler Intervention. 1999 Dec Vol 9(4) 315-330

Steer, Robert A.; Cavalieri, Thomas A.; Leonard, Douglas M.; Beck, Aaron T. Use of the Beck Depression Inventory for primary care to screen for major depression disorders. General Hospital Psychiatry. 1999 Mar-Apr Vol 21(2) 106-111

Storgaard, Heidi; Nielsen, Susanne D.; Gluud, Christian. The validity of the Michigan Alcoholism Screening Test (MAST). Alcohol & Alcoholism. 1994 Sep Vol 29(5) 493-502

Uhlmann, Richard F.; Larson, Eric B. Effect of education on the Mini-Mental State Examination as a screening test for dementia. Journal of the American Geriatrics Society. 1991 Sep Vol 39(9) 876-880

Articles Used-Security

Jayne, Brian C. 1989. A comparison between the predictive value of two common preemployment screening procedures. The Investigator 5(3)

Scoring Cutoffs - Picking The Best

James R. Wygant[1]

Abstract

Various cutoff rules were applied to 200 examination scores, the results of two examiners independently scoring the same 100 verified examinations. Error rate, rate of definite results, false positive results, and false negative results are compared for 13 different decision rules, including several rules devised specifically for this study. The link between error rate and inconclusive rate is considered. Data suggest that asymmetric cutting scores, such as -7/+5, may achieve better results than the more traditional +/-6.

Keywords: cutoffs, cutoff score, scoring, scores, DoDPI, inconclusive, errors, false positive, false negative

Introduction

Although polygraph examiners pride themselves on the level of standardization achieved in chart evaluation, the only real constant is that nearly any method of numerical evaluation can be shown to achieve better results than simple global evaluation of charts. Beyond that, there continues to be debate about a variety of scoring practices, including the minimum scoring levels or "cutoffs" that permit conclusions.

It is not surprising that scoring of any kind improves accuracy, because any scoring method is little more than a means of keeping track of what is observed across a series of charts. Scoring is a polygraph examiner's means of combating the human inability to accurately remember and mentally tabulate the details of the large number of separate decisions that are needed to form the basis of any chart evaluation. A three-chart examination that includes three relevant questions about the issue requires that the examiner make at least 27 different decisions in evaluating respiration, electrodermal response, and cardio response. Examiners use scoring because they can't accurately remember all of what they're seeing as they advance through a detailed analysis of the physiological changes recorded on their charts.

Scoring was popularized by Cleve Backster. Although he teaches several different test formats, he has reserved scoring largely for his "you phase" test, the "did you do it" specific-issue examination. Other formats taught by Backster were expressly intended to lead toward a final "you phase" test.

Backster has taught the "you phase" test as a single-issue procedure. His multiple-issue procedure, an "exploratory" test, is suggested as a means of deciding which issue should become the target in a final "you phase" test. As a consequence, Backster's scoring method, as applied to the "you phase" test, relies upon a total score to make a decision. Backster recognized that it made no sense to add together scores from various questions if the questions did not ask essentially the same thing. If the examinee could be lying to one relevant question while simultaneously telling the truth on another, adding their scores together could theoretically cause them to cancel each other out.

[1] James R. Wygant is a private examiner in Portland Oregon, in practice since 1976. He is the author of several articles in *Polygraph*. Since 1993 he has written and published *Polygraph News and Reviews*, a newsletter for polygraph examiners. He also instructed at the former Western Oregon University School of Polygraph. He may be contacted at 7505 SE Reed College Pl., Portland OR 97202-8362; email jrwygant@earthlink.net.

Although Backster initially used the same cutoffs for findings of either truthfulness or deception, in approximately 1979 he reduced the cutoff levels for truthfulness (Backster, 1979). For instance, the minimum score for a finding of truthfulness on three charts was dropped to +7, while the minimum for deception remained at its previous level of -13. The asymmetric cutting scores addressed concerns about a possible bias toward findings of deception when the weaker comparison question was used. In recent years, other examiners have suggested that asymmetric cutting scores might be a remedy for other non-Backster methods, possibly reducing false positive results, wrongly identifying a truthful person as a liar.

When a comparison question appears on either side of a relevant question, Backster uses the *weaker* comparison question for scoring unless there is no reaction on the relevant question. The Department of Defense Polygraph Institute (DoDPI) uses a test format similar to Backster's, and originally devised by him, but DoDPI uses the *stronger* comparison question when a choice is available, which is only on the first of three relevant questions in the DoDPI format. DoDPI uses equal cutoffs, +/-6, for findings of either truthfulness or deception, but has added its own exceptions to that simple rule.

Many examiners in field practice routinely regard the +/-6 cutoff levels as carved in stone, assuming that those cutoffs are perfect for all examiners under all conditions. In fact, the manner in which polygraph charts are scored suggests that absolute cutoff values at any level probably ought to be considered a myth.

Manual scoring of polygraph examinations, as opposed to computer evaluation, the only other method widely used today, relies upon several variables.

- How experienced is the examiner? Research indicates that more experienced examiners achieve more accurate results.

- What does the examiner regard as a physiological reaction that ought to be scored? Different schools teach different and sometimes contradictory phenomena. For instance, one school may teach that a drop in the cardio tracing, reflecting a brief loss of blood volume at the measurement site, is a reaction, while another school teaches that only rises in that tracing can be scored. One school may teach that any change in respiration from the norm is a reaction, while another teaches that only some form of suppression is a reaction.

- How aggressively does the examiner score? Some examiners are willing to score even slight observable differences. Others are more conservative and will assign a score of zero to what a more aggressive examiner regards as a positive or negative value.

- How does the examiner treat distortion or artifacts? Some examiners regard any apparent distortion, no matter how slight, as reason to ignore a segment of chart, being particularly concerned that distortion might indicate countermeasures. Other examiners might score a slightly distorted chart segment, arguing that the distortion is too slight to significantly impact the tracing.

- How much was the examiner's scoring influenced by any pretest bias? Bias can arise from case information that is highly persuasive, or from either a sympathetic or antagonistic attitude toward the examinee. Examiners are subject to these influences as much as anyone else. They are trained to ignore them, but the degree of success in that effort for any particular examination remains unknown.

Those factors, and others, account for the fact that a room full of examiners scoring the same set of charts will not all arrive at the same total score. A scoring exercise at the 2002 American Polygraph Association (APA) seminar had about half the room of approximately 100 examiners advocating a

score of +1 for one particular comparison, while the other half wanted to assign a value of -1.

For most examinations, differences of a few points in total score will not change an ultimate determination that the charts indicated either truthfulness or deception, or were inconclusive. Most tests produce strong enough indications of either truthfulness or deception that an examiner scoring the same charts a few points lower will still arrive at the same final determination. In other words, if one examiner scores +13 on an examination and another examiner scores +7 on the same examination, the final opinion in either case is truthfulness.

However, if the first examiner scores +6 and the second examiner scores +4, and they are both using +/-6 as cutoffs, there is a difference of opinion -- even though the two-point difference between the two scores in this case is less than the six-point difference in the first example. One examiner concludes truthfulness and the other finds that the charts are inconclusive. If the second examiner were using +4 as his or her minimum for a finding for truthfulness, the final opinions would again agree.

This leads to consideration of the purpose of any cutoffs, which is generally presumed to be a means of keeping accuracy high. There is, however, a link with an examiner's inconclusive rate. If an examiner used cutoffs of +/-20 he would probably achieve higher accuracy than he gets with +/-6. He might move from an average accuracy rate of 90% to an average of 97%, but his inconclusive rate would become unacceptably high, perhaps changing from 12% to as much as 80% (these are hypothetical figures used only for illustration). The slight gain in accuracy is generally not considered to be worth the substantial loss of useable results. On the other hand, if examiners were to use +/-1 as cutoffs, accepting any score greater than zero as an indication of either truthfulness or deception, accuracy would plummet but the inconclusive rate would improve dramatically, probably dropping to around 1%. We must conclude that cutoffs serve two masters, both accuracy and inconclusive rate. The question we tried to resolve with this study was whether there was any system of cutoffs that was better than any other.

Method

This study used 100 verified examinations. Verification was established by the agreement of a panel of examiners, independent of the present work. Fifty of the examinations were verified truthful results, and fifty were verified results of deception. These were the same examinations used in the study previously done by Abrams, Leutwyler, and Wygant (2000). All of these examinations used the DoDPI zone format, which consists of an initial three non-scored questions, followed by a comparison question in the fourth position, relevant question in the fifth position, then another comparison-relevant pair, followed by a non-scored question in the eighth position, and a final comparison-relevant pair. In the DoDPI format, the last relevant question is often a "do you know" type.

Two examiners independently scored the examinations. The verified results were not available to them until they had completed their scoring. They used a seven-position scale in scoring, meaning that for any particular comparison, the values assigned could range from 0 to +/-3. When the first relevant question presented the option of scoring to two adjacent comparison questions, the examiners scored against the stronger comparison question. Both examiners were experienced, each having more than 25 years work in the profession.

The individual question totals obtained by the two examiners in the Abrams et al. (2000) study were preserved, so that a variety of different scoring methods could be applied. Several different methods were reported in the original study and will be reviewed here in greater detail. Other methods not originally considered have been added to this study.

The total scores for each question on each examination were entered into a computer spreadsheet. A variety of decision rules were then devised that reproduced the cutoff rules that examiners might use. These ranged from simple symmetrical values (e.g.,

+/-6) to rules such as DoDPI's that have special exceptions. Those rules were then applied to the scores that the two examiners got from all 100 examinations. The results were then compared to determine which cutoffs produced results that most closely matched the verified results, and what levels of inconclusive results accompanied them. Tabulations were done separately for each examiner, 100 results each, and also by combining their work into a total of 200 results.

Cutoff Rules

The following cutoff rules were considered. Some of these rules have never been seen anywhere else. They were derived from analysis of existing rules and disparities between false positives and false negatives. In other words, if a particular rule had a relatively high error rate and produced many more false negatives than false positives, a modification was added by the author to try to address that, resulting in a new rule.

+/-6
This simple rule regards any total score of +6 or higher as indicating truthfulness and any total score of -6 or lower as indicating deception. There is no consideration given to individual question scores. This is the decision rule for the Utah technique (Bell, et al., 1999).

+/-5
Same as above but with lower minimum values.

+/-4
Same as above but with lower minimum values.

Full DoDPI
This rule uses +/-6 as minimum total scores, but adds some important exceptions. It additionally requires that a finding of truthfulness can only be made if all three relevant questions had positive values (greater than zero), and it requires that a finding of deception be made if any one relevant question had a score of -3 or lower, regardless of total score. This is the cutoff rule taught by DoDPI.

+/-6 or any -3
This is a modified version of "full DoDPI", eliminating the requirement of a positive value for each relevant question. Like "full DoDPI", this rule uses +/-6 for minimum total scores and mandates deception if any one relevant question had a score of -3 or lower, regardless of total score.

Senter A
This is one configuration of a rule proposed by Stuart Senter, Ph.D. (presented at the 2002 APA seminar). He has proposed that inconclusive results could be reduced and accuracy improved slightly if an examiner first used one set of cutoffs and then applied a second set to those that produced inconclusive results on the first trial. In this configuration, an examiner who typically used +/-6 cutoffs would apply that rule first and then apply the "full DoDPI" rule to any inconclusive results.

Senter B
This is the second configuration of Stuart Senter's proposal. In this case an examiner who would ordinarily use the "full DoDPI" cutoffs described above would then apply the simple +/6 cutoffs to any inconclusive results. This is a reversal of the process used in Senter A.

+/-6 or any -4
This is another modification of the DoDPI rule. In this case, +/-6 is used as a primary cutoff, but any individual relevant question having a score of -4 or lower would result in an opinion of deception.

+/-4 or any -3
In this case +/-4 is used as a primary cutoff, but any individual relevant question having a score of -3 or lower would result in an opinion of deception.

-6/+4
These are asymmetric cutting scores. A total score of -6 or lower would indicate deception, but a score of +4 or higher would indicate truthfulness.

-6/+4 or any -3
This is a combination of asymmetric cutting scores with the addition of a

finding of deception if any individual relevant question scored -3 or lower.

-7/+5
This is another asymmetric cutting score, based only a total score.

-5/+6 or any –4
This is the only asymmetric cutting score in which the minimum value for deception is *lower* than the minimum for a finding of truthfulness, but it also requires a finding of deception if the score on any individual relevant question is -4 or lower.

Results

There are a total of 13 different rules in this assortment, all applied to the same 200 examination scores to determine if any one rule achieved significantly better results for these two examiners than any of the others -- a consideration that must weigh inconclusive rate against error rate. Error rates throughout this study apply only to definite results and do not include inconclusive results.

Table 1
Examiner 1 Results Sorted by Errors (N = 100).

	"T" Result	"D" Result	"I" Result	% Definite	False Neg.	False Pos.	**% Errors**
+/-5	46	31	23	77.0	5	1	**7.8**
-7/+5	46	30	24	76.0	5	1	**7.9**
+/-6	43	30	27	73.0	5	1	**8.2**
-6/+4	50	30	20	80.0	7	1	**10.0**
+/-4	50	33	17	83.0	7	2	**10.8**
+/-6 or any –4	42	39	19	81.0	4	5	**11.1**
-5/+6 or any -4	42	39	19	81.0	4	5	**11.1**
full DoDPI	26	45	29	71.0	1	7	**11.3**
Senter A	43	43	14	86.0	5	6	**12.8**
Senter B	41	45	14	86.0	4	7	**12.8**
+/-6 or any –3	41	45	14	86.0	4	7	**12.8**
+/-4 or any –3	47	46	7	93.0	6	7	**14.0**
-6/+4 or any -3	47	45	8	92.0	6	7	**14.1**

Table 1 is derived from the scores of Examiner 1 and it is sorted according to the percentage of errors. The simple cutoff of +/-5 produced the fewest errors when applied to this examiner's scores, but it also achieved definite results with only 77% of the examinations. At the other end of the scale is the more complex rule of -6/+4 for total score, but deception indicated when any one relevant question had a score of -3 or lower. That rule produced the greatest number of errors, 14.1% of the definite decisions, but it also had more definite decisions than all but one other rule.

Table 2 shows Examiner 1 scores now sorted according to inconclusive rate ("% definite"). For this examiner, the "full DoDPI" rule produced the lowest percentage of definite results, 71.0% (or 29.0% inconclusive). The method producing the fewest inconclusive results was cutoffs of +/-4 coupled with a mandatory finding of deception if any relevant question had a score of -3 or lower. Unfortunately, that method also produced the second-worst percentage of errors. At least the errors were nearly evenly distributed, 6 false negatives and 7 false positive.

Table 3 represents the same information for Examiner 2 as shown for Examiner 1 in Table 1. The results are sorted by error rate and they differ slightly from those of Examiner 1. Still, the same two rules produced the worst error rates with both

examiners, and four of the same rules appeared in the best five (lowest error rates) for both examiners, although in slightly different order.

Table 2
Examiner 1 Results Sorted by Inconclusive Rate (% definite) (N = 100).

	"T" Result	"D" Result	"I" Result	% **Definite**	False Neg.	False Pos.	% Errors
full DoDPI	26	45	29	**71.0**	1	7	11.3
+/-6	43	30	27	**73.0**	5	1	8.2
-7/+5	46	30	24	**76.0**	5	1	7.9
+/-5	46	31	23	**77.0**	5	1	7.8
-6/+4	50	30	20	**80.0**	7	1	10.0
+/-6 or any -4	42	39	19	**81.0**	4	5	11.1
-5/+6 or any -4	42	39	19	**81.0**	4	5	11.1
+/-4	50	33	17	**83.0**	7	2	10.8
Senter A	43	43	14	**86.0**	5	6	12.8
Senter B	41	45	14	**86.0**	4	7	12.8
+/-6 or any -3	41	45	14	**86.0**	4	7	12.8
-6/+4 or any -3	47	45	8	**92.0**	6	7	14.1
+/-4 or any -3	47	46	7	**93.0**	6	7	14.0

Table 3
Examiner 2 Results Sorted by Errors (N = 100).

	"T" Result	"D" Result	"I" Result	% Definite	False Neg.	False Pos.	**% Errors**
+/-6	41	35	24	76.0	5	2	**9.2**
-7/+5	49	35	16	84.0	8	2	**11.9**
-6/+4	49	35	16	84.0	8	2	**11.9**
+/-4	49	41	17	90.0	8	3	**12.2**
-5/+6 or any -4	41	42	17	83.0	5	6	**13.3**
+/-6 or any -4	41	41	18	82.0	5	6	**13.4**
full DoDPI	23	46	31	69.0	2	8	**14.5**
+/-5	41	46	13	87.0	5	8	**14.9**
Senter A	41	46	13	87.0	5	8	**14.9**
Senter B	41	46	13	87.0	5	8	**14.9**
+/-6 or any -3	41	46	13	87.0	5	8	**14.9**
+/-4 or any -3	47	48	5	95.0	7	8	**15.8**
-6/+4 or any -3	47	46	7	93.0	7	8	**16.1**

Table 4 shows the results for Examiner 2 sorted according to the inconclusive rate. As with Examiner 1, the "full DoDPI" rule produced the worst inconclusive rate while again producing the greatest imbalance between false positives and false negatives.

The fewest inconclusive results were again produced by the same two rules, which again also had the worst rate of errors.

Table 4
Examiner 2 Results Sorted by Inconclusive Rate (% definite) (N = 100).

	"T" Result	"D" Result	"I" Result	% **Definite**	False Neg.	False Pos.	% Errors
full DoDPI	23	46	31	**69.0**	2	8	14.5
+/-6	41	35	24	**76.0**	5	2	9.2
+/-6 or any -4	41	41	18	**82.0**	5	6	13.4
-5/+6 or any -4	41	42	17	**83.0**	5	6	13.3
-7/+5	49	35	16	**84.0**	8	2	11.9
-6/+4	49	35	16	**84.0**	8	2	11.9
+/-5	41	46	13	**87.0**	5	8	14.9
Senter A	41	46	13	**87.0**	5	8	14.9
Senter B	41	46	13	**87.0**	5	8	14.9
+/-6 or any -3	41	46	13	**87.0**	5	8	14.9
+/-4	49	41	17	**90.0**	8	3	12.2
-6/+4 or any -3	47	46	7	**93.0**	7	8	16.1
+/-4 or any -3	47	48	5	**95.0**	7	8	15.8

Table 5
Combined Examiner Results Sorted by Errors (N = 200).

	"T" Result	"D" Result	"I" Result	% Definite	False Neg.	False Pos.	**% Errors**
+/-6	84	65	51	74.5	10	3	**8.7**
-7/+5	95	65	40	80.0	13	3	**10.0**
-6/+4	99	65	36	82.0	15	3	**11.0**
+/-5	87	77	36	82.0	10	9	**11.6**
+/-4	99	74	34	83.6	15	5	**11.6**
-5/+6 or any -4	83	81	36	82.0	9	11	**12.2**
+/-6 or any -4	83	80	37	81.5	9	11	**12.3**
full DoDPI	49	91	60	70.0	3	15	**12.9**
Senter A	84	89	27	86.5	10	14	**13.9**
Senter B	82	91	27	86.5	9	15	**13.9**
+/-6 or any -3	82	91	27	86.5	9	15	**13.9**
+/-4 or any -3	94	94	12	94.0	13	15	**14.9**
-6/+4 or any -3	94	91	15	92.5	13	15	**15.1**

Of note with both examiners is that there were generally more false negative results, identifying a liar as a truthful person, than the reverse. The standout exception to that was the "full DoDPI" rule, which produced significantly more false positives than false negatives for both examiners. That disparity disappeared, the error rate remained about the

same, and the rate of definite results improved with the "+/-6 or any -3" rule. That is the same rule as "full DoDPI" but without the requirement that a truthful result have a positive value -- no zero and no negative -- for each relevant question.

In the next two tables the scores from the two examiners were combined, producing a total of 200 evaluations, 100 from each of the two examiners.

The results from the combined 200 scores featured in Table 5 show that the simplest rule with the highest cutoffs, +/-6, produced the fewest errors. Not surprisingly +/-4 and +/-5 produced slightly worse error rates, but they were accompanied by about an 8% improvement in the inconclusive rate. This is an expectable demonstration of the link between inconclusive rate and error rate. Of particular note is that a asymmetric cutting score rule, -7/+5, had the second best error rate, coupled with a 5.5% improvement over the inconclusive rate of +/-6. If these results held true for most examiners, the profession would probably achieve better results from the asymmetric -7/+5 than from the simple +/-6. "Better" in this instance means slightly more errors (which in this study were false negatives), but a significantly lower inconclusive rate.

Table 6
Combined Examiner Results Sorted by Inconclusive Rate (% Definite) (N = 200).

	"T" Result	"D" Result	"I" Result	% **Definite**	False Neg.	False Pos.	% Errors
full DoDPI	49	91	60	**70.0**	3	15	12.9
+/-6	84	65	51	**74.5**	10	3	8.7
-7/+5	95	65	40	**80.0**	13	3	10.0
+/-6 or any -4	83	80	37	**81.5**	9	11	12.3
-6/+4	99	65	36	**82.0**	15	3	11.0
+/-5	87	77	36	**82.0**	10	9	11.6
-5/+6 or any -4	83	81	36	**82.0**	9	11	12.2
+/-4	99	74	34	**83.6**	15	5	11.6
Senter A	84	89	27	**86.5**	10	14	13.9
Senter B	82	91	27	**86.5**	9	15	13.9
+/-6 or any -3	82	91	27	**86.5**	9	15	13.9
-6/+4 or any -3	94	91	15	**92.5**	13	15	15.1
+/-4 or any -3	94	94	12	**94.0**	13	15	14.9

The results in Table 6 closely match those for the two individual examiners, with inconclusive rates running from 30% for "full DoDPI" to 6% for +/-4 with mandatory deception if any relevant question scored -3 or lower. The error rates generally got worse as the inconclusive rates got better, although "full DoDPI" was a notable exception. For that rule, the error rate was worse than expected for comparable inconclusive rates.

These data leave no easy solution to the search for the "best" cutoff rule. The generalization of these results to other examiners also cannot be assumed, although the cutoff rules at the extremes of error percentages and inconclusive rates would likely perform similarly for most examiners. In other words, those cutoffs would probably be poor choices for anyone in most circumstances. "Full DoDPI" would probably produce more inconclusive results than most other methods, and +/-6 would probably produce among the most accurate results. But those considerations by themselves are not meaningful. High accuracy coupled with few definite results is not desirable, especially if a slight drop in accuracy can produce a substantial improvement in the inconclusive rate.

All of the 13 methods in this study were plotted on a scatter chart in which the Y-axis (vertical scale) represents the error rate (without inconclusives) and the X-axis (horizontal scale) represents the rate of definite results. The combined data, 200 results, were used for the chart.

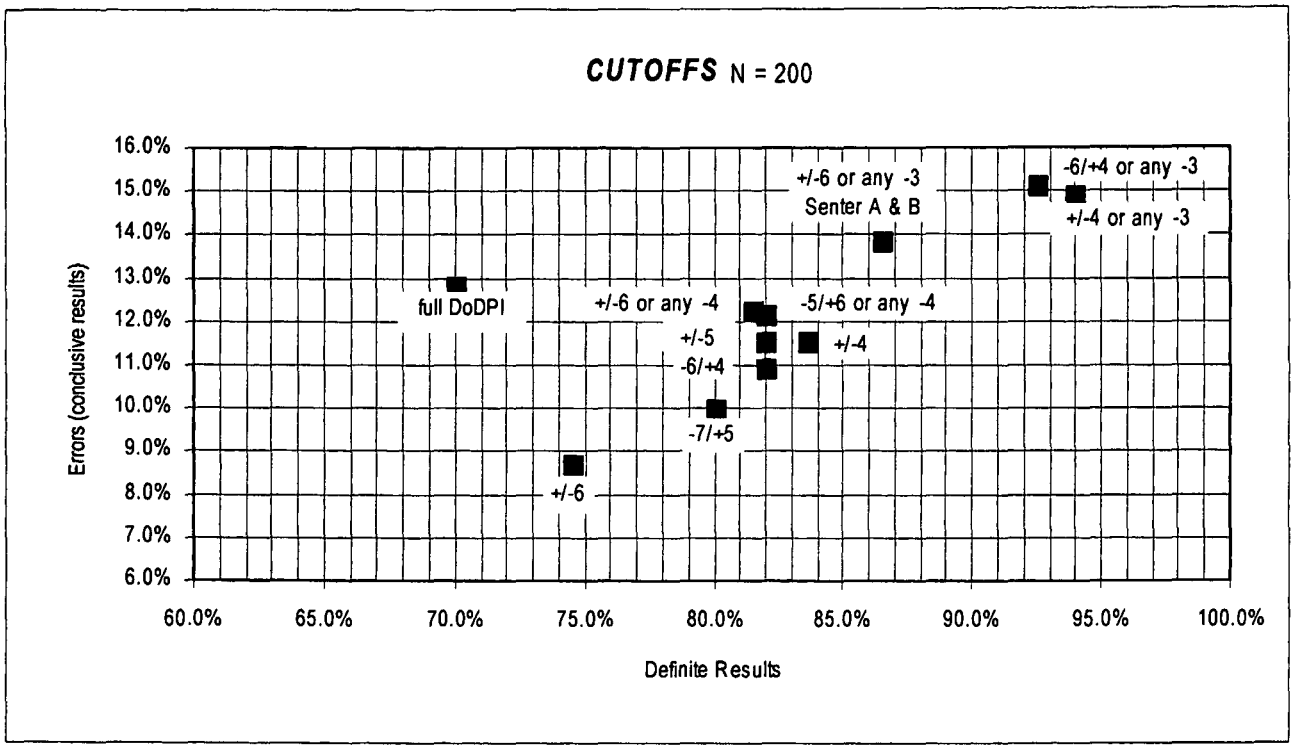


Figure 1 *Errors plotted against definite results as a function of cutoff rule.*

In Figure 1, theoretical best results would be plotted low on the chart and to the right. The bottom right-hand corner of the chart would represent the ultimate best combination, 100% definite results (no inconclusives) and the lowest possible error rate. The error scale on this chart does not go below 6% because no methods in this survey achieved even that. When the actual study data are plotted they mostly lie along a diagonal line running from the lower left toward the upper right-hand corner. This supports the prevailing theory that reductions in inconclusive rate are automatically accompanied by proportional increases in error rate. The obvious exception on this chart is "full DoDPI", which lies well outside the diagonal track formed by the other rules. The cause of that may be revealed by examining another rule, +/-6 with mandatory deception if any relevant is -3 or lower. That rule is identical to "full DoDPI" except that it omits the requirement that a truthful result must be supported by a positive (non-zero and non-negative) score for each relevant question. With that exception removed from the rule, the error rate got only 1% worse, while the inconclusive rate got 16.5% better. In addition, results using that rule fall along the diagonal line.

Discussion

Of the cutoff methods examined in this study, "best" depends on how much accuracy an examiner is willing to sacrifice for the sake of fewer inconclusive results. For instance, consider the differences between +/-4 and +/-6. The percentage of definite results was only 74.5 for +/-6, one of the poorest in the study, but was 83.6 for +/-4. The change in error rates went from 8.7% for +/-6 to 11.6% for +/-4. In other words, a 9% reduction in inconclusive rate is balanced against a 3%

increase in error rate. Many examiners might consider that an acceptable trade-off.

Further study with a larger group of examiners might help establish a better common cutoff than the prevailing +/-6 used by most examiners. But among any group of examiners, individual differences in scoring are likely to cause these rules to shift up and down any scale of errors and definite results. For instance, when +/-4 was applied to the scores obtained by Examiner 1, that rule produced 83.0% definite results with a 10.8% error rate. Applied to Examiner 2, the same rule produced 90.0% definite results with a 12.2% error rate. The two examiners got approximately the same proportions of false negative and false positive errors, but Examiner 2 got 7% more definite results and slightly more errors with the same cutoff rule.

There are several obvious problems with this kind of study. First, the verified results for the examinations are largely determined from confessions, either by the examinee, verifying a result of deception, or by someone else, confirming the examinee's truthful result. This is a standard practice in polygraph validity and reliability studies that utilize real cases, and it is the closest we can usually come to ground truth. For the examinations used in this study, an independent panel of examiners had previously concluded that sufficient verification existed for each examination. Still, this is a fallible method. People occasionally make false confessions, so some "verified" results may be wrong. There is no way to know and no better way to conduct a study with a large number of real cases. Only a laboratory paradigm with a mock crime can guarantee ground truth, but that method has its own set of problems.

Second, the questions used in these examinations must have varied considerably even when addressing similar issues. The two reviewing examiners did not have access to the questions. Given personal preferences and differences in training, there were probably questions in some examinations that the two reviewing examiners would have preferred not to use. Out of 100 examinations there may have been a few that the two reviewing examiners would have excluded from consideration if they had known the case facts and the questions.

Third, because the reviewing examiners did not know the questions they could not judge how appropriate it was to regard these examinations as single-issue, which was what they were purported to be, as opposed to mixed issue. That is a critical consideration, because cutoffs for total scores only apply to single-issue tests. Mixed issue tests dictate scoring each relevant question separately, rather than combining the scores. The reviewing examiners also did not know whether the last relevant question in each examination was the weaker "do you know" type, a common configuration in DoDPI format but one which would not ordinarily fit strict requirements for a single issue. In the face of that possibility, the reviewing examiners still presumed that the last relevant question was part of a single issue. By default, the reviewing examiners regarded all examinations as single-issue.

Fourth, the quality of the charts varied considerably. All were done with computerized instruments, but some had amplitude settings and artifacts that made the charts difficult to read. Fifth, two examiners do not necessarily match the "average" examiner in chart evaluation, so generalization of the results of this study may be poor. Finally, the two reviewing examiners (one of whom was the author) probably did not approach the scoring of these charts with the same level of attention that they would have brought to a real test they had conducted themselves. Examiners give consideration to what the examinee has at stake in an examination, but in this kind of study it cannot be denied that the examiner himself has less at stake than in his own real work.

Although scoring is an attempt to enforce objectivity and to standardize chart evaluation methods, it remains as much art as science. In an ideal world, every examiner would have an opportunity to score 100 verified examinations -- truly verified by some means better than confession -- so each could determine his or her own best cutoffs. Computer software might be devised that presents the charts, allows the examiner to score them, and applies a variety of cutoff rules, finally suggesting to the examiner the

best personal cutoffs to achieve the optimal combination of accuracy and definite results. Until that day arrives, we will continue to try to establish methods that work best for the hypothetical "average" examiner.

References

Abrams, S. & Leutwyler, G. & Wygant, J. (2000). Polygraph validity in the new millenium. *Polygraph, 4*, 344-356 (and *erratum* 2001, *2*, 118).

Backster, C. (1979). *Standardized polygraph notepack and technique guide, 1979 edition.* Backster School of Lie Detection.

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

Bibliography

Barland, G.H. & Raskin, D.C. (1976). *Validity and reliability of polygraph examinations of criminal subjects.* Report No. 76-1, Contract No. NI-99-0001 (Washington, D.C. National Institute of Justice, Dept. of Justice).

Capps, M.H. & Ansley, N. (1992). Comparison of two scoring scales. *Polygraph, 1*, 39-43.

Capps, M.H. & Ansley, N. (1992). Analysis of federal polygraph charts by spot and chart total. *Polygraph, 2*, 110-131.

Capps, M.H. & Ansley, N. (1992). Analysis of private industry polygraph charts by spot and chart total. *Polygraph, 2*, 132-142.

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

Capps, M.H. & Ansley, N. (1992). Strong control versus weak control. *Polygraph, 4*, 341-348.

Elaad, E. (1999). The control question technique: A search for improved decision rules. *Polygraph, 1*, 65-73.

Krapohl, D.J. (1998). A comparison of 3- and 7- position scoring scales with laboratory data. *Polygraph, 3*, 210-218.

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

Krapohl, D.J. (2001). An assessment of the total chart minutes concept. *Polygraph, 4*, 228-241.

Light, G.D. (1999). Numerical evaluation of the Army Zone Comparison test. *Polygraph, 1*, 37-45.

Swinford, J. (1999). Manually scoring polygraph charts utilizing the seven-position numerical analysis scale at the Department of Defense Polygraph Institute. *Polygraph, 1*, 10-27.

Detection of Deception

Jennifer M. C. Vendemia

Abstract

Several modern technologies are being applied to the study of deception. This article will discuss the theoretical application of event-related potential measures, functional magnetic resonance imaging, voice stress analysis, thermal imaging, and pupillometry. The future of the field will be determined by how well developing technologies co-exist with current polygraph techniques, as well as the overall strength of the supporting theoretical framework. It is important that our theoretical understanding of the process of deception become as sophisticated as the techniques that we are using to assess it.

Introduction

Polygraph is currently the most accurate measure of deception. However, modern technologies are being studied. Future techniques will relate to existing polygraph testing methodologies in one of two ways: 1) They will provide another channel of information within the standard polygraph measuring paradigm, or 2) they will provide an alternate methodology to existing exams. These differences are primarily theoretical, and most technologies can be implemented using either strategy. At the present time, thermal imaging, voice recognition, and pupillometry operate within the existing polygraph methodology, because these measures assess peripheral nervous system activity which is easily manipulated by the standard detection of deception exam. However, event-related potential (ERP), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET) measure central nervous system activity, and do not rely on emotional changes necessary for the standard PDD exam. These secondary measures can assess the cognitive aspects of deception such as attention, workload, memory, and salience. Thus, each of these technologies may be better suited for the detection of deception, as opposed to polygraph which emphasizes emotion and arousal.

Voice

The most disheartening of the future technologies has been voice stress analysis. The simple theory behind this measure states that anxiety related to deception will be detectable by slight fluctuations in the vocal recording. Voice stress devices have even been advertised in catalogues such as Sharper Image.

Examinations of voice for the purpose of detection of deception began as early as 1971 (National Academy of Sciences, 1979). Voice examination was conceived and promulgated by marketers and salesman rather than by scientists. The claims of such marketers are often fantastical. Recent demands for detection of deception technology have created an environment wherein many earnest law enforcement agencies and legitimate government funding organizations are falling for this scam. Who wouldn't want a magic wand that could instantly detect deception while a suspect spoke? Imagine the possibilities of being able to detect a lie from a tape of a person's voice or while they spoke on television. Political debates would never be the same! Unfortunately none of the claims of voice stress marketers has ever been substantiated.

The voice stress devices that have been marketed include the Psychological Stress Evaluator (PSE), the Hagoth, the Mark II Voice Stress Analyzer (VSA), and the Computerized Voice Stress Analyzer (CVSA). The CVSA is the most recent of these devices and has been heralded as a new dawn in voice stress detection. However, the only difference between CVSA and earlier devices is that it presents recorded vocal stimuli on a computer screen rather than on paper. Of the 15 university grade publications on voice stress, only one found any evidence that voice stress

study was unsuccessful. Table 1 presents a summary of studies presented in Krapohl (2001). The basic conclusions of these studies are that voice detection of deception is not valid, it is not reliable, and it does not work.

Table 1
Studies that Evaluated Voice Stress in the Detection of Deception.

Researchers	Year	Results	Conclusions
Brenner et al.	1979	-	Validity for deception detection poor
Cestaro	1995	-	Unable to detect deception
Meyerhoff	1995	-	Unable to detect stress
Fuller	1984	-	Validity of stress detection poor
Janiro et al.	1986	-	Unable to detect deception
Hollien et al.	1987	-	Unable to detect stress or deception
Horvath	1978	-	Unable to detect deception
Horvath	1979	-	Unable to detect deception
Kubis	1973	-	Unable to detect deception
Lynch et al.	1979	-	Unable to detect stress
O'Hair et al.	1985	-/+	Deception detected in one subgroup of study
O'Hair et al.	1990	-	Deception findings from previous study not replicated
Suzuki et al.	1973	-	Validity of stress and deception detection poor
Timm	1983	-	Unable to detect deception
Waln et al.	1987	-	Reliability of deception detection poor

Note. From Krapohl, D. (2001). Tech talk: Voice Stress Analysis Research. American Polygraph Association.

In 1999, a lawsuit was brought against the National Institute for Truth Verification and Charles Humble, one of the manufacturers of the CVSA, by a man falsely charged with sexual assault on the basis of CVSA technology. Although the lawsuit was defeated on other grounds, the fact remains that CVSA has not been deemed valid or reliable on scientific grounds. It does not detect deception.

Thermal

Thermal imaging measures changes in regional facial blood flow, particularly around the eyes. Sometimes a change in blood flow to the face is obvious such as when a person blushes. However, the goal of thermal imaging is to capture changes in blood flow related to the fright/flight response mediated by the sympathetic nervous system (Pavlidis, Eberhardt, & Levine, 2002). The clear advantage of this system is that individuals can be tested for deception without their awareness, because measurement takes place through a camera that is sensitive to changes in temperature. The major drawback of thermal imaging is the processing demand. Results from a camera are recorded on a computer, and those files must undergo substantial computer processing before they can be interpreted.

A test with 20 volunteers at the Department of Defense Polygraph Institute (DoDPI) were randomly assigned to stab a mannequin and rob it. They were also instructed to later assert their innocence during a thermal imaging exam. In the test, 83% of the participants were correctly categorized as innocent or guilty (Pavlidis et al., 2002). Pollina and Ryan (2002) conducted a feasibility study combining polygraph measures such as blood volume, respiration, and electrodermal activity with facial skin surface temperature changes. The frequencies of accurate determinations made using polygraph measures, thermal measures, and a combination of polygraph and thermal measures were compared using binary logistic regression. The highest accuracy was obtained using a combination of polygraph and thermal measures (R^2 = .52, p = .001), suggesting that facial thermal measures may

be successfully combined with polygraph measures during a psychophysiological detection of deception examination.

This research is still in its infancy, but is very promising. At this time, the underpinning theory is still unknown. Pollina and Ryan (2002) propose that the orienting response (Sokolov, 1963) may be an underlying component. Sokolov reported decreases in forehead blood volume in response to threatening stimuli that may reflect a defensive self-protective response. Novel stimuli produce increases in forehead blood volume which reflect an orienting response that might improve perceptual ability. Taken together these two components may drive the thermal response.

Pupillometry

Pupillometry, the study of changes in pupil size and movement, is not a modern technique. Pupil dilation can result from sympathetic nervous system stimulation or suppression of the parasympathetic nervous system, and it is this phenomenon that has been most useful in lie detection research. Pupillometry has been used for measuring a wide variety of phenomena including heroin withdrawal (Robinson, 1974), cognitive workload (Taylor, 1981), and memory (Headly, 1981; Krueger, 2001). It was first studied in conjunction with deception in the early 1940s (Berrien, 1942; Berrien & Huntington, 1943; Harney, 1943). These early studies determined that deception was paired with a change in the size of the pupil. Changes in pupil size occurred whenever a crime-relevant question was asked, but the change was more pronounced when the participant intended to be deceptive.

In more recent Concealed Information Test (CIT) studies, larger dilations were identified in "guilty" than "innocent" participants (Janisse & Bradley, 1980; Lubow & Fein, 1996). Other factors besides deception can influence pupil size. A participant's uncertainty about test outcome can cause greater relative changes in pupil size than certainty (Bradley & Janisse, 1979), and the more effective a participant believes the test to be, the greater the change in pupillary response (Bradley & Janisse, 1981). The cognitive processes involved in deception can also influence pupillary dilation (Dionisio, Granholm, Hillix, & Perrine, 2001; Heilveil, 1976). Participants have also shown demonstrably greater pupil dilations when they respond deceptively to learned episodic or semantic information (Dionisio et al., 2001), as well as to their autobiographical information (Heilveil, 1976).

Brain Waves

Brain waves related to specific stimuli or event-related potentials (ERPs) have been used to detect deception for several decades. In 1994, Larry Farwell was named Inventor of the Year by *Time Magazine* for his work on the identification of brain-waves associated with deception. However, there are debates on the methodology appropriate for ERPs, and accurate underlying theoretical mechanisms.

Based on the brain processes known to elicit ERPs, conflicting theories of lying have been developed (Boaz, Perry, Raney, Fischler, & Shuman, 1991). The process of deception may involve attentional capture (Allen & Iacono, 1997), working memory load (Dionisio, Granholm, Hillix, & Perrine, 2001; Stelmack, Houlihan, & Doucet, 1994), or perceived conflict with meaning and a person's memories (Boaz et al., 1991). Attentional capture refers to the directing of attention, generally towards a threat. For example, loud noises capture our attention because they could be threatening. For similar reasons, questions to which one is prepared to lie grab attention because of the threat of potentially being caught (Marston, 1917; Vendemia, 2002). Working memory load refers to how many unique ideas an individual can attend to at one time. An individual telling the truth does not need to keep ideas in active awareness, but someone who is being deceptive needs to keep track of deceptive answers as well as truthful answers. Telling a lie is a far more complex activity than telling the truth.

Three waveforms have been reported in deception research, the P3b, P3a, and N4. The waves vary in the way they are generally produced and in the way they are studied in relation to deception. The P3b is by far the most frequently reported component of the three, and is typically studied in the context of

the "concealed information" oddball paradigm. In the general oddball paradigm, an infrequently occurring stimulus is presented in a sequence of frequently occurring stimuli. For example, a high pitched tone amongst a group of low pitched tones would grab attention because it is different (Fabiani, Gratton, & Coles, 2000). The attention is related to the P3b.

The "oddball" stimulus produces a large positive ongoing peak with a latency of 350-600 milliseconds and a distribution whose maximum amplitude is at parietal sites and whose minimum amplitude is at anterior sites (Verlager, 1997). Similarly, the CIT/oddball consists of low probability stimuli that involve guilty knowledge presented among a series of high probability stimuli that do not involve guilty knowledge. In this paradigm, the low probability guilty knowledge item elicits a larger P3 component than the non-targets (Allen, Iacono, & Danielson, 1992). Although, researchers reporting ERPs from the CIT/oddball in this area do not explicitly describe this waveform as a P3b, its spatio-temporal characteristics suggest that it matches those of the P3b (Rosenfeld, Ellwanger, Nolan, Wu, Bermann, & Sweet, 1999).

The CIT/oddball effect has been demonstrated across multiple design permutations with visual and auditory stimuli. Across these studies, the P3 component of the ERP reliably and accurately indicates the presence of concealed knowledge (Allen & Iacono, 1997; Allen et al., 1992; Bashore & Rapp, 1993; Ellwanger, Rosenfeld, Sweet, & Bhatt, 1996; Farwell & Donchin, 1991; Rosenfeld, 1995, 1998; Rosenfeld, Sweet, Chuang, Ellwanger, & Song, 1996). However, the P3b is involved in many types of higher cortical functions including stimulus evaluation (Gevins, Cutillo, & Smith, 1995; Ruchkin, Johnson, Canoune, Ritter, & Hammer, 1990; Verleger, 1997), attention resource allocation (Comerchero & Polich, 1999), and updating of information held in working memory (Donchin & Coles, 1988; Ruchkin, Johnson, Canoune, & Ritter, 1990).

Precisely which of these underlying processes are involved in deception is unclear, and in the CIT/oddball task an often criticized confound of episodic memory further obscures the findings (Allen & Iacono, 1997). It is important to remember that this is a test of recognition memory of the concealed knowledge rather than a test of deception.

Larry Farwell has patented a form of the CIT/oddball called "Brain Fingerprinting." The goal of the test is to measure memory traces related to the crime. The limitations of the CIT, are the limitations of the Brain Fingerprinting technique. Additionally, Brain Fingerprinting is based on two fallacious assumptions. The first is that brain waves are stable over time. They are not. The shape and type of an individual's brain waves varies with a huge number of variables including time of day, age, alertness, brain trauma, and organic brain syndromes.

The second issue related to the technique revolves around the nature of memory itself. It used to be believed that memory was indelible; that it was like a film that could be replayed over and over. Modern researchers know that there are several factors known to distort memory. These include decay, interference, and misinformation. Decay refers to memory loss related to time, while interference refers to memory loss due to the presentation of new information. Misinformation, which is something that an interrogator understands, refers to the altering of memory by the inadvertent presentation of inaccurate information.

Because memory is rarely entirely accurate and prone to greater distortions over time, and because human brain waves are not stable over time, the Brain Fingerprinting technique is not really a form of fingerprinting at all. Fingerprinting implies a level of reliability not present in this technique. The technique would be best utilized as an additional channel on a polygraph rather than a stand alone method. In 2001, the Brain Fingerprinting technique was found to be acceptable according to the Daubert Rules of Evidence; however, in the particular case, the Harrington case, evidence gained by the Brain Fingerprinting technique did not warrant a new trial.

Not all brain wave measures utilize this technique. However, moving forward with an

accurate brain wave measure of deception requires a better understanding of what is actually occurring in the human brain during deception. Two main theories of deception, the attention theory and the working memory load theory, suggest different patterns of response for the P3b generated in the CIT based on the antagonistic effects of attention and workload (Kok, 2001). Attention theorists argue that attentional capture of the low frequency CIT items increases the amplitude of the P3b while working memory load theorists argue that the increased working memory demands required for deceptive processing suppresses the P3b. Both of these effects can be generated by manipulating task demands. In tasks with an attention-grabbing concealed information item, the P3b is larger, while in tasks with no oddball the P3b is suppressed. In order to examine the actual effects of deception, other waveforms must also be studied.

Like the P3b, the P3a is elicited by an oddball paradigm. In one variant of the oddball, the three-stimulus paradigm, the P3a occurs in response to novel-infrequent stimuli presented in addition to the "typical" oddball stimuli. The P3a can be elicited by shifts in attention (Comerchero & Polich, 1999), switching from difficult to easy task demands (Comerchero & Polich, 1999; Harmony et al., 2000), and alerting (Katayama & Polich, 1998). Across studies reporting the P3a in an oddball paradigm, it is alerting stimuli combined with initial attentional allocation that produce the phenomenon (Katayama & Polich, 1998). The term "P3a" is applied to an assortment of early P3 components with anterior distributions, and the exact conditions necessary to evoke a P3a vary across paradigm and stimulus demands (Katayama & Polich, 1998). In general, the waveform is characterized as a positive going peak with an anterior distribution, and a latency of 250-350 ms (Comerchero & Polich, 1999; Harmony et al., 2000; Spencer, Dien, & Donchin, 1999).

Two ERP studies of deception reported an early positivity with spatio-temporal characteristics similar to the P3a (Matsuda, Hira, Nakata, & Kakigi, 1990; Pollina & Squires 1998). Neither of the reported studies involved the oddball paradigm: (a) Pollina and Squires (1998) employed graded judgments of true and false sentences and (b) Matsuda et al., (1990) used a two-stimulus target detection task in which the first stimulus involved participant related information. Although the findings were mixed, Pollina and Squires (1998) suggested that the P3a occurred in probably true conditions.

Unlike P3b and P3a, the last component reported in studies of deception, the N4 component, is sensitive to semantic incongruity. Researchers argue that deception represents an incongruity between internal truth and external response (Bashore & Rapp, 1993). The N4 is a large negative deflection at around 400 ms with maximum amplitude in anterior and temporal regions. It is produced by stimuli that are incongruent in relation to the preceding context and is predominantly limited to linguistic information (Grunwald et al., 1998; Elger et al., 1997; Hahne & Friederici, 2002; McCarthy, Nobre, Bentin, & Spencer, 1995; Stuss, Picton, & Cerri, 1986). The N4 component has been elicited by the possession of concealed knowledge in sentence completion tasks involving false sentence completions (Boaz et al., 1991), and in a two-stimulus target detection task (Matsuda et al., 1990). Bashore and Rapp (1993) suggest that the N4 is reactive to anomalies in semantic and episodic memory as well as to inconsistencies in language semantics. Pollina and Squires (1998) conducted a study that did not share language inconsistence, but did share anomalies in semantic and episodic memory. No differences were found in N4 amplitude. In that study, participants made graded truth-value judgments that were sometimes inconsistent with memory, and these failed to alter N4 amplitude or latency (Pollina & Squires, 1998). In a two-stimulus task, the N4 was not found to be sensitive to deception, although it was found to be sensitive to response congruity with the second stimulus (Stelmack, Houlihan, & Doucet, 1994; Stelmack, Houlihan, Doucet, & Belisle, 1994).

Research in our lab has revealed a combination of the P3a, P3b, N4, and late positive potential combine during deception (Buzan, Sasine, Spade, & Vendemia, 2002; Vendemia & Buzan, 2002, 2003). The early positive component, the P3a was localized to the anterior cingulate gyrus, a region of the

brain involved in attention. The P3b was associated with activity in many different brain regions, and seems to be involved in decision making (Vendemia & Buzan, 2003). The late occurring negativity (N4) was predominantly localized to the inferior frontal gyrus, and seems to be related to congruity of the response. Finally the late positive complex was associated with regions of the temporal gyrus and anterior cingulate, and may be related to a final reanalysis of the response (Buzan et al., 2002; Vendemia & Buzan, 2003).

Functional Magnetic Resonance Imaging

When a human being engages in a cognitive activity such as subtraction, reading, or deceiving various parts of the brain become active. When neurons in these areas are active their metabolism increases, and they need more blood for nourishment. Brain mapping is achieved by setting up an MRI scanner is such a way that increased blood flow to the active areas of the human brain shows up on a functional MRI image. In a typical fMRI experiment, a participant will lie in a magnet while they perform a particular task. In the earliest fMRI studies, participants watched patterns of grids, such as checkerboards, while scientists measured the output from the visual cortex.

First, an MRI image is taken of the individual's brain. Each person's brain is unique in shape and size, and so these first images are very important. Later, the images of brain activity will be overlaid on the structural image. Next, a series of low resolution scans are taken over time. Some are taken during the task and some are taken when the individual is not engaged in the task. For example, some scans might be taken while an individual is lying, while others might be taken while an individual is telling the truth. Later the two sets of scans are compared to see which is more active. Early fMRI studies of deception have shown activity in the ventrolateral prefrontal cortices, a region of the brain associated with higher order processing (Spence et al., 2001). Research in our lab has shown a relationship between fMRI activity in the anterior cingulate, middle and superior frontal region, and medial temporal gyri and deceptive responses (Vendemia & Buzan, manuscript under review).

Conclusion

Polygraph is still the best measure of deception, but other techniques also exist. These techniques are only as good as the theories that support them. Gall designed phrenology to assess personality characteristics by measuring the protrusions and indentations on an individual's skull (Cleeton, 1927). He developed a sophisticated technique for making these measurements. He designed a complicated anthropometric apparatus quite ahead of its time for measuring the cranial surface. His measurements were excellent, but they didn't have the slightest thing to do with personality. It is important that our theoretical understanding of the process of deception become as sophisticated as the techniques that we are using to assess it.

References

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology, 29*, 504-522.

Allen, J. J., & Iacono, W. G. (1997). A comparison of methods for the analysis of event-related potentials in deception detection. *Psychophysiology, 34*, 234-240.

Bashore, T. R., & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin, 113*, 3-22.

Berrien, R. E. (1942). Pupillary responses as indicators of deception. *Psychological Bulletin, 39*, 504-505.

Berrien, R. E., & Huntington, G. H. (1943). An exploratory study of papillary responses during deception. *Journal of Experimental Psychology, 32*, 443-449.

Boaz, T. L., Perry, N. W., Raney, G., Fischler, I. S., & Shuman, D. (1991). Detection of guilty knowledge with event-related potentials. *Journal of Applied Psychology, 76*, 788-795.

Bradley, M.T., & Janisse, M.P. (1979, 1980). Pupil size and lie detection: The effect of certainty on detection. *Psychology: A Journal of Human Behavior, 16*, 33-39.

Bradley, M. T., & Janisse, M. P. (1981). Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and papillary measures. *Psychophysiology, 18*, 307-315.

Comerchero M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology, 110*, 24-30.

Cleeton, G. U. (1926). Estimating human character. *Scientific monthly, 23*, 427-431.

Dionisio, D. P., Granholm, E., Hillix, W. A., & Perrine, W. F. (2001). Differentiation of deception using pupllary responses as an index of cognitive processing. *Psychophysiology, 38*, 205-211.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 a manifestation of context updating? *Behavioral and Brain Sciences, 121*, 357-374.

Dutton, D. (2000). Introduction. *Polygraph*, 29, 1-5.

Elger, C. E., et al. (1997). Human temporal lobe potentials in verbal learning and memory processes. *Neuropsychologia, 35*, 657-667.

Ellwanger, J., Rosenfeld, J. P., Sweet, J. J., & Bhatt, M. (1996). Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. *International Journal of Psychophysiology, 23*, 9-23.

Fabiani, M., Gratton, G., & Coles, M. G. H. (2000). Event-related brain potentials: Methods, theory, and applications. In. J. T. Cacioppo, G. G. Tassinary, and G. G. Berntson (Eds.), *Handbook of Psychophysiology, 2nd Ed*, (pp. 53 – 84). Cambridge: Cambridge University Press.

Gevins, A., Cutillo, B., & Smith, M. E. (1995). Regional modulation of high resolution evoked potentials during verbal and non-verbal matching tasks. *Electroencephalography and Clinical Neurophysiology, 94*, 129-147.

Grunwald, T., K. Lehnertz, et al. (1998). Limbic ERPs predict verbal memory after left-sided hippocampectomy. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience, 9*, 3375-3378.

Harmony, T., Bernal., J., Fernández, T., T., Silva-Pereyra, J., Fernández-Bouzas, A., Marosi, R. Rodríguez, M., & Reyes, A. (2000). Primary task demands modulate P3 amplitude. *Cognitive Brain Research, 9*, 53-60.

Harney, J.W. (1943). Pupillary responses during deception. *Journal of Criminal Law and Criminology, 34*, 135-136.

Hahne, A. & Friederici, A. D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research, 13*, 339-356.

Headley, D. B. (1981). Puillometric assessment of retrieval operations in factual long-term memory. *Acta-psychologica, 49*, 109-126.

Heilveil, I. (1976). Deception and pupil size. *Journal of Clinical Psychology, 32*, 675-676.

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

Katayama, J. & Polich, J. (1998). Stimulus context determines P3a and P3b. *Psychophysiology, 35*, 23-33.

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology, 39*, 557-577.

Krapohl, D. (2001). Tech Talk: Voice Stress Analysis Research. *American Polygraph Association Newsletter, 34*, 42-44.

Krueger, F., Nuthmann, & A., van-der-Meer, E. (2001). Pupillometric indices of temporal order representation in semantic memory. *Zeitschrift fuer Pscyhologie, 209*, 402-415.

Lubow, R. E. & Fein, O. (1996). Pupillary size in response to a visual guilty knowledge test: New technique for the detection of deception. *Journal of Experimental Psychology: Applied, 2*, 164-177.

Matsuda, T., Hira, S, Nakata, M., & Kakigi, S. (1990). The effect of one's own name upon event related potentials: Event related (P3 and CNV) as an index of deception. *Japanese Journal of Physiological Psychology and Psychophysiology, 8*, 9-18.

Marston, W. M. (1917). Systolic blood pressure symptoms of deception. *Journal of Experimental Psychology, 2*, 117-163.

McCarthy, G., Nobre, A. C. Bentin, S., & Spencer, D. D. (1995). Language-related field potentials in the anterior-medial temporal lobe: I. Intracranial distribution and neural generators. *Journal of Neuroscience, 15*, 1080-1089.

National Academy of Sciences. (1979). *On the Theory and Practice of Voice Identification.* National Academy of Sciences, Washington, D. C.

Pavlidis, I., Eberhardt, N. I., & Levine, J. A. (2002). Seeing through the face of deception: thermal imaging offers a promising hands-off approach to mass security screening. *Nature, 415*, 35.

Pollina, D.A. & Ryan, A. (2003). *The relationship between facial skin surface temperature reactivity and traditional polygraph measures used in the psychophysiological detection of deception: A preliminary investigation.* Department of Defense Polygraph Institute: Columbia, SC. (DoDPI02-R-0007).

Pollina, D. A. & Squires, N. K. (1998). Many-valued logic and event-related potentials. *Brain and Language, 63*, 321-345.

Robinson, M.G. (1974). Assessment of pupil size during acute heroin withdrawal in Viet Nam. *Neurology, 24*, 729-732.

Rosenfeld, J. P. (1995). Alternative views of Bashore and Rapp's (1993). Alternatives to traditional polygraphy: A critique. *Psychological Bulletin, 117*, 159-166.

Rosenfeld, J. P. (1998). Event-related potentials in detection of deception. *International Journal of Psychophysiology, 30*, 27.

Rosenfeld, J. P., Ellwanger, J. W. Nolan, K., Wu, S., Bermann, R.G., & Sweet, J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *International Journal of Psychophysiology, 33*, 3-19.

Rosenfeld, J. P., Sweet, J. J., Chuang, J., Ellwanger, J., & Song, L. (1996). Detection of simulated malingering using forced choice recognition enhanced with event-related potential recording. *The Clinical Neuropsychologist, 10*, 163-179.

Ruchkin, D. S., Johnson, R., Canoune, H. L., Ritter, W., & Hammer, M. (1990). Multiple sources of the P3b associated with different types of information. *Psychophysiology, 27*, 157-176.

Sokolov, E. (1963). *Perception and the Conditioned Reflex.* New Your: Macmillan.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P.W.R. (2001). Behavioral and functional anatomical correlates of deception in humans. *Neuroreport: For Rapid Communication of Neuroscience Research, 12*, 2849-2853.

Spencer, K. M., Dien, J., & Donchin, E. (1999). A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology, 36*, 409-414.

Stelmack, R. M., Houlihan, M., Doucet, C., & Belisle, M. (1994). Event-related potentials and the detection of deception: A two-stimulus paradigm. *Psychophysiology, 7*, s94.

Stelmack, R. M., Houlihan, M., & Doucet, C. (1994). *Event-related potentials and the detection of deception: A two-stimulus paradigm.* Ottawa: University of Ottawa. (NTIS No. AD-A318 987/5INZ).

Taylor, J. S. (1981). Pupillary response to auditory versus visual mental loading: A pilot study using super 8-mm photography. *Perceptual and Motor Skills, 52*, 425-426.

Vendemia, J. M. C. (2002). Hobson's choice: the relationship between consequences and the comparison question. *Polygraph, 31*, 20–25.

Vendemia, J. M. C. (2003). Neural Mechanisms of Deception and Response Congruity to General Knowledge Information and Autobiographical Information in Visual Two-Stimulus Paradigms with Motor Response (in press). Department of Defense Polygraph Institute (DoDPI99-P-0010).

Vendemia, J. M. C. & Buzan, R.F. (2003). *Neural mechanisms of deception and response congruity in a visual two-stimulus paradigm with motor response.* Manuscript submitted for publication.

Vendemia, J. M. C. & Buzan, R. F. (2002). Deception and response congruity in visual two-stimulus paradigms involving motor response. *International Journal of Psychophysiology, 45*, 28-29.

Vendemia, J. M. C. & Buzan, R. F. (2003). *The Effects of response predictability on HD-ERP measures across studies of deception.* Poster session presented at the 10th Annual Convention of the Cognitive Neuroscience Society, New York, NY. Manuscript submitted for publication.

Vendemia, J. M. C., Buzan, R. F., & Simon-Dack, S. L. (2003). *Reaction time of motor responses in two-stimulus paradigms involving deception and congruity with varying levels of difficulty.*

Verleger, R. (1997). On the utility of P3 latency as an index of mental chronometry. *Psychophysiology, 34*, 131-156.

Objective Assessment of Comparison Question Polygraphy

Vance V. MacLaren & Donald J. Krapohl

Abstract

A system for quantifying psychophysiological information obtained in Comparison Question polygraph Tests (CQTs) was developed on a sample of 181 confirmed deceptive and 150 nondeceptive field cases. The proposed system uses permutation tests to arrive at separate estimates of the likelihood of deception and nondeception for individual cases. These probabilities are then combined with base rate information to allow an overall probability of guilt to be calculated for each case. Assuming a 50% base rate of deception and conservative cutoff points for assigning deceptive (p=.90) and nondeceptive (p=.10) outcomes, 91% of deceptive cases (N=152) and 98% of nondeceptive cases (N=128) with conclusive outcomes were correct. The system was cross validated on four groups of field cases. Conclusive outcomes were correct in 90% of deceptive cases verified by confession (N=97), in 92% of deceptive cases verified by urinalysis tests for drug infractions (N=49), and in 64% of deceptive cases independently verified by physical evidence or subsequent confession not associated with the polygraph test (N=11). Conclusive outcomes were correct in 88% of independently verified nondeceptive cases (N=16). The system was also cross validated on a sample of laboratory cases drawn from three separate experiments. In the laboratory samples, 88% of conclusive outcomes were correct for both deceptive (N=33) and nondeceptive (N=25) groups. The proposed scoring system circumvents several logical problems with traditional approaches to quantification in CQT polygraphy, and also provides an effective means of arriving at accurate classifications in approximately 90% of cases.

Forensic polygraph testing involves the monitoring of physiological indicators of psychological stress during questioning about suspected criminal activity. This family of techniques has been in existence since the 1930s (Lykken, 1998), having grown into an application of psychology with important implications for the criminal justice system and various other facets of modern society. Polygraph testing is now routinely applied to problems as diverse as criminal investigation, employee selection, offender rehabilitation, and counterintelligence. Along with this widespread application comes an ongoing debate among practitioners and academics. Professionals in the polygraph industry assume the role of protectors of community safety and national security, while a majority of academics reject the adequacy of available scientific evidence validating forensic polygraphy (National Academy of Sciences, 2002). The debate reflects a wider ideological clash between the values associated with defense of individual rights versus the maintenance of law and order. Such perspectives have slanted past discussions concerning the validity of the most common polygraph test, known as the Comparison Question Technique (CQT; Reid, 1947). Critics point to methodological flaws in the available research literature and argue that the CQT lacks scientific credibility (e.g. Ben-Shakhar & Furedy, 1990; Iacono, 2000; Lykken, 1998). A minority of authors believe that the available field studies (Honts, 1996; Honts & Raskin, 1988; Horvath, 1977; Kleinmuntz & Szucko, 1984; Patrick & Iacono, 1991) and laboratory simulations of CQT polygraphy (e.g. Kircher, Horowitz, & Raskin, 1988) do provide valid estimates of accuracy and assert that the technique has been validated empirically (e.g. Raskin, Honts & Kircher, 1997). In the absence of consensus regarding the standard of evidence that should be accepted as sufficient to provide a reasonable estimate of polygraph test validity, the factual bases supporting either side of this argument remain contentious. Consequently, the volume of rhetoric goes far beyond the availability of hard evidence in support of either position. One major point of concern regards the methods by which practitioners assess the outcomes of individual examinations, a process with the potential to be highly

subjective and open to examiner bias. In this paper, we present a methodology for objectively assessing the physiological information obtained in the CQT. This quantification system was validated using data collected both under field conditions and in laboratory simulations. The proposed system has been designed to address several of the criticisms that have been leveled against CQT polygraphy. While addressing these concerns, the present report also provides the first comprehensive estimates of CQT validity. This work is intended as an initial step toward the emergence of a scientifically based forensic psychophysiology.

CQT Design

The CQT is a dynamic interview procedure in which stress responses to questions are observed. At least two classes of questions are present in a CQT exam. *Relevant questions* (e.g. "Did you steal money from the safe?") may be answered with either deceptive or nondeceptive "yes" or "no" answers, and are compared against a baseline of physiological response established by presenting *comparison questions*. The comparison questions (e.g. "Have you ever borrowed money that you might not be able to pay back?") are phrased in such a way that most examinees are probably deceptive when answering "no" to them. The social situation is manipulated in such a way that examinees feel they must be deceptive to successfully complete the exam. Comparison questions are intended to thereby elicit emotional arousal in both deceptive and nondeceptive suspects. The mechanism underlying the technique is a differential emotional impact that the relevant questions may present to deceptive suspects. These disparate emotional responses are inferred by observing momentary changes in bodily responses mediated by the sympathetic branch of the autonomic nervous system. Although there are numerous physiological channels that could be exploited as indicators of sympathetic activation, forensic polygraph examiners typically focus their observations on diastolic blood pressure, respiratory activity, and electrodermal responses.

In methodological terms, the CQT approach to deception detection may be thought of as a within-subjects experimental design wherein the construct of interest, sympathetic nervous system activation, may be inferred by the measurement of multiple dependent variables. The challenge faced by forensic examiners is that of separating out responses that occur as the result of deception from those that are not. To sift the deception-related signal from the ongoing psychological and physiological noise, an ideal test would have at least two treatment conditions: an experimental condition in which the suspect may respond either deceptively or truthfully, and a control condition that is identical in every way to the experimental condition except that the verbal responses are known to be deceptive. If these conditions could be satisfied, and if sympathetic activation occurs with deception, then observed differences in physiological responding could reasonably be attributed to deception.

In attempting to satisfy the requirements for an adequate and control condition, Horowitz, Kircher, Honts, and Raskin (1997) specifically directed their participants to lie in response to comparison questions. While this ensures that examinees answer the comparison questions deceptively, it does not necessarily provide a sufficient control condition. If the crime relevant questions were to be perceived by a suspect as important to the outcome of their investigation, substantial physiological reactions might be observed, irrespective of the presence or absence of deception. This problem might be even more troublesome in field tests than in laboratory simulations, because criminal suspects have much more at stake than participants in an experiment. Like all simulations of CQT polygraphy, the degree to which these laboratory results may generalize to field conditions is a matter for speculation.

Another attempt at isolating the deception effect was initiated by Furedy and colleagues, who used an experimental procedure known as the Differentiation of Deception Paradigm (DDP). Although not intended as a field polygraph test, the DDP provides experimental support for the notion that increases in sympathetic arousal are associated with deceptive verbal responses. In this design, salience of relevant and comparison questions is held approximately constant by simply asking participants a

series of questions regarding biographical information or general knowledge. Prior to the test, the questions are reviewed with the subject and they are instructed to prepare deceptive replies to a subset of them. Larger autonomic responses to deceptively answered questions have been found in several studies (Dionisio, Granholm, Hillix, & Perrine, 2001; Furedy, Davis, & Gurevich, 1988; Furedy, Gigliotti, & Ben-Shakhar; Furedy, Posner, & Vincent, 1991; Godert, Rill, & Vossel, 2001; Vincent & Furedy, 1992). While this design allows the effect of deception on sympathetic activation to be observed experimentally, the DDP remains tangential to the CQT as currently practiced. However, these studies do provide evidence suggesting that a quantifiable relationship between deception and physiological arousal may exist. It is quite possible that the CQT might have such a deception effect as its principal mechanism of action.

It is generally accepted that CQT crime relevant questions differ qualitatively from comparison questions and that the two may possess different levels of salience for deceptive and nondeceptive examinees. Examiners typically attempt to equalize the perceived importance of the questions by reviewing the test with the examinee before the physiological recording stage and carefully emphasizing the comparison questions. If successful, examinees may respond physiologically to the comparisons, thus providing a crude approximation to the ideal control condition. In so far as this objective is accomplished, physiological responses to the two question types may be interpreted as indicating the presence or absence of deception. If an adequate control condition can not be established, then the test loses its internal validity.

CQT Quantification

A corollary of the interpersonal nature of the CQT is that skilled examiners may be able to obtain important admissions from suspects in the course of pre-test preparation and continued interviewing after the physiological data recording phase is completed. Examinees who believe that their veracity is being assessed objectively may be more likely to confess their guilt than they would be if the polygraph were not a part of the interrogation. This bogus pipeline effect (Jones & Sigall, 1971) may be an important contributor to the utility of forensic polygraphy. Even if the CQT fails to provide a scientifically based method of detecting lies on the basis of physiological responses, it may nevertheless have value as a stageprop in the hands of an experienced interrogator. However, the CQT is purported to be a method of psychophysiological deception detection and the present treatment of CQT quantification is restricted to the assessment of deception solely on the basis of physiological responses.

The traditional approach to assessment in CQT polygraphy (Backster, 1969) does not meet the rigorous standards of scientific psychophysiology. To assess the magnitude of differences between responses to relevant and comparison questions, examiners typically assign numerical scores ranging from -3 through +3 to all three physiological response channels in each relevant / comparison (R/C) question pair. Larger scores indicate greater differences in the magnitude of responses. A negative score is assigned when a relevant question evokes a response greater than its associated comparison question. Positive scores are given when the comparison question is largest. These scores are tallied and cutoffs (typically +/- 6) are then applied to the total in order to render classifications of deceptive, nondeceptive, or inconclusive. This pseudo-quantification amounts to a 7-point subjective rating scale. Because examiners may have access to details about a case under investigation that go far beyond any physiological information gathered during questioning, the assignment of numerical scores in this fashion may be prone to bias and errors in judgment. Subtle factors, like a suspect's demeanor and non-verbal behavior, might influence the subjective interpretation of physiological data. A proper method of quantification should allow decisions to be made using only the physiological information.

A more modern approach to CQT evaluation is the use of multivariate statistical classifiers in conjunction with computerized data acquisition systems (Kircher & Raskin, 1988). While more systematic than the traditional numerical approach, this strategy has its own set of limitations. Because discriminant analysis is a parametric

statistical technique, the accuracy of test outcomes may be degraded to the extent that actual data differ from assumed population characteristics. While these assumptions can be tested and appropriate transformations can sometimes be applied to correct violations, several multivariate approaches have proven less than ideal. In a direct comparison of five multivariate classifiers, the proportion of correct classifications ranged from 71% to 77% (Dollins, Krapohl, & Dutton, 2000). There are several reasons that may contribute to this poor performance. Physiological responses to psychological stimuli often differ widely between individuals and this may impair the efficiency of discriminant functions in making accurate classifications on an individual basis. Also, the generalizability of any statistical classifier depends on the original data used to develop it. A multivariate classifier that is created using laboratory simulation data can hardly be expected to perform well in field cases if the laboratory situation fails to adequately mimic those conditions. Finally, in attempting to increase the variance accounted for by a statistical classifier, developers may capitalize upon spurious relationships that may occur by chance in a data set. This 'overfitting' of data can exaggerate the apparent accuracy of a multivariate classifier, particularly when the classifier is not adequately cross validated on independent samples.

In developing a quantification system for the CQT, careful attention must be paid to the characteristics of the technique. Because the CQT contains a dynamic interview component, no two sessions are exactly alike. Variations in the level of skill possessed by examiners and in the approaches that they take in conducting an examination can lead to differences in test reliability and accuracy. Also, individual examinees may differ in terms of their psychological and physiological reactions under questioning. These problems may be daunting, but they do not necessarily rule out the possibility of developing an objective scoring protocol. Indeed, a proper assessment system should be able to compensate for these sources of error. If one assumes repeatable procedures for test administration and a specifiable range of reaction to the relevant and comparisons, then these sources of error may be accounted for by a flexible scoring system. The present strategy was to create a set of scoring criteria from actual field data, assuming a finite level of error. This error comes both from within-subject variability in physiological responses to repeated questioning, and from between-subject variability that stems from the differences between examiners and examinees.

The Proposed Quantification Strategy

Forensic polygraph examiners must test a bi-directional hypothesis: is the examinee deceptive or nondeceptive when they deny committing a crime? Conceptually, to conclude that an examinee is deceptive is equivalent to a conclusion that there is a high posterior likelihood that the suspect belongs to a wider population of deceptive test-takers. Conversely, a low likelihood of deception implies that the suspect is nondeceptive. In the actuarial approach taken here, the examiner assumes a prior probability of deception (i.e. based on information other than the physiological responses to the polygraph test) and then uses physiological information gained through the test process to revise that estimate accordingly. This revised estimate is referred to as the posterior probability. To estimate the prior probability of deception is to admit and quantify the degree of bias that is always present in human judgement. Failure to deal with examiner bias in an orderly way has been a major criticism of CQT polygraphy (e.g. Ben-Shakhar & Furedy, 1990)

In the traditional numerical approach to CQT evaluation, no provision is made for the differences in subjective importance held by the comparison and crime relevant questions. This is problematic because there is no way of guaranteeing that an examiner has been successful in attempting to equalize a given suspect's psychological impression of the questions. Without accounting for the likely event that the questions differ in terms of salience, there is a very real possibility of error. To circumvent this obstacle, the present approach to CQT quantification uses decision criteria that are asymmetrical. We assume that the levels of physiological response to comparison and relevant questions may be different, but specifiable. To accept the proposition that both deceptive and

nondeceptive suspects ascribe differing levels of salience to the comparison and relevant questions does not necessarily rule out the possibility that some increment in physiological response might also be associated with deception. Under these assumptions, a proper test of the likelihood of deception in an individual case would involve estimating the similarity of the relevant / comparison differences observed in that case to distributions of such differences believed to be typical of the wider populations of deceptive and nondeceptive test takers. Following such a strategy, the incremental difference associated with deception may be isolated.

Another concern addressed by the present method regards the issue of base rates. Base rates are a perennial source of difficulty in any situation in which individuals must be classified on the basis of a test result (Meehl & Rosen, 1955). When the population base rate of occurrence of members of a category is very small, even modest rates of error can render a test useless, or even counterproductive on a system-wide basis. The proposed scoring system was therefore designed to account for population base rates in rendering decisions. If the population base rate of deception in a given application can be estimated, then this information may be incorporated into the decision making process for an individual case. Murphy (1987) provided the following formula for calculating the posterior probability of deception, in odds form, when base rate information is available:

$$\frac{P(D \mid F)}{P(T \mid F)} = \frac{P(F \mid D)}{P(F \mid T)} \times \frac{P(D)}{P(T)}$$

To convert into percentage form, one can divide the value of the term to the left of the equals sign by that value plus one. The result is the posterior likelihood that the suspect is deceptive, given the base rate of deception in that population, and the individual's test performance. The terms *P(D)* and *P(T)* refer to the population base rates of deception and nondeception, respectively. The terms *P(F|D)* and *P(F|T)* refer to the probabilities of true positive and false positive classification, respectively. These latter two components make up the *likelihood ratio* for a given case. The present method provides a means of estimating the likelihood ratio in individual examinations. This allows the posterior probability of deception to be calculated as a revision of the assumed prior base rate of deception in a given application.

Statistical decision making requires an individual's test performance to be systematically compared against the population of tests taken by others who are known to have been deceptive or nondeceptive. Although population statistics are unavailable, this distribution may be estimated using a large sample of tests taken by known deceptive and nondeceptive individuals. One way of comparing an individual test result against a population distribution is to consider the test performance observed in the examinee as the minimum level of performance that would be considered a positive test outcome. Using this point as a cutoff, the proportion of other test takers in the population who would be classified as deceptive may be calculated. By comparing individual test results against large samples of known deceptive and nondeceptive cases, the present method provides a means of estimating *P(F|D)* and *P(F|T)* in individual cases. Using these estimates, and Murphy's (1987) formula, the posterior likelihood of deception may be calculated.

The posterior probability estimates obtained using the proposed system are made while compensating for differences in the perceived salience of crime relevant and comparison questions that may exist in both deceptive and nondeceptive examinees. Examiner bias, characteristics of individual examinees, differential salience of relevant versus comparison questions, and base rates are all uncontrollable factors. Rather than change the CQT technique to accommodate these things, which may not even be possible, the present strategy was to neutralize their influence using a scoring system that takes them into account. The system was assessed on samples of confirmed deceptive and nondeceptive field cases, as well as data obtained in laboratory simulations.

Original Field Data

Field cases

A sample of 150 confirmed nondeceptive and 181 known deceptive

polygraph records was drawn from the database of confirmed field cases maintained by the Department of Defense Polygraph Institute (DoDPI).[1] Guilt was established on the basis of confession of the suspect or irrefutable physical evidence. Innocence was defined either by the confession of another person exonerating the suspect or proof that the crime never occurred. In some cases, only circumstantial evidence suggesting innocence was available. Determination in such cases required consensus among a panel of three polygraph examiners who reviewed both the circumstantial evidence and the polygraph charts. Inclusion in the database depended not on the test outcome, but on the availability of evidence that definitively resolved the suspect's deceptive or nondeceptive status. Selection criteria did not include whether the examination was of a suspect, witness, or victim. Demographic variables (e.g. age, gender, education, income, race) were also not used as selection criteria. The examinations were conducted by a variety of federal, state, and local law enforcement agencies throughout the United States.

In each selected case, the CQT polygraph test followed the Zone Comparison Technique format (ZCT; DoDPI, 1992). These examinations contained three crime relevant and three exclusionary probable lie comparison questions (Horowitz, Kircher, Honts, & Raskin, 1997), and each set of questions was repeated a minimum of three times. In cases where more than three repetitions were made, only the first three were included in the analysis. Each question series also contained one sacrifice relevant, one irrelevant, and two symptomatic questions (Matte, 1996), all of which were not considered in scoring. A commercially available computerized polygraph system (Axcition Systems, Houston, TX) was used to record the physiological data in all cases. Cases with missing data were excluded from the sample.

Physiological measures

A special software package developed for DoDPI (Extract, ver 3.0; Johns Hopkins Applied Physics Laboratory, 1999) was used to measure key physiological responses to questions from the continuous polygraph records. Those features were respiration line length (RLL; Timm, 1982), and electrodermal and blood volume amplitudes (Kircher, Raskin, Honts, & Horowitz, 1988). Respiration line length, as used in this study, is the linear distance of the respiration tracing for 10 seconds, beginning at stimulus onset. Electrodermal and blood volume amplitudes are rises in the tracings above the level recorded at stimulus onset.

Numerical score assignment

To reduce individual differences in physiological responsivity, all measures were within-subject. Each reaction to a crime relevant question was divided by the response evoked by the adjacent comparison question, creating a ratio. This use of relevant / comparison (R/C) ratios put the responses observed across all suspects and across all response channels into a common metric. Each suspect produced 9 ratios for each of the physiological response channels, for a total of 36.

Following Krapohl and McManus (1999), numerical scores were assigned to individual R/C ratios by comparing the ratios observed in each individual to a larger distribution of ratios obtained across suspects. The R/C ratios obtained in the deceptive and nondeceptive groups were segregated into separate distributions and arranged in ascending order. The eight distributions were each divided into seven subgroups, each containing an approximately equal number of R/C ratios. In the nondeceptive group, either 192 or 193 R/C ratios were contained within each *septile*. The deceptive group's septiles each contained either 232 or 233 R/C ratios.

Each R/C ratio was compared to the septile thresholds to assign a numerical score ranging from –3 through +3, in increments of 1. The septile ranges and their associated numerical scores are shown in Tables 1 and 2.

[1] Portions of these data have been used previously by developers of commercially available computerized polygraph systems.

Table 1
Score Thresholds for Comparison Against the Distribution of Confirmed Deceptive Cases.

		Scores (double for EDA)						
Channel	Threshold	-3	-2	-1	0	1	2	3
RLL1	upper	0.725	0.823	0.902	0.967	1.042	1.152	999
	lower	0.000	0.725	0.823	0.903	0.967	1.042	1.152
RLL2	upper	0.741	0.836	0.906	0.969	1.046	1.153	999
	lower	0.000	0.741	0.836	0.906	0.969	1.046	1.153
EDA	upper	999	3.759	2.393	1.814	1.427	1.125	0.780
	lower	3.772	2.408	1.815	1.429	1.128	0.780	0.000
BV	upper	999	2.304	1.639	1.302	1.054	0.837	0.600
	lower	2.307	1.641	1.306	1.054	0.838	0.601	0.000

Table 2
Score Thresholds for Comparison Against the Distribution of Confirmed Nondeceptive Cases.

		Scores (double for EDA)						
Channel	Threshold	-3	-2	-1	0	1	2	3
RLL1	upper	0.813	0.904	0.979	1.044	1.125	1.261	999
	lower	0.000	0.813	0.905	0.980	1.045	1.126	1.262
RLL2	upper	0.815	0.906	0.968	1.039	1.130	1.262	999
	lower	0.000	0.815	0.906	0.968	1.040	1.130	1.263
EDA	upper	999	1.670	1.136	0.884	0.687	0.539	0.357
	lower	1.678	1.136	0.886	0.687	0.541	0.358	0.000
BV	upper	999	1.656	1.266	1.024	0.849	0.689	0.518
	lower	1.659	1.268	1.025	0.850	0.690	0.521	0.000

Lower RLL ratios (i.e. when RLL was shorter after crime relevant questions than comparison questions) resulted in lower numerical scores. Higher BV and EDA ratios (ie when BV increase or EDA amplitude was greater after crime relevant than comparison questions) resulted in lower numerical scores. The EDA channel was weighted twice as strongly as the other channels by doubling the numerical scores from +/-3 to +/-6.

For each case, two sets of numerical scores were assigned. One set was made by sorting the suspect's R/C ratios into the septile ranges derived from the distribution of deceptive cases; the other set was made using the septiles derived from the nondeceptive distribution. The two sets of numerical scores for each case were tallied separately.

Only one RLL score for each R/C comparison was included in the totals. If the two RLL channels produced different numerical scores that were of the same sign, the RLL score furthest from 0 was included. If the RLL scores were on the opposite sides of 0, a score of 0 was assigned. The total scores had a possible range of −108 through +108. These two total numerical scores provided independent estimates of the similarity of each suspect's physiological responses to those observed in the samples of deceptive and

nondeceptive cases.

Likelihood ratio estimation

To provide estimates of the likelihood ratio for each case, a permutation test was conducted on each of suspect's two numerical score totals. Because numerical scores were assigned on the basis of septiles containing equal numbers of R/C ratios, we assumed that the numerical scores each had an equal chance of occurrence within the populations of deceptive and nondeceptive test-takers. To obtain the likelihood ratio for each case, $P(F|D)$ and $P(F|T)$ were estimated by comparing the numerical score totals to a distribution of possible outcomes. To obtain this distribution (see Table 3), the frequency of each possible score was calculated and divided by the total number of possible outcomes.

Table 3
Probabilities of Total Scores for Tests Having Between 6 and 9 Relevant / Comparison Contrasts.

Total Score	Number of Comparisons				Total Score	Number of Comparisons			
	6	7	8	9		6	7	8	9
-38	0.00	0.00	0.00	0.00	0	0.52	0.52	0.51	0.51
-37	0.00	0.00	0.00	0.01	1	0.55	0.55	0.54	0.54
-36	0.00	0.00	0.00	0.01	2	0.58	0.58	0.57	0.57
-35	0.00	0.00	0.01	0.01	3	0.61	0.61	0.60	0.59
-34	0.00	0.00	0.01	0.01	4	0.64	0.63	0.63	0.62
-33	0.00	0.01	0.01	0.01	5	0.67	0.66	0.65	0.64
-32	0.00	0.01	0.01	0.02	6	0.70	0.69	0.68	0.67
-31	0.00	0.01	0.01	0.02	7	0.73	0.72	0.70	0.69
-30	0.01	0.01	0.02	0.02	8	0.76	0.74	0.73	0.72
-29	0.01	0.01	0.02	0.03	9	0.78	0.77	0.75	0.74
-28	0.01	0.02	0.02	0.03	10	0.81	0.79	0.77	0.76
-27	0.01	0.02	0.03	0.04	11	0.83	0.81	0.79	0.78
-26	0.02	0.02	0.03	0.04	12	0.85	0.83	0.81	0.80
-25	0.02	0.03	0.04	0.05	13	0.87	0.85	0.83	0.82
-24	0.02	0.03	0.04	0.05	14	0.88	0.87	0.85	0.84
-23	0.03	0.04	0.05	0.06	15	0.90	0.88	0.87	0.85
-22	0.04	0.05	0.06	0.07	16	0.91	0.90	0.88	0.87
-21	0.04	0.06	0.07	0.08	17	0.93	0.91	0.90	0.88
-20	0.05	0.07	0.08	0.09	18	0.94	0.92	0.91	0.90
-19	0.06	0.08	0.09	0.10	19	0.95	0.93	0.92	0.91
-18	0.07	0.09	0.10	0.12	20	0.96	0.94	0.93	0.92
-17	0.09	0.10	0.12	0.13	21	0.96	0.95	0.94	0.93
-16	0.10	0.12	0.13	0.15	22	0.97	0.96	0.95	0.94
-15	0.12	0.13	0.15	0.16	23	0.98	0.97	0.96	0.94
-14	0.13	0.15	0.17	0.18	24	0.98	0.97	0.96	0.95
-13	0.15	0.17	0.19	0.20	25	0.98	0.98	0.97	0.96
-12	0.17	0.19	0.21	0.22	26	0.99	0.98	0.97	0.96
-11	0.19	0.21	0.23	0.24	27	0.99	0.98	0.98	0.97
-10	0.22	0.23	0.25	0.26	28	0.99	0.99	0.98	0.97
-9	0.24	0.26	0.27	0.28	29	0.99	0.99	0.98	0.98
-8	0.27	0.28	0.30	0.31	30	1.00	0.99	0.99	0.98
-7	0.30	0.31	0.32	0.33	31	1.00	0.99	0.99	0.98
-6	0.33	0.34	0.35	0.36	32	1.00	0.99	0.99	0.99
-5	0.36	0.37	0.37	0.38	33	1.00	1.00	0.99	0.99
-4	0.39	0.39	0.40	0.41	34	1.00	1.00	0.99	0.99
-3	0.42	0.42	0.43	0.43	35	1.00	1.00	1.00	0.99
-2	0.45	0.45	0.46	0.46	36	1.00	1.00	1.00	0.99
-1	0.48	0.48	0.49	0.49	37	1.00	1.00	1.00	1.00

The total number of combinations was 6.57 x 10^{22}. In creating this distribution, the frequency of each of the possible numerical score totals was calculated, taking into account the different weights assigned to the EDA channel and the score selection method applied to the RLL component. Because the numerical scores for each case were calculated using the septiles derived from the samples of known deceptive and nondeceptive cases, the two permutation tests provide an estimate of the likelihood that the observed outcome would occur if the case were drawn from the population of other deceptive cases, as well as an estimate of the likelihood that the outcome would occur if the suspect were nondeceptive. These permutation tests therefore allow estimation of $P(F \mid D)$ and $P(F \mid T)$, respectively.

Because physiological responses tend to be statistically noisey variables, the permutation tests were conducted under the assumption that the amount of dependency between responses to repeated questions is not substantial. To test this notion, the 9 R/C ratios available for each physiological response channel were correlated with one another. This was done across the entire sample of 331 cases.

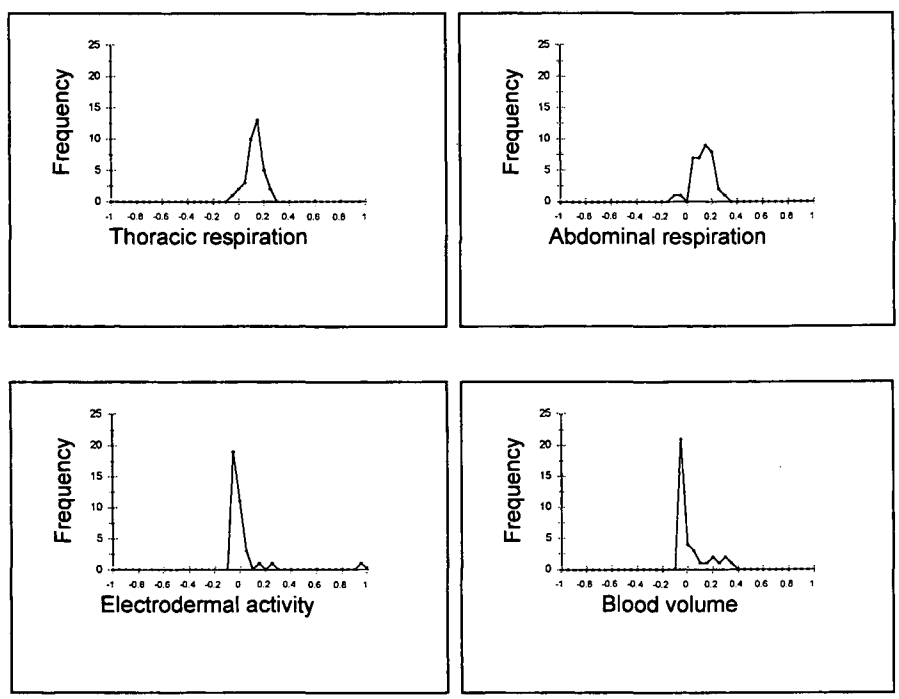


Figure 1.
Distributions of correlations between repeated questions for four response channels.

Figure 1 shows the distribution of correlations observed in the four physiological response channels. Although 66 of these correlations were statistically significant (assuming alpha = .05), there were a total of 144 correlations across all four response channels, and each correlation was made using all 331 cases. The fact that most of these correlations are small (i.e. less than r = .30) supports our hypothesis that the repeated observations were largely independent of one another. In addition, 55 of the significant correlations occurred in the two respiratory channels, and 22 of these were common to both. This reflects the fact that the thoracic and abdominal respiration channels may often provide redundant information.

Posterior probability estimation

Once *P(F | D)* and *P(F | T)* were calculated for each case, these values were entered into Murphy's (1987) formula, along with estimates of the base rates of deception and nondeception ranging from 10% through 90% in increments of 10. These calculations resulted in an estimate of the posterior probability of deception for each case, under each of the assumed base rates. The results are shown in Tables 4 and 5.

Table 4
Outcomes of Confirmed Deceptive Cases (N=181) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	6	7	7	7	7	8	10	12	14
10 to 19%	1	1	1	1	3	3	2	2	6
20 to 29%	0	0	2	3	1	2	1	4	1
30 to 39%	1	1	1	1	1	1	3	3	2
40 to 49%	0	2	1	1	1	3	3	0	1
50 to 59%	1	1	1	1	3	3	1	2	2
60 to 69%	2	2	1	3	3	1	2	1	3
70 to 79%	1	1	4	3	1	2	1	3	4
80 to 89%	2	6	3	2	3	2	4	6	9
over 90%	86	80	80	77	77	75	72	68	59

Table 5
Outcomes of Confirmed Nondeceptive Cases (N=181) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	73	77	79	82	83	85	89	93	96
10 to 19%	4	5	4	3	5	7	5	3	1
20 to 29%	3	2	3	5	5	2	3	0	0
30 to 39%	2	2	3	3	1	2	0	1	1
40 to 49%	1	3	3	2	2	0	1	0	1
50 to 59%	2	3	1	1	0	1	0	1	1
60 to 69%	3	1	2	0	1	0	0	1	0
70 to 79%	4	3	0	1	0	1	1	1	1
80 to 89%	3	1	1	1	1	1	1	1	0
over 90%	4	3	3	3	2	1	1	1	1

Classification

In this sample of 331 confirmed field cases, 150 (45.3%) were confirmed as being nondeceptive and 181 (54.6%) were known to be deceptive. In making classifications in this sample, we first assumed a base rate of deception that most closely reflects the characteristics of this sample, namely 50%. When the posterior likelihood of deception obtained in each case is compared against this criterion, we find that the polygraph information increased the estimated likelihood of deception above 50% in 156 of the 181 deceptive cases (86%). The polygraph

information reduced the estimated likelihood to below 50% in 144 of the 150 nondeceptive cases (96%). Using these simple criteria, one might conclude that the polygraph information could have contributed toward making a correct decision in approximately 91% of these cases.

In criminal investigations, more stringent criteria are required for making classifications. Because forensic evidence, including polygraph test results, can weigh heavily in judicial outcomes that have serious ramifications for the accused, the choice of conservative cutoff points would be justified in most cases. Additionally, assumption of a base rate other than 50% may be justifiable in many criminal cases. If one were to categorize the cases in our sample assuming a 50% base rate of deception and call deceptive the cases in which the posterior likelihood of deception is greater than 90%, the result would be that 139 of 152 (91%) deceptive cases would be correctly classified. Using the 50% base rate and a 10% posterior probability cutoff for determining nondeception would result in 125 of 128 (98%) innocent cases being correctly classified. Although 29 (16%) deceptive and 22 (15%) nondeceptive cases would receive indeterminate outcomes, those with conclusive results show reasonable levels of sensitivity and specificity when these criteria are applied.

Discussion

The accuracy of classifications made in this system depends on the criteria used in decision making. These criteria should not be chosen arbitrarily, but carefully considered and appropriate for the circumstances of a specific test. In many cases, a 50% base rate assumption might not be appropriate. One must assume a base rate of deception that is justified in a given application, as well as reasonable cutoff points for making outcome decisions. When proffering polygraph evidence to courts of law, it might be preferable not to provide dichotomous classifications for cases; rather a set of prior and posterior probabilities which the court would use at its discretion in deciding the outcome of the case. This would place the decision in the hands of the judge or jury, thus avoiding a tendency to usurp their proper role as trier of fact.

In circumstances where categorical classifications are required, the assumed base rate of deception chosen by an examiner must depend upon the circumstances of the case. To the extent that extra-physiological information suggests that the examinee is more or less likely to be guilty of the crime, the assumed base rate might be adjusted accordingly. Similarly, the administrators of law enforcement agencies that use the polygraph might recommend rates of deception believed to occur on a program-wide basis.

The choice of cutoff points must also be chosen carefully, with due consideration of the risk of error. More or less stringent criteria for assigning an outcome of deceptive, nondeceptive, or inconclusive must be chosen to reflect the possible consequences of false positive or false negative error. In a majority of situations in which the polygraph is deployed, the ramifications of the test outcome are so important that conservative cutoff points should be used. In the present sample, the use of conservative cutoffs resulted in many inconclusive results, but low rates of error among the cases with conclusive outcomes.

The rates of correct classification in the first sample of confirmed field cases are encouraging. However, replication of these results is a necessity if decisions regarding the admissibility of polygraph evidence or administrative policies are to be made in consideration of these findings. No such decisions would be justifiable on the basis of a single sample of data, particularly when there is reason to believe that the sample might be biased in some way. In the present case, the use of panel decisions to determine nondeceptive status could have exaggerated the specificity estimate by including only cases having physiological records that appeared nondeceptive to members of the panels. Also, the use of confessions in determining deceptive status could have raised the apparent sensitivity of the test, becasue efforts to elicit the confessions may have been initiated by examiners whenever suspects' physiological records appeared deceptive. These two potential sources of bias could have distorted the results in the first sample, so a second sample of confirmed field data was collected and subjected to the same quantification procedures.

Field Cross Validation

CID Field cases

A sample of confirmed field cases was collected from the United States Army Criminal Investigation Detachment (CID). The CID Polygraph Division is staffed by approximately 20 field examiners at any given time, and two quality control supervisors. All are federally trained and certified, and meet continuing education requirements. The CID polygraph division maintains uniform procedures, high standards, and multiple levels of quality control that make its examinations among the best in the federal services.

All of CID's polygraph cases from January, 1995 through March, 1998 were reviewed along with the investigative files for those cases, which are maintained separately from the polygraph division files. The dates on which the selected tests were conducted spanned the period from January, 1995 through February, 1997. During this period, 3,349 polygraph examinations were conducted on criminal cases. All selected tests were conducted on criminal suspects. The types of crimes under investigation included larceny, possession or distribution of drugs, sexual assault, child abuse, hate crimes, fraud, assault, damaging property, arson, false accusation, and receiving stolen property.

Ground truth confirmation was decided on the basis of the following: an unrecanted confession of the examinee, an unrecanted confession from someone which exculpated the examinee, evidence that the crime under investigation was never committed (e.g. when missing property was discovered to have been innocently misplaced instead of stolen), or forensic evidence such as urinalysis or surveillance tapes. Eyewitness testimony, prosecutorial decisions, or judicial outcomes alone were not considered sufficient to establish ground truth.

Seventeen of the selected cases were confirmed as nondeceptive. All of these cases were confirmed by the confession of someone other than the examinee. Eleven of these confessions were corroborated with physical evidence.

A total of 215 cases were confirmed as being deceptive. Of these, 125 were confirmed solely by the examinee's confession. Because confessions may be elicited as part of a polygraph examination post-test interview, we consider these confessions not to be independent of the polygraph test results.

Fifteen deceptive cases were confirmed independently of the polygraph test, with corroborating physical evidence in 5 cases. The other 10 were confirmed by confession, but the confession occurred following a second polygraph test. We consider these confessions to be independently confirmed because the confession is known to have been separate from the first polygraph test. In such cases, only data from the first test were included in the sample.

A third group of deceptive cases consisted of 75 people who had failed urinalysis tests for drug offenses and requested a polygraph test in an attempt to garner exculpatory evidence. These people would have undergone considerable investigation prior to their polygraph tests, including an interview with a field investigator and a search for physical evidence. The pre-polygraph investigations present many opportunities for exculpatory evidence to be found. Because polygraph tests are typically deployed at the later stages of such investigations, it is likely that most cases proceeding to a polygraph examination involve suspects who are actually guilty. However, it is not possible to know the extent to which deceptive polygraph results in such cases are attributable to the efficacy of the test in detecting guilt, or to a possible effect of examiners' awareness of strong evidence suggesting guilt. We therefore must consider the polygraph results in these cases not to have been independent of the confirmatory evidence.

The final sample of confirmed CID cases consisted of four subgroups: independently confirmed nondeceptive cases (N=17), independently confirmed deceptive cases (N=15), deceptive cases verified by post-polygraph confession (N=125), and deceptive cases confirmed by pre-polygraph urinalysis tests (N=75).

Results

All cases were assessed using the scoring procedures described above and the values from Tables 1, 2 and 3. The results are shown in Tables 6 through 9.

If we assume a 50% base rate of deception in this sample and adopt conservative cutoff points (i.e. 10% and 90%) for deciding test outcomes based on the posterior likelihood estimates, we find that 14 of 16 (88%) independently verified nondeceptive cases were correctly classified. Eighty-seven of 97 (90%) confession verified cases were correctly identified. Forty-five of 49 (92%) urinalysis cases were correctly identified. Among the 11 independently confirmed deceptive cases with conclusive outcomes, 7 (64%) were correct.

Table 6
Outcomes of Confession Verified Cases (N=125) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	4	4	5	8	8	9	12	18	21
10 to 19%	0	2	3	2	4	6	7	3	2
20 to 29%	1	2	2	4	6	5	2	2	3
30 to 39%	3	2	3	4	2	2	1	1	2
40 to 49%	0	2	4	2	1	0	2	4	2
50 to 59%	2	2	2	1	0	2	3	2	1
60 to 69%	2	5	2	1	2	3	2	2	2
70 to 79%	6	2	2	2	4	2	2	1	2
80 to 89%	3	2	4	6	3	2	2	3	4
over 90%	79	77	74	71	70	69	67	66	62

Table 7
Outcomes of Urinalysis Verified Deceptive Cases (N=75) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	5	5	5	5	5	8	11	17	21
10 to 19%	0	0	0	3	4	7	9	4	12
20 to 29%	0	0	3	4	7	7	1	7	1
30 to 39%	0	3	4	5	5	0	7	7	4
40 to 49%	0	3	5	4	0	7	5	0	1
50 to 59%	3	4	4	0	7	5	1	4	0
60 to 69%	4	5	0	7	5	1	4	1	3
70 to 79%	5	1	7	7	1	4	1	1	3
80 to 89%	4	12	7	4	5	1	3	4	3
over 90%	79	67	65	61	60	60	57	55	52

Table 8
Outcomes of Independently Verified Deceptive Cases (N=15) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	7	7	13	13	27	27	33	40	40
10 to 19%	0	7	13	13	7	7	7	0	7
20 to 29%	7	13	0	7	0	7	0	7	0
30 to 39%	0	0	7	7	7	0	7	0	7
40 to 49%	13	7	7	0	0	7	0	0	0
50 to 59%	0	0	0	0	7	0	0	7	7
60 to 69%	7	7	0	7	0	0	7	0	0
70 to 79%	7	0	7	0	0	7	0	7	7
80 to 89%	0	7	0	7	7	7	7	7	0
over 90%	60	53	53	47	47	40	40	33	33

Table 9
Outcomes of Independently Verified Nondeceptive Cases (N=17) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	71	82	82	82	82	82	88	88	88
10 to 19%	12	0	0	0	6	6	0	0	0
20 to 29%	0	0	0	6	0	0	0	0	0
30 to 39%	0	0	6	0	0	0	0	0	0
40 to 49%	0	6	0	0	0	0	0	0	0
50 to 59%	0	0	0	0	0	0	0	0	6
60 to 69%	6	0	0	0	0	0	0	0	0
70 to 79%	0	0	0	0	0	0	0	6	0
80 to 89%	0	0	0	0	0	6	6	0	0
over 90%	12	12	12	12	12	6	6	6	6

Discussion

Using the present criteria, 92% of urinalysis cases and 90% of confession verified cases were correctly identified as deceptive, but only 64% of independently confirmed deceptive cases were correctly classified. Although the present sample of independently confirmed deceptive cases is the largest one in existence, it would be imprudent to make unqualified generalizations on the basis of 15 cases. Despite the apparent differences in hit rates among these groups, it is difficult to elaborate on the source of these differences. One possibility is that the small number of independently confirmed cases produced a sensitivity estimate that is inaccurate because of sampling error. That sample might also have been biased by the inclusion of cases that required a subsequent polygraph test to clear equivocal results on the earlier test.

However, another possibility is that sampling bias may have exaggerated the hit rate found in the confession verified group, and that examiner bias may have indirectly adulterated the results in the urinalysis sample. Each of these mechanisms may have been at play in some cases.

The sensitivity estimate from the sample of 17 independently confirmed nondeceptive cases was 88%. We do not know the extent to which sampling error, sampling bias, or examiner bias may have affected these results, but we do know that the examinees in all of the groups were all criminal suspects and there was sufficient evidence against them to warrant a polygraph exam. Therefore, it would seem that the correctly classified nondeceptive cases were so identified despite the possible availability of inculpatory evidence. This is not consistent with the notion that examiner bias is sufficient to result in positive polygraph test results in nondeceptive suspects.

With no satisfactory way of establishing ground truth in each case, the potential effects of contamination and sampling bias can not be estimated adequately in field studies of polygraph test validity. The available samples of independently verified cases are so small that the results should be viewed with skepticism. The findings may indeed represent accurate estimates of CQT accuracy when the present scoring methods are applied to field data, but they might also be anomalous results obtained using desperately small samples. To provide some convergence of evidence, a third study was done using laboratory data. Although laboratory simulation data have the disadvantage of poor generalizability, ground truth may be securely established for each case in the sample. A concurrent examination of both field and laboratory data is necessary for a comprehensive estimate of accuracy to emerge.

Laboratory Cross Validation

Laboratory Cases

A sample of 76 laboratory cases was drawn from three laboratory simulation studies in which the DoDPI Zone Comparison version of the CQT was used. Participants in one study (DoDPI Research Division Staff, 2000) were instructed either to steal a diamond ring from an office (N=16) or to wait in the lobby of the building for 15 minutes. The innocent participants (N=16) were told that the ring had been stolen, but they were innocent of the crime. Mock criminals and innocent participants were each given a CQT examination by an experimentally blind qualified examiner using the same recording equipment and test format as in the two field samples described above. Participants were recruited from a temporary employment agency and offered a monetary bonus for being found nondeceptive on the polygraph test.

In the second sample (Pollina, Pavlidis, Levine, & Ryan, 2001), mock criminals were instructed to murder a mannequin by stabbing her with a screwdriver and steal $20.00 from her purse (N=12). Mock criminals and innocent participants (N=12) were each given a CQT examination by an experimentally blind qualified examiner using the same recording equipment and test format as in the two field samples.

In the third sample (Bradley, Cullen, & Carle, 1993), student volunteers wrote a brief description of a traumatic event from their past. The students were rewarded with research participation bonus points toward their introductory psychology course, plus a monetary award for being found nondeceptive on the polygraph test. Deceptive participants (N=10) were given a Zone Comparison CQT to determine whether they were the author of the embarrassing story. Those in the nondeceptive group (N=10) each read a story written by a participant in the deceptive group and were given an identical test. The tests were administered by experienced polygraph examiners from a local police department using a non-computerized field polygraph (Lafayette Instruments, Lafayette, IN). The paper charts were numbered, removed from their associated files, shuffled and given to the first author to be manually scored without his knowing the status of each case. Electrodermal and blood pressure increases were measured from a 1 second pre-stimulus baseline to the maximal amplitude observed within 10 seconds of question onset. Respiration line length was measured for 10 seconds following onset using a mechanical planimeter. Although both thoracic and abdominal respiration channels were recorded, many of the records were of poor quality, so only the more interpretable of the two respiration channels was scored for each case. Although funded by DoDPI, this research was conducted by independent investigators at the University of New Brunswick.

Results

Determinations were made using the values from Tables 1, 2, and 3, an assumed 50% base rate of deception, and the 90% and 10% posterior probability points as cutoffs for classifications of deception or nondeception, respectively. The results are shown in Tables 10 and 11.

Table 10
Outcomes of Confirmed Deceptive Laboratory Cases (N=38) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	5	8	11	11	11	11	11	11	16
10 to 19%	3	3	0	0	0	0	0	5	0
20 to 29%	3	0	0	0	0	3	5	0	3
30 to 39%	0	0	0	0	5	3	0	3	3
40 to 49%	0	0	0	5	0	0	0	0	3
50 to 59%	0	0	3	0	0	0	3	3	0
60 to 69%	0	0	3	0	0	3	3	3	0
70 to 79%	0	5	0	3	3	3	3	0	3
80 to 89%	5	0	3	3	5	3	0	3	3
over 90%	84	84	82	79	76	76	76	74	71

Table 11
Outcomes of Nondeceptive Laboratory Cases (N=38) with Various Assumed Base Rates.

Posterior Likelihood of Deception	Prior Probability of Deception								
	90	80	70	60	50	40	30	20	10
0 to 9%	45	45	50	50	58	63	66	68	74
10 to 19%	0	5	8	16	8	5	5	5	8
20 to 29%	5	8	8	0	3	3	3	5	3
30 to 39%	0	8	0	3	3	3	3	5	5
40 to 49%	8	0	3	3	3	3	5	0	3
50 to 59%	8	3	3	3	3	5	3	5	0
60 to 69%	3	3	5	8	3	5	3	3	0
70 to 79%	3	3	5	8	3	5	3	3	0
80 to 89%	5	8	5	5	8	3	3	0	0
over 90%	26	18	16	11	8	8	5	5	5

Using these criteria, 29 of 33 (88%) deceptive participants and 22 of 25 (88%) nondeceptive participants were correctly identified. Although the rates of inclusive outcomes (13% and 34% respectively) were somewhat higher than in the field samples, the error rates appear similar to those observed in the field samples.

Discussion

There is an apparent convergence toward consensus in the results reported here. In the field data used to develop the scoring protocol, 91% of deceptive and 98% of nondeceptive cases with conclusive outcomes were correct classifications. In the CID field cases used to cross validate the system, 90%

of confession verified cases and 92% of urinalysis verified cases were correct, although only 64% of independently verified deceptive cases were correctly classified. Of the independently verified nondeceptive CID cases, 88% were true negatives. When cross validated using laboratory data, 88% of both deceptive and nondeceptive cases were correct calls. Across all of these samples, 307 of 342 (90%) deceptive cases with conclusive outcomes were true positives and 161 of 169 (95%) nondeceptive cases with conclusive outcomes were true negatives. On the basis of these findings, we conclude that the present scoring system results in correct classifications approximately 90% of the time.

This system appears to be fairly robust in the face of some possible sources of bias. First, many of the cases reported here were verified on the basis of confessions made after the polygraph tests. It is possible that confession inducement attempts may have been differentially associated with cases in which physiological responses suggestive of deception were present. This is not a concern in laboratory simulations, so we can use the laboratory data as a benchmark against which to compare the field data. The 88% hit rate found in the laboratory simulations is strikingly similar to the 91% hit rate in the original field sample and the 90% detection rate in the CID field sample. It would therefore seem that this kind of bias did not have a substantial impact on the present findings. It might be a factor in some individual cases, but it does not appear to be a problem sufficient to distort the results by more than a few percentage points across a large sample. Second, it is possible that evidence strongly suggestive of guilt could sway examiners' conduct in an examination in such a way as to inflate the true positive rate. However, if we contrast the 92% hit rate observed in the urinalysis verified sample against the laboratory sample, it again becomes apparent that this source of bias may be a minor one. Third, the fact that almost all of these results are clustered around the 90% point suggests that even the major differences between field tests and those conducted under controlled laboratory conditions are not sufficient to greatly affect the outcomes. Further, the field tests were conducted by a variety of agencies and with suspects in many different types of crimes. Likewise, the laboratory data were collected by three different groups of researchers using both computerized and traditional equipment. The fact that consistent results are obtained in all of these conditions suggests that the present system may be able to accommodate disparate conditions, some of which may be less than ideal.

The one finding that stands out among these samples is the relatively poor hit rate in the independently confirmed deceptive group. The 64% hit rate is markedly lower than the others, but this result should be interpreted carefully. Only 11 of the 15 cases had conclusive outcomes and it is quite possible that this anomalous result could be a reflection of this very small sample size. Still, there is a possibility that this may be the true sensitivity of the scoring system and the other estimates could be in error. Until the present results are re-tested experimentally and replicated, we can not rule out this possibility. However, it would seem most parsimonious to tentatively accept a 91% hit rate in 331 cases, rather than a 64% rate in 11 cases.

The scoring system described here was designed to circumvent several problems in CQT polygraphy. The use of ratios was intended to make the system platform independent, in the sense that a variety of physiological recording systems could be used in conjunction with the scoring method. Neither the computerized systems nor the analog system used here are ideal machines for recording physiological information, but they were sufficient to allow the scoring protocol to reliably classify individual cases accurately. A more pressing concern is the questionable basis of relevant / comparison question contrasts upon which the CQT traditionally depends. The present system uses ratios that are compared against distributions observed in samples of known deceptive and nondeceptive cases, so the need for direct within subject contrasts between responses to relevant and comparison questions is avoided. Using permutation tests to arrive at discrete estimates of the probabilities of deception and nondeception, the results of individual tests may be presented as a posteriori likelihoods that are revisions of assumed a priori base rates. As a

package, this system circumvents problems associated with existing approaches to CQT polygraphy.

The logical advantages of the scoring system would be meaningless if it were unable to produce accurate classifications. This system has been tested on a large assortment of both field and laboratory cases and found to provide accurate assessments approximately 90% of the time. We recommend this system for its logical basis and for its empirically tested ability to provide accurate classifications. For practitioners, it can serve as a robust and empirically validated classification system. For researchers, it may provide a standard format for reporting the results of future field validity studies and laboratory research into the factors that affect validity. It should be remembered, however, that inconclusive and erroneous results do occur, so results of individual tests should be interpreted cautiously in the course of criminal investigations.

References

Backster, C. (1969). *Technique Fundamentals of the Trizone Polygraph Test.* New York: Backster Research Foundation, Inc.

Ben-Shakhar, G. & Furedy, J.J. (1990). *Theories and Applications in the Detection of Deception: A Psychophysiological and International Perspective.* New York: Springer Verlag.

Bradley, M.T., Cullen, M.C. & Carle S.B. (1993). *Control Question Tests by Police and Laboratory Polygraph Operators on a Mock Crime and Real Events (DoDPI93-R-0012).* Fort Jackson, SC: Department of Defense Polygraph Institute.

Dionisio, D., Granholm, E., Hillix, W.A. & Perrine, W.F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology, 38,* 205-211.

DoDPI Research Division Staff (2000). *Test of a Mock Theft Scenario for use in the Psychophysiological Detection of Deception: IV (DoDPI00-R-0002).* Fort Jackson, SC: Department of Defense Polygraph Institute.

Dollins, A. B., Krapohl, D. J., & Dutton, D. W. (2000) Computer algorithm comparison. *Polygraph, 29*(3), 237-247.

Furedy, J.J., Davis, C. & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology, 25,* 683-688.

Furedy, J.J., Gigliotti, F. & Ben-Shakhar, G. (1994). Electrodermal differentiation of deception: The effect of choice versus no choice of deceptive items. *International Journal of Psychophysiology, 18,* 13-22.

Furedy, J.J., Posner, R.T. & Vincent, A. (1991). Electrodermal differentiation of deception: Perceived accuracy and perceived memorial content manipulations. *International Journal of Psychophysiology, 11,* 91-97.

Godert, H.W., Rill, H.G. & Vossel, G. (2001). Psychophysiological differentiation of deception: The effects of electrodermal lability and mode of responding on skin conductance and heart rate. *International Journal of Psychophysiology, 40,* 61-75.

Honts, C.R. (1996). Criterion development and validity of the control question test in field application. *The Journal of General Psychology, 123,* 309-324.

Honts, C.R. & Raskin, D.C. (1988). A field study of the validity of the directed lie control question.

Journal of Police Science and Administration, 16, 56-61.

Horowitz, S.W., Kircher, J.C., Honts, C.R. & Raskin, D.C. (1997). The role of comparison questions in physiological detection of deception. *Psychophysiology, 34*, 108-115.

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

Iacono, W.G. (2000). The Detection of Deception. In J.T. Cacioppo, L.G. Tassinary & G.G. Berntson (eds.) *Handbook of Psychophysiology, 2nd ed.* Cambridge University Pres.

Iacono, W.G. & Lykken, D.T. (1999). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology, 82*, 426-433.

Jones, E.E. & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin, 76*, 349-364.

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

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

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

Kleinmuntz, B. & Szucko, J.J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature, 308*, 449-450.

Lykken, D.T. (1998). *A Tremor in the Blood: Uses and Abuses of the Lie Detector.* New York: McGraw-Hill.

Matte, J.A. (1996). *Forensic Psychophysiology using the Polygraph.* New York: JAM Publications.

Meehl, P.E. & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Murphy, K.R. (1987). Detecting infrequent deception. *Journal of Applied Psychology, 72*, 611-614.

National Academy of Sciences (2002). *The Polygraph and Lie Detection.* Washington, DC: national Academies Press. Available online at *http://www.nap.edu/books/0309084369/html/*

Patrick, C.J. & Iacono, W.G. (1991). Validity of the control question polygraph test: The problem of sampling bias. *Journal of Applied Psychology, 76*, 229-238.

Pollina, D.A., Pavlidis, I., Levine, J. & Ryan, A. (2001). *The Relationship Between Facial Skin Surface Temperature Reactivity and Traditional Polygraph Measures Used in the Psychophysiological Detection of Deception: A Preliminary Investigation (DoDPI01-R-0007).* Fort Jackson, SC: Department of Defense Polygraph Institute.

Raskin, D.C., Honts, C.R. & Kircher, J.C. (1997). The scientific status of research on polygraph techniques: The case for polygraph tests. In D.L. Faigman, D. Kaye, M.J. Saks & J. Sanders (eds.) *Modern Scientific Evidence: The Law and Science of Expert Testimony.* St. Paul, MN: West.

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

Timm, H.W. (1982). Analyzing deception from respiration patterns. *Journal of Police Science and Administration, 10*(1), 47-51.

Vincent, A. & Furedy, J.J. (1992). Electrodermal differentiation of deception: Potentially confounding and influencing factors. *International Journal of Psychophysiology, 13*, 129-136.