REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
June 1996

3. REPORT TYPE AND DATES COVERED
Final Report (May 95 - Jun 96)

4. TITLE AND SUBTITLE
POLYSCORE: A Comparison of Accuracy

5. FUNDING NUMBERS
DoDPI94-P-0006

6. AUTHOR(S)
N. Joan Blackwell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Defense
Polygraph Institute
Building 3165
Fort McClellan, AL 36205-5114

8. PERFORMING ORGANIZATION REPORT NUMBER
DoDPI95-R-0001

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of Defense
Polygraph Institute
Building 3165
Fort McClellan, AL 36205-5114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DoDPI95-R-0001
DoDPI94-P-0006

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Public release, distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Using data collected under a mock crime scenario paradigm, four versions of the John Hopkins University Applied Physics Laboratory (APL) algorithm-based scoring system were evaluated for consistency in scoring accuracy. The four versions were: (a) PASS 2.0, (b) POLYSCORE 2.3, (c) POLYSCORE 2.9, and (d) POLYSCORE 3.0. The algorithm's rates of agreement/disagreement with ground truth were examined, and the same evaluations were made for the psychophysiological detection of deception (PDD) examiners who collected the data. The PDD examiners in this evaluation had an overall accuracy rate of 72.27% when compared to ground truth. The overall rate of accuracy generated by the algorithm (edited dataset) was: (a) PASS 2.0, 63.03%; (b) POLYSCORE 2.3, 67.72%; (c) POLYSCORE 2.9, 72.27%; and (d) POLYSCORE 3.0, 68.91%. With the inconclusive decisions eliminated, the recomputed accuracy rate for the PDD examiners was 79.63%, while each version of the algorithm was comparable (PASS 2.0, [78.95%]; POLYSCORE 2.3, [79.21%]; POLYSCORE 2.9 [83.50%]); POLYSCORE 3.0, [82.83%], with both POLYSCORE 2.9 and POLYSCORE 3.0 exceeding the examiners' level of accuracy. In addition to overall accuracy and accuracy based on the test format used, the effects of subjective manipulation of the data were discussed, and information was provided on the occurrence of decision reversals and statistical outliers.

14. SUBJECT TERMS
POLYSCORE, Axciton, computerized scoring algorithms, Zone Comparison Test (ZCT), polygraph, forensic psychophysiology, psychophysiological detection of deception (PDD)

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Polyscore: A Comparison of Accuracy

N. Joan Blackwell, M.S.

June 1996

Department of Defense Polygraph Institute
Fort McClellan, Alabama 36205-5114
Telephone: 205-848-3803
FAX: 205-848-5332

Report No. DODPI95-R-0001

POLYSCORE: A Comparison of Accuracy

N. Joan Blackwell, M.S.

June 1996

Department of Defense Polygraph Institute
Fort McClellan, Alabama 36205

Director's Foreword

As the discipline of Forensic Psychophysiology evolves, computer hardware and software have become increasingly important to the administration and evaluation of psychophysiological detection of deception (PDD) examinations. Such automation decreases the physical and mental demands placed on examiners by reducing the effort required to operate the instrument. This allows the examiners to concentrate their efforts on the interview and interrogation portions of the examination. While automation can increase examiner efficiency there are potential penalties. As PDD examinations become more automated the examiner surrenders some control over the examination to the hardware and software manufacturers. Those practicing Forensic Psychophysiology must remain vigilant to ensure that the hardware and software used accurately record and evaluate physiologic responses.

This report describes the results obtained when the same data were evaluated by examiners and four versions of a computer program designed to assess physiologic responses recorded during PDD examinations. It is essential that such comparative studies be completed and reported to validate our increasing reliance on computer software. It should be noted that the reported comparisons were made using data collected following a laboratory mock-crime while the computer program was designed using the results of actual criminal examinations. This difference may have influenced the overall accuracy rates if there are intrinsic differences between data collected following actual and mock crimes.

Michael H. Capps
Director

Acknowledgments

The author wishes to express appreciation to the following organizations for providing manpower and resources in support of this research: the Department of Defense Polygraph Institute (DoDPI); the Air Force Office of Special Investigations, (AFOSI); the Naval Investigative Service (NIS); the Defense Investigative Service, (DIS). Special thanks are extended to the individuals who performed tasks associated with project coordination and execution.

This study was supported by funds from the DODPI as project DoDPI94-P-0006. The views expressed in this report are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Abstract

BLACKWELL, N. J. POLYSCORE: A comparison of accuracy. June 1996, Report No. DoDPI95-R-0001. Department of Defense Polygraph Institute, Ft. McClellan, AL 36205.--Using data collected under a mock crime scenario paradigm, four versions of the John Hopkins University Applied Physics Laboratory (APL) algorithm-based scoring system were evaluated for consistency in scoring accuracy. The four versions were: (a) PASS 2.0, (b) POLYSCORE 2.3, (c) POLYSCORE 2.9, and (d) POLYSCORE 3.0. The algorithm's rates of agreement/disagreement with ground truth were examined, and the same evaluations were made for the psychophysiological detection of deception (PDD) examiners who collected the data. The PDD examiners in this evaluation had an overall accuracy rate of 72.27% when compared to ground truth. The overall rate of accuracy generated by the algorithm (edited dataset) was: (a) PASS 2.0, 63.03%; (b) POLYSCORE 2.3, 67.72%; (c) POLYSCORE 2.9, 72.27%; and (d) POLYSCORE 3.0, 68.91%. With the inconclusive decisions eliminated, the recomputed accuracy rate for the PDD examiners was 79.63%, while each version of the algorithm was comparable (PASS 2.0, [78.95%]; POLYSCORE 2.3, [79.21%]; POLYSCORE 2.9, [83.50%]; POLYSCORE 3.0, [82.83%]), with both POLYSCORE 2.9 and POLYSCORE 3.0 exceeding the examiners' level of accuracy. In addition to overall accuracy and accuracy based on the test format used, the effects of subjective manipulation of the data were discussed, and information was provided on the occurrence of decision reversals and statistical outliers.

Key-words: POLYSCORE, Axciton, computerized scoring algorithms, Zone Comparison Test (ZCT), polygraph, forensic psychophysiology, psychophysiological detection of deception (PDD).

Table of Contents

List of Figures

List of Tables

A diagnostic technique which relies on human interpretation of test data is immediately suspect (Nunnally, 1978). Rater bias, inexperience and even incompetence are problems that plague any field in which humans are asked to make interpretive judgments. For more than fifty years, the data resulting from psychophysiological detection of deception (PDD) examinations has essentially relied on human interpretation. Accordingly, much of the scientific community considers such data suspect. That contention, along with the ever present need to accurately decipher the complex physiological tracings generated during a PDD examination, are the driving forces behind the development of automated, algorithm-based scoring systems. In 1993, the Department of Defense Polygraph Institute (DoDPI) conducted a full-scale study to evaluate one such system (Blackwell, 1994). In this paper, the data from that original study have been analyzed using three subsequent upgrades of that system in order to determine what effect the ensuing revisions and refinements have had on the overall level of algorithm scoring accuracy.

POLYSCORE (formerly known as the Polygraph Automated Scoring System [PASS]) is one of the most recent ventures designed to eliminate subjectivity from the process of interpreting PDD examinations. Categorized as a personal computer software package, POLYSCORE implements a scoring algorithm developed by the Johns Hopkins University Applied Physics Laboratory (APL) under contract to the National Security Agency (NSA).

The algorithm uses a logistic regression model, and during processing the data is detrended, mathematically filtered, and then standardized. POLYSCORE currently works in conjunction with the Axciton Computerized Polygraph (Axciton Systems, Incorporated, Houston, TX), and the Lafayette Computerized Polygraph (Lafayette Instrument Company, Lafayette, IN). Both are stand alone PDD systems which record the physiological data (i.e., respiration, electrodermal and cardiovascular) collected during a PDD examination. POLYSCORE then, in turn, uses that physiological data to produce an overall "probability of deception" for the examination (Polygraph Automated Scoring System User's Guide, Version 2.0, 1993a).

The test scoring criterion currently taught at the DoDPI, along with other APL selected criterion, were used as a starting point in the development of the POLYSCORE algorithm. A list of approximately 1500 analysis "features" was generated by creating combinations of the scoring criteria along with varying the sampling interval (i.e., the number of seconds in the scoring window). Systematically, the list was distilled to include only those features, or criteria which contributed to the highest level of accuracy when used in the evaluation of PDD examinations (Capps, 1993). As a result, expanded scoring windows were established for each of the PDD components, and the cardiovascular channel was split into a pulse channel and a blood volume rate of change (derivative) channel (Fig. 1). (Note: The respiration response interval was expanded from 16 to 18 seconds in subsequent versions of the algorithm.) The physiological signals were also assigned scoring weights as shown in Table 1.

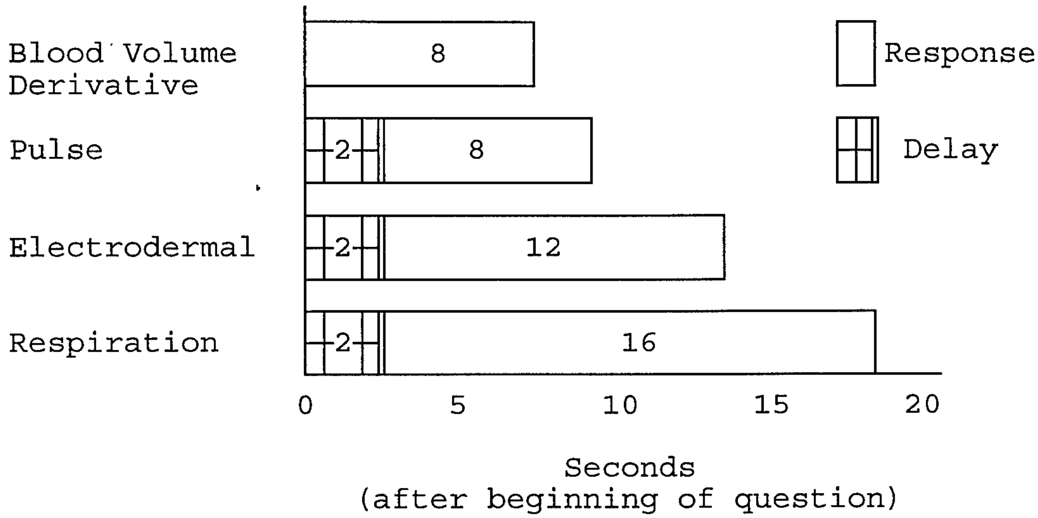

Figure 1. Response intervals, or scoring windows for the four PDD channels generated by Polygraph Automated Scoring System, Version 2.0. From "Polygraph Automated Scoring System, Version 2.0" by M. Capps, 1993. Adapted with permission.

Table 1
PASS Signal Scoring Weights

Channel	%
Blood Volume Derivative	21
Pulse	14
Electrodermal	49
Respiration	16
Total	100

Note. From "Polygraph Automated Scoring System, Version 2.0" by M. Capps, 1993. Adapted with permission. PASS = Polygraph Automated Scoring System.

Having been unable to produce acceptable results using mock crime data during the early stages of the algorithm's development, APL later decided to use "live" PDD examinations (i.e., field cases) collected by various law-enforcement agencies. Field cases were used, not only to develop the algorithm, but also to later assess its accuracy. Use of field cases rather than the laboratory-generated mock crime data presented a distinct problem, however; the ground truth information necessary for accuracy assessments was not readily available in the field cases.

As a result, a two-component guideline was established for algorithm development. The first component facilitated the use of cases which had been resolved, either through the confession of the examinee or someone else. The second component allowed for the inclusion of cases which, when evaluated, had been assigned the same decision by the original examiner and two other examiners appointed to blind score the tests (Capps, 1993). Therefore examinations judged either deception indicated (DI), no deception indicated (NDI), or inconclusive (INC) were incorporated into the algorithm, as long as all three examiners arrived at the same decision.

The algorithm's level of accuracy was initially defined as POLYSCORE's rate of agreement with the combined decisions from both resolved cases and cases evaluated by the three examiners. Subsequently, as information regarding case resolution filtered in from the field examiners, the rate of agreement with confirmed ground truth was continually reexamined.

Of the 374 cases, or subjects used to develop the prototype (PASS 2.0), the probability generated by the algorithm supported ground truth (actual case resolution or the decision of the three examiners) on 93.3% of them, disagreed on 0.5%, and resulted in 6.2% of the cases erroneously being labeled INC. That is, using one of the two methods for determining ground truth, the developers labeled each case either DI, NDI or INC and the algorithm agreed with the respective decision on 349 cases and disagreed twice. On the remaining 23 cases (all deemed to be either DI or NDI), a decision probability of INC was generated.

When APL eliminated those 23 cases from the analysis (as would be done in PDD field accuracy reporting) the prototype's rate of agreement with ground truth was 99.4%, and the rate of disagreement was 0.6% (Capps, 1993). Further analysis of the figures indicated that the algorithm was 100% effective in clearing innocent individuals and 98.8% effective in detecting guilty individuals.

This high level of accuracy sparked great excitement in the PDD field--primarily because, for the first time, there was reportedly an objective tool available to examiners which would consistently enable them to provide an accurate decision for their PDD examinations.

The same immediate and widespread acceptance of the prototype, PASS 2.0, did not occur in the research community, however. This was due, in part, to the concern that the sample of physiological data used to establish the algorithm was seriously biased by the methods used when selecting cases for inclusion in the database. For example, only 374 of the available 750 cases were selected for use--and initially, the decision for only 60 cases of those had been confirmed. In a letter (dated 8 February 1993) to Mr. Ray Pollari, then Acting Deputy Assistant Secretary of Defense (CI&SCM), Dr. William J. Yankee, Director, DoDPI expressed a number of reservations regarding the claims that the algorithm was an objective, reliable, and valid PDD tool. As a result, a full-scale study was conducted at DoDPI to determine the effectiveness of PASS 2.0 in discriminating programmed innocent and programmed guilty participants in a controlled laboratory setting using a mock crime scenario (Blackwell, 1994).

Though there are inherent difficulties with generalizing the results of mock crime cases to those cases collected under field conditions, the effectiveness of the algorithm-based scoring system, as judged by its accuracy on the study data, was far below the expected level (62.5% overall; 79.0% overall with INCs eliminated). A recommendation made by the author of the report called for that same data to be analyzed using subsequent, versions of the algorithm. By reexamining the original data any change in accuracy occurring as a result of the various system upgrades would be easily discernible. This paper reports the results of those comparisons (using PASS 2.0 and POLYSCORE 2.3, 2.9, and 3.0), but first provides an overview of the procedures used during the original data collection effort.

Methods Overview (original study)

Data collection during the original research effort required 24 days during a consecutive 5-week period and occurred on-site at the Department of Defense Polygraph Institute (DoDPI), Fort McClellan, AL. Personnel involved in the data collection process included two PDD examiners, one subject handler, one role player designated as the "deliberate intruder," and one scenario setter. All examinations conducted during this project were administered in standard configuration PDD suites maintained by DoDPI and were videotaped using wall and ceiling mounted video cameras and commercial videotape recorders.

Research Design

The study used a Zone Comparison Test (ZCT) format, and employed a counterbalanced 2 X 2 design as depicted in Figure 2. Subjects were programmed as innocent or guilty in a mock crime scenario involving the theft of $124.00. Two types of control question tests (CQTs) were used during the study: (a) an experimental version of the directed lie control (DLC), and (b) the conventional probable lie control (PLC) currently in use throughout the PDD community. (Note: The question list for both CQTs was identical--only the pretest procedures were different.)

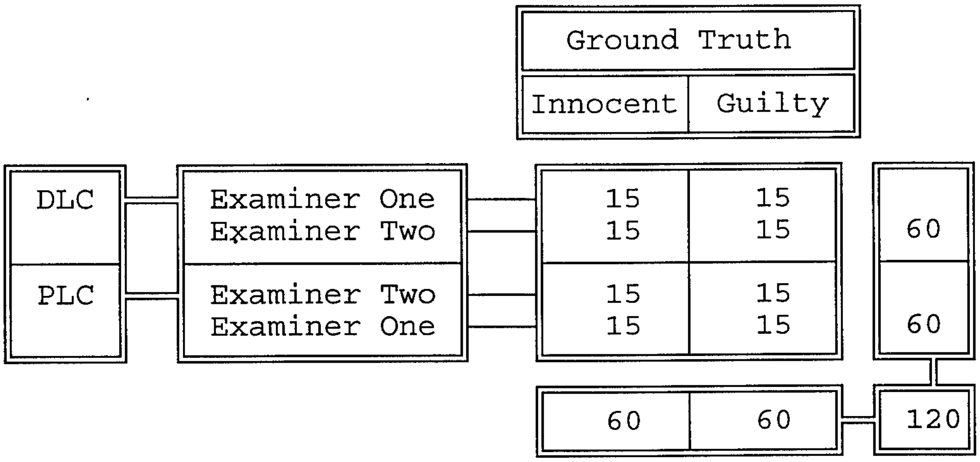

Figure 2. Diagram showing experimental design. DLC = directed lie control - experimental version; PLC = probable lie control.

Subjects

Data from a total of 120 male (42.5%) and female (57.5%) subjects were used in the original study. All were civilians from the local community and were provided by an employment agency contracted for the recruitment of subjects. None had undergone a PDD examination prior to the study. The predominantly white (63.9%) group ranged in age from 19 to 60, and the majority of them (75.4%) had at least the equivalent of a high school education. Based on self report, the individuals were in good to excellent health (93.3%), and the majority were well rested, having had six or more hours of sleep the previous night (78.4%). As a group, they reported experiencing little pain or discomfort (98.4%), and a relatively small percentage (9.2%) indicated the use of medications prior to the examination.

PDD Examiners

All of the examinations were conducted by two certified PDD examiners assigned to DoDPI. Both had over 18 years cumulative experience, having served first as criminal investigators, field PDD examiners and finally as DoDPI instructors. They were proficient in operational procedures of the Axciton Computerized Polygraph equipment used during data collection. After three days of procedural refinements, each of the two PDD examiners conducted three examinations per day until he completed his required number of subjects. The research design required that each examiner test 30 individuals who had been randomly programmed innocent, and 30 who had been randomly programmed guilty.

Apparatus

Hardware. Two Axciton Computerized Polygraph Systems (Version 48-I; 16 bit parallel format) were used. The specific channels consisted of: (a) two pneumograph channels utilizing convoluted tubes to measure changes in thoracic and abdominal areas during expiration and inspiration, (b) one electrodermal channel utilizing fingerplate electrodes to measure changes in sweat gland activity on two fingers of the subject's non-dominant hand, and (c) one cardiograph channel utilizing a standard blood pressure cuff, pump bulb assembly and sphygmomanometer to indicate changes in relative blood pressure and blood volume.

Software. During the original study, PASS 2.0, (1993b) developed by the Johns Hopkins University APL, was used to analyze the physiological data collected and stored by the Axciton Computerized Polygraph. Crunch 4.0, (1991), was used to calculate the various data analysis comparisons.

Crime Scene

Space within a typical office lounge was used as the crime scene. The area, referred to as the "Country Store," consisted of two small tables, one of which contained a display of snack food (i.e., candy, etc.). The other table held a small plastic cash box which contained $124.00 in paper currency (four $1s; two $5s; one $10; and one $100), and $3.00 in assorted coins.

Scenario

The mock crime was defined as a theft of $124.00 from the Country Store cash box. Care was taken to prevent the PDD examiners from discerning guilt or innocence via the individual's knowledge of the building, therefore, both innocent and guilty subjects traveled the same route from briefing area to exam room.

Innocent Subjects. The scenario setter took each subject from the waiting area individually, and escorted him/her to the designated briefing area. All were told that a theft of money from the Country Store had occurred, and that each was a suspect in the case due to having been in the area at the time. Each subject was assured that he/she was obviously innocent of the crime and that the task at hand was simply to be honest and cooperative with the PDD examiner. Following programming, each subject was escorted to a holding area, and shortly thereafter, taken to the PDD examination room and introduced to the PDD examiner.

Guilty Subjects. The scenario setter took each subject from the waiting area individually, and escorted him/her to the designated briefing area. In turn, all were told that they were going to steal money from the Country Store and then undergo a PDD examination regarding the theft of the money. It was explained to each subject that the primary goal was to convince the PDD examiner that he/she was innocent of committing the crime. Each individual was then escorted into the area referred to as the Country Store, and shown the cash box and store merchandise.

The scenario setter explained and demonstrated the steps each subject was to follow in committing the theft: (a) take only the paper currency out of the box, (b) count it, (c) conceal it on their person or in a purse, and then (d) immediately leave the room through the designated doorway. It was stressed to all guilty participants that it was vitally important not to be seen stealing the money, otherwise the PDD examiner could inadvertently be informed of the circumstances prior to the examination, effectively rendering the examination unnecessary. (Note: All subjects had been informed that the employment agency would not pay them if they did not take the PDD examination.)

If seen with the money, each subject was instructed to act as if he/she had just purchased a candy bar and were simply making change. Each individual was then to conceal the money and leave the room immediately with a candy bar in hand. Following a review of the steps, each subject was then left alone in the room to carry out the scenario.

Approximately 10-15 seconds after the scenario setter left the room an individual designated as the "deliberate intruder" entered the crime scene through another doorway in order to surprise the subject committing the theft. The deliberate intruder was instructed to remain in the room, making small talk, cleaning the sink counter, etc., until the individual completed the task and left. This was done to heighten the arousal level of the subject by forcing him/her to conceal his actions while committing the crime. The deliberate intruder was not there to confront the individual, but rather to make him/her nervous by obstructing a "clean" getaway following the commission of the crime.

When the subject exited the crime scene he/she was led to a designated area, where the scenario setter made sure the individual both had the money, and knew the amount of money stolen. If the subject had not completed counting the money prior to being interrupted he/she was instructed to confirm the amount at that time. The money was again concealed on the subject's person, and the candy bar was hidden in the room to prevent the examiner from ascertaining that the individual had been programmed guilty.

In order to reinforce details regarding the commission of the theft, the subject was then escorted to a holding area where he/she completed a multiple-choice questionnaire which asked for details regarding the crime and the crime scene. Shortly thereafter, each subject was escorted to the PDD examination room and introduced to the PDD examiner.

Procedure

Upon arrival, the subjects were welcomed to DoDPI, provided with a general briefing on the purpose of the study, and informed that his/her participation was completely voluntary. Subjects were then provided with a packet containing a copy of a project briefing form, a volunteer agreement affidavit, and a background information form. After completing the required paperwork those individuals who chose to participate in the study were escorted to a designated waiting area.

Each subject was randomly assigned to either the innocent group or the guilty group and was then programmed individually as described in the Scenario section of this report. After being introduced to a PDD examiner all subjects underwent the pretest and in-test portions of a ZCT PDD examination.

Utilizing a prescribed ZCT question list, the PDD examiner conducted three tests, with a fourth test authorized only if necessary, (i.e., as a result of excessive movement distortions, etc.). Following the PDD examination, each subject was debriefed and released.

PDD Examination Scoring Criteria

Each examination was scored by the original examiner, and later blind scored by two other certified and similarly qualified PDD examiners. A standard seven position scale (+3, +2, +1, 0, -1, -2, -3) was used in conjunction with the DoDPI criteria for ZCT spot analysis, and numerical evaluation (Department of Defense Polygraph Institute, 1992, August; Department of Defense Polygraph Institute, 1992).

POLYSCORE Analysis Overview (current study)

Four versions, or upgrades of the algorithm-based scoring system were examined in this comparison of overall accuracy: (a) PASS 2.0, (b) POLYSCORE 2.3, (c) POLYSCORE 2.9, and (d) POLYSCORE 3.0. According to the developer, major changes have been made to the scoring system since the initial release of the prototype, PASS 2.0, which occurred in 1992. As of this writing, POLYSCORE 3.0 is the latest version of the software.

From an operational standpoint, the upgrades have essentially been transparent to the user. However, each has resulted in a change in the way the data are processed. Some of the changes which can potentially affect the decision, or call are discussed below.

Cutoff Score

Each version of the algorithm has been programmed by the developer to provide a 90-10 cutoff score as a decision making guideline. Therefore, any examination during this study which received a probability score of 0.90 or higher was categorized as DI. Any score of 0.10 or lower was labeled NDI, and all other scores were considered INC. In short, this simply means that, 9 out of 10 times, an individual displaying the given physiological responses would be correctly identified as either guilty or innocent. In the field, adherence to this cutscore is a matter of individual agency policy.

Subjective Manipulation of the Data

All versions of the scoring system have been designed to recognize, tag, and exclude from the analysis any segment within the physiological tracings which it identifies as containing an artifact. However, to varying degrees each version of the algorithm also permits the subjective manipulation of the data.

In PASS 2.0, the examiner can override the algorithm's decision to exclude an area from scoring consideration, and can also select areas to be eliminated which were not tagged by the system. The same is true for POLYSCORE 2.3, 2.9, and 3.0, with one exception--the examiner can override only selected tagged areas. Any segment defined by the algorithm as a statistical outlier will automatically be excluded from the analysis. (Note: The APL specifies that for the score to be valid, any control or relevant question which contains artifacts must be excluded from the analysis.)

As a result, for the purpose of comparison each data file in this study was scored twice. On the first run, the algorithm was allowed to interpret the data without benefit of subjective manipulation, i.e., tagging only those areas which violated its definition of acceptable. On the second run, problem areas identified by the original examiners were eliminated. In addition, artifacts falling outside the examiner scoring window, but within the algorithm scoring window were identified and eliminated by the researcher. In the Results sections those analyses will be labeled "unedited" and "edited," respectively.

Results

The objective of the original study was to determine the effectiveness of the PASS in detecting deception in a controlled laboratory study using mock crime data. In the current report, however, the emphasis has been shifted to a simple evaluation of the consistency of the decisions made by the various versions of the same algorithm. Throughout this report, the accuracy findings have been presented both with INC decisions included, and also with INC decisions excluded, as would be done in field data reporting.

Overall Accuracy

Table 2 shows the overall percentage of accuracy (both with and without INCs) generated by: (a) the original examiner; (b) both blind scorers; (c) PASS 2.0, with and without examiner edits; and (d) POLYSCORE 2.3, 2.9 and 3.0, with and without examiner edits. (Note: Though the original dataset contained 120 subjects, only 119 are used in these calculations due to a loss of data caused by a damaged computer diskette.) The figures are grouped by correct, incorrect and INC decisions.

Comparing the original examiner's level of accuracy with that generated by each version of the algorithm (using the edited dataset) there is a trend towards greater accuracy from PASS 2.0 to POLYSCORE 2.3, and from POLYSCORE 2.3 to POLYSCORE 2.9. In fact, POLYSCORE 2.9 equals the accuracy level attained by the original examiner (72.27%). However, accuracy drops when the data are analyzed by POLYSCORE 3.0 (68.91%).

The original study showed that the algorithm was more accurate in clearing innocent individuals, while the PDD examiners were more accurate in detecting guilty individuals. As can be seen in the Innocent and Guilty breakouts in Table 2, the same statement is true when comparing the level of examiner accuracy to that of POLYSCORE 2.3, 2.9, and 3.0.

Also during the original study, the examiners had a lower overall INC rate. The same is true when considering the overall rates and the guilty rates for each subsequent version of the algorithm. However, the INC rate for the innocent group was 13.56% for the examiners, and ranged between 11.86% and 16.95% for the edited datasets scored by the algorithm.

The bottom portion of Table 2 shows the same categories of information with the INCs having been eliminated from the calculations. POLYSCORE's trend of increased accuracy overall is still evident here, with POLYSCORE 2.9 and POLYSCORE 3.0 exceeding the original examiner's 79.63% level of accuracy (83.50% and 82.83%, respectively).

Tables 3 and 4 also show the results for the original examiner, both blind scorers and the four versions of the algorithm. However, for the purposes of comparison, the information is grouped according test format; either, DLC or PLC. Table 3 presents the data with the INC decisions included and Table 4 presents the data with the INC decisions eliminated.

Table 2
POLYSCORE Comparison

Accuracy	Examiner Decision	Blind Scorer #1	Blind Scorer #2	PASS 2.0 U	PASS 2.0 E	POLY 2.3 U	POLY 2.3 E	POLY 2.9 U	POLY 2.9 E	POLY 3.0 U	POLY 3.0 E
With inconclusive decisions											
Overall (n=119)											
% Correct (TP + TN)	72.27	62.19	64.71	61.35	63.03	65.55	67.72	70.59	72.27	72.27	68.91
% Incorrect (FP + FN)	18.49	15.13	18.49	19.33	16.81	14.29	17.65	14.29	14.29	15.97	14.29
% INC	9.24	22.69	16.81	19.33	20.17	20.17	15.13	15.13	13.45	11.77	16.81
Innocent (n=59)											
% Correct (TN)	62.71	55.93	57.63	71.19	72.88	81.36	74.58	83.05	77.97	79.66	74.58
% Incorrect (FP)	23.73	22.03	22.03	8.47	10.17	8.47	13.56	11.86	10.17	10.17	11.86
% INC	13.56	22.03	20.34	20.34	16.95	10.17	11.86	5.08	11.86	10.17	13.56
Guilty (n=60)											
% Correct (TP)	81.67	68.33	71.67	51.67	53.33	50.00	60.00	58.33	66.67	65.00	63.33
% Incorrect (FN)	13.33	8.33	15.00	30.00	23.33	20.00	21.67	16.67	18.33	21.67	16.67
% INC	5.00	23.33	13.33	18.33	23.33	30.00	18.33	25.00	15.00	13.33	20.00
With inconclusive decisions eliminated											
Overall											
% Correct (TP + TN)	79.63	80.44	77.78	76.04	78.95	82.11	79.21	83.17	83.50	81.90	82.83
(n)	(86)	(74)	(77)	(73)	(75)	(78)	(80)	(84)	(86)	(86)	(82)
% Incorrect (FP + FN)	20.37	19.57	22.22	23.96	21.05	17.90	20.79	16.83	16.51	18.10	17.17
(n)	(22)	(18)	(22)	(23)	(20)	(17)	(21)	(17)	(17)	(19)	(17)
Innocent											
% Correct (TN)	72.55	71.74	72.34	89.36	87.76	90.57	84.62	87.50	88.46	88.68	86.28
(n)	(37)	(33)	(34)	(42)	(43)	(48)	(44)	(49)	(46)	(47)	(44)
% Incorrect (FP)	27.45	28.26	27.66	10.64	12.25	9.43	15.39	12.50	11.54	11.32	13.73
(n)	(14)	(13)	(13)	(5)	(6)	(5)	(8)	(7)	(6)	(6)	(7)
Guilty											
% Correct (TP)	85.97	89.13	82.69	63.27	69.57	71.43	73.47	77.78	78.43	75.00	79.17
(n)	(49)	(41)	43)	(31)	(32)	(30)	(36)	(35)	(40)	(39)	(38)
% Incorrect (FN)	14.04	10.87	17.31	36.74	30.44	28.57	26.53	22.22	21.57	25.00	20.83
(n)	(8)	(5)	(9)	(18)	(14)	(12)	(13)	(10)	(11)	(13)	(10)

Note. E = edited test; FN = false negative; FP = false positive; TN = true negative; TP = true positive; U = unedited test.

Table 3

POLYSCORE Comparison Using Directed Lie Control and Probable Lie Control Test Formats

Accuracy	Examiner Decision	Blind Scorer		PASS 2.0		POLY 2.3		POLY 2.9		POLY 3.0	
		#1	#2	U	E	U	E	U	E	U	E
				Directed lie control							
Overall (n=59)											
% Correct (TP + TN)	71.19	66.10	59.32	64.41	62.71	66.10	66.10	71.19	71.19	71.19	61.02
% Incorrect (FP + FN)	15.25	13.56	20.34	11.86	15.25	11.86	16.95	11.86	11.86	15.25	15.25
% INC	13.56	20.34	20.34	23.73	22.03	22.03	16.95	16.95	16.95	13.56	23.73
Innocent (n=29)											
% Correct (TN)	51.72	55.17	51.72	75.86	68.97	79.31	65.52	82.76	72.41	75.86	65.52
% Incorrect (FP)	24.14	17.24	24.14	3.45	13.79	10.34	20.69	10.34	10.34	10.34	17.24
% INC	24.14	27.59	24.14	20.69	17.24	10.34	13.79	6.90	17.24	13.79	17.24
Guilty (n=30)											
% Correct (TP)	90.00	76.67	66.67	53.33	56.67	53.33	66.67	60.00	70.00	66.67	56.67
% Incorrect (FN)	6.67	10.00	16.67	20.00	16.67	13.33	13.33	13.33	13.33	20.00	13.33
% INC	3.33	13.33	16.67	26.67	26.67	33.33	20.00	26.67	16.67	13.33	30.00
				Probable lie control							
Overall (n=60)											
% Correct (TP + TN)	73.33	58.33	70.00	58.33	63.33	65.00	68.33	70.00	71.67	73.33	76.67
% Incorrect (FP + FN)	21.67	16.67	16.67	26.67	18.83	16.67	18.33	16.67	16.67	16.67	13.33
% INC	5.00	25.00	13.33	15.00	18.83	18.33	13.33	13.33	11.67	10.00	10.00
Innocent (n=30)											
% Correct (TN)	73.33	56.67	63.33	66.67	76.67	83.33	83.33	83.33	83.33	83.33	83.33
% Incorrect (FP)	23.33	26.67	20.00	13.33	6.67	6.67	6.67	13.33	10.00	10.00	6.67
% INC	3.33	16.67	16.67	20.00	16.67	10.00	10.00	3.33	6.67	6.67	10.00
Guilty (n=30)											
% Correct (TP)	73.33	60.00	76.67	50.00	50.00	46.67	53.33	56.67	60.00	63.33	70.00
% Incorrect (FN)	20.00	6.67	13.33	40.00	30.00	26.67	30.00	20.00	23.33	23.33	20.00
% INC	6.67	3.33	10.00	10.00	20.00	26.67	16.67	23.33	16.67	13.33	10.00

Note. E = edited test; FN = false negative; FP = false positive; TN = true negative; TP = true positive; U = unedited test.

Table 4

POLYSCORE Comparison Using Directed Lie Control and Probable Lie Control Test Formats With Inconclusive Decisions Eliminated

Accuracy	Examiner Decision	Blind Scorer #1	Blind Scorer #2	PASS 2.0 U	PASS 2.0 E	POLY 2.3 U	POLY 2.3 E	POLY 2.9 U	POLY 2.9 E	POLY 3.0 U	POLY 3.0 E
Directed lie control											
Overall											
% Correct (TP + TN)	82.35	82.98	74.47	84.44	80.44	84.78	79.59	85.71	85.71	82.35	80.00
(n)	(42)	(39)	(35)	(38)	(37)	(39)	(39)	(42)	(42)	(42)	(36)
% Incorrect (FP + FN)	17.65	17.02	25.53	15.56	19.57	15.22	20.41	14.29	14.29	17.65	20.00
(n)	(9)	(8)	(12)	(7)	(9)	(7)	(10)	(7)	(7)	(9)	(9)
Innocent											
% Correct (TN)	68.18	76.19	68.18	95.65	83.33	88.46	76.00	88.89	87.50	88.00	79.17
(n)	(15)	(16)	(15)	(22)	(20)	(23)	(19)	(24)	(21)	(22)	(19)
% Incorrect (FP)	31.82	23.81	31.82	4.35	16.67	11.54	24.00	11.11	12.50	12.00	20.83
(n)	(7)	(5)	(7)	(1)	(4)	(3)	(6)	(3)	(3)	(3)	(5)
Guilty											
% Correct (TP)	93.10	88.46	80.00	72.73	77.27	80.00	83.33	81.82	84.00	76.92	80.95
(n)	(27)	(23)	(20)	(16)	(17)	(16)	(20)	(18)	(21)	(20)	(17)
% Incorrect (FN)	6.90	11.54	20.00	27.27	22.73	20.00	16.67	18.18	16.00	23.08	19.05
(n)	(2)	(3)	(5)	(6)	(5)	(4)	(4)	(4)	(4)	(6)	(4)
Probable lie control											
Overall											
% Correct (TP + TN)	77.19	77.78	80.77	68.63	77.55	79.59	78.85	80.77	81.13	81.48	85.19
(n)	(44)	(35)	(42)	(35)	(38)	(39)	(41)	(42)	(43)	(44)	(46)
% Incorrect (FP + FN)	22.81	22.22	19.23	31.37	22.45	20.41	21.15	19.23	18.87	18.52	14.82
(n)	(13)	(10)	(10)	(16)	(11)	(10)	(11)	(10)	(10)	(10)	(8)
Innocent											
% Correct (TN)	75.86	68.00	76.00	83.33	92.00	92.59	92.59	86.21	89.29	89.29	92.59
(n)	(22)	(17)	(19)	(20)	(23)	(25)	(25)	(25)	(25)	(25)	(25)
% Incorrect (FP)	24.14	32.00	24.00	16.67	8.00	7.41	7.41	13.79	10.71	10.71	7.41
(n)	(7)	(8)	(6)	(4)	(2)	(2)	(2)	(4)	(3)	(3)	(2)
Guilty											
% Correct (TP)	78.57	90.00	85.19	55.56	62.50	63.64	64.00	73.91	72.00	73.08	77.78
(n)	(22)	(18)	(23)	(15)	(15)	(14)	(16)	(17)	(18)	(19)	(21)
% Incorrect (FN)	21.43	10.00	14.82	44.44	37.50	36.36	36.00	26.09	28.00	26.92	22.22
(n)	(6)	(2)	(4)	(12)	(9)	(8)	(9)	(6)	(7)	(7)	(6)

Note. E = edited test; FN = false negative; FP = false positive; TN = true negative; TP = true positive; U = unedited test.

Subjective Manipulation of the Data

Tallying the four versions of the algorithm, there were 38 separate cases (32% of total) affected by subjective manipulation of the data (i.e., the edited data produced a different decision from the unedited data). Each version of the algorithm generated a different number of cases (PASS 2.0, [16]; POLYSCORE 2.3, [14]; POLYSCORE 2.9, [13]; POLYSCORE 3.0, [19]), but the case overlap was minimal. There were only two cases which, when edited, resulted in a different decision on all four versions of the algorithm. There were an additional four cases which resulted in a different decision when analyzed by three of the four versions of the algorithm.

Decision Reversals

While there were numerous definitive (NDI or DI) decisions which shifted to INC, or vice versa, there were only eleven cases overall in which the decision actually reversed from one definitive call to the other when scored with a later version of the algorithm. Two innocent individuals who had been correctly labeled NDI initially (by PASS 2.0 and POLYSCORE 2.3), were later labeled DI by POLYSCORE 2.9 and 3.0. Another had been labeled DI by PASS 2.0 and POLYSCORE 2.3, correctly relabeled NDI by POLYSCORE 2.9, and then mislabeled again by POLYSCORE 3.0. The original examiner erred on two of the cases and judged the other one INC.

There were also two cases where guilty individuals were correctly labeled DI (by PASS 2.0 and POLYSCORE 2.3) only to be reversed by a later version of the algorithm (2.9 and 3.0, respectively). The PDD examiner correctly labeled both cases DI. The remaining six cases involved guilty individuals who were initially called NDI, but correctly deemed DI by a subsequent release of the software (Note: In each case, the correct decision was rendered by POLYSCORE 2.9 and/or POLYSCORE 3.0). By comparison, the original examiner correctly identified four as DI, mislabeled one as NDI and called one INC.

Algorithm/PDD Examiner/Blind Scorer vs. Ground Truth

There were 61 cases (51.26% of total dataset) where the original examiner and all four versions of the algorithm (using edited data) agreed upon the decision. Of those, there were 51 cases, or 42.86% of the total dataset, where the original examiner and all four versions of the algorithm also agreed with ground truth.

Statistical Outliers

As mentioned previously, later releases of the algorithm have been equipped with the capability to identify and exclude statistical outliers. When the dataset was analyzed by POLYSCORE 2.3, a total of four cases were identified as having statistical outliers. That number rose to 55 cases when POLYSCORE 2.9 was used and then dropped to 35 when the same data were analyzed by

POLYSCORE 3.0. There was only one occasion when POLYSCORE 2.3 and 2.9 identified a statistical outlier in the same case. There was an overlap on 24 cases when comparing POLYSCORE 2.9 and POLYSCORE 3.0. In addition, there were two cases where a statistical outlier was identified by all three versions of the software.

Discussion

There are inherent difficulties with generalizing the results of mock crime cases to those cases collected under field conditions. Due to an inability to create sufficient stress-inducing "consequences" for everyone involved in a controlled laboratory mock crime scenario, the occurrence of a certain percentage of false decisions is to be expected. Therefore, neither the algorithm, nor the examiners were expected to attain the level of accuracy normally associated with the performance of their respective tasks--neither did.

As a result, this analysis focused on consistency of performance, rather than concentrating exclusively on accuracy. Many of the same findings generated by Blackwell (1994) were upheld in the current evaluation. For example, all versions of the algorithm are more accurate at clearing innocent individuals than detecting guilty individuals. The reverse is true of the examiners.

Having caveated the emphasis on accuracy, it was interesting to find that the overall percentage of correct decisions increased with each successive release of the algorithm--except POLYSCORE 3.0, which actually decreased in accuracy. More interesting was the fact that accuracy rates calculated for innocent subjects remained constant, never varying more than a few percentage points, while the decisions for guilty individuals improved as much as 13 percentage points (13.34% with INCs included; 9.6% with INCs excluded).

Considering the effects of subjective manipulation of the data, it was disconcerting to see the same "edits" produce a different decision on one-third of the cases examined. However that may have resulted, in part, from the algorithm's processing of statistical outliers.

In PASS 2.0 the tests could be edited exactly as the examiner chose to edit them. Later versions did not offer that option. Since there was no consistent pattern of cases which produced the statistical outliers, there was no way for the researcher to score all the tests in exactly the same way from version to version. When scored, the tests included not only the set of edits made by the examiners, but also any additional edits, due to statistical outliers, made by the respective version of the algorithm. POLYSCORE 2.9 tagged statistical outliers in 55 cases. It is noteworthy that the combination of examiner edits and algorithm generated statistical outliers used by POLYSCORE 2.9 produced the highest overall accuracy rate of any version examined.

Decision reversals, were found to have negligible relevance. Only eleven definitive calls were actually reversed by subsequent versions of the algorithm. Of a greater concern were the number of cases which moved from a definitive decision to a decision of INC. Generally, all versions of the algorithm demonstrated a higher INC rate than the PDD examiners, particularly when assessing guilty individuals.

Blackwell (1994) found that the decisions agreed upon by both the examiner and the algorithm were much more likely to be accurate than when either decision was considered alone. The same was true in this analysis when judging the accuracy rates which include INC decisions in the calculation. However, there was little difference between the accuracy rates generated by the unilateral decisions of later versions of the algorithm (POLYSCORE 2.9 and 3.0), when compared to the combination of algorithm and examiner decisions.

In summary, the APL algorithm-based scoring system is a user friendly software package which, in this study, provided moderately high levels of accuracy (±80.0%) on mock crime data. According to the manufacturer, one of the primary advantages POLYSCORE offers the PDD field is scoring consistency. Due to the fact that POLYSCORE is a computer-based system, it does offer scoring consistency--but only when the same artifacts are edited with consistency by all the scoring examiners. As shown in this report, modest variability in subjective editing can and does impact the outcome. Differences in interrater agreement also point to a need for a truly consistent method for evaluating PDD examinations (Blackwell, 1994).

Additionally, throughout this report, the accuracy findings have been presented in terms of overall accuracy (including INC decisions), and accuracy with the INC calls eliminated, as done in field data reporting. As was pointed out in Blackwell (1994), this was done in order to provide PDD managers with a means for computing the cost effectiveness of using POLYSCORE, in addition to the expected benefits of improved scoring accuracy.

Blind scorers in this study produced inconclusive rates of up to 23%. POLYSCORE performance on inconclusives was no better. Unless the algorithm can demonstrate increased accuracy and/or decreased inconclusive rates as compared to examiner evaluation, it is of limited value as an augmentation to the current method of physiological analysis.

References

Blackwell, N. J. (1994). An evaluation of the effectiveness of the polygraph automated scoring system (PASS) in detecting deception in a mock crime analog study (Report No. DoDPI94-R-0003). Fort McClellan, AL: Department of Defense Polygraph Institute.

Capps, M. (1993, January). Polygraph automated scoring system (PASS), version 2.0. Paper presented at the Department of Defense Polygraph Institute training program on PASS Operations, Fort McClellan, AL.

Crunch 4.0 [Computer program]. (1991). Oakland, CA: Crunch Software Corporation.

Department of Defense Polygraph Institute. (1992). Chart Evaluation. (Available from Department of Defense Polygraph Institute, Building 3195, Fort McClellan, AL 36205)

Department of Defense Polygraph Institute. (1992, August). Zone Comparison Technique (ZCT). (Available from Department of Defense Polygraph Institute, Building 3195, Fort McClellan, AL 36205)

Polygraph Automated Scoring System (PASS) user's guide, version 2.0 [Computer program manual]. (1993a). Laurel, MD: The Johns Hopkins University Applied Physics Laboratory.

Polygraph Automated Scoring System (PASS) 2.0 [Computer program]. (1993b). Laurel, MD: The Johns Hopkins University Applied Physics Laboratory.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.