CPRC
CANADIAN POLICE RESEARCH CENTRE

CCRP
CENTRE CANADIEN DE RECHERCHES POLICIÈRES

TR-07-94
Field Validity Study of the Canadian Police College Polygraph Technique

Charles Robert Honts, Ph.D.
Associate Professor of Psychology
University of North Dakota

TECHNICAL REPORT

March, 1994

NOTE: Further information about this report can be obtained by calling the CPRC information number (613) 998-6342

SUMMARY

A field study of the Canadian Police College Polygraph Technique was conducted. Data from the files of 41 field cases were considered, and 32 of them were found to have at least some information that provided independent confirmation of the polygraph examination outcome. Numerical scores and decisions from the original examiners and an independent evaluation based only on the polygraph charts were analyzed. The results indicated that the numerical scoring approach taught at the Canadian Police College is very reliable, with all estimates of inter-rater reliability exceeding 0.91. The results also indicated that the Canadian Police College Polygraph technique was highly valid in discriminating truthtellers from deceivers. Excluding inconclusives, the decisions of the Original Examiners were correct 96% of the time. When only those cases which were confirmed at the highest level (all included a confession by the perpetrator of the crime) were considered, no errors were made by either the Original Examiners or by the Independent Evaluator. Although these results are very supportive of the Canadian Police College Polygraph Technique, they are somewhat limited by the small number of cases in the study. Additional research with a larger number of cases is highly recommended, and specific recommendations for future research projects are provided.

SOMMAIRE

Au cours d'une étude in situ des techniques polygraphiques enseignées au Collège canadien de police (CCP), on a examiné des données tirées de 41 dossiers, dont 32 contenaient au moins quelques renseignements qui confirmaient de façon indépendante les résultats des tests polygraphiques. On a également analysé les notes numériques attribuées aux suspects, les décisions rendues par les examinateurs et les conclusions tirées à la suite d'une évaluation indépendante fondée uniquement sur les imprimés polygraphiques. L'étude a révélé que la méthode de notation numérique enseignée au CCP est très sûre : tous les coefficients d'objectivité s'élèvent à plus de 0,91. On a également démontré que les tests polygraphiques employés au CCP distinguaient avec beaucoup de précision les personnes qui disaient ou dissimulaient la vérité. Si l'on ne tient pas compte des tests non concluants, l'on constate que les examinateurs arrivaient à la bonne conclusion dans 96 % des cas et que les évaluations indépendantes fondées uniquement sur des facteurs physiologiques faisaient mouche dans 93 % des cas. Si l'on ne tient compte que des conclusions confirmées avec la certitude la plus totale (dans chacun des cas en question, le suspect avait avoué sa culpabilité), on voit que les examinateurs et les évaluateurs indépendants n'ont commis aucune erreur. Bien que ces résultats soient très positifs, ils sont quelque peu limités en raison du petit nombre de dossiers examinés. Il est donc fortement conseillé d'approfondir les recherches en étudiant plus de dossiers, et le rapport contient des recommandations précises sur des projets de recherche qu'on pourrait entreprendre à l'avenir.

Background

Worldwide, the control question test (CQT) is the most commonly used polygraph test in law enforcement today. Furthermore, a version of the CQT is the only polygraph technique currently being taught at the Canadian Police College's Polygraph Training Unit. Thus, the validity of the CQT is a matter of considerable importance to Canadian law enforcement and to law enforcement in general. However, the actual validity of the CQT has been, and continues to be, the subject of a polemic debate in the scientific community (Honts & Perry, 1992).

Two recent comprehensive reviews arrive at very different conclusions about the accuracy of the control question test (Iacono & Patrick, 1988; Raskin, 1989). Iacono and Patrick conclude, "the best defense one can offer for the continued use of the CQT is that its accuracy is indeterminate" (p. 233), while Raskin concludes, "The voluminous scientific literature indicates that they [control question tests] can be highly accurate when properly employed in appropriate circumstances" (p. 290). These strikingly different conclusions are based on differing opinions regarding the value of laboratory experiments and on which field studies are considered to have adequate methodology.

Critics (e. g., Kleinmuntz & Szucko, 1982; Iacono & Patrick, 1988; Lykken, 1981) of polygraph tests generally dismiss the results of all laboratory simulation studies as useful for estimating field accuracy. They argue that the qualitative context produced by the threat of criminal sanctions in the real world cannot be simulated in the laboratory. Others (Kircher, Horowitz, & Raskin, 1988; Raskin, 1989) have suggested that the differences between laboratory and field settings may not be that great. They have argued that if simulation studies use representative subject populations, realistic polygraph practices, and include some motivation to deceive, then they can provide useful information for estimating field accuracy. Kircher et al. (1988) conducted a meta-analysis on 14 laboratory studies of the control question test. They reported an average accuracy of 87% for the five studies (Ginton, Netzer, Elaad, & Ben-Shakhar, 1982; Kircher & Raskin, 1982; Podlesny & Raskin, 1978; Raskin & Hare, 1978; Rovner, Raskin & Kircher 1979) they rated as most realistic in the above three characteristics. The four studies (Barland & Raskin, 1975; Bradley & Ainsworth, 1984; Bradley & Janisse, 1981; Szucko & Kleinmuntz, 1981) they rated as least realistic on the above three characteristic produced an average accuracy rate of only 73%. Given that high quality laboratory studies have demonstrated that the control question test can be highly accurate, it is very reasonable to then focus on field studies. Only field studies can provide evidence that a technique shown to be accurate in the laboratory can be accurately implemented in the field.

Accuracy estimates for the CQT based on field studies have varied wildly, ranging from chance to near perfection. In considering the field studies, the scientific arguments have generally centered on the methodology used in the various experiments. Honts and Perry (1992) have recently reviewed the literature concerning field validity studies of the CQT. They note that the adequacy of field studies have generally been evaluated on the following four factors: subjects, evaluations, sampling strategy, and criterion development. More specifically, it seems to be generally agreed that:

- **If the primary target for generalization is the application of polygraph testing in law enforcement, then the subjects of field validity studies should be suspects in real-life criminal cases.** Questions about the validity of the CQT with victims and other types of subjects are of interest, but such questions require separate examination.
- **Evaluations should be based on the physiological data alone. Moreover, the evaluations should be conducted by persons trained and experienced in doing blind chart evaluation.**
- **The sampling of cases should be by some acceptable scientific basis. Cases must not be selected on the basis of the quality of the charts or on the accuracy of the outcome of the original examiner. Patrick and Iacono (1991) have argued persuasively that an exhaustive sampling strategy is likely to produce the minimum amount of sampling error.**
- **A criterion for who is innocent and who is guilty must be developed that is independent of the polygraph test. Generally, confessions have been considered to be the only acceptable criterion. Unfortunately, the use of a confession criterion introduces a number of problems of sampling bias that, in turn, raise questions about the usefulness of confession studies (Patrick & Iacono, 1991). However, for the present, confession-based criteria appear**

to be the best available. It is clear that developmental work is needed to see if viable alternatives to confession criteria are possible and useful.

The United States Congress' Office of Technology Assessment (OTA, 1983) conducted an extensive review of the field studies existing at that time. They found 10 studies that met their minimal standards and reported an accuracy rate of 90% on criterion-guilty subjects and an accuracy rate of 80% on criterion-innocent subjects. Lamentably, none of those studies adequately satisfies all of the criteria specified above. Since the OTA study, three new field studies (Honts & Raskin, 1988; Patrick & Iacono, 1991; Raskin, et al., 1988) have been reported in the literature. Honts and Perry (1992) argue that all of these studies appear to satisfy the above methodological criteria. Across those three studies, when inconclusive outcomes were excluded, the independent evaluators correctly classified 78% of the criterion-innocent subjects and correctly classified 98% of the criterion-guilty subjects. It is interesting to note that the independent evaluations in the Honts and Raskin and the Raskin et al. studies were as accurate as, or more accurate than, the original examiners. However, the independent evaluators in the Patrick and Iacono study, who were Royal Canadian Mounted Police (RCMP) examiners, were much less accurate with innocent subjects than were the original examiners. The original examiners were also RCMP examiners, and they correctly classified 90% of the innocent subjects. The reasons for the differences between these studies are not apparent, and deserve study.

The Patrick and Iacono study may have suffered from a problem known as criterion contamination (Muchinsky, 1993). That is, they may have been measuring something other than just the validity of the CQT. For example, recent laboratory (Barland, Honts, & Barger, 1989) and field (Raskin, et al., 1988) research has suggested that the CQT does not have very good specificity. That is, the CQT can determine if the subject is attempting deception to some issue, but the technique may not be very good at determining which issue is being responded to deceptively when more than one issue is addressed. The Honts and Raskin and the Raskin et al. studies directly addressed this problem methodologically, while there is no indication that Patrick and Iacono took such issues into account.

There is another issue that might contribute to criterion contamination. RCMP polygraph examinations usually address the most serious level of involvement under investigation. If the suspect being tested was not guilty of the most serious level of involvement, but was involved in the crime, an issue arises about how such a subject should be classified. Consider the following scenario. A diamond ring is stolen. Suspect A takes a polygraph test wherein the relevant questions are of the form, "Did you steal the diamond ring?" The suspect fails the polygraph test, and subsequently confesses that while he did not steal the diamond ring, his brother did and he sold the stolen property. Is Suspect A, truthful or deceptive with regard to the polygraph examination? The position taken in this study was that Suspect A would have been considered deceptive, since his or her intention was to deceive the polygraph examiner. However, others may not have taken this approach, and in fact, anecdotal accounts from the RCMP (Kaster, personal communication, 1991) suggest that such criterion contamination may have been a problem with the Patrick and Iacono study.

In order to address some of the problems associated with conducting field studies in the detection of deception, and to obtain an estimate of the validity of the Canadian Police College Polygraph Technique, a research contract was issued in October of 1991. The present document is the final report of that project.

Goals Proposed for the Original Project

Given the background described in the previous section, the present study was proposed and then subsequently funded in 1991. The Goals and Design of the study were as follows:

Primary Goals:

1. To conduct a field study of the validity of the Canadian Police College Polygraph Technique that would meet scientific standards.

   That is, the primary portion of the present study was to use:

   - Exhaustive sampling, without reference to chart quality or the original examiners' outcome.
   - Independent evaluations based only on the polygraph charts.
   - Independent evaluations conducted with the Canadian Police College Polygraph Technique numerical scoring system by examiners experienced with that system.
   - A confession criterion of ground truth to establish the basic Innocent and Guilty Conditions.

2. To avoid problems of criterion contamination through the use of explicit operational definitions of Guilt and Innocence.

3. To explore the utility of the use of confession criteria. This was to be accomplished through the development of a Strength of Confirmation rating scale. Some subjects were to be selected with the standard and very conservative confession criterion. Other subjects were to be sampled with progressively less conservative criteria. Differences between these categories of strength of classification were to be examined statistically. It was hoped that the outcome of this analysis would have important implications for the conduct of future studies by possibly providing a method for broader sampling of cases from field samples.

Secondary Goals:

1. To examine the validity of the Canadian Police College Polygraph Technique with other types of subjects. As criminal suspect cases were sampled, those cases with Victims, Informants, and Other types of cases that had extra-polygraphic confirmation information were to be sampled as they were found. If sufficient numbers of such cases were found, statistical analyses were to be conducted.

2. To collect demographic information across all subjects. Little is known about the effects of demographic variables on the validity of polygraph tests. If sufficient variability is obtained across demographic variables additional analyses were to be conducted.

Synopsis of the Proposed Research Design

The investigation was to have been an archival field study of the validity of the Canadian Police College control question polygraph technique as it was administered to criminal suspects. Data from 200 subjects was to be obtained from the RCMP files. Sampling was to be sequential and exhaustive working back through the files from the end of fiscal year 1991 until the specified number of confirmed cases from criminal suspects were obtained. As a secondary goal, the validity of the Canadian Police College control question polygraph test was to be examined with subjects other than criminal suspects. Confirmed cases from victims, informants, and other subject types were also to be selected for analysis as they were found during sampling for confirmed criminal suspect cases. All of these data were to be independently evaluated by an experienced RCMP examiner appointed by the Scientific Authority. The charts were also to be independently evaluated by the Contractor. After the independent evaluations, the Contractor was to conduct reliability and validity analyses on the independent evaluations of the RCMP examiner, the Contractor, and on the data provided by the Original Examiners. The Contractor was to provide a final report of these analyses with recommendations for future research and with suggestions for possible changes in current techniques and practices.

Progress of the Project

In the first wave of data collection, the Scientific Authority provided 23 case files, containing 29 polygraph examinations. Strength of Confirmation Rating Forms were filled out by the Scientific Authority and were included in the case files, as were the polygraph charts, all case information, and the Original Examiner's score sheets. However, neither the Case Information Forms nor the Polygraph Information Forms were provided by the Scientific Authority. The Contractor performed blind independent numerical analyses of the polygraph charts, and he also performed Strength of Confirmation evaluations of the materials in each of the case files. The Contractor also performed a number of statistical analyses on the cases on hand at that time and the results of those analyses were presented in the first *Interim Report* (Honts, 1992). The initial data analysis indicated that the Strength of Confirmation Ratings were very reliable and they suggested that the polygraph techniques were both highly reliable and highly valid. In the conclusions of that report, the Contractor made a number of suggestions for the second wave of data collection.

The project received reauthorization on 14 June 1993, and a second wave of data collection commenced. The Contractor received 12 additional cases in mid-October of 1993, and was told at that time by the Scientific Authority (Canadian Police College) that this would be all of the data that would be provided in this project. The Contractor proceeded with evaluation and data analysis and, according to the terms of the contract, submitted a second *Interim Report* (Honts, 1994) describing the analyses of the data from the second wave of data collection. The remainder of this report provides a final report of the methods, analyses, and results from the combined data from the two waves of data collection.

Method and Results

Primary Design and Cell Size

The overall design of the primary portion of this study was a Guilt (2: Guilty, Innocent) by Confirmation (4: Strong, Moderate, Weak, None) factorial design. The original proposal called for sampling to continue sequentially back through the files until 75 criminal suspect cases with extra-polygraphic confirmations of guilt and 75 criminal suspect cases with extra-polygraphic confirmations of innocence were obtained for independent scoring. Fifty cases were also to be sampled from the pool of unconfirmed cases. Sampling of unconfirmed cases was to be random, except that 25 cases were required to come from cases the Original Examiner called truthful and 25 cases were required to come from cases the Original Examiner called deceptive.

Obtained Cases

Very early in the research process the Polygraph Training Unit of the Canadian Police College encountered administrative difficulties in obtaining case information from the field. The sampling procedure had to be drastically modified for the project to progress. The sampling that was finally achieved was an exhaustive sample from one examiner. All polygraph tests conducted by that examiner between December, 1990 and February, 1992 were provided to the Polygraph Training Unit for use in this study. In total, files were provided for 23 cases. Those cases included 29 polygraph tests.

The second wave of data collection provided materials from 12 polygraph examinations. Those cases were obtained from a single examiner (not the one used in the first wave of data collection). Those cases represented an exhaustive sample of all cases by that examiner that contained any confirmatory information other than the polygraph result. The sample covered the period between July, 1992 and June, 1993, inclusive.

Strength of Confirmation Ratings

The Scientific Authority and the Contractor both rated the Strength of Confirmation provided by the respective case files from the first wave of data collection on a 7-point scale that ranged from 1 = Very Weak confirmation to 7 = Very Strong confirmation. In practice, the "1" category of this scale functioned as the operational equivalent of the design category of No Confirmation. The "1" anchor of the Strength of Confirmation rating scale should probably be renamed to reflect its correspondence to the experimental design. These two ratings were correlated, and they were found to be very similar, $r = 0.94$, $p < 0.0001$. This result indicates that the procedures used to estimate the Strength of Confirmation were very reliable and they greatly exceeded the original design requirement that the obtained r between the Contractor and the Scientific Authority exceed a value of 0.60.

Given the strong results in the first data wave, only the Contractor evaluated the Strength of Confirmation in the second wave. The levels of confirmation assigned by the Contractor to all cases are illustrated in Figure 1.

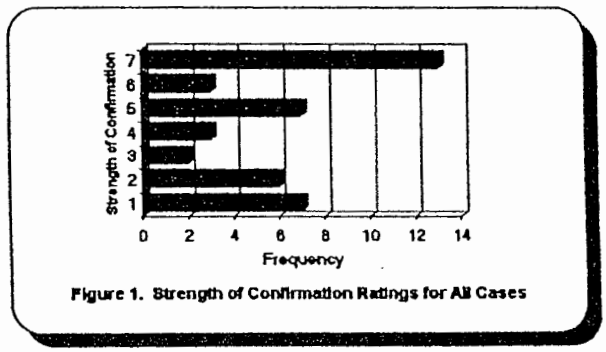


Figure 1. Strength of Confirmation Ratings for All Cases

Criteria for Cell Inclusion

Criteria for the initial inclusion of cases in the various categories of confirmation was as follows:

- **Strong Confirmation. To be entered into the strong confirmation conditions, the case must have been rated by the Scientific Authority at a value of 5 or greater on the Strength of Confirmation scale. Further, to be included in the Strong Confirmation condition there must have been a confession by the perpetrator of the crime under investigation.**
- **Moderate Confirmation. To be entered into the Moderate Confirmation conditions the case must have been rated at a Strength of Confirmation of at least 5 on the 7-point scale. However, for the Moderate Confirmation category no confession was necessary.**
- **Weak Confirmation. Cases that have some confirmation information, but ratings on the Strength of Confirmation scale of less than 5,**

were to be included in the Weak Confirmation category.

- **No Confirmation. Cases without confirmatory information were selected into the No Confirmation category. This category was equivalent to a Strength of Confirmation rating of 1.**

Using the Contractor's ratings, 7 subjects were selected into the No Confirmation Category, 11 subjects were selected into the Weak Confirmation category. Ten subjects were selected into the Moderate Confirmation category and 13 subjects were selected into the Strong Confirmation category (a rating greater than 5 plus a confession). Of the 7 subjects selected into the No Confirmation category, the original examiner classified 6 of those subjects as truthful, and the remaining examination produced an inconclusive outcome.

Of the 11 subjects selected into the Weak Confirmation category, the Contractor felt that 3 of those cases contained information that indicated the subject of the respective examinations was truthful, while 8 of the cases contained information that indicated that the subject of the examination was deceptive. Interestingly, there were two subjects in this category that the Contractor thought were confirmed deceptive, while the Scientific Authority thought the cases were confirmed truthful. However, the Scientific Authority seems to have taken the polygraph examination outcome into account in his assessment, while the Contractor did not. In any event, these were the only substantial disagreements between the two evaluations of confirmation. Of the 6 subjects selected into the Moderate Confirmation category, the Contractor found evidence in the case files that 4 were deceptive while the evidence indicated that 2 were truthful. Of the 15 subjects in the Strong Confirmation category, the evidence plus a confession indicated that 6 of the subjects were truthful while 9 were deceptive. The final content of the cells of the design are illustrated in Table 1. Because the Contractor's Confirmation estimates were made without reference to the polygraph tests, the Contractor's values were used for all analyses reported here.

Table 1. Distribution of Cases by Confirmation Category After the End of the First Wave of Data Collection.

	Confirmation			
Guilty Status	Strong	Moderate	Weak	None
Innocent	6	2	3	6
Guilty	9	6	8	1(inc.)

Numerical Evaluations

Numerical evaluations were performed by the Original Examiners and the Contractor according to the techniques taught at the Canadian Police College (CPC). Numerical evaluations assigned by an instructor of the CPC's Polygraph Training Unit were also available for the data from the second wave of data collection. The CPC numerical scoring techniques are based on the numerical scoring system developed and validated at the University of Utah (e.g., Kircher & Raskin, 1988). That numerical scoring system has been shown to be both highly reliable and highly valid in a number of laboratory and field studies (for a review see Raskin, 1989).

In the CPC numerical scoring system the physiological responses of the subject are evaluated at each relevant and control question pair for the presence of the following criteria: Respiration, decrease in amplitude, slowing of rate, and increase in baseline; Electrodermal Response, increase in amplitude, increase in duration, and increase in complexity; Cardiovascular Response, increase in amplitude of the slow wave, and increase in duration of the slow wave response. Additional detail of these criteria can be found in Kircher and Raskin (1988).

At each relevant and control question pair each physiological system is evaluated independently. At each comparison point a score is assigned on a 7-point scale that varies between -3 and +3. If a criterion-defined response to the relevant question in the pair is stronger, a negative score is assigned. If a criterion-defined response to the control question in the pair is stronger, a positive score is assigned. Equivalent response to both questions, including no response to both questions, results in a score of zero. After all relevant and control question pairs have been scored, all of the scores are summed. The total numerical score is then evaluated to make a decision regarding the subject's credibility. Total numerical scores greater than 5 result in a decision of truthful. Total numerical scores less than -5 result in a decision of deceptive. Total numerical scores between -5 and +5, inclusive, result in an inconclusive outcome.

The numerical scores generated in this study were analyzed to examine their reliability and validity. The validity of decisions based on those numerical scores was also examined.

Reliability

Total numerical scores generated by the Original Examiners and the Contractor were available for all 41 subjects. Those total numerical scores were correlated. The resulting correlation was significant, $r = 0.91$, $p < .0001$. The total numerical scores from the second wave of data collection from the Original Examiners, the Contractor, and an Instructor from the CPC's Polygraph Training Unit were also correlated. The results of those analyses are illustrated in Table 2. These results indicate that the numerical scoring system taught at the CPC's Polygraph Training Unit is highly reliable. These results are comparable to the best result obtained in highly controlled laboratory studies.

Table 2. Inter-Rater Reliability Coefficients for the Three Numerical Evaluations (2nd Data Wave Only).

	CPC Instructor	Contractor
Original Examiner	0.91	0.96
CPC Instructor		0.92

Subject Types

Two case types were represented in the data set. Thirty-five of the examinations were of criminal suspects. Six of the examinations were of persons who could have been victims of crime. Possible differences between these categories of cases were explored by conducting a Guilt (Innocent, Guilty) X Subject Type (Suspect, Victim) Analysis of Variance (ANOVA) of the numerical scores generated by the Original Examiners and the independent scoring by the Contractor. Subjects for whom there was no confirmatory evidence were eliminated from these analyses as were the two subjects over whom there was disagreement over the direction of confirmation. The means of these analyses are shown in Table 3. The ANOVAs revealed a significant main effect of Guilt in the numerical scores of the Original Examiner, $F(1, 28) = 37.81$, $p < .001$, and in the numerical scores of the Contractor, $F(1, 28) = 28.44$, $p < .001$. There were no significant main effects of Subject Type in either of the data sets. Similarly, the interaction of Guilt and Subject Type was not significant in either data set. Since Subject Type does not appear to affect the numerical scores, the design was collapsed across that variable for additional analyses.

Table 3. Mean Numerical Scores and Standard Deviations by Guilt, Case Category, and Evaluation.

Evaluation / Guilt Status	Subject Type: Suspect	Subject Type: Victim
Original Examiner		
Innocent	8.20 (3.55) n=10	10.00 (0.00) n=1
Guilty	-8.41 (6.30) n=17	-8.25 (15.64) n=4
Contractor		
Innocent	5.00 (7.10) n=10	-3.00 (0.00) n=1
Guilty	-13.12 (7.08) n=17	-11.50 (16.03) n=4

Note: Standard deviations are shown in parentheses.

Level of Confirmation

The numerical scores of the Original Examiners and the Contractor were examined with a Guilt (Innocent, Guilty) by Level of Confirmation (High, Medium, Low) ANOVA. The means for this analysis are presented in Table 4. Subjects for whom there was no confirmatory information were not included in this analysis nor were the two subjects on whom the evaluators disagreed about the direction of confirmation. The ANOVA revealed a main effect of Guilt in the scores of the Original Examiners, $F(1, 26) =$

Table 4. Mean Numerical Scores and Standard Deviations by Guilt, Level of Confirmation, and Evaluation.

Evaluation / Guilt Status	Level of Confirmation: Low	Medium	High
Original Examiner			
Innocent	8.67 (1.53) n=3	9.00 (2.82) n=2	8.00 (4.52) n=6
Guilty	-5.00 (10.97) n=6	-6.62 (7.58) n=8	-13.28 (4.15) n=7
Contractor			
Innocent	-4.67 (1.52) n=3	6.00 (5.66) n=2	8.17 (5.27) n=6
Guilty	-11.33 (12.13) n=6	-12.00 (9.38) n=8	-15.00 (5.45) n=7

Note: Standard deviations are shown in parentheses.

45.68, $p < .001$, and in the scores of the Contractor, $F(1, 26) = 30.55$, $p < .001$. In neither data set were the main effects of Level of Confirmation or the interactions involving Level of Confirmation significant. It is interesting to note that in the Contractor's scores the interaction of Guilt and Level of Confirmation approached significance, $F(2, 26) = 2.57$, $p = .095$. This marginal effect is due to a striking difference in the numerical scores assigned by the Contractor to subjects in the Innocent-Low Level of Confirmation cell. This may be noteworthy since this is the only cell where differences were noticeable between the Contractor and the Original Examiners.

Validity of the Numerical Scores for the Detection of Deception

Since there were no significant Level of Confirmation effects, the data were collapsed across Levels of Confirmation for the initial validity analysis. Mean numerical scores generated by the Original Examiners and the Contractor for guilty and innocent suspects are shown in Table 5. The validity of the numerical scores was tested in two ways. Initially, ANOVA was used to test for differences between the numerical scores assigned to innocent and guilty subjects. Then, correlations were calculated between the numerical scores and the guilt criterion to index the discriminative power of the numerical scores.

Table 5. Mean Numerical Scores, Standard Deviations, and Detection Efficiency *r* values for Innocent and Guilty Subjects for the Two Evaluations.

Evaluation	Innocent	Guilty	Detection *r*
Original Examiners	8.36 (3.41) n=11	-8.38 (8.27) n=21	0.76
Contractor	4.27 (7.16) n=11	-12.81 (8.89) n=21	0.71

Note: Standard deviations are shown in parentheses.

The ANOVAs revealed significant main effects for Guilt for the Original Examiners, $F(1, 31) = 40.88$, $p < .001$, and the Contractor, $F(1,31) = 30.18$, $p < .001$. The validity correlations were very strong for both the Original Examiners and the Contractor, accounting for 58% and 50% of the criterion variance, respectively. This is very good performance, and is on a par with the strongest results reported in high quality laboratory studies.

To provide a reasonable comparison to other field studies of the detection of deception, separate analyses were performed on only those examinations that were confirmed by confession. The means from those analyses are shown in Table 6. ANOVA revealed a significant main effect for Guilt in the numerical scores of both the Original Examiners, $F(1, 11) = 78.38$, $p < .001$, and the Contractor, $F(1, 11) = 60.20$, $p < .001$. The correlations assessing the discriminative power of the numerical scores were stronger in this subset, having values of 0.94 and 0.92, for the Original Examiners and the Contractor, respectively.

Table 6. Mean Numerical Scores, Standard Deviations, and Detection Efficiency *r* values for Innocent and Guilty Subjects (Highest Level of Confirmation Only) for the Two Evaluations.

Evaluation	Innocent	Guilty	Detection *r*
Original Examiners	8.00 (4.52) n=6	-13.29 (4.15) n=7	0.94
Contractor	8.17 (5.27) n=6	-15.00 (5.45) n=7	0.92

Note: Standard deviations are shown in parentheses.

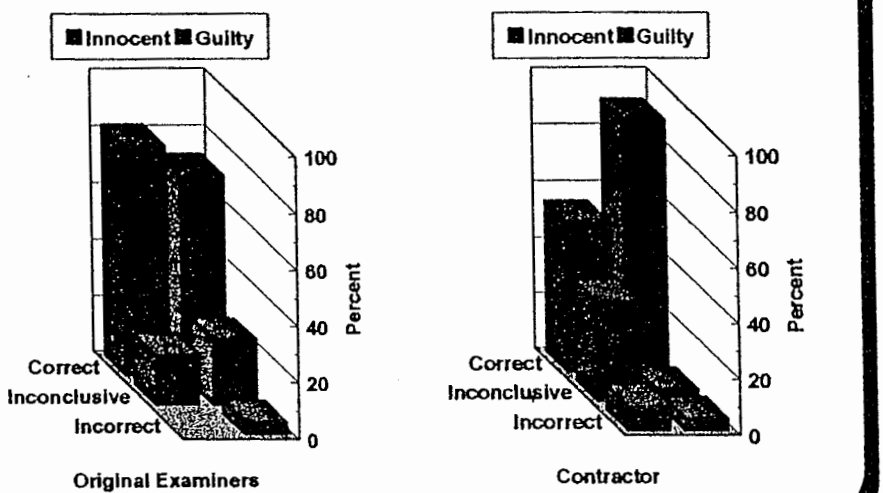
Decisions

Decisions were derived from the numerical scores with the standard field decision rule described earlier. As with the numerical scores, the decisions of the Original Examiners and the Contractor were initially analyzed without reference to Level of Confirmation. The outcomes for the Original Examiners and the Contractor are illustrated in Figure 2. With Innocent suspects, the Original Examiners correctly classified 81.8% (9/11) and called 18.2% (2/11) of the suspects inconclusive. No Innocent suspects were incorrectly classified. With Guilty suspects, the Original Examiners correctly classified 71.4% (15/21) of the suspects, incorrectly classified 4.8% (1/21) and call 23.8% (5/21) of the suspects inconclusive. Excluding inconclusives, 96% of the Original Examiners' decisions were correct.

The classification performance of the Contractor is also illustrated in Figure 2. With Innocent suspects, the Contractor correctly classified 54.5% (6/11), incorrectly classified 9.1% (1/11) called 36.4% (4/11) of the Innocent suspects inconclusive. With Guilty suspects, the Contractor correctly classified 90.5% (19/21) of the subjects, incorrectly classified 4.8% (1/21) and called 4.8% (1/21) of the suspects inconclusive. Excluding inconclusives, 93% of the Contractor's Decisions were correct.

The power of the Original Examiners' and the Contractor's decisions as discriminators of

Figure 2. Outcomes for All Cases With Some Confirmation



truthtellers and deceivers was assessed by coding the guilt criterion (0,1) and the decisions (1=truthful, 2=inconclusive, 3=deceptive) and then correlating the resulting data vectors. This analysis produces a detection efficiency coefficient that is useful in making comparisons of discriminative power across studies (Kircher, Horowitz, & Raskin, 1988). The detection efficiency coefficient for the Original Examiners was 0.81. The detection efficiency coefficient for the Contractor was 0.76. These are very strong detection efficiency coefficients and they indicate that the decision produced by both the Original Examiners and the Contractor were good discriminators of truthtellers and deceivers.

To provide measures of decision accuracy that are comparable to other field studies of the detection of deception, the decision accuracy of the Original Examiners and the Contractor were also analyzed for only those cases with the strongest Level of Confirmation. The confirmation materials for all of these cases contained a confession by the perpetrator of the crime under investigation. The decision results with these 13 cases are presented in Figure 3.

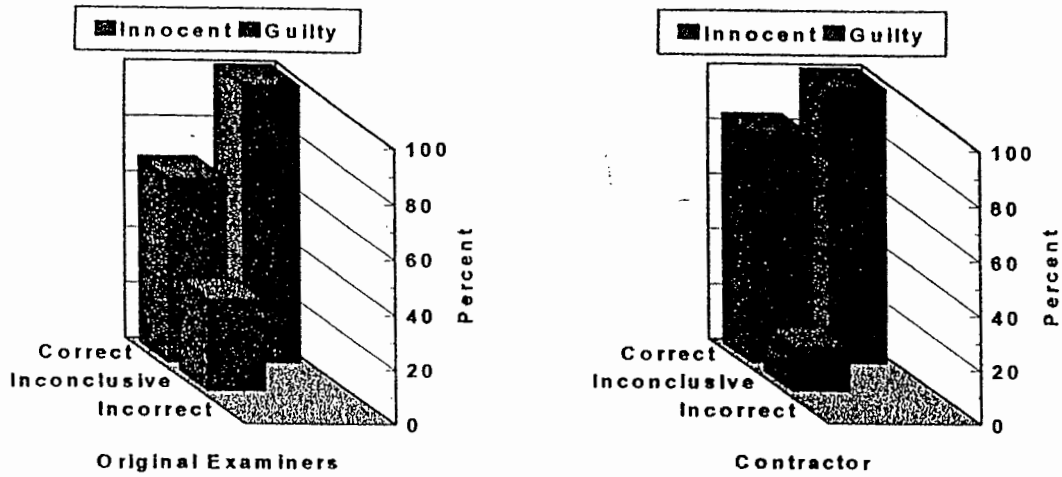
With Innocent suspects, the Original Examiners correctly classified 66.7% (4/6) and called 33.3% (2/6) of the suspects inconclusive. No Innocent suspects were incorrectly classified. With Guilty suspects, the Original Examiners correctly classified 100% (7/7) of the suspects. There were no inconclusive or incorrect outcomes. Excluding inconclusives, 100% of the Original Examiners' decisions were correct.

The classification performance of the Contractor for those subjects with strongest Level of Confirmation is also illustrated in Figure 3. With Innocent suspects, the Contractor correctly classified 83.3% (5/6), and called 16.7% (1/6) of the Innocent suspects inconclusive. There were no incorrect classifications of Innocent suspects. With Guilty suspects, the Contractor correctly classified 100% (7/7) of the subjects. There were no incorrect or inconclusive outcomes. Excluding inconclusive outcomes, 100% of the Contractor's decisions were correct.

Detection efficiency coefficients were also calculated for the Original Examiners' and the Contractor's decisions. The detection efficiency coefficient for the Original Examiners was 0.93. The detection efficiency coefficient for the Contractor was 0.96. This is very strong performance, indicating that performance in this study was as strong as that seen in any study of the Control Question Test for the Psychophysiological Detection of Deception.

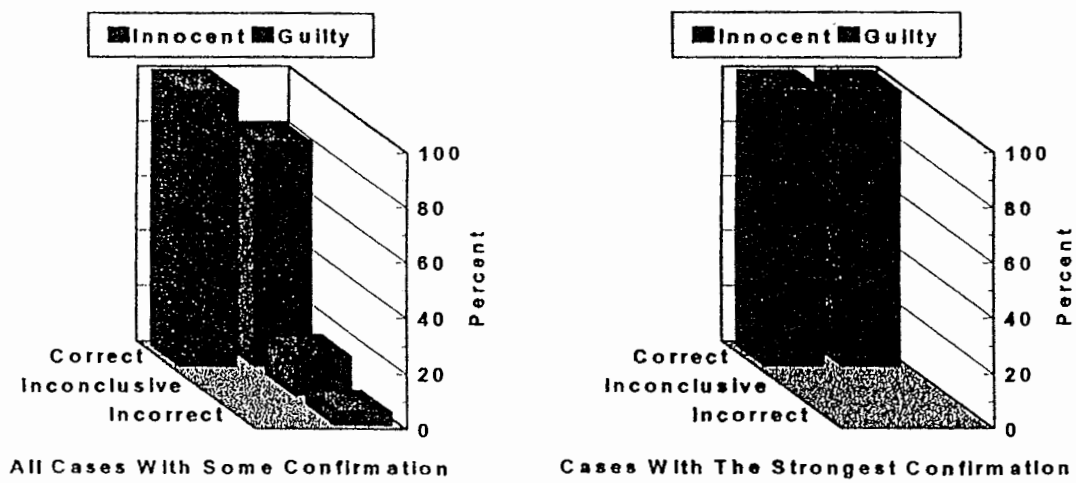
Finally, it is interesting to note that in their field calls the Original Examiners in this study did not always follow the standard decision rule described in this text. In four cases involving two innocent and two guilty suspect, while the total numerical score for the case was inconclusive, the examiner went ahead and rendered a decision. In all four cases those decisions were correct. Performance of the

Figure 3. Outcomes for Cases With the Strongest Level of Confirmation



Original Examiners for their actual field calls is illustrated in Figure 4. With all cases where there was some confirmation, the Original Examiners' field calls were 100% correct with the Innocent, and with the Guilty suspects, they correctly classified 81% (17/21), incorrectly classified 4.8%% (1/21), and called 14.3% (3/21) inconclusive. Excluding inconclusives, 97% of the Original Examiners' decisions were correct. When just the 13 cases with the strongest Level of Confirmation were considered, the field calls of the original examiners were 100% correct with all subjects. There were no incorrect or inconclusive outcomes. For all cases with some confirmation, the detection efficiency scores for the field calls was 0.89. For just the cases with the strongest Level of Confirmation, the detection efficiency coefficient was 1.00.

Figure 4. Accuracy of Field Decisions Made By the Original Examiners



Discussion

The present study was successful in achieving a number of its primary goals. A field study was conducted of the Canadian Police College Polygraph technique that met the following scientific standards:

- Exhaustive Sampling was used without reference to chart quality or the original examiners' decision.
- Independent evaluations based only on the physiological data were made by an experienced evaluator who used the Canadian Police College Polygraph Technique numerical scoring method.
- Confessions were used to provide a group of cases with the strongest possible level of confirmation.
- Explicit Operational Definitions were used to avoid criterion contamination.

However, the present study was only partially successful in exploring the utility of an alternative to the confession criterion.

The secondary goals of the study were only partially addressed. It had been hoped that there would be sufficient data to address the question of the validity of the control question test with victims and other categories of examinees. However, because of the limited number of cases made available for analysis, this secondary goal was not adequately addressed. Although analyses failed to reveal significant differences between suspects and victims, there was only one case in the truthful/victim condition. The lack of cases makes meaningful analysis of this question impossible. Nevertheless, It is interesting to note that there was a 13-point difference in the total numerical scores assigned to this case by the Original Examiner and the Contractor. This may be an interesting difference considering how similar the two sets of scores were in general. Clearly additional research needs to be conducted with categories of subjects other than suspects, and this study only raises additional questions in that area. Due to the limited number of cases, no analyses based on demographic data were attempted.

Four interesting findings were made in this study. First, strong evidence was obtained that the Canadian Police College Polygraph Technique numerical scoring system is very reliable. Inter-rater correlation coefficients were all above 0.9, indicating excellent reliability for the numerical scoring system. This finding is not surprising since the numerical scoring technique taught at the CPC Polygraph Training Unit is based on the numerical scoring system developed at the University of Utah (Podlesny & Raskin, 1978; Raskin & Hare, 1978; Kircher & Raskin, 1988). That system has consistently been found to be highly reliable (Raskin, 1986). However, the present study demonstrates the fact that the reliability of the technique extends to field settings in Canada. This finding is similar to, and complements the results of, Raskin, et al. (1988) with the United States Secret Service. These results indicate that the instructional unit at the Polygraph Training Unit of the Canadian Police College is doing a good job in training examiners to do numerical scoring.

The second finding of interest in this study concerns the validity of polygraph tests conducted in the field by examiners using the Canadian Police College Polygraph Technique. The present study suggests that this technique is very accurate in discriminating truthtellers from deceivers in field settings. Both the numerical scores and the decisions based on them provided strong validity coefficients. The results in this field study are as strong as the best results seen in high quality laboratory studies and in other adequately controlled field studies. When only the cases with the strongest Level of Confirmation were considered, neither the Original Examiners nor the Contractor made any errors of classification. These results suggest that the Canadian Police College Polygraph Technique is a highly valid tool for use in the field for the detection of deception.

It is interesting to consider why the results of this study are in such striking contrast to the results of the study reported by Patrick and Iacono (1991). In the Patrick and Iacono study there were many more false positive errors than there were in the present study. These differences may be due to differences in the respective operational definitions of guilty and innocence, or they may be due to criterion contamination problems as discussed in the introduction of this report. Without examining the actual data from the Patrick and Iacono study, it will not be possible to determine the exact nature of the methodological differences between these studies. Unfortunately, the data from the Patrick and Iacono study are not available to the Canadian Police College Polygraph Training Unit.

The other major difference between this study and the Patrick and Iacono study concerns the difference between the accuracy of the original examiners and the independent evaluators. In the Pat-

rick and Iacono study there was a tremendous loss of accuracy between Original Examiners and the independent evaluators, with the false positive rate being many times higher in the independent evaluations. In the present study the independent evaluations were only slightly less accurate than the Original Examiners. Moreover, when only the cases with the highest level of confirmation were considered, the independent evaluations by the Contractor were slightly more accurate than the evaluations by the original examiners. This result is in sharp contract to Patrick and Iacono's results, but is consistent with the two other field studies that have employed similar methods. In the field studies reported by Honts and Raskin (1988) and by Raskin et al. (1988), the independent evaluations were nearly as accurate as the evaluations of the Original Examiners. It is not possible to know what happened in the Patrick and Iacono study to cause the independent evaluations to be of such low accuracy, but in comparison to other high quality field studies, the Patrick and Iacono study can be seen to be an outlier in this regard, and its value should be assessed accordingly.

The third interesting finding of this study concerns the cases when the Original Examiners in this study chose to break the scoring rules and render a decision when the numerical scores supported only an inconclusive outcome. In such cases in this study, they were always correct in their calls. Admittedly this finding covers only four cases and is of limited generalizability, but it poses some interesting questions. What is it about those four cases that led the examiners to break the rules? Can the information used by the examiners to make that decision to break the rules be objectified and used in a systematic way? The present study has no data to offer to answer these questions, but they are of interest and deserve study.

The fourth interesting finding of the present study concerns the Strength of Confirmation ratings. The results of this study suggest that the process of rating the Strength of Confirmation may be a useful way to approach criterion development in field studies of the detection of deception. The present approach to providing such ratings was highly reliable. Although this does not indicate validity for the approach, it is an important and necessary first step in this direction.

No significant differences were found in numerical scores across the Levels of Confirmation. This finding must be qualified by the fact that the present study is of relatively low power to find such effects, and a null finding under such circumstances has minimal weight. Nevertheless, it is interesting to note that the means for the Medium and High Levels of Confirmation were almost identical. This finding suggests that it may be possible to combine such categories without a loss of accuracy in the criterion. This would be of great benefit in that it would make the acquisition of data from the field much easier and might help to avoid some of the sampling problems that have plagued field studies in this area.

Conclusions and Recommendations

The results of this study are very supportive of the Canadian Police College Polygraph Technique. High estimates were obtained for the both the reliability and validity of the technique. Acceptable scientific methods were followed and significant results of large magnitude were obtained. However, the study is somewhat limited by the small sample size. A larger number of cases would have allowed for additional analyses and for stronger statements about the results.

Recommendations for the CPC Polygraph Training Unit

The present results do not suggest any major changes for the training program at the CPC Polygraph Training Unit. Examiners trained at the CPC Polygraph Training Unit appear to be able to produce highly reliable and valid results with the techniques currently in use. However, the independent evaluations by the Contractor for all cases with some confirmation (see Figure 2) indicate some weakness with innocent subjects, particularly in terms of inconclusives. Other research has indicated that the Directed Lie Control Test offers some advantages over the Control Questions Test (Honts & Raskin, 1988). The CPC Polygraph Training Unit might consider the Directed Lie Control as a supplement to their already strong program. Similarly, computerized evaluations of polygraph charts have shown some advantages with innocent subjects (Kircher & Raskin, 1988; Raskin, et al. 1988). Accelerated application of computerized scoring in the field might also be considered.

Recommendations for Future Research Projects

Based on the present project, the following research projects should be considered for possible support:

- The present study should be replicated with a much larger sample size. The original goal of this study for a sample of 200 cases is reasonable, and could be obtained if full administrative support were available for the project.
- The Level of Confirmation approach to criterion development appears worthwhile and should receive additional study.
- Research should be undertaken to specifically address the testing of victims.
- Research should be undertaken to determine what information examiners use when they decide to break the rules and render opinions that cannot be supported by the obtained numerical scores.

References

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). The validity of detection of deception for multiple issues. *Psychophysiology, 26*, S13. (Abstract)

Barland, G. H., & Raskin, D. C. (1975). An evaluation of field techniques in detection of deception. *Psychophysiology, 12*, 321-330.

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

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

Ginton, A., Netzer, D., Elaad, E., & Ben-Shakhar, G. (1982). A method for evaluating the use of the polygraph in a real-life situation. *Journal of Applied Psychology, 67*, 131-137.

Honts, C. R. (1992). *Interim report: Field validity study of Canadian Police College polygraph technique.* Interim report filed on Contract No. M9010-F107/01ST, Science Branch, Supply and Services Canada. Grand Forks, ND: C. Honts Consultations.

Honts, C. R. (1994). *Research Report: Termination of Data Collection and Statistical Analyses.* Interim report on Science and Services Canada Contract No. M9010-1-F107/01ST, *Field validity study of Canadian police college polygraph technique.* Grand Forks, North Dakota: C. Honts, Consultations.

Honts, C. R., & Perry, M. V. (1992). Polygraph admissibility: Changes and challenges. *Law and Human Behavior, 16*, 357-379 .

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.

Iacono, W. G., & Patrick, C. J. (1988). Assessing deception: Polygraph techniques. In R. Rogers (Ed.), *Clinical assessment of malingering and deception.* New York: Guilford. (205-233).

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. (1982). Cross-validation of a computerized diagnostic procedure for detection of deception. *Psychophysiology, 20,* 568-569 (Abstract).

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. (1982). On the fallibility of lie detection. *Law and Society Review, 17*, 85-104.

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

Munchinsky , P. M. (1993). *Psychology applied to work: An introduction to industrial and organizational psychology, 4th Ed.* Pacific Grove, CA: Brooks/Cole.

Office of Technology Assessment (1983). *Scientific validity of polygraph testing: A research review and evaluation – A technical memorandum (OTA-TM-H-15).* Washington, D. C.: U. S. Government Printing Office.

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.

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

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

Raskin, D. C. (1989). Polygraph techniques for the detection of deception. In D. C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence.* New York: Springer. (247 - 296).

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

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, S. W. (1988). *A Study of the Validity of Polygraph Examinations in Criminal Investigations.* Final Report to the National Institute of Justice, Grant Number 85-IJ-CX- 0400, Salt Lake City: University of Utah, Department of Psychology.

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

Szucko, J. J., & Kleinmuntz, D. (1981). Statistical versus clinical lie detection. *American Psychologist, 36,* 488-496.

Acknowledgments

The author wishes to acknowledge and thank S/Sgt. John W. Kaster of the Polygraph Training Unit of the Canadian Police College for his considerable effort, assistance, and support throughout this project. The project would not have been possible without him. The author also wishes to thank Mary Devitt for her help in the preparation of this report. Finally, the author wishes to thank David Raskin and John Kircher for listening and for their comments on this project.

The comments and conclusions in this report are those of the author. The do not necessarily reflect the official position or policy of the Canadian Government, the Royal Canadian Mounted Police, the Canadian Police College, or the Polygraph Training Unit.

Correspondence regarding the Contractor's role in this project should be directed to the author at the address on the back cover. EMAIL inquires are welcomed.

About the Author

The author is presently an Associate Professor of Psychology at the University of North Dakota, Grand Forks, North Dakota, USA. He has been involved in the psychophysiological detection of deception as either a polygraph examiner or a research scientist since 1976. He is the author of over 100 publications and scientific presentations in the area of credibility assessment and has been recognized as an expert witness in the courts of both Canada and the United States. He has been guest lecturing at the Canadian Police College's Polygraph Training unit periodically since 1986.