A Non-Standard Method for Estimating Accuracy of Lie Detection Techniques, Demonstrated on a Self-Validating Set of Field Polygraph Examinations

Avital Ginton- March 2011

Horowitz-Ginton Credibility Assessment Academy, Bnei-Brak, Israel

Author Note

At the time this research took place Avital Ginton (Police Commander, *retired*) was the head of the Scientific Interrogation Labs and Behavior Section (Investigative Psychology) in the Criminal Identification and Forensic Science Division of Israel Police. Over the years he was also a faculty member in the Department of Psychology and the Law Faculty of Tel-Aviv University and in the Department of Criminology of Bar-Ilan University, in Israel.

A portion of the data presented here was presented by him at the NATO advanced study institute of credibility assessment in Maratea, Italy, June 14-21, 1988.

Correspondences concerning this article should be addressed to Avital Ginton
E-mail: ginton@zahav.net.il

A pre-print version

The final version was published in Psychology, Crime & Law, 2012

Abstract

A unique method for estimating the accuracy of the Comparison Question Test (CQT) - a psychophysiological detection of deception method ("polygraph") - in the field is presented based on a combined probabilistic and algebraic model. It is built on paired examinations in criminal cases in which two opposing versions per case have been subjected to polygraph tests. The developed model is ground-truth free, thus there was no need to rely on external criteria of deception (e.g. confessions or physical evidence) in estimating the accuracy of the CQT. Results indicate an accuracy rate of 0.94 in detecting guilty examinees (Sensitivity) with a 0.06 False Negative rate and an accuracy rate of 0.835 (Specificity) with False Positive of 0.165 for the innocents. These figures excluded 20% of the cases that were ruled inconclusive. When no inconclusive calls were allowed, the accuracy rate dropped down to 0.8 with 0.2 error rates for both the guilty and the innocent examinees. The importance of this research stems from its being a field study that due to the unique developed methodology was not subjected to the weaknesses usually found in other polygraph field validity studies.

This method is applicable for estimating accuracy in other techniques of deception detection and with some necessary adaptations probably also to some eyewitnesses situations.

Key words: Polygraph, CQT, Field paired examinations, Accuracy rates, Self-validating examinations, Detection of Deception..

A Non-Standard Method for Estimating Accuracy of Lie Detection Techniques, Demonstrated on a Self-Validating Set of Field Polygraph Examinations

The use of the polygraph for psychophysiological detection of deception has been controversial within the scientific community for many years, and in particular the Comparison Question Test (CQT; previously known as the Control Question Test; Ben-Shakhar,2002; Raskin & Honts, 2002; NRC, 2003). The CQT is based on comparing psychophysiological responses to relevant questions concerning a crime or behavior of interest with responses to other, mostly provocative questions that are intended to generate physiological reactions. It is presumed that those who are telling the truth to the relevant questions will typically show a greater response to these provocative questions than to the relevant questions, whereas liars will show a greater response to the relevant questions. These provocative comparison questions usually, but not necessarily, relate to some probable past misbehavior of the examinee and are designed to induce deceptive answers from him/her even in case they are telling the truth about the relevant issue.

A main issue in this controversy is the CQT criterion validity, better known as the estimated accuracy of the test. While some have estimated that the test has about 90% accuracy (e.g., Raskin & Honts, 2002), others have suggested a much lower number, or claimed that it is not possible to have any serious estimation due to the lack of a good body of research (Iacono & Lykken,1997; Ben-Shakhar & Furedy, 1990; Ben-Shakhar, 2002). A comprehensive review by the National Research Council of the USA National Academy of

Sciences (2003) pointed out that the general quality of the research regarding this matter is relatively low, and “This situation partly reflects the inherent difficulties of doing high-quality research in this area” (NRC, 2003, p.120). A main aspect of these difficulties relates to the hard-to-bridge dichotomy of laboratory versus field studies found in the body of polygraph research.

Validity studies of polygraph examinations can be grouped into two categories, field studies and analogue studies. Field studies are based on real life examinations, and thus have high ecological validity (Brewer, 2000; Shadish, Cook, & Campbell, 2002). However, they suffer from lack of scientific control over the situation. This is crucial in particular when it comes to the difficulties experienced in obtaining a reliable, objective criterion against which the polygraph results can be validated. Even when it seems that relatively reasonable objective criteria exist - for instance, confessions (Raskin, Kircher, Honts, & Horowitz, 1988) or convictions,[1] - they are prone to various sampling biases, mainly due to a substantial dependency between the polygraph results and the investigation and/or the judicial processes (Patrick & Iacono, 1991, but see Honts, 1996; Krapohl, Shull, & Ryan, 2002). That happens because the actions, steps and directions taken by the investigators on their way to reaching ground truth and solving the case are heavily influenced by polygraph results (NRC, 2003, p. 113-115). For instance, suppose the practice in an investigation unit is to heavily interrogate any suspect who has failed a polygraph test, and to release any suspect who has passed the test. These passing examinees might very well include false negative results in cases that

[1] In order to convict a suspect the guilt must be proven beyond a reasonable doubt. Thus, although not perfect, one can count on convictions to a very high degree to be correct. This is not the case with acquittals, since a reasonable doubt is enough to acquit a suspect and obviously this approach results in quite a number of guilty suspects with false negative error acquittal verdicts.

might never have been solved and therefore no ground truth could have been established for them. Thus, taking only solved cases in which the ground truth have been established would miss all these false negative examinations and of course will bias the sample and the potential consequences drawn from it. Furthermore, let us assume that the interrogators in this unit are very competent and got almost all guilty suspects to confess and even managed to reveal additional evidence to confirm the confessions. Thus, most of the cases in which a suspect has failed a polygraph test, eventually got solved and become a potential case in a pool of known ground truth examinations to be used in validation research. Alas, hypothetically, a chance detection rate of guilty people, which results in Deceptive Indication outcomes in 50% of the guilty examinees might still hide 50% False Negative that will not be shown in the sample.

Moreover, these criteria for ground truth are not immune from error - for instance, in the last two decades the Innocence Project has detailed over 250 wrongful convictions, many of them due to false confessions[2].

On the other hand, analogue studies are conducted in experimental settings (e.g. a mock crime). The subjects are not as psychologically involved in the events under investigation as they would be in the real world. This and especially the lack of realistic fear of failure on the test questions the usefulness of analogue studies for assessing the validity of polygraph examinations in real criminal investigations (Orne, 1975; Ginton, Daie, Elaad, & Ben-Shakhar, 1982). The methodological handicaps found in analogue studies are almost impossible to overcome within the conventional behavioral science approach (but, see Ginton, et al., 1982), especially if one considers the limitations imposed by ethical standards.

[2] The Innocence Project is an US national litigation and public policy organization dedicated to exonerating wrongfully convicted people through DNA testing and reforming the criminal justice system to prevent future injustice. http://www.innocenceproject.org/

In order to apply any analogue polygraph research to real world situations, all examinations must have been perceived by the examinees as real life exams with real consequences. This also applies to the examiners, but to a lesser degree. This perspective suggests that the most informative work will derive from actual field examinations. Alas, actual field research suffers from the lack of confidence that the ground truth is fully established independent of the polygraph results and their effects, and from probable sampling bias, suggesting that the most informative work will derive from examinations that have been conducted under a good laboratory set up. This internal-external validity tradeoff cannot be advanced in one direction or the next unless some ethically problematic steps are taken. For instance, one can run field polygraph tests when there is no real need to find out whether the examinee is telling the truth or lies, because that has already been fully established by other means (unaware of by the examinees and the examiners). Or, one could deceive the examinees into believing that they are taking a real polygraph exam -when in fact it is a laboratory experiment in which the subject has not given their consent in advance to participate in the research (Ginton, et al., 1982).

In a study presented at a conference in 1985[3], Ginton introduced a new approach which made it possible to estimate the general accuracy of the CQT using a set of real life polygraph examinations, while avoiding the above mentioned substantial flaws (Ginton, 1986). The distinctive novel characteristic of that study was that it was able to deduce a valid accuracy estimate of the CQT from the polygraph outcomes of a set of real life paired polygraph examinations despite not knowing and not relying on the actual objective truth of any

[3] IDENTA 85 - The international congress on techniques for criminal identification & counter terrorism. Jerusalem, Israel 1985.

particular case. The figures computed in that study, were an accuracy of 0.805 and an error rate of 0.195 (Ginton, 1986; Ginton, 2009).

In that study the accuracy rate was an overall estimation with no differentiation between the estimated accuracy in detecting liars or identifying truth tellers. However, it is well documented that there might be a difference between the rates of false positive and false negative errors (e.g. see OTA, 1983).

The present study, of which some preliminary results were presented elsewhere (Ginton 1988)[4], elaborates upon the methodology of the Ginton (1986) study by differentiating the accuracy rates for detecting deceptive examinees from the accuracy for detecting truth telling examinees.

The study is based on a pool of CQT field examinations that have been conducted in cases where two examinees presented contradicting versions related to the same criminal event. The contradicting versions could take the form of accusing each other of doing a certain act (e.g. driving the car during a hit and run accident when two people who were in the car got caught later.) or that one party accused the other of doing something that was totally denied by him/her (e.g. “I saw him stealing my bicycle.”)

In the process of selecting the proper cases for the study, the background of each case was carefully analyzed to eliminate cases in which it was conceivable to assume that the two parties might have told their subjective truth or that both of them were lying. Thus, only cases that instilled strong confidence *a priori* that each pair consisted of one liar and one truthteller were kept for further analysis. Following this procedure it was possible to estimate the

---

[4] In 1988 the research model was presented with some preliminary results, in a close forum, however, it has never been published in the scientific literature, and remains unheard and unknown to most of the relevant community.

accuracy of the CQT by mathematical computations based upon the mere proportions of the various results obtained by the polygraph examinations per-se. Thus, there was no need to rely on criteria such as confessions, convictions or any other sort of verification that may be prone not only to substantial sampling bias but may also suffer from incongruity with the real truth (e.g. false confessions).

Method

Selection of cases

All the records of CQT examinations conducted in two main polygraph laboratories of the Israel National Police, during the years 1982-1985, were screened for cases in which two individuals presented opposite versions concerning an incident under polygraph examination. There were about 250 cases that had two examinations per case. It was important to avoid cases in which the two opposite versions could present the subjective truth of the two individuals, or that both of them were deceptive in their versions. Therefore all the cases went through a selection procedure carried out by two experienced polygraph experts. These experts were blind to the results of the polygraph test. They independently analyzed all the facts involved in each incident carefully, including any full detailed version provided by each individual that could be found in the files (excluding of course the results of the polygraph tests). The experts were instructed to select only cases which it was clear enough from the background material that one and only one individual must have been telling the truth on the relevant questions in the test, while the other individual had been lying, - although it was not possible to know which individual had told the truth or lied. Only cases that had been selected independently by both experts were kept for further analysis. That is, if one of the two experts could have imagine a reasonable scenario in which both parties could had been told their subjective truth in their versions or that both had been lying, the case was deselected for further analysis.

To further clarify the screening procedure the following example is given: In a case of a hit and run accident the involved car was found a few hours later. Three eyewitnesses gave statements that at the time of the accident two people were in the car, the driver and another person sitting next to him. The owner of the car was located and admitted being in the hitting car, however he denied being the driver. According to his account, because he was tired his friend took over and drove the car. Contrary to that, the friend, while admitting being in the car when the accident took place denied being the driver when the accident occurred and insisted that the owner of the car drove it by himself at that point. The 3 eyewitnesses were not sure who the driver was. The two guys took a polygraph test in which they were asked "Did you drive the car at the time of the accident?" and "Was your friend (x) driving the car at the time of the accident?".

Based on the above, assuming the improbability of a situation in which an unknown third person drove the car and committed the accident, or the occurrence of a pathological twist in the memory of one of them, one can safely deduce that in this pair of examinees <u>one and only one</u> individual must have been telling the truth to the relevant questions on the test, while the other individual had been lying - <u>although it was not possible to know which individual had told the truth or lied.</u>

The experts were very conservative in their decisions in the screening process and disqualified most of the cases by hypothesizing reasonable scenarios in which both parties could had been told their subjective truth when answering the relevant questions on the polygraph test or that both could have been lying. Only 64 cases (128 examinations) survived the selection and were used in the present research. These cases featured sexual offenses, thefts, frauds, police brutality and a few miscellaneous issues.

Scoring

Overall the 128 polygraph records were blindly numerically scored by eight experienced examiners, using the standard seven-point scale per comparison point, ranging from +3 to -3. This scoring technique is routine at the Polygraph Unit of Israel police (e.g. Elaad, Ginton, & Ben-Shakhar, 1998). Each particular record was scored by three examiners. One examiner scored all the 128 records, whereas the number of records scored by the others varied from 22 to 72, with a mean of 36.5 and a median of 28. The scorers were completely blind to the paired nature of the tests as well as the details of cases involved. For detailed explanation of the numerical scoring technique the reader is referred to Barland and Raskin (1975). Nevertheless, it is important to mention that in the end of the scoring, each record receives a final numerical score. A negative score indicates that the physiological responses given to the relevant questions were on the whole stronger than those given to the comparison questions, thus indicating deception to the relevant issue. On the other hand, a positive final score indicates that the physiological responses given to the comparison questions were on the whole stronger than those given to the relevant questions, thus indicating truthfulness. One should note that a negative score means a positive result – that is, because the test is about detecting deception so a positive result indicates that the person is likely deceptive, whereas the positive score means a negative result (the person is likely truthful). It should be mentioned that unlike the present study, an inconclusive zone is usually applied by examiners as a measure of precaution. Thus, in order to reach a conclusion, the score must exceed a certain cut-off point or else it is considered an inconclusive result.

In these paired cases, 100% accuracy should show an outcome such that each pair should feature one positive and one negative score (opposite outcomes). However, since we might expect errors, some pairs might produce either two negative or two positive scores (similar outcomes). It is obvious that the proportion of pairs with similar outcomes is negatively related to the accuracy of the CQT. Thus, the more accurate the technique differentiates those who are lying from those who are telling the truth, the smaller the proportion of pairs with similar outcomes. Hence, given the proportion of the various outcomes of pairs, it is possible to calculate the accuracy of the CQT, even in the absence of any ground truth information concerning the cases.

The computational model

Let us define the probability of detecting a guilty person, [True Positive], as P(TP)=X and the probability of classifying a guilty person as truthful, (False Negative), as P(FN)=W.

When a decision is made regarding a guilty person, it could be either a correct decision or a wrong one. In probability terms it means that P(TP)+P(FN)=X+W=1. In a similar manner, we can define the probability of classifying an innocent person as truthful, [True Negative], P(TN)=Y and the probability of falsely classifying an innocent person as deceptive, [False Positive], P(FP)=Z . For the innocent examinees it follows that P(TN)+P(FP) =Y+Z=1.

Assuming the examinations of the two individuals in each pair were conducted independently, the probability of having correct decisions for both parties equals the multiplication product of their two separate probabilities i.e. X * Y or XY, and the probability of having incorrect decisions for the two of them equals W*Z or WZ. Following the same logic, it is clear that the probability of having pairs with two deceptive outcomes equals XZ, (the probability of True Positive – X, multiplied by the probability of False Positive – Z.), while YW expresses the probability of getting pairs with two non-deceptive results.

A pair of examinations with opposite outcomes means either two correct decisions or two mistakes. Hence, the proportion of pairs with opposite outcomes equals to the united probabilities of making two correct decisions or two mistakes, i.e. XY+WZ. Similarly, a pair with two negative numerical scores [two deception indicated - DI results] means one true and one false deception results. The proportion of such pairs in the present sample, equals the probability of having such pairs, i.e. XZ. Following the same logic, YW is equal to the proportion of pairs with two positive numerical scores (i.e. two non-deception indicated, NDI, results). For further clarification of this model, see table 1.

Probability of getting a certain outcome of pairs (pairs with one deceptive and one non deceptive, two deceptive or two non-deceptive outcomes) should be manifested in the proportion of such pairs in the sample.

From this model it follows that, given the percentage of pairs of each of the three possible combinations of outcomes found in the sample, one can extract the values of X,Y,W and Z, which indicate the accuracy and the error rates of CQT in detection of deception.
A hypothetical example will illustrate how this works. Supposed we know that the accuracy rate in detecting liars- True Positive - is 0.9 with a False Negative error rate of 0.1, while the accuracy of correctly identifying truth tellers – True Negative - is 0.8 with a False Positive error rate of 0.2. When dealing with paired examinations, it is expected that the probability of correctly identifying both parties (one deceptive and one truth teller) is the intersection of the

Table1

Set of probabilities for yielding various outcomes in paired polygraph examinations of opposing versions.

Within Guilty Examinees	Within Innocent Examinees	Within Pairs of Exams: Opposite Outcomes	Within Pairs of Exams: Similar Outcomes
Probability of detecting a guilty person (True Positive) P(TP) = X Correct outcome	Probability of classifying an innocent person as non-deceptive (True Negative) P(TN) = Y Correct outcome	Probability of getting pairs with two correct outcomes (one true deceptive & one true non-deceptive) (Two True) P(2T) = XY	Probability of getting pairs with two deceptive outcomes (one true & one false positive) (True & False Positives) P(TP&FP) = XZ
Probability of falsely classifying a guilty person as innocent, (False Negative) P(FN) = W Incorrect outcome	Probability of falsely classifying an innocent person as deceptive, (False Positive) P(FP) = Z Incorrect outcome	Probability of getting pairs with two incorrect outcomes (one false deceptive & one false non-deceptive) (Two False) P(2F) = WZ	Probability of getting pairs with two non-deceptive outcomes (one true & one false negative) (True & False Negatives) P(TN&FN) = YW
P(TP) + P(FN) = = X+W=1	P(TN) + P(FP) = = Y+Z=1	P(2T) + P(2F) + P(TP&FP) + P(TN&FN)= = XY+WZ+XZ+YW = 1	

two probabilities, i.e. 0.9 * 0.8 = 0.72. Thus, only 72% of the pairs are expected to result in two correct outcomes. Also it is expected that in 2% (0.1*0.2=0.02) both outcomes will be errors. The expected percentage of pairs with one True and one False Positive is 18% (0.9*0.2=0.18) and with one True and one False Negative is 8% (0.8*0.1=0.08).

The proportions of these four types of pairs are the results of the four specific values of True or False Positive and Negative rates. Thus, it is possible to unfold backward from given proportions of the four categories of pairs within a certain sample and estimate the accuracy as well as the error rates of the test used in the sample. The way to do it is by using algebraic equations as describe above in Table 1.

Preparing the data for analysis

Since each record was scored by three examiners it was not likely that reliability would be 100%. It was also expected that for some records the three scores might not even agree on a positive or negative score. . To deal with this, the data was prepared for the analysis in two alternative ways to enable two separate parallel analyses:

1) Mean scores were computed for each record by averaging the three scores given to them. The signs of these arithmetic means served for indicating whether the records should be considered as deception indicated (DI) or non-deception indicated (NDI) records. Whenever the mean was zero (6 cases), the direction of the record was established by the majority (5 cases) or by adding another scorer (1 case). For example, a record which had been scored by three different scorers as +1; -3; +2, produced a 0 mean score, however since two of the scores were positive, the record were treated as a positive one. The extra scorer was employed for the one record that had been scored +1; -1; 0.

2) Records in which the direction pointed by two scorers contradicted each other (plus vs. minus), were considered to be non-indicative records, and therefore they were excluded from further computations. These "non-indicative" exams are in a way the equivalent of the Inconclusive term usually applied in the numerical scoring technique when the total final score of an exam doest not exceed a certain cutoff point.

There were 26 such examinations belong to 25 criminal cases, (in one case the examinations of both parties were non-indicative).

Results

As mentioned above, each record was scored blindly by three examiners. Reliability of the scoring was established by computing the inter-correlations among these three sets of scores across the 128 records. As can be seen in Table 2, the scoring was highly reliable with reliability coefficients around 0.9 and percentage of agreements regarding the direction pointed by the scores around 84%.

Table 2
Inter-correlations* between 3 sets of scores given to the same 128 records. In brackets the percentage of agreements regarding the direction pointed by the scores.

	1st set of scores	2nd set of scores	3rd set of scores
1st set of scores	-	.919(86%)	.87(83%)
2nd set of scores	.919(86%)	-	.88(84%)
3rd set of scores	.87(83%)	.88(84%)	-

*. Scores were grouped into two categories, negative vs. non-negative scores and Tettrachoric correlation method for fourfold table, was applied (Ferguson 1966).

Table 3 presents the distribution of the various pairs of records, according to their outcomes. The data presented in the upper row is based on the arithmetic means of the three scores given to each record. That resulted in 44 pairs of records with opposite outcomes, [one positive and one negative], and 20 pairs with similar outcomes – of them 10 pairs with two positive outcomes and 10 pairs with two negative results. The proportion of opposite to similar outcomes is significantly different from that expected by chance fluctuations alone (Binomial Test of Hypothesis p=0.004, two-tail). This indicates that the CQT did make some distinctions between deceptive and non-deceptive examinees. However in order to estimate the accuracy of the test, a computation based on the aforementioned algebraic model is needed.

Table 3
Distribution of the various pairs of records, for the two alternative ways of analysis, according to their outcomes

	Pairs with Opposite Outcomes	Pairs with Similar Outcomes		Total
		Two DI Outcomes (One F.P.)	Two NDI Outcomes (One F.N.)	
All records	44	10	10	64
Indicative Records only	31	6	2	39*

* 26 records (20.3%) belong to 25 different pairs were considered as non-indicative records. See text for further explanation.

Those values derived from the data based on the 1st method of analysis, were then plugged into the model, producing the following equations :

$$XY + WZ = 44/64$$

$$XZ = 10/64$$

$$YW = 10/64$$

A simple algebraic computation led to the following solution:

$$X = 0.8 \; ; \; W = 0.2$$

$$Y = 0.8 \; ; \; Z = 0.2$$

A detailed, step by step solution in shown in Appendix 1.

The model thus predicts an accuracy rate of 80% for both the guilty and the innocent examinees, with an error rate of 20%.

As shown above, the inter judge reliability was not perfect and on occasion there were records that the three scorers produced scores in different directions. Practically it meant that those records were hardly indicative, and better treated as inconclusive. There were 26 such records or 20.3% of the sample. The 26 records belong to 25 different pairs of examinations (in one case the examinations of both parties were non-indicative).[5] All the pairs with non–indicative records (one or two per pair) were removed from further analysis (2nd method of analysis). Excluding these pairs that contained "non-indicative" records, only 39 pairs were left. The data presented in the lower row of Table 3, gives the distribution of these remaining pairs of records according to their outcomes. Thirty-one of them consisted of pairs with opposite outcomes, and 8 pairs with similar outcomes – of them, 6 pairs with two deceptive records and 2 pairs with two non-deceptive records. Here too, the proportion of opposite to

[5] The probability of getting independently 26 records belong to 25 different pairs or more, is about 0.21. So, though it is relatively a rare event it is still within acceptable limits.

identical outcomes is significantly different from that expected by chance fluctuations only (Binomial Test of Hypothesis, p=0.0004, two-tail).

For estimating the accuracy the following equations were derived from these data:

$$XY + WZ = 31/39$$

$$XZ = 6/39$$

$$YW = 2/39$$

The solution of these equations indicated that:

$$X = 0.94 \ ; W = 0.06$$

$$Y = 0.835 \ ; Z = 0.165$$

These results indicate that when excluding the tests that were not indicative (20.3%), the accuracy rates increased for both the guilty and the innocents, but not symmetrically. The accuracy rate for guilty examinees rose to 94%, but the accurate rate for innocents was only 83.5%, with 6% false negatives and 16.5% false positives respectively.

The internal validity of the study is dependent upon the assumption that the two examinations in each pair were conducted independently of each other. Otherwise, one cannot exclude the possibility that knowledge of the result of the first examination in each pair might have influenced the manner in which the second examination is conducted. This would produce an artificial increase in the probability of obtaining an opposite outcome. This is especially acute when the two parties are tested by the same examiner. An analysis of the cases indicates that this concern was not founded. In the present study 35 pairs of examinations had been originally conducted by two different polygraph examiners per pair and only 29 pairs had been conducted by a single examiner per pair. Table 4 shows the break down of all the pairs and their blind scoring classification according to these two categories. It is evident that no real difference exists between them (χ^2 = .09, *df* = 1, *p* > .75). In addition, in many of the cases that involved two examiners, the examinations took place parallel in time, eliminating any potential influence of one outcome on the other. It should be mentioned however that since both examiners had been exposed to the same background information, the existence of some sort of interdependency can not be totally ignored. On the other hand, there are some indications that the polygraph examiners of Israel Police are relatively resistant to bias caused by background information (Elaad, Ginton & Ben-Shakhar 1998). So, for any practical matter it seems that the 2 examinations that have been conducted by two examiners per pair should be treated as if they were conducted totally independent of each other.

Table 4
Distribution of outcomes in paired examinations (mean of 3 scores per examination- see 1st method of analysis), conducted by a single examiner per pair or two different examiners.

	Single examiner per pair	Two examiners per pair	Total
Opposite outcomes	21	23	44
Similar outcomes	8	12	20
Total	29	35	64

Note. The distributions of outcomes is not significantly different between pairs conducted by single or two different examiners (χ^2 with Yates correction = .09, *df* = 1, *p* = .76)

Discussion

Another important question, given the specific procedure that was applied in choosing the subset of polygraph examinations for the analysis is whether the accuracy rates found in the present study can be generalized to all CQT polygraph field examinations? Can this subset be treated as a valid representation of the CQT polygraph examinations as a whole with regard to their accuracy? Actually two questions are hidden here. The first is - can we consider the cases of paired examinations as a proper representation of the total volume of CQT examinations *of which most are* single examinee's cases? The other question is - can we accept the specific sample of paired cases used in the study as an unbiased sample of the paired examinations cases in general, or is it possible that the screening procedure within the paired cases has resulted in a systematic bias that affected the estimated accuracy rates?

It is the author's belief that in the paired cases the tension or anxiety felt by the examinees are greater than usual because these individuals are confronted by another person's direct accusation that they are the guilty or deceiving party. Such a direct accusation usually bears a heavier emotional load compared to being a suspect due to mere circumstantial evidence. That in turn may increase the odds that the innocent persons would produce false positive outcomes.[6]

This elevated emotional load when it comes to a guilty person, increases the probability of having a confession before a polygraph examination takes place. Thus the guilty people who took these examinations (i.e. did not confess before the test) have already proved their

[6] One can raise the point that also in a large proportion of single examinee cases, the suspect is accused by someone who is very unlikely to be a suspect, providing the same anxiety as in the paired cases. But as a matter of fact all these cases are potentially paired examinees cases, because the accused suspect's denial of being guilty, might means that the accusing person become an alleged suspect of giving false testimony or complain.

resistance to interrogation. They should probably be considered "better liars" on average compared to a typical person, and therefore we might expect a higher rate of false negatives. Taken together, these suggest that the estimates derived from the above model are conservative estimates of the CQT accuracy in general, in real life situations.

With regard to the concern that the screening procedure within the paired cases has been resulted in a systematic bias that affected the found accuracy rate, it should be stressed that the subset of polygraph examinations that were chosen for the analysis, comprised of all the cases which based on the background material, seemed to be safe enough to assume that in each pair there must have been one truth-teller and one liar to the relevant questions. Although this effort has reduced the number of cases used for estimating the accuracy to approximately 25% of the total number of paired cases, the author sees no reason to believe that this has created a biased subset of cases resulting in a biased accuracy rates and the findings of the present study can be generalized to CQT polygraph field examinations as a whole.

As mentioned above, there is an inherent difficulty in getting valid estimations of the actual accuracy of polygraph examinations. The two sets of common research paradigms, namely field studies and analogue (laboratory) studies suffer from different inherent weaknesses that limit the ability to deduce from them a valid estimation of the actual accuracy rate of the CQT (Orne, 1975; Ginton, Daie, Elaad, & Ben-Shakhar, 1982; OTA, 1983; NRC, 2003). In light of these limitations, the importance of the present research stems from its being a field study that, due to the unique developed methodology, was not subjected to the main weaknesses usually found in other field validity studies, namely, the difficulties in obtaining a reliable and unbiased objective criterion against which the polygraph results can be validated. Interestingly enough, the accuracy estimates derived from the present analysis converge nicely with meta-analysis of traditional confession criterion field studies (Honts & Peterson, 1997; Raskin & Honts, 2002; NRC 2003; Honts & Schweinle, 2009), and by that reinforces the strength of the outcomes.

Assuming it is possible to generalize the accuracy rates found in the present study to the more common situation in which a polygraph examination does not have a counterpart examination of a person holding the opposite version, the probability of a certain polygraph outcome to be correct is estimated by using a conditional probability computation (Bayes formula). In most cases that the prior probability of an examinee being guilty or innocent is completely unknown, one should treat it as an equal probability (i.e. P = 0.5). Statistically this is similar to the cases in the present study, in which the prior probability of each party in a pair, to be the guilty or the innocent person is equal, and by definition 50% of the examinees are guilty and of course 50% are innocent. In that situation, based on the accuracy rate found in the present study, the conditional probability that a given single polygraph result of NDI is correct (excluding inconclusive/nonindicative outcomes) is 0.93 , and for a DI result it is 0.85.[7]

However, whenever there is a reasonable way to estimate the prior probability of guilt vs. innocence, one should incorporate it into the computation. For example, for years in the polygraph laboratories of the Israeli Police, the estimated prior probability of being guilty has

[7] A NDI outcome can be a True Negative, P = 0.835, or a False Negative, P = 0.06. Since the prior probability of telling the truth or being deceptive is 0.5, the weight given to both options are the same. Thus, the conditional probability that this specific outcome is actually correct equals 0.835/0.835+0.06 = 0.835/0.895 = 0.93. A DI outcome can be a True Positive, P = 0.94, or a False Positive, P=0.165. Thus, the conditional probability that this specific outcome is actually correct equals 0.94/0.94+0.135 = 0.94/1.105 = 0.85

been approximately 0.25 and being innocent 0.75.[8] Incorporating these estimations into the Bayes formula would yield a probability of 0.977 for a NDI result to be correct, and of only 0.655 for a DI result to be correct.

From a practical point of view, the idea to treat the polygraph result of one party in a pair as a sort of verification test for the polygraph result of the other party, is suitable not only for research purposes but also for the polygraph usage in the field. Thus, whenever the polygraph outcomes of the two parties point at the same direction (both deceptive or both non-deceptive), one should treat it as a warning signal against taking any clear cut measure based on any of the polygraph outcomes in that case, particularly if the case analysis can not see any possibility that both parties are lying or telling their subjective truth. On the other hand whenever the two independent results indicate opposite directions, one can rely on them to a higher degree than in the case of a single test. This rational has been manifested in Israeli Police Polygraph laboratories practice and guidance for years (Ginton & Zoltak, 1991) and lately, in a similar manner, in the Marin Protocol adopted by the American Polygraph Association (Marin, 2000; Krapohl, 2005; APA-Model Policy, 2007).

In line with this rational, based on the estimated accuracies presented in the study, if we don't use any inconclusive 'non-indicative' decisions as in the first method of analysis, the probability of two opposite results to be correct is 0.94[9] and based on the second method of

[8] Personal knowledge based on internal annual reports of Israel Police Polygraph Labs.

[9] Two Opposite Outcomes can either be two correct results or two mistakes. Probability of getting two correct results is $0.8^2 = 0.64$. Probability of getting two incorrect results is $0.2^2 = 0.04$. Probability of two correct results, given two Opposite Outcomes is $^{0.64}/_{0.64+0.04} = ^{0.64}/_{0.68} = 0.94$

analysis that eliminates the 20% “non-indicative” results from computation, this probability rises to 0.987[10].

Finally, the presented genuine method of criterion validity analysis that was not dependent upon an external criterion of guilt or innocence is not limited to polygraph. The same kind of analysis could be applied easily to other areas in deception detection and perhaps even to widely disparate areas like eyewitness identification.

[10] Two Opposite Outcomes can either be two correct results or two mistakes. Probability of getting two correct results is 0.94 * 0.835 = 0.785. Probability of getting two incorrect results is 0.06 * 0.165 = 0.0099. Probability of two correct results, given two Opposite Outcomes is $^{0.785}/0.785+0.0099 = ^{0.785}/0.7949 = 0.987$

References

APA-Model Policy, (2007). Model policy for paired testing ,*Polygraph* ,36(2), 117-120.

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

Ben-Shakhar, G. (2002). A critical review of the control questions test (CQT). In M. Kleiner (Ed.), *Handbook of Polygraph Testing* (pp. 103-126). New York: Academic Press.

Ben-Shakhar, G., & Furedy, J. J. (1990). *Theories and applications in the detection of deception*. New York: Springer-Verlag.

Brewer, M. (2000). Research design and issues of validity. In H. Reis & C. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 3-16). Cambridge: Cambridge University Press.

Elaad, E., Ginton, A., & Ben-Shakhar, G. (1998). The role of prior expectations in polygraph examiners' decisions. *Psychology, Crime and Law*, 4, 1-16.

Ferguson, G. A. (1966). *Statistical analysis in psychology and education.* New York: McGraw-Hill Companies.

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

Ginton, A. (1986). A built-in validity in polygraph field examinations. In: *Anti-Terrorism, Forensic Science & Psychology in Police Investigations,* (pp. 148-154). Boulder, CO: Westview Press; Jerusalem: Heiliger and Co. (Proceedings of IDENTA 85. The international congress on techniques for criminal identification & counter terrorism. Jerusalem, Israel, 1985.)

Ginton, A. (1988). Estimating accuracy of the control-questions-test, using a self validating set of real-life polygraph examinations. *Preliminary oral report at the NATO advanced study institute of credibility assessment*, Maratea, Italy, June 14-21, 1988.

Ginton, A. (2009). A validity study of the Comparison Question Test, based on internal criterion in field paired examinations. *Polygraph*, 38(4), 275-280.

Honts, C. R. (1996). Criterion development and validity of the CQT in field application. *Journal of General Psychology*, 16, 56-61.

Honts, C. R. & Peterson, C. F. (1997). Brief of the Committee of Concerned Social Scientists as Amicus Curiae, U.S. Vs. Scheffer, in the Supreme Court of the United States..

Honts, C. R., & Schweinle, (2009). Information gain of psychophysiological detection of deception in forensic and screening settings. *Applied Psychophysiology and Biofeedback*, *34*, 161-172. (Available online July 2009).

Iacono,W. G., & Lykken,D. T. (1997). The scientific status of research on polygraph techniques: The case against polygraph tests. In D. L. Faigman, D.Kaye, M. J. Saks & J. Sanders (Eds.), *Modern Scientific Evidence: The law and science of expert testimony* (pp. 582-618). St Paul, Minnesota: West Publishing.

Krapohl, D. (2005). Polygraph decision rules for evidentiary and paired-testing (Martin Protocol) applications. *Polygraph*, 34(3), 184-195.

Krapohl, J. D., Shull, K.W., & Ryan, A. A. (2002). Does the confession criterion in case selection inflate polygraph accuracy estimates?. *Forensic Science Communications*, 4(3), 234-244.

Martin, J., (2000). He said/ She said: Polygraph Evidence in Court. *Polygraph*, 29(4), 299-304.

Orne, M. T. (1975). Implications of laboratory research for the detection of deception. In N. Ansley (Ed.), *Legal Admissibility of the Polygraph* (pp. 94-119). Springfield, IL: Charles C. Thomas Publishers.

OTA, U.S. Congress. (1983). *Scientific validity of polygraph testing: A research review and evaluation*. Washington D.C: Office of Technology Assessment.

NRC – National Research Council. (2003). *The Polygraph and Lie Detection, Committee to Review the Scientific Evidence on the Polygraph*. Washington, D. C.: The National Academic Press.

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

Raskin, D. C., Kircher, J. C., Honts, C. R., & Horowitz, M. S. (1988). *A study of the validity of polygraph examinations in criminal investigations*. Final report to the National Institute of Justice. Salt Lake City: Department of Psychology, University of Utah.

Raskin, D.C., & Honts, C. R. (2002). The comparison question test. In M. Kleiner, (Ed.) *Handbook of Polygraph Testing* (pp. 1-48). New York: Academic Press.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Company.

Appendix 1

A step by step solution of the algebraic equations.

The model involved 6 algebraic equations. Three basic equations present the algebraic model and another 3 present the application of the model to the specific data derived from the sample.

The 3 basic equations are:

Within Guilty Examinees - Equation (a) $P_{(TP)} + P_{(FN)} = X+W=1$

Within Innocent Examinees - Equation (b) $P_{(TN)} + P_{(FP)} = Y+Z=1$

Within Pairs of Examinations - Equation (c) $P_{(2T)} + P_{(2F)} + P_{(TP\&FP)} + P_{(TN\&FN)}$

$= XY+WZ+XZ+YW = 1$

The 3 equations based on the specific data derived from the 1st way of analysis, are:

Equation (d) $XY + WZ = 44/64=0.6875$

Equation (e) $XZ = 10/64 = 0.15625$

Equation (f) $YW = 10/64= 0.15625$

When based on data derived from the 2nd way of analysis, the 3 equations are:

Equation (d)' $XY + WZ = 31/39=0.7948718$

Equation (e)' $XZ = 6/39= 0.1538461$

Equation (f)' $YW = 2/39= 0.051282$

Solution in steps:

1. Equation (a) $X+W=1$
   $W=1-X \quad ; X=1-W$

2. Equation (b) $Y+Z=1$
   $Z=1-Y \quad ; Y=1-Z$

Hence:

3. $WZ=(1-X)(1-Y); \qquad XY=(1-W)(1-Z)$

Based on the 1st way of analysis:

4. Equation (d)

$$XY + WZ = 44/64= 0.6875$$
$$XY + (1-X)(1-Y) = 44/64= 0.6875$$
$$XY+1-Y-X+XY = 0.6875$$
$$2XY-X-Y+0.3125=0$$
$$X+Y-2XY-0.3125=0$$

And

$$WZ+(1-W)(1-Z)=44/64=0.6875$$
$$WZ+1-Z-W+WZ=0.6875$$
$$2WZ-W-Z+0.3125=0$$
$$W+Z-2WZ-0.3125=0$$

5. Equation (e) $XZ=0.15625$
   $X=0.15625/Z \quad ; \quad Z = 0.15625/X$

6. Equation (f) $YW = 0.15625$
   $Y = 0.15625/W \; ; \quad W = 0.15625/Y$

Hence

$$XY = (0.15625/Z)(0.15625/W)$$

And

$$WZ = (0.15625/X)(0.15625/Y)$$

8. Back to Equation (d):

$XY + WZ = 44/64= 0.6875$

$XY+ (0.15625/X)(0.15625/Y) =0.6875$

$XY+ 0.024414/XY = 0.6875$

Multiplying both sides of the equation by XY:

$XY(XY+0.024414/XY) = 0.6875XY$

$XY^2+0.024414 = 0.6875XY$

$(XY)^2-0.6875(XY)+0.024414 = 0$

9. Quadratic Equation Solution:

$(XY)_1 = 0.6499363199099528$

$(XY)_2 = 0.037563680090047125$

Note- Algebraic solutions of Quadratic Equations produce two alternative solutions with two different figures. In our case it means that the figures of the accuracy and error rates can be switched over and still produces the same specific distribution of the paired polygraph outcomes in the sample. Namely, from the algebraic perspective it does not matter whether the accuracy is for instance 0.8 with an error rate of 0.2, or the opposite way around, the accuracy is only 0.2 and the error rate is 0.8.

Obviously, from our perspective we reject the second option and stay with the first solution:

$(XY)_1 = 0.6499363199099528$

10. From the distribution of the outcomes it is clear that in this sample the False Positive and False Negative rates were the same (same proportion of pairs with negative similar outcomes and positive similar outcomes), which means that W=Z and X=Y.

Thus when:

X=Y

And

$(XY)_1 = 0.6499363199099528$

The square root of 0.6499363199099528 is value of X and Y

$X = 0.8062$

$Y = 0.8062$

$Z = 0.1938$

$W = 0.1938$

<u>Based on the second method of analysis:</u>

11. Equation (d)' $XY + WZ = 31/39 = 0.7948718$

$XY + (1-X)(1-Y) = 31/39 = 0.7948718$

$XY+1-Y-X+XY = 0.7948718$

$2XY-X-Y+0.2051382=0$

$X+Y-2XY-0.2051382=0$

12. Equation (e)' $XZ = 0.1538461$
$Z = 0.1538461/X$; $X = 0.1538461/Z$

13. Equation (f)' $YW = 0.051282$
$W = 0.051282/Y$; $Y = 0.051282/W$

Hence:

14. $WZ = (0.051282/Y)(0.1538461/X)$

And

$XY = (0.1538461/Z) (0.051282/W)$

15. Back to Equation (d)':

$XY + WZ = 31/39 = 0.7948718$
$XY+ (0.051282/Y)(0.1538461/X)= 0.7948718$
$XY+ 0.0078895/XY = 0.7948718$

Multiplying both sides of the equation by XY:

$(XY)^2- 0.7948718 (XY)+ 0.0078895= 0$

16. Quadratic Equation Solution:

$(XY)_1 = 0.78481916578313\,89$
$(XY)_2 = 0.010052634216861145$

Based on the first solution:

17. $(XY)_1 = 0.7848191$; $Y = 0.78481916578/X$

18. Back to Equation (a) ; $X+W =1$

19. Back to Equation (f)' : $YW =$ **2/39= 0.051282**
$W =$ **0.051282/Y**

Hence:

20. $X+$ **0.051282/Y** $= 1$

Multiplying both sides of the equation by Y

21. **Y(X+0.051282/Y) = Y**

$XY + 0.051282 = Y$
$Y - XY – 0.051282 = 0$

$Y (1-X) –0.051282 = 0$

$Y = 0.051282/1-X$

22. Back to the quadratic equation's solution

$(XY)_1 = 0.7848191;$

$Y = 0.78481916578/X$

23. **$0.051282/1-X = 0.78481916578/X$**
24. **$0.051282X = (1-X)\ 0.78481916578$**
25. **$0.051282X = 0.78481916578 - 0.78481916578X$**
26. **$0.051282X + 0.78481916578X = 0.78481916578$**
27. **$0.83610116578X = 0.78481916578$**

28. **$X = 0.78481916578/0.83610116578 = 0.93866531695$**

29. **$Y = 0.78481916578/0.93866531695 = 0.83610116578$**

30. **$X = 0.93866531695$**
$Y = 0.83610116578$
$W = 0.06133468305$
$Z = 0.16389883422$