EUROPEAN
POLYGRAPH

Volume 1 • Summer 2007 • Number 1

JERZY KONIECZNY[*]
Andrzej Frycz Modrzewski
Krakow University College
(POLAND)

Selected aspects of uncertainty in polygraph examination

Introduction

The diagnostic process in polygraph testing involves a comparison between the intensity of the response registered to one type of question – so-called relevant questions – with the intensity of response registered to another type of question, such as control questions, probable lie questions, neutral questions, etc., depending of the technique employed.

It may be shown that this situation is typical for scientific evidence construed as an assessment of comparison. To this end, consider the following (Aitken, Taroni, 2004):

[*] jerkonieczny@wp.pl

> The interpretation of scientific evidence may be thought of as the assessment of a comparison. This comparison is that between evidential material found at the scene of a crime (denote by *M_c)* and evidential material found on a suspect, a suspect clothing or around his environment (denote this by *M_s).* Denote the combination by *$M = (M_c, M_s)$.* (…) Qualities (…) or measurements (…) are taken from *M*. Comparisons are made of the source form and the receptor form Denote these by E_c and E_s, respectively, and let *$E = (E_c, E_s)$* denote the combined set. Comparison of E_c and E_s is to be made and the assessment of this comparison has to be quantified. The totality of the evidence is denoted by *E* and is such that *$E_v = (M, E)$.*

In the case of polygraph examination, material M_c is created within the psyche of the individual who has perpetrated an act, while material M_s exists in the psyche of each individual. "Qualities" are constituted from the aforementioned types of questions and responses to the questions, whereas "measurements" are the intensities of the responses. In the case of relevant questions, we obtain E_c, and other questions E_s. Evidence from polygraph testing E_v,, i.e. *$E_v = (M, E)$* comprises: the questions used in the examination and the intensity of the responses registered after these questions. To reiterate from Aitken and Taroni: „Comparison of E_c and E_s is to be made and the assessment of this comparison has to be quantified".

While the foregoing observations seem to be fully in accordance with elementary intuition concerning polygraph testing, they are worth restating, since Aitken and Taroni's comments are relevant to the evaluation of evidence in forensic science in general, while to date the interpretation of polygraph examinations have remained outside the mainstream of forensic science. This does not benefit the discipline.

Problem

Let us consider the following:

(1) X_i Є (DI)

where DI denotes the quality of deception indicated, (DI) denotes the set of individuals designated by this quality, and X_i denotes a specific individual. Let us assume that (1) was formulated as a result of a polygraph examination. The meaning of (1) is naturally such that the individual X_i was diagnosed as deceptive. Accepting (1) leads to acceptance of further statements, such as: (a) X_i was presented with the suggestion of undergoing an examination, (b) X_i agreed to undergo the examination, (c) the examiner conducted the test, (d) the examiner interpreted the charts generated during the examination and drew a conclusion, and (e) the examiner is convinced that (1) is true.

Note that analogous comments concern the statement:

(2) X_j € (NDI)

General, investigative and juridical expertise, as well as elementary knowledge of scientific methods lead to the conclusion that not all statements of the nature of (1) and (2) are true, despite the examiners' conviction. Stated briefly, some conclusions drawn from examinations are false, and the question of what is the proportion of true to false conclusions is as old as polygraph testing itself. A massive body of literature is devoted to this issue and it is not the purpose of this paper to cite or analyse it. Those interested may find it useful to review current literature, such as for example *The Polygraph and Lie Detection* (2003).

Therefore, in a sense, an element of uncertainty is present in the results of any examination. For convenience sake, and at the risk of oversimplifying the matter, let us assume that a statement of the following nature would be more realistic than (1) and (2):

(3) X_i probably € (DI)
(4) X_j probably € (NDI)

the sense of which is that X_i was as a result of the examination diagnosed as an individual who is probably deceptive, and X_j as probably truthful.

The aim of this paper then is to analyse selected aspects of uncertainty in polygraph testing.

Selected sources of uncertainty in the diagnostics process

Three methods of interpreting test results are used: the visual method (global, qualitative), the numerical method, and the computer method.

The first method is considered the weakest and most subjective, since it is based on a general impression, the strength and consistency of the responses registered and on informal evaluations of facts, examinee utterances and his/her attitude during the test. The accuracy of diagnostic decisions made using the global method is significantly lower than those made using the numerical method (Kirchner, Raskin, 2003).

In the global method, practically all of the elements involved in the interpretation may constitute a source of uncertainty, given the ambiguity of the terms employed in their description, such as the aforementioned "general impression," "strength and consistency," "evaluation of facts," "subject's attitude," etc.

Numerical methods, although definitely constituting major progress in the diagnostic procedures, are not flawless either, as indicated in the term by which some authors to refer to them: "semi-objective". Some approaches suggested for use in numerical methods are highly precise. The 7-position scale adopted by the Department of Defence Polygraph Institute (Swinford, 1999) is an instance of such precision. Even such solutions, however, are not – because they cannot be – free from ambiguous expressions, such as: "…the most common physiological response (…) is an increase (…) from the baseline level – **usually** beginning at or **near** stimulus onset and lasting for a **few** seconds …" (Swinford, 1999; own emphasis). This is not in criticism of Swinford, because in the description of individual and unique psycho-physiological phenomena such expressions are inevitable.

Other ambiguities appear as well in connection with the numerical method, sometimes of a very basic nature. Matte (1996), in presenting a chart of the distribution of points for the examinations of "guilty" and "innocent" individuals, does not include the legend for the vertical axis, inaccurately indicates the mean (which should be marked with a point on the horizontal axis and not a vertical section), calculates for unknown reasons the mean and standard deviation up to 4 decimal places while the scoring chart based on whole integers, and does not indicate whether the chart is asymptotic towards the horizontal axis. Also, the priory assumption of threshold values of +3 and -5 (although probably empirically justified) is another source of uncertainty.

The probabilistic nature of polygraph examination is even more visible when computer methods are employed in diagnostics. In their extensive analysis of this subject matter, Kirchner and Raskin (2003) explicitly use type (3) and (4) sentences. At the same time, it seems unquestionable that automated computer systems of diagnosis have an advantage over other methods; current research shows this even in relation to the relevant-irrelevant test, which is often considered rather outdated (Honts, Amato, 2007). Thus, although it is still admissible to use various methods of diagnostics (for instance, the APA Standards of Practice advice in section 3.10.1: "Examiners' conclusions and opinions are required to be based on quantitative or numerical scoring…". Analogously, the Standard Practices for Interpretation of Psychophysiological Detection of Deception (Polygraph) Data, ASTM International, Designation: E2229 – 02, indicates only “global evaluation” in section 4.1 and “numerical evaluation” in section 4.2). The dominance of automated computer algorithms is imminent.

Naturally, however, it may not be conceded that scoring methods are the only source of uncertainty in polygraph examinations. Yet, the foregoing brief reminder will suffice for the following discussion.

Uncertainty and notions of probability

When the qualitative method of evaluation of the polygraph examination is used, the examiner’s uncertainty is expressed in terms of subjective probability, also known as psychological probability. The premise for using this notion of probability is the high level of difficulty in computing calculations using notions of frequency, and in particular, of combinations thereof; in fact, it is practically impossible to do so. The likelihood contained in the diagnosis (for instance, X_i probably Є (DI), X_j very probably Є (NDI)…) is a measure of the examiner’s conviction, who – using his/her common sense and experience – overcomes (or rather bypasses) the calculation problems. While in many cases there is nothing inappropriate in it, stepping beyond the traditional trio of outcomes (DI, NDI, inconclusive) may be useful. This is so because using an expression of the expert’s conviction on a scale enables an attempt to include a kind of sum of observations resulting from the subject’s behaviour, subject’s attempts to employ countermeasures to affect examination outcomes, and the combination of results of various tests (e.g., control question tests and peak of tension tests), etc. The term “probably” may also mean that the expert, using a specific diagnostic algorithm, did not find in the available material complete grounds that he/she requires, but sufficient information that he/she considers to be important, and states that “X_i is rather DI than NDI”, or “more arguments exist to consider X_i as DI than otherwise”. It is important to note here that instead of using such expressions, the use of the “examination of X_i jest inconclusive” formula may lead to a loss of important information.

Polygraph examiners in Poland often use this manner of expressing uncertainty. One must bear in mind, however, that using subjective probability necessitates taking into account the fact that the assessment of its value may vary considerably depending on the person who undertakes the assessment, which is a major weakness of this approach.

Kirchner and Raskin (2003) analysed frequency probability using Bayes' Theorem in the context of the numerical method (7-position scale) and using computer techniques. Their comments are worth quoting:

> Numerical evaluators use cutoffs of +/-6 to classify polygraph outcomes as truthful, deceptive or inconclusive. A score of +6 or greater is considered a truthful outcome, a score of -6 or less is a deceptive outcome, and scores between the cutoffs are inconclusive. In contrast to the categorical decisions by the polygraph examiner, the probabilities output by the computer are continuous. (…) Our research suggests that the optimal cutoffs are .70 and .30 for truthful and deceptive decisions respectively. They are optimal in the sense that they produce relatively few inconclusive outcomes and relatively high accuracy rates.

It is clear therefore that computer techniques make it possible to attain a scientific standard of uncertainty, i.e. a way of expressing it in terms of frequency probability. It seems, however, that an approach in terms of significance probability is also possible.

Significance probability approach

The basic operation enabling the use of statistical induction in diagnosing the results of a single examination is the automation of ranking of the intensity of responses to test questions. Response ranking surfaced in the research on numerical evaluation of records and have been described in detail (Honts, Driscoll, 1988; Miritello, 1999). Without going into technical details, it is worth noting that producing a ranking using a computer algorithm is very simple and may be performed automatically immediately upon completion of the examination.

Let us assume that the subject X_m was examined using a test including N_c relevant questions and N_s other questions (control questions, probable lie questions, neutral questions), where N_c, $N_s \geq 4$ (if the test included buffer questions, they should be disregarded in the calculation). The diagnosis, i.e. the decision to find the subject X_m among either the (DI) set or the (NDI) set involves a comparison of intensities of responses to N_c and N_s questions. If the distribution of response intensities for responses from both sets appear to be "similar" or if the intensity of responses to N_s questions higher, this constitutes grounds for including X_m in the (NDI) set; if the intensities are higher for the responses to N_c questions, X_m will be placed in the (DI) set. The decision may be taken on either a global or a numerical basis. It is also possible, however, to assume a null hypothesis that the intensities of responses to questions from both groups come from a population of identical response intensities. Thus, the following null hypothesis:

H_0: intensities of reactions after N_c and N_s questions may be treated as coming from a joint general population.

Once an automated joint ranking of response intensities for questions from one test is compiled (e.g., RIT, CQT, PLT…) it becomes clear that in order to test the null hypothesis, a non-parametric statistical tool for an ordinal scale must be used. The Wilcoxon rank sum test is a classical tool of this kind and it is considered to be a very good alternative to the *t* test (Ferguson, Takane, 1989). If these do not produce grounds for rejecting the null hypothesis, the subject may be found (NDI), while rejecting the null hypothesis indicates either an (NDI) or a (DI) result depending on the value of the sum of the ranks in both groups of questions.

By designating the responses to relevant questions as E_c, and to other questions as E_s, we can see that the aforementioned Aitken and Taroni requirement is satisfied, since this procedure makes it possible (in an objective manner, assuming that the ranking algorithm is correct) to achieve a quantitatively comparative assessment of E_c and E_s.

While the procedure outlined above appears formally correct, it nonetheless raises a number of fundamental questions. First, is it permitted to count neutral questions, control questions, probable lie questions, etc., to one sample? Second, should a directional or a non-directional test be employed to test the null hypothesis? Third, what criteria should be used to assume a particular significance level in testing the null hypothesis? Fourth, what rules should be adopted to reach a conclusion on the basis of a number of invariant tests (for instance, mixed question test, silence answer test, yes test) in a single polygraph examination? This is not an exhaustive list of the issues.

One may venture a guess that the answer to the first question is “yes”. As for the second and third questions, the answers will depend on the acceptable proportion of type one and type two errors and on the expected restrictiveness of the tests. The fourth and most difficult question might be answered if consideration is given to the possibility of employing a correlation coefficient or an ANOVA-type test for an ordinal scale (such as the Kruskal-Wallis test, which, however, would require a continuous ordering of E_c and E_s during the entire examination), or to other, different statistical tools. There are no doubts that only experimental research will bring answers to these (and other) questions within reach. Such research is currently underway.

The utility of a probable opinion

The ultimate goal of a polygraph examination is to supply a premise (in the form of scientific proof) in a logical argument aimed at reaching a legal decision. In fact, scientific proof is not "independent" in the sense that an expert's opinion is based not only on the observations from the examination but also on certain theoretical grounds. This theoretical background is referred to in this context as indirectly relevant evidence, ancillary evidence, or auxiliary evidence. Its role in the construction of the framework of proof is outlined below, using David Schum's concept (Schum, 2000) and adopting his approach to the circumstances of polygraph examination.

Let us assume that subject X_i, is suspected of having perpetrated an act, has undergone a polygraph examination, as has consequently been designated as (DI). Does this statement, i.e. X_i € (DI) allow us to conclude that X_i is, in careful terms, associated with the act? Schum claims that it does, provided that we are in possession of a generalisation that supports or licenses such reasoning. Such a generalisation might in this case assume the following form: "Whenever something like the opinion "Xi € (DI)" (event A) happens, then something like "X_i is associated with the offence" (event B) *probably* happens". It is not surprising that the author immediately adds: "There is never any guarantee that an asserted generalisation does apply in a particular instance. How strongly ancillary evidence supports generalisation (...) also bears upon the strength of the probabilistic linkage between events..." (Schum, 2000).

This "strength of the probabilistic linkage" constitutes at the same time a measure of uncertainty of the examination results. If the global or semi-numerical methods were used to interpret the charts, the estimate of the degree of uncertainty of the generalisation, and consequently of the examination as scientific proof, will remain qualitative.

A qualitative estimate of the level of certainty/uncertainty of the polygraph examination does not of course preclude its usefulness, particularly as a basis for action. The estimate of "very likely" may correspond with the legal standard of "clear and convincing evidence", "likely" – "clear showing", "medium likelihood" – "preponderance of the evidence", etc., as suggested by C. Weiss (2003). Weiss, however, admits that – as any subjective scale – such scales are only capable of expressing the subjective belief as to the degree of uncertainty in a given situation.

The presence of the subjective element in the interpretation of polygraph examination has one more aspect that should be counted among extra-legal factors and placed in the sphere of cultural context. Namely, one may not disregard the fact that there are individuals and whole communities – and not just in the legal profession – who object to the use of the polygraph, for example on moral grounds. It is worth remembering that in the post-Soviet countries for example public opinion was for decades indoctrinated against "lie-detection", presented as an abomination of American capitalism; traces of such attitudes are still evident today, despite advances in research. M. Damaška (2003) pointed out such phenomena connected with the changes in fact-finding technology. It appears that the subjectivity present in the interpretation of polygraph examinations is conducive to such attacks against the method.

Conversely, attaining standards for quantitative estimation of uncertainty, which is increasingly common for identification methods in forensic sciences, will work to the advantage of polygraph testing in terms of social attitudes, particularly among practising lawyers.

Finally, let us consider the argument that is perhaps the most important one in favour of reducing subjectivity as the generator of uncertainty in polygraph testing. Namely, this subjectivity factor may become a reason for a generally negative evaluation of polygraph expertise. In the European discussion – which is of such great importance today – of the quality of forensic expertise such a comment was made: "A final notable aspect of forensic science is that many forensic science techniques call for large degrees of subjective judgement. (...) This is not a criticism of those techniques, but we should note that an implication of it is that, where techniques rely on subjective judgement rather than articulable and testable principles, they require careful empirical validation in order to substantiate their proponents' claims" (Redmayne, 2000).

Conclusions

Bringing the methodology of polygraph examination closer to the quality standards of other areas of forensic science definitely seems useful. Diagnoses of "Deception Indicated", "No Deception Indicated", and "Inconclusive" are becoming obsolete, finding declining support in methodology and, more importantly, do not account well for uncertainty. The introduction and spread of computer methods creates myriad new possibilities to use inferential statistics and this direction of research into interpreting the results of polygraph examinations seems to be the most promising.

References

Aitken, C., Taroni, F. (2004), *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons.

Damaška, M. (2003), *Epistemology and legal regulation of proof*, Law, Probability and Risk 2, 117–130.

Ferguson, G. A., Takane Y. (1989), *Statistical Analysis in Psychology and Education*, McGraw-Hill.

Honts, C. R., Driscoll, L. N. (1988), *A Field Validity Study of the Rank Order Scoring System (ROSS) in Multiple Issue Control Question Tests*, Polygraph 17 (1), 1–13.

Honts, C. R., Amato S. (2007), *Automation of a screening polygraph test increases accuracy*, Psychology, Crime & Law, 13 (2), 187–199, Abstract.

Kirchner, J. C., Raskin, D.C. (2003), *Computer Methods for the Psychophysiological Detection of Deception*, in: Kleiner, M. (ed.), *Handbook of Polygraph Testing*, Academic Press.

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

Miritello, K. (1999), *Rank Order Analysis*, Polygraph 28 (1), 74–76.

The Polygraph and Lie Detection (2003), Committee to Review the Scientific Evidence on the Polygraph, The National Academies Press.

Redmayne, M. (2000), *Quality and Forensic Science Evidence: an Overview*, in: Nijboer, J. F., Sprangers W. J. (eds.), *Harmonisation in Forensic Expertise. An inquiry into the desirability of and opportunities for international standards*, Thela Thesis.

Schum, D. A. (2000), *Singular Evidence and Probabilistic Reasoning in Judicial Proof*, in: Nijboer, J. F., Sprangers W. J. (eds.), *Harmonisation in Forensic Expertise. An inquiry into the desirability of and opportunities for international standards*, Thela Thesis.

Swinford, J. (1999), *Manually Scoring Polygraph Charts Utilising the Seven-Position Numerical Analysis Scale at the Department of Defense Polygraph Institute*, Polygraph, 28 (1), 10–27.

Weiss, C. (2003), *Expressing scientific uncertainty*, Law, Probability and Risk 2, 25–46.