MARCIN GOŁASZEWSKI

https://doi.org/10.4467/20801335PBW.23.014.17664
https://orcid.org/0000-0003-1653-1669

The issue of multiple testing in polygraph screening – implications for internal security

Abstract

Screening polygraph tests are most often conducted before a candidate is hired. Bayesian statistics help to better understand the real meaning of the test results obtained. Retesting also happens - usually when a candidate fails on the first attempt. In the article, the author considers various scenarios related to retesting and their consequences, including for the internal security of individual institutions and countries. He seeks optimal solutions so that testing procedures promote tightness in systems while remaining efficient and useful.

Keywords

polygraph, testing, screening, retesting, results, interpretation, security.

Screening in various areas of life is carried out on people in whom no specific signs of abnormality or specified problems, doubts for verification have been revealed. Such examinations make it possible to determine the risk of occurrence or detection of a specific disease, disorder or undesirable behaviour. They involve the preliminary identification of a problem before the appearance of its visible symptoms. They are part of prevention, early detection and prognosis of risks.

The essence and importance of polygraph screening in personnel procedures

Among other things, psychophysiological polygraph examinations are conducted as part of recruitment procedures for work or service, in counterintelligence and anti-sabotage checks on high-risk groups (e.g., among agents, informants, persons scheduled for special assignments). Although originally polygraph examination methods were designed for investigative purposes, in the modern world (including the USA, China, Israel, the Russian Federation, many European countries, including Poland) most examinations are carried out as part of personnel procedures - both in state institutions and in the private sector. Above all, they are performed in government security agencies, law enforcement agencies and other entities whose activities are relevant to security and public order.

Polygraph screening conducted for personnel purposes simultaneously serves several functions: a) selection - starting with self-selection, i.e. deterring unsuitable candidates, who know or anticipate what the recruitment criteria may be and are aware that they do not meet them; b) preventive - by preventing and discouraging inappropriate activities - this applies both to those who seek engagement and those already employed; c) informational, detection - expanding information about the person being screened (including information that is not available by other methods), pointing to potential problems for further verification.

The literature indicates that among the most appropriate topics that candidates, for example, for officers in the uniformed services, are asked about during polygraph examinations are: issues related to intolerance (use of domestic violence, ethnic and racial slurs), involvement in criminal activities in adult life, drug use in the last few years, disciplinary actions in previous jobs[1]. The selection of issues depends on the specifics of the institution and its internal policies. It may take into account, among other things, concealment of contacts with high-risk groups, materials of possible pressure or handling of classified information (if access to such information was previously granted).

In Poland, polygraph examinations are one of the stages of qualification proceedings for all special services (the Internal Security Agency (ABW), the Foreign Intelligence Agency, the Military Counterintelligence Service, the Military Intelligence Service, the Central Anti-Corruption Bureau), the State Protection Service (SOP), the Border Guard, as well as in selected cells of the National Revenue Administration (KAS), where the activities specified in the Act on the KAS are

[1] M. Handler et al., *Integration of Pre-Employment Polygraph Screening into the Police Selection Process*, “Journal of Police and Criminal Psychology” 2009, vol. 24, no. 2, pp. 69–86.

performed. As for normative acts at the level of a law or regulation, only in the case of the SOP, Border Guard, KAS and ABW is there an indication of the obligatory nature of this stage (see Article 72(1)(2) of the *Act of 8 December 2017 on the State Protection Service*; Article 31(1a)(5) of the *Act of 12 October 1990 on the Border Guard*; Article 153(2) of the *Act of 16 November 2016 on the National Revenue Administration*; § 5(1)(6)(d) of the *Ordinance of the Prime Minister of 29 November 2002 on the template of the personal questionnaire and the detailed principles and procedure for conducting the qualification procedure for candidates for service in the Internal Security Agency*). In the Military Police, polygraph examinations are subjected to persons considered for "positions requiring special aptitude" (see Article 8a(8)(1)(d) of the *Act of 24 August 2001 on the Military Police and Military Ordnance Authorities*). Similar provisions, theoretically limiting the number of candidates subject to such examinations, can be found in the legal acts regulating the functioning of the listed special services. In reality, however, the list of positions requiring special predispositions or skills covers all posts of officers, i.e. the so-called uniformed posts (as a rule, it does not apply to civilian employees). At the same time, the legislation contains certain 'loopholes' allowing the polygraph examination to be omitted, e.g. with regard to officers transferring between services (this is the case, for example, in the SOP - see Article 72(1)(1)(b) of the aforementioned Act on the SOP; see also Article 50(3) of the *Act of 9 June 2006 on the Central Anti-Corruption Bureau*).

The optional pre-employment polygraph examination is also allowed in the police force. However, experience to date does not indicate that this instrument is routinely used, except for internal recruitment in the Bureau of Internal Affairs.

Each of the aforementioned institutions also has a legal basis for subjecting not only candidates but also officers to polygraph tests. Although the nature of these examinations implies that they are voluntary (or at least non-coercive), in practice, an unjustified refusal to undergo polygraph tests will be treated as a disciplinary offence, since the referral to the examination (usually within the competence of the head of the service concerned) is a form of service order. In the case of applicants, the examination is voluntary in the sense that the application to the service itself is the result of free will. On the other hand, the qualification procedure of an applicant who refuses to participate in the examination will result in a decision to refuse admission to the service.

A separate issue is the need for cooperation between the expert conducting the polygraph test and the person being tested. Obtaining reliable test results requires following the instructions of the examiner. Revealing symptoms of deliberate non-cooperation on the part of the examinee (including through the use of so-called distractions) leads to an opinion equivalent in effect to test results indicating deception. In these circumstances, it is usual to stop at one test, without the possibility of retesting. Furthermore, if the rejection of a candidate for service is a consequence of the result of a polygraph examination, it is either because of disqualifying statements he or she has made during the examination or because of unremovable doubts about the reliability of the answers given during the tests that ended with unfavourable results.

The polygraph examination as a stage of the uniformed service qualification procedure is intended to help assess a candidate's aptitude, but above all it provides recruiters and security division officers with a valuable tool in managing risk in the institution. It is not the only and sufficient tool for making specific personnel decisions, and is used in parallel with other procedures. Nevertheless, due to its effectiveness in eliciting hitherto unknown information from the candidate, as well as the high diagnostic value of the tests themselves, it is an extremely important stage, and any doubts that remain afterwards can hardly be simply ignored.

There is probably no test in any part of life that has perfect sensitivity and specificity and thus gives one hundred per cent accurate indications and is perfectly usable (never produces inconclusive results). The tests used in polygraph examinations also have their limitations. The idea, however, is to accept a level of probability that minimises the percentage of false identifications, both incriminating innocent, honest persons (type one error) and relieving guilty, lying persons (type two error)[2].

Policy-makers at an institution are unlikely to be aware (or rarely are) that they can work out rules with polygraphers that define testing priorities, the limits of error tolerance, i.e. how thick the mesh of the sieve can be to make it effective on the one hand and to make the whole process run smoothly on the other. These could be policies that apply to all testing, or could be specific to particular subject areas covered in testing. A more restrictive policy would need to be considered for counter-intelligence issues, for example. Polygraphers would continue to look for the same diagnostic features of polygraph records, the only change would be in the decision rules (numerical grade thresholds) that translate into final test results. This is a problem that requires answering two questions that go hand in hand: 1) what percentage of insincere people (and in what subjects) can be accepted to pass imperceptibly with more liberal criteria, and in return not deprive a larger group of the sincere ones of the chance they deserve, and increase the number of people who make it to the next stage of recruitment?; 2) how much can the procedure be sealed so as to more effectively reject undesirable, insincere

[2] Cf. R. Eggleston, *Sixth Wilfred Fullagar Memorial Lecture: Beyond reasonable doubt*, "Monash University Law Review" 1977, vol. 4, no. 1, pp. 1–2.

individuals and consequently misjudge a proportionally larger number of blameless candidates who would be unlucky enough to have a false alarm, and somewhat block the admission; 2) how much can the procedure be sealed so as to more effectively reject undesirable, insincere individuals and consequently misjudge a proportionally larger number of blameless candidates who would be unlucky enough to have a false alarm, and somewhat block the permeability of the whole procedure? Statistics point to a golden rule in this respect, but does it correspond to the real interests of the institution and, more broadly, to the interests of the state?

When all scientifically validated tests - both screening and diagnostic (single-strand) - are considered, their average accuracy was estimated at 87.1% (screening tests are 85% accurate and single-strand tests 92.1%, with inconclusive result percentages of respectively: 12,5% i 8,8%). A meta-analysis published by the American Polygraph Association in 2011 provides detailed data on the parameters of individual tests meeting the criteria for scientific validation[3]. No similar analyses have been carried out since then, and the results of the few new empirical studies have been published either separately or as a supplement to the aforementioned 2011 report[4].

In turn, the statistical reference data collated within the available material on recognised test data analysis systems allows for a mathematical representation of the meaning of specific test results - determining how strong a result we are dealing with and expressing the degree of confidence of the investigator[5]. In this way, *p*-values are determined, i.e. the probability of a particular outcome or an even more extreme outcome than the null hypothesis, which would assume that for this outcome there is no difference between people answering critical questions honestly and dishonestly. Alternatively, one can use a chance category that describes how many times more likely a given hypothesis is to be true than false. But beware - even if such data are available and, for example, indicate misleading questions about the commission of an act, they are not equivalent to the actual probability of such

[3] American Polygraph Association, *Meta-Analytic Survey of Criterion Accuracy of Validated Techniques*, "Polygraph" 2011, vol. 40, no. 4.

[4] See eg.: *Addendum to the 2011 Meta-analytic Survey – the Utah Four-Question Test ("Raskin Technnique") / ESS*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2020, vol. 49, no. 2, pp. 73–81.

[5] Cf. R. Nelson et al., *Using the Empirical Scoring System*, "Polygraph" 2011, vol. 40, no. 2, pp. 67–78; R. Nelson, *Multinomial Reference Distributions for the Empirical Scoring System*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2017, vol. 46, no. 2, p. 115; the same, *Multinomial Cutscores for Bayesian Analysis with ESS and Three-Position Scores of Comparison Question Polygraph Tests*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2020, vol. 49, no. 1, pp. 61–72.

an event occurring. We learn only and up to how often a test result obtained occurs in a specific population (e.g. misleaders). Alternatively, how the odds - the ratio of the probability of a given hypothesis to the opposing hypothesis - change after a given test result.

It is impossible to ignore the so-called baseline indicator, which tells us what initial chances can be assigned to the occurrence of a given event or trait in the population under study. The likelihood of a psychologist accurately diagnosing a personality disorder when examining a group of service candidates will be different from that in a population of psychiatric hospital patients. More virus infections are to be expected in the symptomatic patient group than among the asymptomatic patients. The chance of a spy in the population of anglers is presumably lower than in the population of people applying for employment in places involving access to classified information. It is worth bearing in mind that even the most accurate test that purports to detect something that is scarce in a given sample group will yield a higher number of false positives than the same test (or even a less accurate test) performed in an analogous group of people, many of whom have the trait sought.

It is not possible to perfectly translate into numbers all the circumstances that support a given hypothesis. Therefore, even if such an attempt is made, it will only be a rough estimate. For example, let us imagine that three castaways arrive on a deserted island in a raft, one of whom is murdered by a stabbing. None of the other arrivals admits guilt. Assuming that no circumstances are revealed that would support a particular personal version, there is an equal chance of guilt (1:1) for both subjects prior to the examination - the base ratio is therefore 0.5. Generally in polygraph examinations, this is the ratio assumed at the start of the reliability diagnosis. This is the only assumption that can be fairly applied in everyday practice, but it is also purely theoretical, even utopian. In real life, there are many different circumstances that will affect how likely a hypothesis is. In the example under discussion, these would include, for example, possible traces of the crime (other than memory traces).

If a person subjected to polygraph testing responds to critical questions in a significant way, as people who lie or conceal usually do, and other reasons are unlikely - then by reductive inference he or she is indeed deemed to be lying. Once it is accepted that one of the castaways is lying, and other evidence does not provide a compelling case for his innocence and does not incriminate the other, the former is most likely to be found guilty. This can be an unreliable approach - a true corollary sometimes leads to a false rationale (the opposite of deduction, where a prior set of true premises always leads to an unquestionable conclusion). In everyday life, however, much of our reasoning looks similar, subjectively, although we may not literally be recalculating the strength of arguments every time. We need resolutions, decisions, even when we rely on information that is to some extent uncertain. A court that determines a particular state of facts is giving expression to its sufficient conviction that other versions are improbable in relation to the one it has finally accepted. If absolute certainty were required, it is true that no innocent would ever be harmed (the actual number of miscarriages of justice remains unknowable and unknowable anyway), but neither would any criminal be punished.

Limitations of multiple testing

The legislation for the various institutions approaches the issue of repetition in the qualification procedure in different ways. For some entities, there are no specific provisions that refer to polygraph examinations in this aspect. In turn, for example, in § 14 of the *Ordinance of the Minister of Finance of 28 March 2018 on conducting a psychophysiological examination, physical fitness test and psychological examination of officers of the Customs and Fiscal Service*, it can be read that (...) *if the result of the conducted psychophysiological examination is inconclusive and it is not possible to draw up a report on its basis,* (...) *the examination may be repeated once, on the basis of the same application, within a period not exceeding 30 days from the date of obtaining this result*. There also appears to be no formal impediment to ordering a further examination with a separate request.

Based on the exchange of experience between polygraphers at national and international seminars, it can be concluded that in some institutions, only tests completed with inconclusive results are repeated and when new circumstances arise about the candidate that require verification by polygraph examination. In others, it is the practice to repeat the examination as a 'second chance', i.e. it is also performed if the first one yielded unfavourable indications in the tests, while at the same time it did not elicit statements from the person examined that the examining body would consider disqualifying after reading the report. It would appear that such repetition is not permitted everywhere. For example, in the provision contained in § 5(5) of the Ordinance on qualification proceedings for service in the ABW introduced in the amendment of 31 January 2022 (*Ordinance of the Prime Minister of 31 January 2022 amending the Ordinance on the template of the personal questionnaire and the detailed principles and procedure for conducting the qualification procedure for candidates for service in the Internal Security Agency*) indicates that the repetition of one of the stages of the procedure (including the psychophysiological examination) is possible (...) *in the event of the emergence of new circumstances concerning the candidate that may affect the outcome* (...), *in order to clarify the doubts that have arisen*. Without an indication of new circumstances (which could include any additional statements made by the candidate or information learned about him/her from other sources), a repeat examination should therefore be considered pointless. To ignore this requirement would not only be contrary to the law, but also questionable on methodological grounds, as discussed later in this article.

Undoubtedly, each case requires an individual approach and is also subject to multidimensional analysis at several levels of the hierarchical structure at the top of which is the head of the office or institution concerned. It is not possible to apply a single measure to all, as each person is inherently different. It is understandable and justifiable, and even in some situations worthy of support, that there is an unequal approach to candidates for service and to officers and staff who are already in service (and often have a lot of seniority). This can be explained with an example. In both cases, polygraph screening was carried out (in the candidate's case, it was a typical test performed before admission, and in the officer's case, it was a follow-up test). Both provided doubts related to some issue (even an identical one, e.g. regarding the use of illegal psychoactive substances). Decision-makers considering next steps have the hypothetical candidate on one side and the officer on the other. In the case of the former, a possible rejection of the application means not taking a risk that need not materialise, but also the consequences, the losses are small. The only labour and financial resources invested are related to carrying out the earlier stages of the selection procedure. The candidate himself, apart from his own time for holding a few meetings and travelling, has also not managed to invest much. The situation is different for an officer for whom a specific investment has been made for training. This is a person who already has specific knowledge, knows the secrets and has probably contributed more or less to the development of the institution. In the case of a candidate, it is necessary to assess whether a given circumstance signalled after a polygraph examination may translate negatively into his or her functioning in the workplace or service. In the case of an employee, whether a similar circumstance is also likely to cause serious problems, but more importantly, whether it has so far caused them. If not, is the employee promising improvement and is it possible to apply supportive or sanctioning measures. To better illustrate these differences, one can refer to a partnership relationship. When, at the first meeting, one person does not like the behaviour of the other, it is very likely that this will also be the last meeting. On the other hand, in a longer-lasting relationship, with closer formal ties and shared experiences, an attempt is usually made to draw attention to a problem or to force a change before saying goodbye. If this is successful, there is a strengthening of the relationship. However, the repetition of undesirable behaviour means that the motivation to seek agreement decreases and there is room for more radical solutions.

This differentiated approach can be seen in the assessment criteria set for the medical panels. They probably take into account the fact that diseases can be acquired in the course of service, including in connection with service. The *Ordinance of the Prime Minister of 15 April 2003 on the assessment of physical and mental capacity for service in the Internal Security Agency* states that "unilateral hearing loss in the low and medium frequency band" disqualifies a candidate, but does not disqualify an officer scheduled for further service. The same is true for, inter alia, "neurotic disorders under treatment, promising improvement" or "harmful use of alcohol and drugs".

Different institutions, both in Poland and internationally, have different approaches to polygraph retesting with the substantive scope of the examination unchanged. The related decisions are either routine, procedural (regardless of the result of the first examination) or ad hoc (depending on the result - indicating the candidate's sincerity or insincerity, or possibly the lack of a clear conclusion). In the first case, the polygraph examination initiates and crowns the recruitment process. In the second, inconclusive examinations are generally repeated - without much controversy. In some subjects, tests with a result of "NDI" (no deception indicated) are repeated - as if to be sure - and the opposite results end the procedure. In others, on the contrary, a repeat is ordered in the event of a "DI" (deception indicated) result, and a result favourable to the subject allows the procedure to continue. What are the implications of the above solutions? Each has its advantages and disadvantages. Depending on the specifics of the institution, some may be more optimal, others less so.

The average polygraph screening carried out as part of recruitment procedures consists (omitting other stages of the examination, such as the pre-test interview) of a multi-problem test (containing questions related to a list of key issues when assessing a candidate's suitability for a job or service). Once significant changes in physiological responses have been identified, the polygrapher usually decides on a second test (in which case a de facto retesting is already carried out during the first examination). This second test is usually chosen from the group of diagnostic (single-threaded) tests, with higher accuracy, specificity (however, still - due to the broad scope of the critical issue, e.g. the commission of a crime, without indicating the category, even more so without referring to the act at a specific time and place - the test maintains its screening character).

It is recommended that the repeat test should be different from the first (in an ideal world, not only the test, but also the instrument, the analytical method and the investigator himself should be different). This is to prevent the same non-random cause that may have previously been responsible for the error from occurring again under unchanged conditions[6]. The desire to ensure objectivity and immunity from the influence of previous information should also prompt the next polygrapher not to be familiar with the results of previous tests. There remains the practical problem of adequate preparation for such testing. Sometimes it is not possible to replace the person who conducted the previous examination. In such cases, it is best for the examiner to strictly follow a structured interview procedure, to use techniques that are less dependent on the personal characteristics of the polygrapher, e.g. "DLC" (directed lie comparisons) control questions, to be assisted by computer-assisted algorithmic analysis of the recorded records, to undergo quality control with blind interpretation of the charts.

After a test result indicating insincerity is obtained, sometimes further auxiliary tests are carried out, guiding on more specific aspects of the revealed problem. It should be noted that in many institutions, particularly in the United States, the examination is limited to one screening test only, and a second test is performed exceptionally. This may happen if, in the course of this examination, the candidate makes a new statement justifying a follow-up verification (on the principle of - is this really all that should have been added, or does the problem remain and lie in yet another memory or memories that the examinee has not provided).

During screening, it is not known what the initial probability of a given trait of the population being tested is. For this reason, in polygraph examinations a 1:1 chance is assumed as standard (it is equally likely that the respondent will be sincere and will be insincere). We do not know in advance (however, we are not completely powerless when making estimates[7]), During screening, it is not known what the initial probability of a given trait of the population being tested is. For this reason, in polygraph examinations a 1:1 chance is assumed as standard (it is equally likely that the respondent will be sincere and will be insincere). It is impossible to know in advance (however, we are not completely powerless when making estimates) what proportion of the recruitment population is concealing, for example, drug use or contacts with high-risk groups. Intuitively, however, one can assume that the former will be more numerous than the latter. Consequently - if positive (i.e. these bad) test indications are obtained, the probability that the indication is true will be higher with regard to drugs than, for example, espionage. This can

[6] Cf. R. Nelson, F. Turner, *Bayesian Probabilities of Deception and Truth-telling for Single and Repeated Polygraph Examinations*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2017, vol. 46, no. 1, p. 62.

[7] Approximate numbers can sometimes be drawn from various sources of information - historical data, scientific studies, surveys, police statistics, court statistics and others. For comparison - an interesting method is to study the concentration of viruses in wastewater as an attempt to monitor the actual epidemiological situation.

be shown in numbers. The probability of a respondent having a trait signalled by a test (e.g. hiding something about drugs) can be calculated as follows (making the hypothetical assumption that 20% of the population had personal experience with drugs):

$$\text{A posteriori probability after a positive test ("SR"/"DI"):}$$
$$= \frac{\text{initial probability of a lie x index of test sensitivity}}{\begin{array}{c}\text{(initial probability x test sensitivity)}\\ +[(1-\text{initial probability) x false positive rate}]\end{array}}$$

If the screening test was conducted in a standardised "DLST" (Directed Lie Screening Test) format and an "ESS" (Empirical Scoring System) was used to analyse the data, it is known to have a sensitivity of 0.809 and a false positive (FP) rate of 0.146[8]. The analysis of the polygraph records concludes with the opinion: "SR" (significant responses) as a consequence of the respondent's way of responding to a question about hiding drug-related information. When the available figures are substituted under the above formula, one obtains:

$$\text{a posteriori probability} = \frac{0{,}20 \text{ x } 0{,}809}{\begin{array}{c}(0{,}20 \text{ x } 0{,}809)\\ +[(1-0{,}20) \text{ x } 0{,}146]\end{array}} = 0{,}58$$

In the situation described in the example, after one screening test it was estimated that the probability of the hypothetical respondent hiding a drug problem was 58%. Is this low? Seemingly. Although, if only the average accuracy of the "DLST" test of 85.8% has been taken into account so far in drawing conclusions, the difference is apparent.

We will be faced with even less certainty if the 'alert' from the screening test occurs at an incident whose real-life occurrence in the sample population is even rarer than for the drugs used in the example. Let it be contacts with foreign special services. It can be subjectively assumed that 1% of the candidate population was involved in such and the same test ("DLST") was performed, ending with a positive indication. When the data are substituted into the formula, it can be concluded that after the test, the probability that the respondent is a spy is only 5.3%. Stopping at this stage would result in the majority of such indications being realistically false.

[8] American Polygraph Association, *Meta-Analytic Survey*…, p. 244.

This is why, among other things, during the COVID-19 pandemic, those who believed that widespread testing of the entire population did not make sense, as many people would have ended up in isolation for no reason, especially at a time when the assumed incidence rate among the general population was already relatively low. Testing everyone would have been a rational action at most during the peak of the pandemic. The correct approach was to refer people for additional diagnostic, genetic PCR-type tests after a positive indication of a rapid, antigenic screening test. A similar course of action applies to polygraph tests.

Indications of screening tests are rightly treated as a reason for further testing. It makes sense to carry out another, preferably more accurate, test from the range of diagnostic tests after such an initial test. Using the hypothetical study described in the article as an example, it is possible to observe how the drug situation will change after a further test. Before the test, it was estimated that the probability of the subject coming into contact with drugs was 20%. After the screening test, this probability increased to 58%. An additional test - a diagnostic test, e.g. in the standardised "You-Phase" format ("Backster Bi-Zone"), with an average accuracy of 90.4%, is administered. The "ESS" system will again be used to analyse the test data. The test is then characterised by a sensitivity of 0.845 and a false positive rate of 0.138[9]. The analysis concludes with a "DI" resolution. The data must be substituted into the formula for the a posteriori probability (the a priori probability after the first test and before the second test is 0.58).

$$\text{probability after test 2} = \frac{0{,}58 \text{ x } 0{,}845}{\begin{array}{c}(0{,}58 \text{ x } 0{,}845)\\ +[(1-0{,}58) \text{ x } 0{,}138]\end{array}} = 0{,}89$$

At this stage, the degree of certainty that the problem exposed after the first screening test is indeed present is already much higher. After test two, the probability has increased to 89%.

The statistical properties of a given test (e.g. sensitivity, specificity, decision error rate) do not always directly translate into conclusions about the actual strength of the relationship between two variables in a statistical population. As Raymond Nelson and Finley Turner note, Bayes' theorem helps to describe real-world relationships more effectively and accurately, given the initial assumptions that have been made about a research problem[10].

[9] Ibid, p. 240.

[10] See: R. Nelson, F. Turner, *Bayesian Probabilities of Deception*..., pp. 53–80.

For the sake of tidiness, a variant of the screening test, in which no significant changes in physiological responses are recorded, e.g. with questions about crime and the existence of pressure materials on the respondent, should still be presented. Let us assume that only one test in the standardised format "USAF MGQT" will be used throughout the research process and the data analysis will be carried out according to the rules of the "federal system", on a 7-item scale. Under these conditions, the specificity of the test is 0.538. In turn, the percentage of false negative (FN) results is 0.079[11]. The result obtained is "NSR" (no significant responses). No data are known to adequately adjust the initial (a priori) probability, so the standard (0.5) will be adopted. The formula should be modified accordingly:

$$\text{A posteriori probability after a negative test ("NSR"/"NDI"):}$$
$$= \frac{\text{initial probability of sincerity x test specificity index}}{\begin{array}{c}\text{(initial probability x test specificity)}\\ +[(1-\text{initial probability) x false negative rate}]\end{array}}$$

Then substitute the data:

$$\text{a posteriori probability} = \frac{0{,}5 \text{ x } 0{,}538}{\begin{array}{c}(0{,}5 \text{ x } 0{,}538)\\ +[(1-0{,}5) \text{ x } 0{,}079]\end{array}} = 0{,}87$$

After this test, the realistic probability that the respondent was not lying when claiming not to have committed crimes and not hiding circumstances that constitute potential pressure material was 87%. Before the test, the hypothesis of truthfulness was assumed to be 50% likely, and after the test, this probability increased by 37 percentage points.

The encouraging results of the two-stage testing model described above in a hypothetical study with a drug problem disclosure might make one wonder whether doing more tests on the same topic might not be the right thing to do, since the probability of an accurate diagnosis increased so much after the second test. Well, the matter is not so simple, the benefits are only apparent and the risk of incorrect conclusions dangerously increases. Moreover, so far only indications that are consistent with each other (in the direction of indicating insincerity) have been considered. Meanwhile, a situation may arise where the results of two consecutive tests (no matter whether performed as part of one study or during two separate encounters) diverge (usually a "DI"/"NDI" discrepancy). If the subject does

[11] American Polygraph Association, *Meta-Analytic Survey*…, p. 244.

not make significant follow-up statements between these tests (e.g. by admitting to an incident), this will not be a common case. Nevertheless, it is possible for various reasons (these reasons will not be discussed further here, as this is a topic for a separate study; for the purposes of this analysis, only the properties of the tests themselves have been considered).

When the polygraph examination is the last or one of the last stages of the recruitment process, the candidate will usually have had an initial interview, a psychological examination and other tests - e.g. knowledge, fitness tests, may have already received the opinion of the relevant medical panel, and may have been subject to checks (including criminal records or environmental interviews). HR cells (especially in the face of increased staffing needs) will therefore be reluctant to stop at one unfavourable polygraph test result, which would in a way nullify the efforts made so far. This increases the motivation to have a 'second chance' examination. Sometimes, in such a situation, the candidate will complete information about himself/herself that he/she did not provide during the first examination, and this will make it possible to successfully complete the recruitment process. It also happens that the new statements do not appear and the subsequent test results in a particular outcome - more often consistent with the outcome of the basic test, but sometimes the opposite. Which result is then more reliable? Perhaps a third test should be considered necessary in a discrepancy situation? All the scenarios described are controversial - including the first one (which assumes supplementary information), which is theoretically the clearest. In the case of a candidate who has corrected his or her position (i.e. finally admitted something that was hidden at the beginning), there may be a conviction that it is worth taking the risk of lying the first time, keeping quiet about something, because the institution will tolerate it and there will be the possibility of a possible correction, and with a stroke of luck - it will work the first time. The other scenarios, on the other hand, give rise to certain consequences related to the nature of screening tests and the inevitability of a certain number of false results - both positive and negative.

The study by William Yankee and Douglas Grimsley shows that, on the one hand, although no statistically significant differences were found between the accuracy of the test performed during the first test and the repeat test on the same issue (which is a good indication of the reliability of the test), on the other hand, an increase in the number of false negatives was observed in the case of subjects acting as misleaders, and the overall accuracy (after excluding inconclusive results) dropped from 87% to 76% (perhaps a consequence of the phenomenon of habituation, i.e. gradual habituation to repeated test stimuli, perhaps the use of some psychological defence mechanisms) - there is no certain answer in this regard). In contrast, the situation hardly changed at all in the sincere group (100% and 96% correct indications respectively)[12]. One searches in vain for empirical studies showing the opposite trend. This therefore gives food for thought on which result to consider more reliable in a situation of divergence and transformation from "DI" to "NDI" (leaving aside whether there was a change in the research context between the first and second surveys through new statements made by the candidate).

Let's see how a recruitment procedure to an institution might look statistically for, say, the total number of applicants per year. A hypothetical scenario is that 750 candidates take part in the selection procedure. According to the subjective, a priori calculations of the recruiters, 20% of this population has undesirable characteristics. Not counting the demonstration test, with which the essential stage of any polygraph examination begins, everyone was screened in the aforementioned "DLST" format and using the "ESS" analysis system, with the following parameters: sensitivity - 0.809; FP - 0.146; specificity - 0.751; FN - 0.112. What numbers should be expected? In the context of retesting, inconclusives (INC) and indications of insincerity (more precisely, SR - significant reactions) are mainly in the field of attention. After taking into account the mentioned parameters, the following measures will be obtained:

- test sensitivity (0.809) x initial probability of insincerity (0.2) x number of subjects tested (750) = 121 true positives;
- test specificity (0.751) x initial probability of honesty (0.8) x number of subjects tested (750) = 451 true negatives;
- FN (0.112) x initial probability of insincerity (0.2) x number of subjects tested (750) = 17 false negatives;
- FP (0.146) x initial probability of honesty (0.8) x number of subjects tested (750) = 88 false positives;
- the remaining tests qualify as inconclusive (750 - 677 = 73).

If the interview process had been completed at this stage, the employer would have thanked 88 (11.7%) candidates to whom it might have given a chance. At the same time, he would have accepted 17 (2.3%) of those he would have preferred not to employ. 73 (9.7%) would have had to be invited for another test because of an inconclusive result. The rest of the staffing decisions - concerning 572 out of 662 people (86.4%) - would have been correct.

Polygraph screening typically uses a 'successive hurdles' approach. Negative test results (i.e. those that indicate no problem - "NSR", "NDI") close the case and provide a pass for the candidate to proceed. Positive results, on the other hand, require additional testing or other verification activities before they are considered

[12] W. Yankee, D. Grimsley, *Test and Retest Accuracy of Psychophysiological Detection of Deception Test*, "Polygraph" 2000, vol. 29, no. 4, pp. 289–298.

the basis for any action or decision. There is also, of course, repetition of inconclusive tests.

Tests (both first and subsequent) can produce results that are accurate or inaccurate, consistent with each other and divergent. If it is assumed approximately that the average accuracy of a screening test is 85%, the following variants of results are obtained:

- two independent tests with correct and concordant results: 0.85 x 0.85 = 0.72;
- two independent tests with inconsistent and consistent results: 0.15 x 0.15 = 0.02;
- the sum of the expected concordance of the two independent tests will be: 0.72 + 0.02 = 0.74 and vice versa - in 26% of cases a potential discrepancy in the results has to be accounted for;
- combined accuracy of two independent, consistent tests: 0.72 / 0.74 = 0.97.

Let us further extend the variants described with a case study involving two tests during the first examination (screening and diagnostic, accurate at 85% and 90% respectively)[13], followed by a repeat examination expected by the client, with a different polygrapher performing a third test (diagnostic, accurate at 90%):

- three tests with correct and consistent results: 0.85 x 0.90 x 0.90 = 0.69;
- three tests with incorrect and consistent results: 0.15 x 0.10 x 0.10 = 0.0015;
- the sum of the expected concordance of the three tests will be: 0.69 + 0.0015 = 0.6915; on the other hand, in approximately 31% of cases, there may be divergent results. As can be seen, with each successive test, the probability of a different result from the previous one increases, and consequently the risk of error also increases, if only the last test with a change is taken as the basis for the conclusion;
- the combined accuracy of the three consistent tests will be: 0,69 / 0,6915 = 0,998.

Let's go back to the sample selection procedure described, in which 750 candidates took part. After the first test, 209 people scored positive and were referred for a retest. The PPV (positive predictive value), i.e. how many of the positive results are true, is now 0.727. This means that the initial probability of misrepresentation assumed by the recruiters increased from 20% to 72.7% after

[13] A caveat is that tests performed by the same investigator in the same activity - even if the type of test is changed - are not fully independent of each other. Some of the variables that do not affect the results co-occur. The example presented represents a certain simplification - as if we were dealing with independent tests every time. More on this topic: R. Nelson, J. Kircher, M. Handler, *How to Calculate the Expected Agreement and the Combined Accuracy of Two Test Results*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2018, vol. 47, no. 1, pp. 18–25.

the first test in the group with positive results. The NPV (negative predictive value), which indicates the percentage of true negatives, was instead estimated at 0.273. After applying the "You-Phase" / "ESS" test, the result is then:

- 128 true positives = test sensitivity (0.845) x PPV (0.727) x number of subjects tested (209);
- 43 true negative results = test specificity (0.757) x NPV (0.273) x number of subjects tested (209).

The following false results will not be avoided:

- 20 false negatives = FN (0.134) x PPV (0.727) x number of subjects tested (209);
- 8 false positives = FP (0.138) x NPV (0.273) x number of subjects tested (209);
- the remaining results will qualify for the inconclusive set (209 - 199 = 10).

The figures show inexorably that a second round of tests would make it possible to correct falsely incriminating results in 43 people (21% of retests, so quite good), but the side effect would be 20 more false negatives. This unfortunately means that, under the conditions presented, almost one in three tests with a negative result turns out to be false (32%) and, in absolute numbers, even more undesirables will thus be admitted to the institution than during the first tests, but then more than 3.5 times as many candidates took part. In such a situation, it remains to be hoped that the other 'fuses' will not fail. However, if the polygraph test is positioned almost at the end of the proceedings, there may not be many left.

Conclusions and recommendations

The above considerations are intended to encourage not the numerical estimation of everything, but the correct interpretation of the meaning of the results of the various tests (including polygraphs), the attribution of appropriate weight to them, the adjustment of decision-making processes to this, and an awareness of the potential consequences of various actions, e.g. the profit and loss balance of retesting.

The accuracy of testing only increases when successive trials produce consistent results. Repeated testing of unsuccessful candidates only corrects downwards the number of false positives, but - to the disappointment of many or wilful ignorance - does not generally increase testing accuracy with results different from the first time. A similar phenomenon would occur if only those candidates who managed to pass test 1 unchallenged were repeatedly tested. This would succeed in eliminating from recruitment some of the unreliable people who passed the first sieve, but overall it would not look any better either.

The inevitable increase in the number of false negative indications during repeat examinations creates a pressing need to find a solution to this problem. Certainly, the situation is not acceptable in institutions sensitive to national security. So either these ways will be found, or it would be advisable to put an end once and for all to the 'corrective' tests and try to verify doubts with other tools (if they exist and are effective enough). One solution is to question the candidate thoroughly about critical issues, even though he or she has successfully passed the test, or - alternatively - to rely on the examiner's intuition in taking the direction of further conversation. Nelson and Turner conclude that such an approach would, however, reduce polygraph tests to a pseudoscience[14]. Neither researchers nor practitioners therefore have perfect answers.

One can refer once again to the experience of the COVID-19 pandemic and recall periods when exit from isolation had to be preceded by two consecutive negative test results. The pattern was: positive test result 1 (isolation), negative test result 2 (hope for end of isolation), negative test result 3 (recovery status). This analogy is very pertinent to polygraph examinations, when one starts with a positive result of the first test and thinks about the next or subsequent tests, which may give a negative result. Moreover - the accuracy of the testing is also similar, as is the division between screening and diagnostic tests. Thus, if an "SR"/"DI" result is followed by an "NDI", one should consider one more confirmation, one more test.

New statements made by the candidate in relation to a critical issue that threatens them after the first test create different research circumstances. They may be a viable cause, an explanation for the recorded bodily reactions, but they still do not provide assurance that the candidate is still not hiding something. Without the emergence of new circumstances, logical arguments for retesting are not properly seen.

In summary, an additional test is highly recommended after a positive (unfavourable) result of the first screening test. However, if a second test has already been performed at the first screening (and this is most often the case), a further test (meaning a third test and possibly more) is not recommended. The risk of false negative results outweighs the benefit of reducing the indication of false positives.

Since there is no phenomenon in the form of a perfect test, it is impossible to ensure that the interests of the applicant and the host institution are equally protected (especially with regard to ensuring internal security and the protection of classified information). Therefore, something has to be decided.

[14] R. Nelson, F. Turner, *Bayesian Probabilities of Deception*..., p. 71.

In different decision-making procedures, priorities are shaped somewhat differently. In a criminal trial, the principle of *in dubio pro reo* applies (Article 5 § 2 of the Code of Criminal Procedure). The protection of classified information is given priority over other legally protected interests in clearance proceedings (Article 24(4) of the *Act of 5 August 2010 on the Protection of classified information*).

The internal security of the institution undoubtedly takes precedence over the interests of the candidate in selection proceedings for employment or service. In other personnel proceedings - investigations, disciplinary proceedings - the priority is to ensure that an innocent person is not held accountable, and in the case of those who have made a mistake, it is important to assess whether they are likely to improve. By contrast, when recruiting for a job or service, it seems more important that an undesirable person (especially in terms of some relatively riskier aspect than others - e.g. counter-intelligence) is not taken on, even at the cost of some (small) percentage of misjudged people, but still without incurring more expense and mutual obligations on the employer-employee line. The imprudent 'pushing' of such candidates through the commissioning of successive tests until one favourable test result (with an increasing probability of a false negative result) would result in the entry into the structures of the institution concerned of persons who, in extreme circumstances, could pose a threat to state security.

Bibliography

Addendum to the 2011 Meta-analytic Survey – the Utah Four-Question Test ("Raskin Technique") / ESS, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2020, vol. 49, no. 2, pp. 73–81.

American Polygraph Association, *Meta-Analytic Survey of Criterion Accuracy of Validated Techniques*, "Polygraph" 2011, vol. 40, no. 4.

Eggleston R., *Sixth Wilfred Fullagar Memorial Lecture: Beyond reasonable doubt*, "Monash University Law Review" 1977, vol. 4, no. 1, pp. 1–2.

Handler M. et al., *Integration of Pre-Employment Polygraph Screening into the Police Selection Process*, "Journal of Police and Criminal Psychology" 2009, vol. 24, no. 2, pp. 69–86.

Nelson R., *Multinomial Cutscores for Bayesian Analysis with ESS and Three-Position Scores of Comparison Question Polygraph Tests*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2020, vol. 49, no. 1, pp. 61–72.

Nelson R., *Multinomial Reference Distributions for the Empirical Scoring System*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2017, vol. 46, no. 2, pp. 81–115.

Nelson R. et al., *Using the Empirical Scoring System*, "Polygraph" 2011, vol. 40, no. 2, pp. 67–78.

Nelson R., Kircher J., Handler M., *How to Calculate the Expected Agreement and the Combined Accuracy of Two Test Results*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2018, vol. 47, no. 1, pp. 18–25.

Nelson R., Turner F., *Bayesian Probabilities of Deception and Truth-telling for Single and Repeated Polygraph Examinations*, "Polygraph & Forensic Credibility Assessment: A Journal of Science and Field Practice" 2017, vol. 46, no. 1, pp. 53–80.

Yankee W., Grimsley D., *Test and Retest Accuracy of Psychophysiological Detection of Deception Test*, "Polygraph" 2000, vol. 29, no. 4, pp. 289–298.

Legal acts

Act of 8 December 2017 on the State Protection Service (i.e. Journal of Laws of 2023, item 66, as amended).

Act of 16 November 2016 on the National Revenue Administration (i.e. Journal of Laws of 2022, item 813, as amended).

Act of 5 August 2010 on the Protection of classified information (i.e. Journal of Laws of 2019, item 742, as amended).

Act of 9 June 2006 on the Central Anti-Corruption Bureau (i.e. Journal of Laws of 2022, item 1900, as amended).

Act of 24 August 2001 on Military Police and Military Order Authorities (i.e. Journal of Laws of 2021, item 1124, as amended).

Act of 6 June 1997 - Code of Criminal Procedure (i.e. Journal of Laws of 2022, item 1375, as amended).

Act of 12 October 1990 on the Border Guard (i.e. Journal of Laws of 2022, item 1061, as amended).

Ordinance of the Prime Minister of 31 January 2022 amending the Ordinance on the template of the personal questionnaire and the detailed principles and procedure for conducting the qualification procedure for candidates for service in the Internal Security Agency (Journal of Laws of 2022, item 263).

Ordinance of the Minister of Finance of 28 March 2018 on conducting a psychophysiological examination, physical fitness test and psychological examination of officers of the Customs and Fiscal Service (i.e. Journal of Laws of 2022, item 379).

Ordinance of the Prime Minister of 15 April 2003 on the assessment of physical and mental capacity for service in the Internal Security Agency (i.e. Journal of Laws of 2014, item 242).

Ordinance of the President of the Council of Ministers of 29 November 2002 on the template of the personal questionnaire and the detailed principles and procedure for conducting the qualification procedure for candidates for service in the Internal Security Agency (i.e. Journal of Laws of 2014, item 61).

Marcin Gołaszewski, PhD

Doctor of social sciences in the discipline of legal sciences, expert in polygraphic examinations at the Regional Court in Warsaw, President of the Board of the Polish Association of Polygraphic Examinations.