Studia Prawnicze. Rozprawy i Materiały 2017, nr 2 (21) • Studies in Law: Research Papers 2017, No. 2 (21)
ISSN 1689-8052 • e-ISSN 2451-0807

Jan Widacki
prof. dr hab., Wydział Prawa, Administracji i Stosunków Międzynarodowych, Krakowska Akademia im. Andrzeja Frycza Modrzewskiego

Marcin Gołaszewski
doktorant, Wydział Prawa, Administracji i Stosunków Międzynarodowych, Krakowska Akademia im. Andrzeja Frycza Modrzewskiego; prezes Polskiego Towarzystwa Badań Poligraficznych

Subiektywizm w badaniach poligraficznych

Wprowadzenie

Problem subiektywizmu w badaniach naukowych jest interesujący dla metodologii ogólnej nauk, a także dla poszczególnych dyscyplin naukowych. W metodologii badań, subiektywizm oznacza margines swobody interpretacji badacza, niepodlegający obiektywnym kryteriom, a raczej osobistym przekonaniom badacza. Innymi słowy, subiektywizm to pierwiastek mający wpływ na wynik badania, zależący nie od przedmiotu badanego, obiektywnie istniejącego i obiektywnie posiadającego jakieś cechy, także nie od przyrządu pomiarowego, ale od badacza, czyli podmiotu poznającego, jego właściwości i cech, a czasem także przekonań i przyjętych *a priori*, bez koniecznego uzasadnienia założeń. Problem subiektywizmu dotyczy wszystkich nauk, od historycznych począwszy[1], poprzez nauki społeczne, aż po nauki przyrodnicze.

Wydaje się, że im dyscyplina bardziej ścisła, im bardziej zaawansowana metodologicznie, tym mniejszy w niej margines subiektywizmu. Zgodnie z tym założeniem, margines subiektywizmu z zasady będzie większy w socjologii czy psychologii niż w fizyce czy chemii. W naukach społecznych, jak się zdaje, akceptowany przez badacza pogląd – a tym bardziej przyjęta przez niego teoria – mają wpływ zarówno na stawianie problemów, budo-

---

[1] Artykuł powstał w ramach projektu DEC-2013/11/B/HS5/03856 finansowanego ze środków Narodowego Centrum Nauki. Por. B. Miśkiewicz, *Wprowadzenie do badań historycznych*, Poznań 1993, s. 165; J. Topolski, *Jak się pisze i rozumie historię. Tajemnice narracji historycznej*, Warszawa 1996, s. 339.

wanie hipotez, jak i później na interpretację wyników badań empirycznych, które te hipotezy sprawdzają[2]. W naukach przyrodniczych, takich jak fizyka czy chemia, jak wspomniano, margines subiektywizmu w badaniach wydaje się z natury rzeczy mniejszy – w zasadzie daje się sprowadzić do błędu pomiaru lub jego błędnej interpretacji. W metodologii fizyki rozróżnia się „błędy pomiarowe" (błąd przybliżenia, błąd przeoczenia, pomyłki) oraz „niepewności pomiarowe" (niepewność wzorcowania, niepewność eksperymentatora, niepewność przypadkowa)[3]. Szczególnie w „niepewności eksperymentatora" kryć się może pierwiastek subiektywizmu. Problem subiektywizmu w filozofii nauki i metodologii jest też rozważany przy okazji porównywania metod ilościowych i jakościowych[4].

Subiektywizm w naukach sądowych

O ile problem subiektywizmu badacza w naukach podstawowych jest interesujący z metodologicznego, czy nawet filozoficznego punktu widzenia, a w badaniach w naukach przyrodniczych zbyt szeroki margines subiektywizmu prowadzić może do co najwyżej błędnych teorii, to w naukach sądowych (ang. *forensic sciences*) ma on sens jak najbardziej praktyczny – co więcej często istotny dla losów konkretnego człowieka, przeciwko któremu toczy się postępowanie sądowe. W ekspertyzie mamy do czynienia z badaniem (wyjaśnianiem) faktu jednostkowego, w oparciu o ustalenia jakiejś dyscypliny naukowej. Jak zauważył kiedyś Józef Życiński[5]:

> W procedurach badawczych występujących przy klasyfikacji i wyjaśnianiu faktów jednostkowych niemałą role odgrywają elementy pozadyskursywne uzależnione od indywidualnych predyspozycji poszczególnych badaczy, ich intuicji, zdolności kojarzenia faktów, wcześniejszych doświadczeń itp. […] Przecenianie roli czynników pozadyskursywnych w poznaniu może jednak prowadzić łatwo do subiektywizmu, w którym przez odwołanie do osobistych długoletnich doświadczeń usiłowano by wprowadzać interpretacje niemożliwe do racjonalnego uzasadnienia[6].

---

[2] Por. J. Czapiński, *Wartościowanie – zjawisko inklinacji pozytywnej. O naturze optymizmu*, Wrocław 1985; M. Materska, *Psychologiczna i formalna analiza sądów oceniających*, [w:] *Psychologia a poznanie*, red. M. Materska, T. Tyszka,, Warszawa 1997; T. Tyszka, *Psychologiczne pułapki podejmowania decyzji*, Gdańsk 2000; M. Weber, *Obiektywność poznania w naukach społecznych*, [w:] *Problemy socjologii wiedzy*, Warszawa 1985.

[3] Por. H. Szydłowski, *Pracownia fizyczna wspomagana komputerowo*, Warszawa 2012.

[4] Por. np. R.G. Long, M.C. White, W.H. Friedman, D.V. Brazeal, *The „Qualitative" versus „Quantitative" research Debate: A Question of Metaphorical Assumptions?*, „International Journal of Value-Based Management" 2000, t. 13, nr 2, s. 189–197.

[5] J. Życiński, *Problem wyjaśniania w naukach nomotetycznych a poznanie faktu jednostkowego*, [w:] *Z zagadnień teorii opinii biegłego*, red. J. Widacki, Materiały IV Sympozjum Metodologii Kryminalistyki i Nauk Pokrewnych, Chęciny 24–25 VI 1983, Katowice 1983, s. 60–61.

[6] *Ibidem*, s. 61.

Ekspertyzy w naukach sądowych opierają się na opisach, pomiarach i ocenach. Nawet pomiar będący istotą wielu ekspertyz w naukach sądowych obciążony jest, jak każdy pomiar, pewną niedokładnością czy niepewnością (ang. *uncertainty*), która może rzutować na wartość wydanej w oparciu o ten pomiar opinii. Może być przyczyną błędu lub pomyłki, może też rodzić niepewność[7] co do wiarygodności opinii.

Niedokładność pomiaru może być wynikiem niedoskonałości narzędzia pomiarowego lub sposobu jego użycia. Stąd też konstruuje się coraz dokładniejsze narzędzia pomiarowe. Niewłaściwy sposób użycia narzędzi pomiarowych może wynikać z niefachowości osoby dokonującej pomiaru, jej niezręczności czy niedbalstwa. Czasem z innych jeszcze powodów, jak na przykład wymuszony pośpiech czy szczególne, niesprzyjające okoliczności dokonywania pomiaru. Dotyczy to mierzenia, ważenia lub porównywania uzyskanych wyników ilościowych. Ma to znaczenie dla wszystkich ekspertyz opartych na metodach ilościowych, a więc przede wszystkim w chemii sądowej, toksykologii sądowej, biologii sądowej i podobnych dyscyplinach. W większości ekspertyz wielkość błędu pomiaru w pewnych określonych granicach (por. niżej) jest dopuszczalna i nie ma wpływu na ostateczny wynik i oparte na nim wnioski.

Nauka posługuje się między innymi pojęciem błędu standardowego (lub błędu średniego), przez który rozumie się odchylenie średnie wyników pomiarów tej samej wielkości, które otrzymano przy użyciu tego samego narzędzia pomiarowego[8].

Dokonanie pomiaru (lub opisu) przy ekspertyzie jest zwykle dopiero punktem wyjścia do dalszych działań, których finałem jest wydanie opinii. Opinia bowiem to – jak słusznie zauważył kiedyś Kazimierz Jaegermann – na pewno coś więcej niż sam wynik badania (opis czy pomiar)[9]. Wedle tego autora, na opinię składają się działania pomiarowo-opisowe, działania interpretacyjne i wnioskowanie[10]. Opinia końcowa (wnioski z opinii) jest zaś tak naprawdę decyzją biegłego[11].

---

7 S. Bell, *Measurement uncertainty in forensic sciences. A practical guide*, London–New York 2017, s. XIX.

8 Ten błąd standardowy (S) wyraża się wzorem:

$$s_{\bar{x}} = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n\,(n-1)}}$$

x_i – wartość uzyskana w pomiarze numer i, x – wartość średnia wyników pomiarów, n – liczba pomiarów.

9 K. Jaegermann, *Opiniowanie sądowo-lekarskie. Eseje o teorii*, Warszawa 1991, s. 28.

10 *Ibidem*, s. 29.

11 *Ibidem*, s. 33.

Błąd pomiaru nabiera szczególnego znaczenia, gdy przepis prawa lub orzecznictwo sądowe wprowadza progi ilościowe istotne dla kwalifikacji prawnej czynu. Tak na przykład zgodnie z art. 115 § 16 kodeksu karnego[12] stan nietrzeźwości zachodzi, gdy zawartość alkoholu we krwi przekracza 0,5 promila albo prowadzi do stężenia przekraczającego tę zawartość, lub gdy zwartość alkoholu w 1 dm^3 wydychanego powietrza przekracza 0,25 mg albo prowadzi do stężenia przekraczającego tę wartość. Natomiast zawartość alkoholu w wydychanym powietrzu miedzy 0,1 a 0,25 mg na 1 dm^3 lub od 0,2 do 0,5 promila we krwi oznacza „stan po spożyciu alkoholu" (por. art. 46 ust. 2 ustawy o wychowaniu w trzeźwości[13] oraz art. 87 § 1 kodeksu wykroczeń[14]). Podobnie jest z rozdziałem na „znaczną" bądź „nieznaczną" ilość środka psychotropowego (por. art. 62a ustawy o przeciwdziałaniu narkomanii[15]).

Jedną z pierwszych polskich prób sprawdzenia poprawności pomiarów zawartości alkoholu w krwi dokonywanych przez profesjonalne laboratoria, należące do instytucji państwowych (w tym czasie nie było jeszcze prywatnych laboratoriów), a więc laboratoriów akademickich, laboratoriów służby zdrowia i laboratoriów milicyjnych (które rutynowo wykonywały wówczas takie badania dla potrzeb organów ścigania i wymiaru sprawiedliwości, a te przyjmowały ich wyniki za podstawę swych orzeczeń), były badania przeprowadzone przez Instytut Ekspertyz Sądowych w 1966 r.[16]. Trzydzieści takich laboratoriów (pracowni) otrzymało pięć wzorcowych roztworów alkoholowych sporządzonych sztucznie, w warunkach laboratoryjnych w surowicy krwi ludzkiej. Jedna z próbek zawierała 0,17 promila alkoholu (tj. poniżej górnej granicy przyjętej dla alkoholu fizjologicznego), inna zawierała 0,23 promila alkoholu, czyli już powyżej granicy „stanu wskazującego na użycie alkoholu", a inna – 0,48 promila alkoholu, czyli poniżej „progu nietrzeźwości" (który wynosił wtedy 0,50 promila alkoholu). Okazało się, że próbkę zawierającą 0,17 promila jedno z laboratoriów oceniło na 0,28 promila (czyli powyżej granicy „stanu wskazującego na użycie alkoholu"), próbki zawierające 0,23 promila alkoholu oceniano od 0,0 promila aż do 0,36 promila. Próbki zawierające 0,48 promila oceniano od 0,22 promila (czyli poniżej granicy „stanu wskazującego na użycie alkoholu"), aż po 0,57 promila (powyżej granicy

12 Ustawa z dnia 6 czerwca 1997 r. – Kodeks karny, Dz.U. z 1997 r., nr 88, poz. 553.

13 Ustawa z dnia 26 października 1982 r. o wychowaniu w trzeźwości i przeciwdziałaniu alkoholizmowi, Dz.U. z 1982 r., nr 35, poz. 230.

14 Ustawa z dnia 20 maja 1971 r. – Kodeks wykroczeń, Dz.U. z 1971 r., nr 12 poz. 114.

15 Ustawa z dnia 29 lipca 2005 r. o przeciwdziałaniu narkomanii, Dz.U. z 2005 r., nr 179, poz. 1485.

16 J. Markiewicz, J. Nedoma, H. Serda, *Z badań nad dokładnością oznaczeń alkoholu we krwi*, „Problemy Kryminalistyki" 1966, nr 61–62, 1966, s. 472–478.

„stanu nietrzeźwości"). Trzeba pamiętać, że z każdym z tych bezkrytycznie przyjętych wyników wiążą się określone konsekwencje prawne. Niektóre osoby trzeźwe byłyby uznane za nietrzeźwe, niektóre nietrzeźwe – za trzeźwe. Pokazuje to przy okazji, jak ważna jest certyfikacja i atestacja laboratoriów, którego to problemu sądy i prokuratury zupełnie nie doceniają. W czasie gdy przeprowadzano opisywane badania certyfikacji i atestacji laboratoriów jeszcze nie było. Dziś realizowanej w podobny sposób procedurze poddają się dobrowolnie niektóre placówki badawcze z zakresu nauk sądowych. Jednak brak aktualnego atestu, a nawet niepowodzenie przy próbie jego uzyskania niestety w oczach organów ścigania i sądów nie dyskwalifikują takiego laboratorium i wyników jego badań, które są uznawane za dowody.

We wszystkich metodach ilościowych w naukach sądowych każda interpretacja pomiaru bezwzględnie musi brać pod uwagę fakt znanego i określonego błędu przy tego rodzaju pomiarach. Problem subiektywizmu w naukach sądowych zaczyna się wszędzie tam, gdzie wyniki pomiaru (bardziej lub mniej dokładnego) wymagają jeszcze oceny i interpretacji[17].

Jest jednak szereg nauk sądowych, w których w ramach ekspertyzy nie dokonuje się w ogóle pomiarów fizycznych, a poprzestaje na opisie pewnych cech. Są też takie, w których pomiar ma jedynie znaczenie drugorzędne, pomocnicze. Podstawą ustaleń poprzedzających wydanie opinii w takich dyscyplinach jest z zasady ocena dokonywana przez eksperta. Tak jest na przykład w psychologii sądowej, psychiatrii sądowej, klasycznej medycynie sądowej czy grafologii sądowej – i oczywiście w detekcji kłamstwa, zarówno instrumentalnej, jak i nieinstrumentalnej.

Ta ocena eksperta dokonywana jest w oparciu o pewne przyjęte w tej dyscyplinie kryteria. Niewątpliwie podstawowym jest określenie „typowości przypadku". Biegły ocenia badany przez siebie przypadek, porównując go z przypadkiem typowym. Ten ostatni z samej istoty ma charakter statystyczny. „Typowy" znaczy tu bowiem zwykle tyle, co „najczęściej występujący". Już z tego wynika, że istnieją przypadki inne niż typowe. Prawdziwość oceny, czy badany przypadek można porównywać z typowym, a tym bardziej szczegółowość i poprawność tej oceny, zależą od wielu czynników. Wśród nich niewątpliwie istotne są takie jak aktualny stan danej dziedziny wiedzy (wynikający

---

17 Por. Ch. O'Hara, J.W. Osterburg, *An introduction to criminalistics. The application of the physical sciences to the detection of crime*, Bloomington–London, 1972, s. 680 i n.; T.F. Kiely, *Forensic evidence: science and the criminal law*, Boca Raton–London–New York–Washington DC 2001, s. 2 i n.; T. Vosk, *Measurement uncertainty*, [w:] *Encyclopedia of Forensic Sciences*, red. J.A. Siegel, P.J. Saukko, Oxford 2013, s. 322–331; C.G.G. Aitken, F. Taroni, A. Biedermann, *Statistical interpretation of evidence: Bayesian analysis*, [w:] *Encyclopedia…*, s. 292–297; B. Robertson, G.A. Vignaux, Ch. Berger, *Interpreting evidence. Evaluating forensic sciences in the courtroom*, wyd. 2, Oxford 2016, s. 62–63, 102–104.

z jej zaawansowania metodologicznego, liczby przeprowadzonych badań naukowych), liczba podobnych przypadków znanych biegłemu, liczby znanych mu przypadków nietypowych, wreszcie stan literatury naukowej (jej rozległości i jakości) z tej dziedziny, liczba i jakość ośrodków naukowych zajmujących się tą dziedziną. Znaczenie mają też stopień znajomości tego stanu wiedzy przez eksperta (czyli faktyczny zakres jego „wiadomości specjalnych"), jego doświadczenie zawodowe, ale także jego warunki osobiste, zdolności, a na koniec jego motywacje, podlegające najrozmaitszym wpływom – zarówno wewnętrznym, jak i zewnętrznym, w tym także sugestiom z różnych źródeł. Te sugestie płynąć mogą od osób prowadzących śledztwo albo być konsekwencją znajomości wyników innych ekspertyz przeprowadzonych w tej sprawie, sugerujących, że wynik aktualnie realizowanego badania nie powinien być z nimi sprzeczny. Wszystkie te czynniki składają się na margines subiektywizmu eksperta. W dużej mierze od jego doświadczenia zależy, czy trafnie rozpozna „typowość przypadku", która jest później punktem odniesienia przy ocenie konkretnego przypadku. Wydaje się, że powyższe dostatecznie wyjaśnia, dlaczego biegli, dysponujący identyczną wiedzą podręcznikową, opierający się na identycznych danych wydają czasem rozbieżne opinie[18].

Piszący – bodajże jako jedyny w polskiej literaturze – o subiektywizmie w badaniach kryminalistycznych Jarosław Moszczyński zauważa, że: „w badaniach kryminalistycznych istnieją obszary subiektywizmu rozumiane jako obszary nie w pełni sprecyzowane, w których biegły posiada większą lub mniejszą swobodę interpretacji, ocen oraz dokonywania rozstrzygnięć"[19].

Nie wdając się w logiczny rozbiór powyższego zdania (o jakie sprecyzowanie jakich obszarów chodzi? jakich rozstrzygnięć ma dokonywać biegły?), przyjmijmy, że przez element subiektywizmu w ekspertyzie (z zakresu nauk sądowych) rozumieć będziemy, występujący przy ocenie dokonywanej przez eksperta, margines swobody jego interpretacji, niepodlegający żadnym obiektywnym kryteriom.

Szerokość tego marginesu jest różna przy różnych rodzajach ekspertyz. Przy daktyloskopii jest on relatywnie niewielki, a na przykład w ekspertyzie pisma – bardzo szeroki[20]. Zależy to zarówno od stopnia złożoności materii,

---

[18] Por. K. Jaegermann, Z. Marek, *Rozbieżności w opiniach sądowo-lekarskich*, „Archiwum Medycyny Sądowej i Kryminologii" 1979, t. 29, nr 4, s. 249; por. K. Jaegermann, *op. cit.*, s. 67.

[19] J. Moszczyński, *Subiektywizm w badaniach kryminalistycznych*, Olsztyn 2011.

[20] Por. H.H. Harralson, *Developments in handwriting and signature identification in the digital age*, Oxford, 2013, s. 120 i n., por. także: S.N. Srihari, S. Cha, H. Arora, S. Lee, *Individuality of handwriting*, „Journal of Forensic Sciences" 2002, t. 47, nr 4, s. 1–17; S.J. Strach, *Probability conclusions in handwriting comparison*, „International Journal of Forensic Document Examiners" 1998, t. 4, nr 4, s. 313–317.

jak i od zaawansowania metodologicznego danej dyscypliny identyfikacji (badań).

Ostatnio w kryminalistyce zauważalny jest nurt kwestionujący dotychczasowe pojęcie identyfikacji indywidualnej[21]. Faktycznie, zalecenie, by biegły we wnioskach opinii pisał „uważam, że jest tak a tak" zamiast, jak to jest dziś powszechnie przyjęte, „stwierdzam, że jest tak a tak", ma głębokie uzasadnienie metodologiczne i świadomie lub nie, uwzględnia margines subiektywizmu eksperta.

Zakres subiektywizmu przy ewaluacji zapisów na wykresach w badaniach poligraficznych

W badaniach poligraficznych przedmiotem oceny są zapisy krzywych wykreślanych przez poligraf, z których każda przedstawia zmiany w zakresie jednej z funkcji fizjologicznych organizmu osoby badanej. Standardowy poligraf rejestruje zmiany w przebiegu czynności oddychania, w pracy układu krążenia (zmiany częstotliwości tętna, względne zmiany ciśnienia krwi, zmiany reakcji skórno-galwanicznej, czasem dodatkowo inne parametry, jak na przykład zmiany reakcji pletysmograficznej, która zresztą jest konsekwencją modyfikacji w pracy układu krążenia)[22]. Są one wskaźnikiem zmian aktywności organizmu, na które składają się zarówno wywołane pytaniami testu zmiany emocjonalne, jak i zmiany będące efektem wysiłku intelektualnego, koniecznego przy kłamstwie, które nazywamy najczęściej, choć niezupełnie ściśle „fizjologicznymi korelatami emocji"[23].

Zadaniem biegłego jest odróżnienie i ocena, które zmiany na krzywych są symptomatyczne i zostały wywołane treścią pytania testowego, a które są efektem rozlicznych artefaktów, a także porównywanie reakcji na pytania krytyczne (relewantne, związane – ang. *relevant*) i kontrolne (porównawcze – ang. *comparison*) lub obojętne (niezwiązane – ang. *irrelevant*). Biegły dokonuje zatem oceny zaobserwowanej zmiany przebiegu krzywych wykreślanych przez poligraf (rzadziej – odczytuje wielkości zmiany w reakcjach, wyrażone liczbowo w określonych jednostkach miary, np. w omach dla GSR/EDA), uznaje związek tej zmiany z treścią pytania, a na koniec – porównuje ją do jakiegoś wzorca. Jak widać – mamy tu do czynienia z dwoma ważnymi dla oceny elementami: uznaniem czegoś (jakiegoś kształtu krzywej) za wzorzec

---

21 Por. J. Konieczny, *Zmiana paradygmatu czy kryzys kryminalistyki?*, „Państwo i Prawo" nr 1, 2012, s. 3–16; idem, *Przeciwko kryminalistycznej identyfikacji indywidualnej*, „Problemy Współczesnej Kryminalistyki" 2014, t. 17, s. 49–58.

22 Por. *Kryminalistyka*, red. J. Widacki, wyd. 3, Warszawa 2016, s. 418.

23 *Ibidem*, s. 419.

i dokładnym porównaniem zaobserwowanej w konkretnym badaniu zmiany z tym wzorcem.

Wspomniany wyżej wzorzec wraz z rozwojem techniki badań poligraficznych ulegał zmianie. W pierwszych latach stosowania poligrafu za symptomatyczną uznawano każdą zmianę w przebiegu krzywej, z wyjątkiem takiej, której przyczyna wydawała się oczywista i była inna niż pytanie testowe – na przykład: poruszenie się badanego, jego kichnięcie, kaszel – które to fakty odnotowywano zresztą na taśmie poligrafu. Tak więc każda zmiana przebiegu krzywej, jeśli nie była spowodowana wspomnianym poruszeniem się, kichnięciem, kaszlem itp., uznawana była za symptomatyczną. Natomiast wysoką częstotliwość, a zwłaszcza powtarzalność takich oczywistych zaburzeń zapisu oceniano z reguły jako próbę świadomego utrudniania badania, a więc coś, co przy końcowej ocenie wiarygodności badanego nie jest bez znaczenia[24].

Z czasem uznano, że „symptomami kłamstwa" są niektóre tylko zmiany w przebiegu krzywych. Powstawały całe atlasy ilustrujące przykładowe reakcje uznane za typowe. Takim swoistym atlasem były na przykład liczne ilustracje w kolejnych wydaniach książki Johna Reida i Freda Inbaua[25]. Takimi atlasami, opracowanymi w oparciu o polskie badania eksperymentalne, były w pewnym sensie publikacje z końca lat 70.[26]. Podobne powstają i dziś[27]. Obecnie różne techniki badań uznają za symptomatyczne tylko niektóre zmiany w przebiegu krzywych, odmawiając takiego waloru zmianom do niedawna i w innych technikach uznawanych za symptomatyczne[28].

Jak widać, uznanie, co ostatecznie jest wzorcem dla reakcji symptomatycznej – pod wpływem doświadczeń praktyki i eksperymentów naukowych – również podlega ewolucji. Należy pamiętać, że ekspert w każdym konkretnym badaniu uzyskane reakcje porównuje z wzorcem – widać więc, że ta ocena, zazwyczaj nieoparta na dodatkowych pomiarach, jest jedynie interpretacją eksperta. Z praktyki wiadomo, że eksperci oceniający te same zapisy poligraficzne dokonywali niekiedy odmiennych, nieraz sprzecznych ocen. Fakt istnienia takich rozbieżności słusznie przypisywano m.in. subiektywizmowi ekspertów.

---

[24] Por. J. Widacki, *Analiza przesłanek diagnozowania w badaniach poligraficznych*, Katowice 1982, s. 81–84.

[25] J. Reid, F. Inbau, *Truth and deception. The polygraph ("lie-detector") technique*, Baltimore 1977.

[26] Por. J. Widacki, *Z rozważań nad istotą symptomów kłamstwa przy lie-detection test*, „Archiwum Medycyny Sądowej i Kryminologii" 1975, t. 25, nr 1; idem, *Wartość diagnostyczna badania poligraficznego i jej znaczenie kryminalistyczne*, Kraków 1977 (ryc. 7, 8, 9).

[27] Por. np. А.Ю. Молчанов, Н.А. Молчанова, *Атлас полиграмм*, Ярославль [A.J. Mołczanow, N.A. Mołczanowa, *Atlas poligramów*, Jarosław], 2007.

[28] Por. J. Widacki, *Historia badań poligraficznych*, Kraków 2017, s. 126–127.

Począwszy od lat 40. XX w., metodą tzw. ślepej interpretacji (ang. *blind scoring*) ustalano empirycznie zakres rozbieżności ocen ekspertów oceniających te same zapisy poligraficzne. Jak opisuje w niepublikowanej pracy doktorskiej Fabian Rouke[29], dwóch ekspertów, niezależnie od siebie, otrzymało do oceny zapisy reakcji uzyskanych przez innych ekspertów w czasie eksperymentu laboratoryjnego (były to jedynie zapisy reakcji skórno-galwanicznej). Ich zadaniem było dokonanie oceny reakcji i wskazanie „sprawcy". Jeden ekspert wskazał „sprawcę" trafnie w 91% przypadków, drugi – w 88%. Zatem odsetek trafnych wskazań ekspertów różnił się tylko o 3 punkty procentowe[30]. W innym eksperymencie[31], sześciu ekspertów dostało do niezależnej oceny taśmy z zapisami z 25 autentycznych spraw (których wynik był znany), wyselekcjonowanych z archiwum policji w Miami. Rozpiętość wyników była tu większa niż w eksperymencie Rouke'go: od 69 do 81% trafnych ocen reakcji. Zatem rozbieżność trafności ocen wśród sześciu ekspertów wynosiła 12 p.p. Jeszcze większą rozpiętość wyników wykazał eksperyment Josepha Kubisa[32]: od 73 do 92 % trafnych wskazań, czyli rozbieżność wynosiła 19 p.p. W badaniach Franka Horvatha i Johna Reida[33] taka rozbieżność między trafnymi ocenami ekspertów wynosiła aż 27,5 p.p. W kolejnych badaniach Horvatha[34] rozbieżność trafnych ocen między 10 ekspertami oceniającymi ten sam materiał wynosiła aż 31 p.p. W innych badaniach amerykańskich (Stanley Slowik i Joseph Buckley[35], Douglas Wicklander i Fred Hunter[36]) różnica ocen dotyczyła odpowiednio 12,8 i 25 p.p.

---

29 F.L. Rouke, *Evaluation of the indices of deception in the psychogalvanic technique*, 1941 (niepublikowana praca doktorska, Fordham University), [cyt. za:] G.H. Barland, *Reliability of polygraph chart evaluations*, „Polygraph" 1972, t. 1, nr 4, s. 192.

30 Tak liczona rozbieżność ocen jest przybliżona i może mieć znaczenie tylko orientacyjne. Pokazuje w istocie jedynie, że eksperci, oceniając ten sam zbiór zapisów, różnili się w swoich ostatecznych ocenach.

31 W.D. Holmes, *The degree of objectivity in chart interpretation*, [w:] *Academy Lectures on Lie-detection*, t. 2, Springfield 1957, s. 67–70.

32 J.F. Kubis, *Studies in lie-detection. Computer feasibility considerations*, Fordham University, New York, 1962, RADC-TR 62-205, Project No 5534, AF 30(602)-2270, prepared for Rome Air Development Center, Air Force Systems Command, USAF, Griffiss AFB, New York.

33 F. Horvath, J. Reid, *The reliability of polygraph examiner diagnosis of truth and deception*, „Journal of Criminal Law, Criminology and Police Science", t. 62, nr 2, 1971, s. 276–281.

34 F. Horvath, *The accuracy and reliability of Police polygraphic (lie-detector) examiners' judgements of truth and Deception. The effect of selected variables*, 1974 (niepublikowana praca doktorska, Michigan State University).

35 S. Slowik, J.P. Buckley, *Relative accuracy of polygraph examiner diagnosis of respiration, blood pressure and GSR recordings*, „Journal of Police Science and Administration" 1975, t. 3, nr 3, s. 305–309.

36 D.E. Wicklander, F.L. Hunter, *The influence of auxiliary sources of information in polygraph diagnosis*, „Journal of Police Science and Administration" 1975, t. 3, nr 4, s. 405.

W jedynych jak dotąd polskich badaniach[37] tego typu, z końca lat 70. XX w., dwóch ekspertów o identycznym przeszkoleniu i podobnym doświadczeniu zawodowym oceniało zapisy poligraficzne z badań eksperymentalnych 80 osób. Jeden z ekspertów uzyskał 85% trafnych wskazań, drugi – 67,5%. Jak widać, rozbieżność ich wyników wynosiła 17,5 p.p. Tabela 1. przedstawia podsumowanie wyżej opisanych rozbieżności między ekspertami w ocenach reakcji.

Tabela 1. Rozpiętość odsetka trafnych opinii ekspertów w wybranych badaniach naukowych

Autorzy (rok publikacji)	**Rozpiętość ocen trafnych (%)**	**Rozbieżność ocen miedzy ekspertami (p.p.)**
Rouke (1941)	88–91	3
Holmes (1957)	69–81	12
Horvath i Reid (1971)	70–97,5	27,5
Kubis (1973)	73–92	19
Horvath (1974)	–	31
Slowik i Buckley (1975)	–	12,8
Wicklander i Hunter (1975)	70–95	25
Widacki (1977)	67,5–85	17,5

Źródło: opracowanie własne Autorów.

Należy zwrócić uwagę, że wszystkie te badania wykonane były z zastosowaniem jakościowych metod interpretacji zapisów. Pomijając pierwsze z badań – wykonane tylko na zapisach psychogalwanometru (najłatwiejszych do oceny) – rozbieżności trafności interpretacji reakcji w różnych badaniach wynosiły od 12 nawet do 31 p.p. Te różnice trafności ocen różnych ekspertów oceniających ten sam materiał mogły mieć różne przyczyny. Można zasadnie podejrzewać, że jedną z nich był subiektywizm ekspertów, rozumiany tak, jak zdefiniowaliśmy go wyżej.

Badań, które bezpośrednio określałyby, do jakiego stopnia element subiektywny miał wpływ na dokładność wyniku badania poligraficznego i do jakiego stopnia oceny poszczególnych poligraferów dokonane na podstawie tego samego materiału są zgodne, jest stosunkowo niewiele. Jeden z pierwszych takich eksperymentów zrealizowano na zlecenie Departamentu Obrony USA[38]: 30 poligraferów w ślepej interpretacji oceniało po 90 zapisów (po

[37] J. Widacki, *Wartość diagnostyczna badania poligraficznego…*

[38] P.J. Bresh, R.A. Brisentine, *The reliability of blind interpretation of polygraph record for lie-detection purposes (Report prepared for DoD)*, 1968; [cyt. za:] J. Widacki, *Wprowadzenie do problematyki badań poligraficznych*, Warszawa 1981, s. 135–136.

30 wykonanych technikami Backstera, Reida i POT), z których połowa dotyczyła badanych zakwalifikowanych do zbioru DI (nieszczerych – ang. *deceptive*), połowa – do NDI (prawdomównych – ang. *non-deceptive*). Współczynnik korelacji Kappa w zależności od stosowanej techniki wynosił od 0,15 do 0,51[39]. Najwyższą zgodność stwierdzono przy badaniach wykonanych techniką Backstera.

Podobny był zrealizowany kilka lat później eksperyment japoński, przeprowadzony przez Akihiro Suzuki i współpracowników[40]: 26 ekspertów oceniało zapisy z 30 spraw realizowanych przez trzech ekspertów. Zgodność ocen mierzona współczynnikiem rzetelności Spearmann-Browna wynosiła 0,798.

W roku 2003 opublikowano raport opracowany dwa lata wcześniej dla Department of Defence Polygraph Institute[41]. Na podstawie analizy literatury porównuje on m.in. zgodność ocen (diagnoz) w badaniach psychologicznych, medycznych i poligraficznych. W badaniach psychologicznych ocena zgodności dotyczyła diagnoz DSM-III i DSM-IV, a w diagnostyce medycznej – badań rentgenologicznych, ultrasonograficznych, tomograficznych i rezonansu magnetycznego, a zatem zaawansowanych technologicznie metod diagnostyki. Według raportu, mierzona współczynnikiem Kappa zgodność ocen poligraferów wynosiła 0,77 i była porównywalna ze zgodnością ocen psychologów (która wynosiła 0,79) i była znacznie wyższa niż zgodność diagnoz lekarzy (która wynosiła jedynie 0,56).

Wszystkie ukazane wyżej wyniki zdają się potwierdzać tezę, że w badaniach poligraficznych, choć zgodność ocen ekspertów jest relatywnie wysoka – to zwłaszcza, gdy zapisy ocenia się jedynie jakościowo[42], margines subiektywizmu eksperta jest szeroki. Ten margines subiektywizmu z założenia znacznie ograniczyć powinna numeryczna interpretacja zapisu.

Pierwsze próby numerycznej (ilościowej) interpretacji zapisów podjęto na początku lat 60. XX w. W badaniach, które zakończyły się raportem z 1962 r., J. Kubis posłużył się 4-stopniowa skalą oceny reakcji. Reakcji bardzo wyraźnej (ang. *very significant*) przypisywano 3 punkty. Wyraźnej (ang. *significant*)

---

[39] *Ibidem*.

[40] A. Suzuki, S. Watanabe, K. Ohnishi, K. Matsuno, M. Arasuna, *Polygraph examiners' judgements in chart interpretation: Reliability of judgement*, „Kagaku Keisatsu Kenkyusho" (Police Science Report) 1973, t. 26, nr 1, s. 34 i n.

[41] P.E. Crewson, *A comparative analysis of polygraph with other screening and diagnostic tools*, „Polygraph" 2003, t. 32, nr 2, s. 57–85.

[42] Posługujący się dziś tą niewątpliwie przestarzałą już metodyką interpretacji zapisów nazywają ją najczęściej „globalną" lub „holistyczną".

– 2 punkty. Słabej, „wątpliwej" (ang. *doubtfully significant*) – 1 punkt. Gdy reakcji nie stwierdzono (ang. *non-significant*) – stawiano 0 punktów[43].

Jak podaje James Matte[44], przy realizacji eksperymentu Kubisa współpracował Cleve Backster. W 1963 r. Backster dopracował swoją technikę badań poligraficznych[45], której integralną częścią była numeryczna ocena zapisów według 7-stopniowej skali (od +3 do -3). Wedle tej skali Backster zalecał oceniać różnice między reakcjami na pary pytań krytycznych. Ocena miała być dokonywana na każdej krzywej oddzielnie, po czym wyniki podlegały sumowaniu. Jeśli reakcja na pytanie krytyczne była silniejsza niż reakcja na odpowiadające mu pytanie kontrolne, oznaczana była znakiem minus. Wielkość reakcji oceniano podobnie jak u Kubisa: 3 – różnica bardzo wyraźna, 2 – różnica wyraźna, 1 – różnica niewielka, 0 – brak różnicy. W przeciwieństwie do Kubisa, Backster wielkość cyfrową przypisywał nie wprost wielkości reakcji, ale różnicy miedzy reakcją na pytanie krytyczne i odpowiadające mu pytanie kontrolne.

Wedle Backstera oceny należało dokonywać na każdej krzywej oddzielnie. Jeśli więc aparat był 3-kanałowy (pneumo, GSR, kardio), a w teście były trzy pary pytań krytycznych i kontrolnych, maksymalnie wartość reakcji mogła wynieść ∓ 27 pkt (3x3x3). Ze względu na to, że w technice Backstera wykonuje się ten sam test trzy razy, maksymalna liczba punktów mogła wynosić ∓ 81 (27x3). Inaczej mówiąc, wyniki każdego pojedynczego testu mieszczą się na kontinuum od -27 do +27 punktów, a wyniki całego badania – od -81 do +81 punktów. Wedle założeń Backstera, im wynik bliższy pozycji skrajnej, tym jest pewniejszy. Backster przyjął[46], że wynik między -5 a +5 dla pojedynczego testu, a między -15 a +15 dla całego badania, uznać należy za nierozstrzygnięty (INC – ang. *inconclusive*). Wynik między -15 a -81 – pozwala zaliczyć badanego do grupy „nieszczerych" (DI), wynik między +15 a +81 – do grupy „szczerych" (NDI).

Przy metodzie numerycznej oceny zapisów, ostateczna diagnoza (wskazanie) jest prostą konsekwencją wyniku liczbowego. Ten ostatni jest efektem sumowania punktów uzyskanych przy ocenie wielkości reakcji (ściślej: przy ocenie różnic reakcji na pytania krytyczne i kontrolne). Ocenie eksperta podlega tylko wielkość różnicy mię-

[43] J.F. Kubis, *op. cit.*, s. 20–22. Warto przypomnieć, że Kubis wprowadził tę skalę dla ułatwienia oceny zapisów poligraficznych wykonanych w ramach eksperymentu nie przez profesjonalnych poligraferów, ale przez studentów psychologii.

[44] J.A. Matte, *Forensic psychophysiology. Using the polygraph: scientific truth verification – lie detection*, New York 1996, s. 46.

[45] Por. J. Widacki, *Historia badań…*, s. 129.

[46] Wydaje się, że przyjął to arbitralnie, na podstawie własnego doświadczenia, brak bowiem publikowanych prac, z których wynikałoby statystyczne uzasadnienie dla takich założeń.

dzy reakcjami na pytanie krytyczne i odpowiadające mu pytanie lub pytania kontrolne. Ocena ta jest sformalizowana i ograniczona do skali (np. w systemie Utah: dramatyczna różnica; duża różnica; niewielka, ale zauważalna różnica; brak różnicy).

Do najważniejszych skal ocen numerycznych należy zaliczyć[47]:

a) 7- i 3-pozycyjne skale wykorzystywane przy ewaluacji zapisów na poligramach z testów pytań kontrolnych, w tym:
   - 7-pozycyjną Rządu Federalnego Stanów Zjednoczonych (-3, -2, -1, 0, +1, +2, +3),
   - 7-pozycyjną Uniwersytetu w Utah (-3, -2, -1, 0, +1, +2, +3),
   - 3-pozycyjną Rządu Federalnego Stanów Zjednoczonych (-1, 0, +1),
   - 3-pozycyjną w Empirycznym Systemie Oceniania (ESS – ang. *Empirical Scoring System*), zawierającą ten sam przedział ocen jak w federalnej skali 3-pozycyjnej, z wyjątkiem parametru EDA (ang. *electrodermal activity* – aktywność elektrodermalna), gdzie przypisuje się wartości: -2, 0, +2. Modyfikacja ta jest rezultatem badań naukowych, które wykazały, że dane z tego parametru stanowią 50-procentowy wkład w ostateczną wartość decyzyjną i najlepiej korelują z kryterium „winy”;

b) skalę ocen testu ukrytych informacji (CIT – ang. *Concealed Information Test*), z wykorzystaniem systemu Lykkena (0, 1, 2).

W skalach 3-pozycyjnych przeznaczonych dla testów pytań kontrolnych obowiązuje prosta zasada: „większe jest lepsze” (ang. *bigger is better*). Natomiast skale 7-pozycyjne są nieco bardziej skomplikowane i w zależności od relatywnej wielkości reakcji istnieje więcej wariantów oceny. Każdą z powyższych skal uznaje się za potwierdzoną naukowo, ale federalna skala 3-pozycyjna nie spełnia obowiązujących standardów praktyki największego zrzeszenia ekspertów za zakresu badań poligraficznych na świecie – American Polygraph Association (APA). Testy oceniane wyłącznie tą skalą dają wyższy odsetek wyników nierozstrzygniętych niż określony przez APA limit 20%. Federalna skala 3-pozycyjna jest dopuszczalna pod warunkiem wykorzystania dodatkowo skali 7-pozozycyjnej przy braku rozstrzygnięcia po pierwszym liczeniu.

W teście ukrytych informacji, zwanym także testem wiedzy o czynie (GKT – ang. *guilty knowledge test*), oceny przypisuje się na innych zasadach niż w testach pytań kontrolnych. W tym przypadku ukryty w pojedynczej serii testu bodziec krytyczny – „klucz” (np. nazwa lub cecha przedmiotu, sposób wykonania jakiejś czynności, fotografia miejsca czy osoby związanej ze sprawą) porównywany jest z bodźcami neutralnymi – równie prawdopodob-

---

47 Por. *Współczesne standardy badań poligraficznych*, red. M. Gołaszewski, Warszawa 2013, s. 26–43.

nymi z punktu widzenia osoby niewinnej. Dana seria testu (podtest) jest oceniana na 2 – jeśli najbardziej znaczące zmiany reakcji fizjologicznych badanego wystąpiły przy bodźcu krytycznym; 1 – gdy była to druga pod względem istotności reakcja; 0 – w pozostałych przypadkach. Metoda ewaluacji Lykkena opiera się wyłącznie na danych z komponentu EDA. Wynik testu równy liczbie poddanych ewaluacji podtestów oznacza identyfikację pozytywną (RI – ang. *recognition indicated* – stwierdzono rozpoznanie). Natomiast rezultat mniejszy od liczby podtestów zakwalifikowanych do oceny świadczy o braku rozpoznania kluczowego bodźca przez osobę badaną (NRI – ang. *no recognition indicated*). Opinia może być ponadto poparta liczbowo – poprzez określenie prawdopodobieństwa nierozpoznania szczegółów zdarzenia przez badanego.

Współcześnie numeryczna (ilościowo-jakościowa) interpretacja zapisów na poligramach jest immanentną i konieczną częścią wszystkich w zasadzie stosowanych technik badań poligraficznych. Eksperci korzystają głównie z trzech potwierdzonych naukowo systemów numerycznej analizy danych testowych (TDA – ang. *test data analysis*), obejmujących określone kryteria diagnostyczne przy interpretacji zapisów, rodzaje ocen i sposób ich łączenia oraz reguły decyzyjne. Te systemy skrótowo nazywamy: „federalnym", „Utah" oraz „ESS". Wszystkie są modyfikacjami metody opracowanej przez Backstera, a różnice występujące pomiędzy są nieznaczne.

System Rządu Federalnego USA opierał się najpierw na 22 cechach diagnostycznych nauczanych przez Szkołę Policji Wojskowej (USAMPS – ang. *United States Army Military Police School*)[48]. Tak duża liczba kryteriów utrzymywała się od lat 70. XX w. aż do 2006 r. Wówczas Instytut Badań Poligraficznych Departamentu Obrony USA (DoDPI – ang. *Departament of Defense Polygraph Institute*) – biorąc pod uwagę podstawy naukowe oraz ideę uproszczenia systemu – zredukował liczbę cech diagnostycznych do ośmiu głównych i trzech pomocniczych.

W latach 70. XX w. zespół naukowców z Uniwersytetu Utah (Salt Lake City), pod kierunkiem prof. Davida Raskina, rozpoczął pracę nad udoskonaleniem technik pytań kontrolnych. Już wtedy wiadomo było, że ocena numeryczna umożliwia istotnie większą trafność niż inne metody analizy danych z testów poligraficznych. Uznano natomiast, że znane dotąd systemy były niedoskonałe. Niektóre reguły decyzyjne czy cechy przyjmowane za diagnostyczne nie miały dostatecznych podstaw naukowych. Postanowiono więc zmodyfikować system oceny Backstera, który opierał się na zbyt skompliko-

---

48 Zob. R.S. Weaver, *The numerical evaluation of polygraph charts: Evolution and comparison of three major systems*, „Polygraph" 1980, t. 9, nr 2, s. 94–108.

wanych zasadach i był niekorzystny dla osób prawdomównych[49]. W efekcie opracowano podejście „Utah” do testów pytań porównawczych wraz z systemem numerycznej analizy danych testowych, uwzględniającym dziewięć przesłanek diagnozowania. Koncepcja została potwierdzona wieloma badaniami i zrecenzowanymi publikacjami naukowymi w ciągu kolejnych 40 lat.

Najnowszą metodą jest tzw. Empiryczny System Oceniania, po raz pierwszy opisany w 2008 r. przez Raymonda Nelsona, Marka Handlera i Donalda Krapohla[50]. Autorzy wyszli z założenia, że należy opracować system, który zapewni jak największą zgodność między ekspertami i trafność decyzyjną, będzie nieskomplikowany, doskonale udokumentowany naukowo, możliwy do skomputeryzowania i powszechnego stosowania zarówno przez doświadczonych, jak i niedoświadczonych poligraferów. Liczbę cech diagnostycznych ograniczono do sześciu. Do zalet systemu należy możliwość wskazania przez biegłego znaczenia statystycznego konkretnego rezultatu testu[51]. Wyniki końcowe numerycznej ewaluacji poligrafów uzyskuje się poprzez odniesienie się do ustalonych progów decyzyjnych, w zależności od przyjętej tolerancji błędu (standardowo 5–10%), wymaganego poziomu statystycznej istotności i prawdopodobieństwa błędu na podstawie reprezentatywnych danych[52].

Wspólnymi cechami diagnostycznymi dla wszystkich trzech głównych systemów analizy danych testowych są: tłumienie oddechu i podniesienie linii bazowej krzywej oddychania; zmiany w amplitudzie w kanale EDA; podniesie linii bazowej w kanale kardio oraz redukcja amplitudy pulsu w kanale reakcji naczynioruchowych.

---

49 B.G. Bell, D.C. Raskin, Ch.R. Honts, J.C. Kircher, *The Utah Numerical Scoring System*, „Polygraph” 1999, t. 28, nr 1, s. 1–9.

50 R. Nelson, D.J. Krapohl, M. Handler, *Brute-force comparison: A Monte Carlo Study of the Objective Scoring System version 3 (OSS-3) and human polygraph scorers*, „Polygraph” 2008, t. 37, nr 3, s. 185–215.

51 Wartość diagnostyczna testu poligraficznego jest rozumiana jako uzyskana empirycznie (w badaniach naukowych obejmujących potwierdzone rozwiązania spraw) średnia dokładność (trafność) danego testu – czyli liczba prawidłowych identyfikacji, z wyłączeniem wyników nierozstrzygających, skoro te ostatnie nie skutkują żadną decyzją. Natomiast znaczenie statystyczne wyniku testu wylicza się na podstawie danych normatywnych właściwych dla populacji osób odpowiadających szczerze i nieszczerze (ewentualnie rozpoznających i nierozpoznających ukryte w teście szczegóły związane z określonym zdarzeniem). W oparciu o dane normatywne dotyczące poszczególnych rezultatów testów oraz przyjmowaną tolerancję błędu (wartość α – wynoszącą zazwyczaj 0,05 przy kwalifikowaniu reakcji jako typowe dla osoby odpowiadającej nieszczerze) ustala się optymalne progi decyzyjne przy analizie numerycznej zapisów reakcji badanego.

52 M. Handler, R. Nelson, W. Goodson, M. Hicks, *Empirical Scoring System: A cross-cultural replication and extension study of manual scoring and decision policies*, „Polygraph” 2010, t. 39, nr 4, s. 202.

Reguły decyzyjne w poszczególnych systemach ewaluacji danych zależą od typu testu (wieloproblemowy, jednoproblemowy, wieloaspektowy) i jego formatu (np. ZCT – ang. *Zone Comparison Test*; DLST – ang. *Directed Lie Screening Test*). Zostały wyznaczone w sposób arbitralny (jak u Backstera) lub na podstawie badań empirycznych, które dostarczyły danych do analiz statystycznych (jak w przypadku systemu ESS).

Warto pamiętać, że zapisy na poligramach są wizualną reprezentacją danych rejestrowanych przez poligraf. W przeszłości, kiedy stosowano aparaty analogowe, biegły mógł polegać wyłącznie na własnej obserwacji, obarczonej potencjalnymi zniekształceniami związanymi ze zdolnością postrzegania ludzkiego oka i dość dużą przestrzenią do subiektywnej interpretacji. Miał możliwość skorzystania co najwyżej z podręcznego narzędzia pomiaru w postaci linijki. Ponadto, nie mógł proporcjonalnie zwiększyć ani zmniejszyć krzywych na raz zadrukowanej rolce z wykresami. Jeśli w trakcie testu niedostatecznie wyregulował czułość jednego z czujników, mógł nie dostrzec istotnej różnicy w porównywanych reakcjach organizmu badanego. Długość linii oddechu (jedną z cech diagnostycznych) mógł sobie jedynie wizualizować w wyobraźni. Współcześnie, większość badań przeprowadza się z wykorzystaniem poligrafów komputerowych, współpracujących z oprogramowaniem, które eliminuje opisane trudności. Graficzne przedstawienie reakcji można dostosowywać zarówno na bieżąco w trakcie testu, jak i po jego zakończeniu (odbywa się to oczywiście automatycznie, z zachowaniem odpowiednich proporcji zarejestrowanych danych źródłowych, bez manipulacji). Opcje specjalistycznych programów pozwalają m.in. na pomiar amplitudy reakcji (wyrażonej liczbowo lub na podstawie graficznego wykresu) czy długości linii krzywej. Powstały algorytmy (np. do ewaluacji kanału reakcji naczynioruchowych), które dokładnie wyliczają różnice między fragmentami analizowanych zapisów. Bywa to niekiedy pomocne, choć np. reguły systemu ESS, w duchu uproszczenia procedury, zalecają powstrzymanie się od stosowania jakichkolwiek dodatkowych narzędzi pomiaru.

Dzięki komputeryzacji potencjalny subiektywizm w zakresie samej interpretacji zapisów został istotnie ograniczony. Co więcej – stworzono programy dokonujące całościowej automatycznej ewaluacji danych i o ile mają one swoje ograniczenia wynikające np. ze sztywnego pola oceny reakcji i nie mogą zastąpić eksperta przy identyfikacji artefaktów – o tyle w analizowanym tu kontekście gwarantują pełen obiektywizm i perfekcyjną niezawodność. Niektóre z programów (np. OSS-3[53] czy algorytm dołączony do oprogramowania systemu poligrafu CPS) dostarczają wyników na pozio-

---

53 Por. R. Nelson, D.J. Krapohl, M. Handler, *op. cit.*

mie trafności podobnym do manualnych ocen ekspertów, a bywa nawet, że je przewyższają[54].

Wprowadzenie numerycznej oceny zapisów i liczbowych kryteriów zaliczenia badanego do jednej z trzech grup: DI, NDI lub INC, wyeliminowało praktycznie te dodatkowe przesłanki diagnozowania w badaniach poligraficznych, które polegały na szeroko pojętej i z natury rzeczy jedynie jakościowej ocenie zachowania badanego przed badaniem, w jego trakcie i po nim[55]. Poprzednio do tych symptomów przywiązywano dużą wagę i uwzględniano je przy podejmowaniu decyzji o końcowym wyniku badania[56].

Część empiryczna

Hipotezy badawcze

Biorąc pod uwagę przedstawione rozważania teoretyczne, sformułowano następujące hipotezy badawcze:

1) Metody numeryczne pozostawiają mniejszy margines subiektywizmu niż metody jakościowe (globalne, holistyczne).
2) Metody o mniejszej skali są bardziej obiektywne niż te o skalach 7-pozycyjnych. Pozostawiają bowiem mniejszy margines na subiektywną ocenę eksperta.
3) Oceniający metodą ślepej interpretacji – dzięki wyeliminowaniu innych niż zapisy na wykresach czynników (świadomie lub nieświadomie branych pod uwagę przy wydawaniu opinii) – uzyskają większy odsetek trafnych identyfikacji niż eksperci przeprowadzający badania.

Możliwe jest również przyjęcie hipotezy odwrotnej – tzn. że to badający będą dokładniejsi w swoich ocenach niż niezależni recenzenci, którzy nie mieli styczności z osobą badaną i nie znali przedmiotu badania. Oznaczałoby to, że w trakcie ekspertyzy, poza poligramami, biegły kieruje się mimo wszystko jeszcze innymi, subiektywnie odbieranymi przesłankami (np. niewerbalnymi, albo opiniami innych biegłych czy własną oceną faktów poznanych z akt sprawy) i co ważne – odbywa się to z korzyścią dla dokładności diagnozowania.

---

54 Szerzej na ten temat: National Research Council, *The polygraph and lie detection*, Washington DC, 2003, s. 298–322; A.B. Dollins, D.J. Krapohl, D.W. Dutton, *A comparison of computer programs designed to evaluate psychophysiological detection of deception examinations: Bakeoff 1*, „Polygraph" 2000, t. 29, nr 3, s. 237–257.

55 Por. J. Widacki, *Analiza przesłanek diagnozowania…*

56 Por. F. Horvath, *Verbal and nonverbal clues to truth and deception during polygraph examinations*, „Journal of Police Science and Administration" 1973, t. 1, nr 2, s. 138–152.

Na tę drugą możliwość zdaje się wskazywać znaczna część spełniających wysokie standardy naukowe analiz dotyczących trafności wskazań przy badaniach poligraficznych z wykorzystaniem technik pytań kontrolnych w rzeczywistych sprawach. Tabela 2 przedstawia wyniki testów ocenianych przez niezależnych ekspertów („na ślepo”), natomiast tabela 3 – rezultaty uzyskane przez poligraferów, którzy sami przeprowadzali badania. Odsetek prawidłowych wskazań osób winnych był zbieżny dla obu grup, natomiast badający byli trafniejsi od „ślepo” oceniających zapisy w przypadku wskazywania osób niewinnych.

Tabela 2. Wyniki badań terenowych na temat dokładności testów CQT ocenianych przez niezależne osoby

Autorzy* (rok publikacji)	**Winni (%)**				**Niewinni (%)**			
	n	**Identyfikacja prawidłowa**	**Identyfikacja błędna**	**Test nierozstrzygnięty**	**n**	**Identyfikacja prawidłowa**	**Identyfikacja błędna**	**Test nierozstrzygnięty**
Honts (1996)	7	100	0	0	6	83	0	17
Honts i Raskin (1988)	12	92	0	8	13	62	15	23
Patrick i Iacono (1991)	52	92	2	6	37	30	24	46
Raskin i in. (1988)	37	73	0	27	26	61	8	31
Średnia	27	89	1	10	82	59	12	29
Procent opinii		98	2	–		83	17	–

* Ch.R. Honts, *Criterion development and validity of the control question test in field application*, „Journal of General Psychology” 1996, t. 123, s. 309–324; Ch.R. Honts, D.C. Raskin, *A field study of the Directed Lie Control Question*, „Journal of Police Science and Administration” 1988, t. 16, s. 56–61; C.J. Patrick, W.G. Iacono, *Validity of the Control Question Polygraph Test: The problem of sampling bias*, „Journal of Applied Psychology” 1991, t. 76, nr 2, s. 229–238; D.C. Raskin i in., *A study of the validity of polygraph examinations in criminal investigations*, 1988 (raport końcowy dla National Institute of Justice, University of Utah).

Źródło: opracowanie własne na podstawie danych udostępnionych dzięki uprzejmości prof. D. Raskina.

Tabela 3. Wyniki badań terenowych na temat dokładności testów CQT ocenianych przez osoby przeprowadzające testy

Autorzy* **(rok publikacji)**	**Winni (% prawidłowych identyfikacji)**	**Niewinni (% prawidłowych identyfikacji)**
Horvath (1977)	100	100
Honts i Raskin (1988)	92	100
Raskin i in. (1988)	95	96
Patrick i Iacono (1991)	100	90
Honts (1996)	94	100
Średnia	98	97

* Jak w tab. 2 i dodatkowo: F. Horvath, *The effect of selected variables on interpretation of polygraph records*, „Journal of Applied Psychology” 1977, t. 62, nr 2, s. 127–136.

Źródło: opracowanie własne na podstawie danych udostępnionych dzięki uprzejmości prof. D. Raskina.

Opis eksperymentu

W finansowanym przez Narodowe Centrum Nauki projekcie naukowo-badawczym, zatytułowanym *Instrumentalne i nieinstrumentalne metody detekcji nieszczerości – problemy kryminalistyczne, etyczne i prawne*, realizowanym przez Krakowską Akademię im. Andrzeja Frycza Modrzewskiego, wzięło udział 15 ekspertów z zakresu badań poligraficznych. Trzech z nich przeprowadziło łącznie 39 badań osób, z których część odgrywała w eksperymencie rolę „winnych” (osoby, które oddały strzał na uniwersyteckiej strzelnicy sportowej), a pozostali nie mieli bezpośredniego związku z zainscenizowanym zdarzeniem.

Dla wzmocnienia motywacji, każdy z badanych, bez względu na rolę, jaka odgrywał w eksperymencie („winny” czy „niewinny”), otrzymywał banknot 50-złotowy. Osoby, które strzelały na strzelnicy i miały zataić ten fakt przed badającym poligraferem, były instruowane, że jeśli poligrafer wskaże je jako kłamiące, będą musiały zwrócić 50 zł. Jeśli poligrafer nie rozpozna, że w czasie badania kłamały, 50 zł stanie się ich własnością. Z kolei osoby, które nie strzelały na strzelnicy i miały odgrywać role osób „niewinnych” i prawdomównych, były instruowane, że jeśli poligrafer potwierdzi ich prawdomówność, 50 zł stanie się ich własnością. Jeśli zaś poligrafer błędnie wskaże je jako kłamiące, będą musiały zwrócić banknot. Miało to w założeniu przybliżyć motywację uczestników eksperymentu do motywacji osób badanych w autentycznych sprawach. Dla osoby kłamiącej błąd poligrafera jest w nich korzystny, dla osoby prawdomównej – niekorzystny. Badający wykorzystywali test jednoproblemowy[57] w formacie *Utah Zone Comparison Test* (test porównania stref, opracowany przez Uniwersytet Utah), o średniej dokładności wynoszącej 90,2–93% i średnim odsetku wyników nierozstrzygniętych na poziomie 9,8–10,7% (w zależności od rodzaju zastosowanych pytań porównawczych i systemu analizy danych testowych)[58].

Kolejnych 12 ekspertów miało za zadanie dokonać ślepej interpretacji zapisów na poligramach każdego z tych 39 badań. „Ślepi recenzenci" zostali podzieleni na trzy 4-osobowe grupy, z których każda miała posługiwać się inną metodą analizy danych testowych (globalną i dwoma numerycznymi: ESS i Utah). Dodatkowo wykorzystano całkowicie obiektywną metodę ewaluacji w postaci skomputeryzowanego algorytmu (reguły Sentera i analiza prawdopodobieństwa Raskina w programie OSS-3).

Poszczególne 4-osobowe grupy starano się dobrać tak, aby znajdowały się w nich osoby o podobnym poziomie wyszkolenia i doświadczenia zawodowego, choć z obiektywnych przyczyn (przede wszystkim ze względu na relatywnie niewielką populację poligraferów w Polsce) – ten postulat mógł być uwzględniony jedynie częściowo. Najbardziej doświadczeni dokonywali interpretacji metodą Utah, a najmniej doświadczeni – ESS, ale wcześniejsze eksperymenty na świecie potwierdziły, że system ESS daje w przypadku niedoświadczonych poligraferów podobne rezultaty jak u wykwalifikowanych ekspertów, bez istotnych statystycznie różnic[59].

---

57 W teście jednoproblemowym pytania relewantne (krytyczne) są względem siebie zależne znaczeniowo. Nie ma wówczas logicznej możliwości, by badany, kłamiąc w odpowiedzi na jedno pytanie, był równocześnie szczery przy drugim. Odwrotnie jest w teście wieloproblemowym (przesiewowym), gdzie pytania relewantne są od siebie niezależne znaczeniowo, dotyczą różnych zdarzeń i czynów, a zatem badany może jednocześnie na część pytań odpowiadać zgodnie z prawdą, a przy pozostałych próbować wprowadzić w błąd. Z podobną sytuacją mamy do czynienia w przypadku testu wieloaspektowego – gdzie pytania relewantne dotyczą wprawdzie jednego problemu, jednakże różnych jego aspektów (np. poza dokonaniem czynu także planowania, pomocnictwa czy wiedzy o nim – a więc okoliczności, które nie muszą współwystępować).

58 Procedura przeprowadzania tego testu została szczegółowo opisana m.in. w: M. Handler, R. Nelson, *Utah approach to Comparison Question Polygraph Testing*, „Polygraph" 2009, t. 38, nr 1, s. 15–30.

59 Zob. B. Blalock, B. Cushman, R. Nelson, *A replication and validation study on an empirically based manual scoring system*, „Polygraph" 2009, t. 38, nr 4, s. 281–288.

Wyniki i ich omówienie

Tabela 4 przedstawia rozpiętość trafnych identyfikacji między grupami ekspertów posługujących się różnymi metodami analizy danych testowych. Obliczono również współczynnik alfa Krippendorffa (tabela 5), bezpośrednio odnoszący się do zgodności między oceniającymi, z czym z kolei wiąże się zjawisko subiektywizmu. Co nie dziwi – niemal perfekcyjną zgodność uzyskano między wynikami ewaluacji z zastosowaniem algorytmów komputerowych w ramach programu OSS-3. Spośród oceniających manualnie – najbardziej zgodni byli ze sobą eksperci posługujący się systemem numerycznej interpretacji ESS (w skali 3-pozycyjnej). Większe rozbieżności odnotowano przy bardziej skomplikowanym systemie Utah (w skali 7-pozycyjnej) oraz ocenie globalnej, nazywanej szyderczo „metodą na oko".

Dla porównania – zgodność decyzji przy systemie ESS (zastosowanym dla testu jednoproblemowego – takiego jak w opisanym eksperymencie), która została zmierzona współczynnikiem Kappa Fleissa i przedstawiona w opracowaniu M. Handlera, R. Nelsona, Walta Goodsona i Matta Hicksa, wyniosła 0,84 (95% CI = 0,73, 0,95)[60]. Z kolei zgodność szczegółowych ocen numerycznych, przypisywanych każdemu mierzonemu parametrowi, wyniosła 0,56[61]–0,61[62] u niedoświadczonych poligraferów i 0,57[63]–0,61[64] u doświadczonych.

Tabela 4. Rozpiętość trafnych opinii ekspertów w poszczególnych grupach w eksperymencie realizowanym w Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego w latach 2014–2017

Metoda ewaluacji poligramów	**Rozpiętość trafnych identyfikacji (%)**	**Rozbieżność rezultatów między ekspertami (p.p.)**
ESS	73,5–86	12,5
Utah	79–96	17
globalna (jakościowa)	69–83	14
algorytmy OSS-3	75,7–76,3	0,6

Źródło: opracowanie własne Autorów.

60 M. Handler, R. Nelson, W. Goodson, M. Hicks, *op. cit.*, s. 205.

61 B. Blalock, B. Cushman, R. Nelson, *op. cit.*

62 R. Nelson, D.J. Krapohl, M. Handler, *op. cit.*

63 *Ibidem.*

64 R. Nelson, B. Blalock, M. Oelrich, B. Cushman, *Reliability of the Empirical Scoring System with expert examiners*, „Polygraph" 2011, t. 40, nr 3, s. 134.

Tabela 5. Współczynniki zgodności (rzetelności) dla poszczególnych grup poligraferów w eksperymencie realizowanym w Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego w latach 2014–2017

Metoda ewaluacji poligramów	**alfa Krippendorffa [CI]**
ESS	0,57 [0,47, 0,66]
Utah	0,43 [0,33, 0,54]
globalna (jakościowa)	0,43 [0,32, 0,53]
algorytmy OSS-3	0,82 [0,58, 1,00]

Źródło: opracowanie własne Autorów.

Rozbieżności w przypisywaniu przez ekspertów ocen numerycznych poszczególnym reakcjom organizmu (zapisom na poligramach) i przy podejmowaniu decyzji co do opinii końcowych mogły wynikać m.in. z takich czynników jak:

- niedookreśloność niektórych cech diagnostycznych[65] oraz reguł decyzyjnych – ich powiązanie ze zdolnościami percepcyjnymi eksperta (np. przy zasadzie „większe jest lepsze" wiele zależy od tego, w którym momencie oceniający dostrzeże różnicę i co uzna za „wyraźną" różnicę);
- błędne zastosowanie kryteriów diagnostycznych przewidzianych dla danego systemu analizy danych testowych;
- proste błędy matematyczne przy obliczeniach[66];
- kierowanie się przez eksperta innymi względami, niezwiązanymi z obiektywnymi przesłankami diagnozowania (błąd konfirmacji[67], efekt pierwszeństwa[68], efekt aureoli[69] itp.).

65 Chodzi tu np. o kontrowersje dotyczące złożoności reakcji elektrodermalnych – kiedy mówimy o reakcji złożonej, a kiedy już o dwóch odrębnych reakcjach?

66 To może dość zaskakujące, ale przy analizie rozbieżności w przypisywaniu szczegółowych ocen numerycznych, okazało się, że w niektórych przypadkach eksperci prawidłowo zakwalifikowali zapisy jako odzwierciedlające reakcje symptomatyczne i dostrzegli wyraźną różnicę między reakcjami na pytanie krytyczne i kontrolne, lecz pomyłka przy zupełnie nieskomplikowanych obliczeniach powodowała, że wynik testu był inny, niż gdyby ekspert prawidłowo podsumował swoje oceny cząstkowe.

67 Błąd konfirmacji – preferowanie tych informacji, które potwierdzają wcześniejsze oczekiwania – niezależnie od tego, czy są prawdziwe.

68 Efekt pierwszeństwa – większy wpływ informacji otrzymanych wcześniej (np. danych z akt, przypuszczeń śledczych) na tworzenie ogólnego wrażenia o osobie lub zdarzeniu, niż informacji, które przetworzyliśmy później.

69 Efekt aureoli – tendencja do automatycznego przypisywania pozytywnych cech osobowości na podstawie pierwszego wrażenia.

Przykładowy fragment poligramu z zapisami niespójnie ocenianymi przez ekspertów przedstawia rysunek 1. Poligraferzy nie mieli w tym wypadku problemów z komponentem EDA (pierwsza krzywa od dołu). W przypadku kardio część uznała, że bardziej znacząca zmiana wystąpiła przy pytaniu krytycznym (R3) i przyznała ocenę -1, zaś pozostali ocenili widniejącą parę pytań neutralnie – tj. na 0[70]. Natomiast zmiany cyklu oddechowego w strefie pytania R3 – w porównaniu do pytania kontrolnego C2 – były czasem interpretowane jako równorzędne, w innych przypadkach jako mniej znaczące, a w pozostałych – nawet bardziej, co skutkowało pełną gamą ocen numerycznych parametru pneumo wśród poligraferów[71]. Nie rozstrzygając, którzy z oceniających mieli rację, a którzy nie – należy zauważyć, że tego typu sporne sytuacje od czasu do czasu występują i wówczas niezwykle ważne może okazać się doświadczenie eksperta, uwzględnienie specyficznych cech osobniczych badanego czy tendencji w jego reakcjach na przestrzeni całego wykresu.

W tabeli 6 zestawiono odsetki trafnych identyfikacji oraz wyników nierozstrzygniętych w opiniach wydanych w oparciu o określoną metodę analizy zapisów na poligramach. Najlepiej poradzili sobie „ślepo" oceniający metodą numeryczną Utah (ok. 85%), a następnie metodą ESS (79%). Zbliżony poziom dokładności osiągnęli „ślepo" oceniający eksperci posługujący się metodą globalną i algorytmami komputerowymi oraz ci, którzy osobiście przeprowadzili badania (ok. 76–77%). Wyniki okazały się nieco niższe od tych przedstawionych dla testu Utah ZCT w raporcie APA z 2011 r., podsumowującym metaanalizę technik badań poligraficznych[72]. Należy przyjąć, że była to wypadkowa koncepcji eksperymentu, sposobu jego realizacji, kwalifikacji, doświadczenia i jakości pracy zaangażowanych poligraferów.

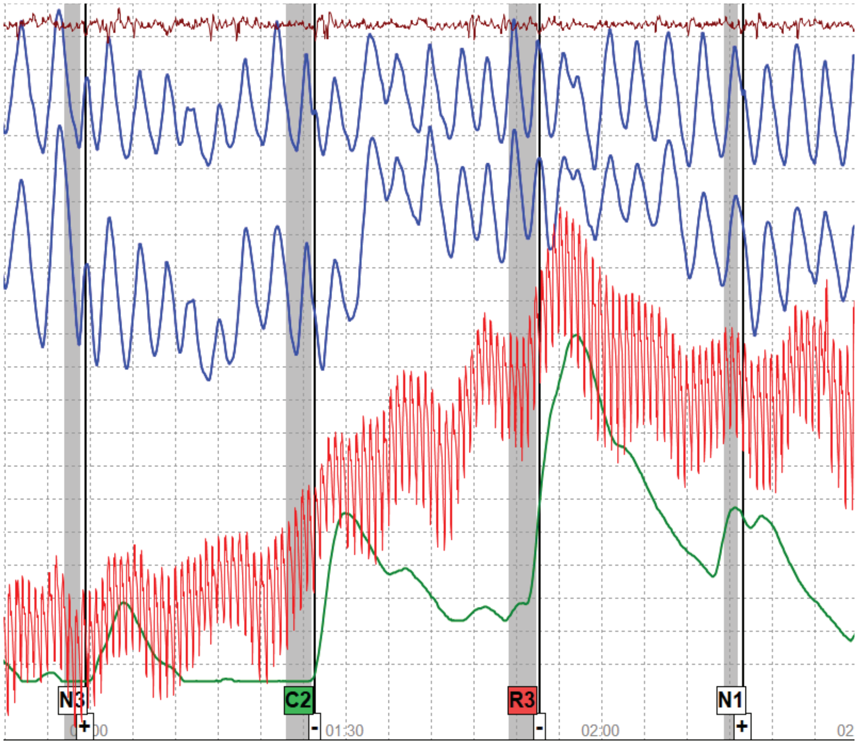
---

70 Poligraferzy decydujący się na ocenę -1 dla parametru kardio argumentowali, że istotna zmiana w ciśnieniu krwi, którą można zaobserwować w okolicy pytania C2, zaczyna się jeszcze przed wprowadzeniem tego bodźca i dalej nie widać nowej reakcji fazowej. Natomiast zmiana przy pytaniu R2 jest bardzo wyraźna i zaczyna się w odpowiednim czasie po zadaniu pytania testowego. Co ciekawe – podobnych obserwacji zapisów dokonali także eksperci, którzy przyznawali ocenę 0. Powoływali się jednak na to, że skoro znacząca zmiana reakcji fizjologicznych nastąpiła jeszcze przed pytaniem C2, to mogło nie być dostatecznych warunków do tego, by organizm istotnie zareagował po wprowadzeniu bodźca, i dlatego pytanie C2 nie powinno być wykorzystane jako miarodajne pytanie porównawcze.

71 Przyznający ocenę +1 dostrzegali zmianę linii bazowej – zwłaszcza dla brzusznego (dolnego) pneumo w C2. Oceniający na 0 uznawali, że nie zaistniały jakieś konkretne cechy diagnostyczne, albo wystąpiły przy obu porównywanych pytaniach. Argumentacja dla oceny -1 opierała się na stwierdzeniu krótszej linii oddychania w strefie pytania R3.

72 American Polygraph Association, *Meta-analytic survey of criterion accuracy of validated techniques*, „Polygraph" 2011, t. 40, nr 4, s. 195–305.

Rysunek 1. Fragment poligramu przedstawiający zapisy reakcji dla pary pytań: R3 i C2. Zapisy w kanałach pneumo i kardio były często odmiennie oceniane przez ekspertów



Źródło: opracowanie własne Autorów.

Odsetek wyników nierozstrzygniętych u sędziów kompetentnych był w eksperymencie wyższy od średniej w przypadku wykorzystania metody ślepej interpretacji zapisów (14,7–19,9%), zaś niższy u badających i w analizie za pomocą algorytmów komputerowych (odpowiednio: 3,8% i 5,1%). Stało się tak być może dlatego, że analiza komputerowa opierała się na precyzyjnych danych liczbowych, natomiast badający – w odróżnieniu do „ślepo" oceniających – usilniej dążyli do rozstrzygnięcia i brali pod uwagę dodatkowe, subiektywne czynniki przy opiniowaniu. Są to jednak jedynie prawdopodobne hipotezy, które wymagałby udowodnienia.

Mając na uwadze wyniki powyższych analiz statystycznych – empirycznie potwierdzono, że metoda numeryczna ESS pozostawia istotnie mniejszy margines subiektywizmu niż metoda globalna (jakościowa). Ponadto, metody o mniejszej skali (w tym ESS) są bardziej obiektywne (z mniejszą przestrzenią na subiektywną ocenę eksperta) niż te o skali 7-stopniowej (w tym Utah).

Tabela 6. Trafność identyfikacji dokonanych przez poszczególne grupy poligraferów w eksperymencie realizowanym w Krakowskiej Akademii im. Andrzeja Frycza Modrzewskiego w latach 2014–2017

Metoda ewaluacji poligramów		**n [sprawca/ świadek]**	**Nierozstrzygnięte (%)**	**Trafność* (%)**	**PPV** [CI]**	**NPV*** [CI]**
Badający	ESS	39 [16/23]	5,1%	75,7%	0,67 [0,34, 0,90]	0,76 [0,64, 0,85]
„Ślepa interpretacja”	ESS		19,9%	79%	0,78 [0,47, 0,94]	0,89 [0,73, 0,97]
	Utah		19,2%	84,5%	0,82 [0,54, 0,97]	0,82 [0,68, 0,89]
	globalna (jakościowa)		14,7%	76,7%	0,62 [0,37, 0,80]	0,81 [0,70, 0,92]
	algorytmy OSS-3		3,8%	76%	0,75 [0,39, 0,95]	0,78 [0,67, 0,83]

* odsetek trafnych identyfikacji z wyłączeniem wyników nierozstrzygających; ** prawdopodobieństwo, że osoba, która uzyskała pozytywny wynik testu, rzeczywiście posiada diagnozowaną cechę; *** prawdopodobieństwo, że osoba, która uzyskała negatywny wynik testu, rzeczywiście nie posiada diagnozowanej cechy.

Źródło: opracowanie własne Autorów.

Natomiast jeżeli chodzi o porównanie odsetka prawidłowych identyfikacji między ekspertami przeprowadzającymi badania a osobami dokonującymi ślepej interpretacji metodami numerycznymi, okazało się, że w ramach eksperymentu to ci ostatni byli dokładniejsi (84,5% dla ślepo oceniających systemem Utah i 79% dla ślepo oceniających systemem ESS wobec 75,7% w przypadku badających). Wydaje się, że mogło to wynikać z dwóch okoliczności:

a) badający oceniali „na poczekaniu”, a „ślepo” interpretujący pracowali bez presji czasu w warunkach biurowych i domowych;
b) nie można też wykluczyć, że badający – mniej lub bardziej świadomie – sugerowali się innymi (poza poligramami) czynnikami związanymi z osobistym odbiorem badanych, lecz wbrew dotychczasowym badaniom amerykańskim[73] – przyczyniło się to do obniżenia trafności identyfikacji.

---

[73] Por. np. W.D. Holmes, *op. cit.*

Należy jednak zaznaczyć, że różnice w trafności analiz nie były na tyle statystycznie istotne, aby można było mówić o wyższości jednej metody nad innymi.

Generalnie w eksperymencie lepiej identyfikowano „niewinnych" niż „winnych". Stało się tak prawdopodobnie m.in. dlatego, że pytania kontrolne odnosiły się do realnego życia, a krytyczne – jedynie do zainscenizowanego zdarzenia, ale w warunkach laboratoryjnych wydaje się to nieuniknione. Wykorzystana motywacja finansowa dla „winnego" badanego, za jego nietrafną identyfikację, mogła okazać się niewystarczająca, by zrównoważyć czy przeważyć ciężar gatunkowy pytań związanych z realnymi przeżyciami, a wykorzystanych w testach jako kontrolne. Problem jest znany i opisywany w literaturze. Powszechnie przyjmuje się, że trafność wskazań w badaniach eksperymentalnych jest niższa niż w badaniach podejmowanych w realnych sprawach, właśnie ze względu na niższą motywacje badanych[74]. Zastosowanie wyższej motywacji badanych w ramach eksperymentu mogłoby się jednak okazać zbyt ryzykowne i niewspółmierne do efektów. Z kolei w warunkach dobrowolnego udziału w eksperymencie nie było możliwe zastosowanie sankcji negatywnej dla „zdemaskowanego kłamcy". Z tego względu, że w badaniach eksperymentalnych pytania krytyczne dotyczą sytuacji wyreżyserowanej (umowne kłamstwo), fikcyjnej, a pytania kontrolne – prawdziwych zdarzeń z życia, tak jak w badaniach w realnych sprawach, emocjonalna waga tych ostatnich jest dla badanego z założenia większa. Dlatego też niektórzy autorzy sugerują, by w badaniach eksperymentalnych nie stosować testów pytań kontrolnych, a jedynie testy szczytowego napięcia (POT – ang. *peak of tension*) lub testy klasyczne[75]. Jednak wielu autorów w badaniach eksperymentalnych z powodzeniem stosowało techniki pytań kontrolnych. Na przykład Gordon Barland w swych eksperymentach wykorzystywał technikę Backstera[76].

---

[74] Por. np. S. Abrams, *Polygraph validity and reliability. A review*, „Journal of Forensic Sciences" 1973, t. 18, nr 4, s. 318; por. także, J. Widacki, *Wprowadzenie do problematyki badań…*, s. 128–130.

[75] Por. F.K. Berrien, *A note on laboratory studies of deception*, „Journal of Experimental Psychology" 1939, t. 24, nr 5, s. 542–546; P.V. Trovillo, *Scientific proof of credibility*, „Tennessee Law Review" 1953, t. 22, s. 743–766.

[76] G.H. Barland, *An experimental study of field techniques in lie-detection*, 1972 (tekst niepublikowany, University of Utah); idem, *Detection of deception in criminal suspects. A field validation study*, 1975 (tekst niepublikowany, University of Utah).

Wnioski

Badania, jak już napisano wyżej, wykazały, że numeryczne metody oceny zapisów poligraficznych są dokładniejsze i zmniejszają zakres subiektywizmu ekspertów. Wśród sposobów na ograniczenie subiektywizmu ekspertów z zakresu badań poligraficznych można wymienić standaryzację możliwie największej części, o ile nie całej procedury przeprowadzania badań. W tym kontekście komplementarne znaczenie ma komputeryzacja (automatyzacja instrukcji przekazywanych badanemu oraz zastosowanie algorytmów do analizy danych testowych).

Perspektywy są obiecujące, ale zalecane jest zachowanie ostrożności. Trzeba pamiętać, ze badanie poligraficzne składa się z kilku integralnych części. Oprócz rejestracji reakcji i oceny uzyskanych zapisów, do zadań eksperta należy też warunkujące powodzenie odpowiednie formułowanie pytań testowych, dobór właściwych testów, sposobu przeprowadzenia wywiadu przedtestowego (może z wyjątkiem standardowych elementów – takich jak zwięzły opis tego, co mierzy poligraf, wyjaśnienie najważniejszych kwestii związanych z psychologią i fizjologią człowieka) i rozmowy po testach. Tak więc samo perfekcyjne zbudowanie algorytmów ewaluacji zapisów poligraficznych, choć szalenie ważne i niewątpliwie ograniczające margines subiektywizmu, nie rozwiązuje problemu do końca. Badanie poligraficzne nie jest i zapewne nigdy nie będzie prostym pomiarem. Zawsze istotna będzie w nim rola eksperta – jego umiejętności, doświadczenie i predyspozycje.

Abstract
Expert's subjectivity in polygraph examination

The element of subjectivity in polygraph examination shall mean a margin of discretion, appearing in evaluation of expert's interpretation, which is not subordinated to any other objective criteria. The existence of this margin of discretion confirms a divergence of evaluations carried out by different experts assessing the same reactions. In assessing the same records of reactions, experts differed in their evaluation but the divergence of correct evaluations of the same reactions was ranging from 3% to 31%. The convergence of expert's evaluations can be measured according to the reliability coefficient (e.g. Spearman-Brown), the correlation coefficient (e.g. Kappa, Fleiss' Kappa). The convergence of different expert's evaluations which is shown in literature (calculated according to Kappa coefficient) is relatively high (0.77).
In the experiment which was conducted by authors, experts was evaluating records of reactions using a qualitative method (global) and a quantitative method (numerical). The consistency coefficient (reliability) calculated by the alfa-Krippendorff coefficient

was significantly higher applying the quantitative methods (numerical) – 0.43 to 0.82, whereas in the qualitative evaluation it was only 0.43.
In view of the above, authors' hypothesis is confirmed by the fact that there is the smaller scale of subjectivity in evaluations of numerical polygraphic records than in the qualitative method (global).
Key words: subjectivity of polygraph examination, subjectivity of experts opinion, qualitative and quantitative methods of polygrams interpretation

Streszczenie
Subiektywizm w badaniach poligraficznych

Przez element subiektywizmu w badaniach poligraficznych rozumiemy występujący przy ocenie dokonywanej przez eksperta margines swobody interpretacji, niepodlegający żadnym obiektywnym kryteriom.
O istnieniu tego marginesu przekonuje rozbieżność ocen dokonywanych przez różnych ekspertów. Oceniając te same zapisy reakcji, eksperci różnili się w ich ocenie, przy czym rozbieżność poprawnych ocen tych samych reakcji wynosiła od 3 do 31%.
Zgodność ocen ekspertów może być mierzona współczynnikiem rzetelności (np. Spearmana-Browna), współczynnikiem korelacji (np. Kappa, Kappa Fleissa).
Wykazywana w literaturze zgodność ocen różnych ekspertów (liczona współczynnikiem Kappa) jest relatywnie wysoka (0,77).
W wykonanym przez autorów eksperymencie, eksperci oceniali zapisy reakcji metodą jakościową („globalną") oraz metodami ilościowymi (numerycznymi). Współczynnik zgodności (rzetelności) liczony współczynnikiem alfa-Krippendorffa był zdecydowanie wyższy przy zastosowaniu metod ilościowych (numerycznych) – 0,43 do 0,82 niż przy ocenie jakościowej – jedynie 0,43.
Potwierdza to hipotezę autorów, że numeryczne oceny zapisów poligraficznych są obciążone mniejszym marginesem subiektywizmu niż metoda jakościowa („globalna").
Słowa kluczowe: subiektywizm w badaniach poligraficznych, subiektywizm opinii biegłego, ilościowe i jakościowe metody interpretacji poligramów