PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 27–47
DOI: 10.31648/przegldpsychologiczny.9456

Czy kryzys wiarygodności w psychologii?

Jerzy Marian Brzeziński[1]
Uniwersytet im. Adama Mickiewicza w Poznaniu
Wydział Psychologii i Kognitywistyki
https://orcid.org/0000-0003-1582-4013

Streszczenie

Wraz z opublikowaniem w prestiżowym *Science* głośnego w środowisku nie tylko psychologów, artykułu prezentującego wyniki – zakrojonych na dużą, międzynarodową skalę (w badaniach brało udział 125 badaczy) – replikacji badań empirycznych z obszaru psychologii (zob. Open Science Collaboration, 2015) znacząco wzrosło zainteresowanie globalnym wynikiem uzyskanym przez zespół B. Noseka. Okazało się bowiem, że o ile w 97% badań oryginalnych uzyskano wyniki istotne statystycznie ($p < 0{,}05$), o tyle w badaniach replikacyjnych było to tylko 36%. Ten wynik stał się, jak uważa autor niniejszego artykułu, podstawą nieuprawnionych uogólnień dotyczących słabości metodologicznej psychologii jako nauki empirycznej. Psychologia jest nauką empiryczną, ale ma też swoje osobliwości związane z jej przedmiotem i metodą (Orne, 1962/1991, 1973/1993; Rosenzweig, 1933; Rosenthal, 1966/2009). Nie jest też uprawiana w izolacji społecznej, kulturowej. Wreszcie podlega także ostrym nakazom/zakazom etycznym. Zaś psychologowie publikujący wyniki badań empirycznych, poddawanych analizom statystycznym, są ograniczani przez zwyczaje panujące w redakcjach czasopism naukowych (zresztą, nie tylko psychologicznych). Te są one zainteresowane drukowaniem wyłącznie prac przestawiających wyniki istotne statystycznie (co oznacza: „$p < 0{,}05$"!), a co prowadzi do powstawania tzw. efektu szuflady (*file drawer effect*, Rosenthal, 1979). Nie można też, co autor mocno podkreśla, ograniczać dyskusji tylko do spełniania, przez badania prowadzone przez psychologów, do zagadnień statystycznych (zwłaszcza problemu *mocy statystycznej testu istotności* – co stało się ostatnimi laty modne w badaniach psychologicznych. W tym artykule autor omawia, prezentując też własny punkt widzenia, następujące problemy: 1) osobliwości metodologiczne psychologii jako nauki empirycznej; 2) triada: istotność statystyczna (problematyczne kryterium „$p < 0{,}05$"), *effect size*, *power of statistical test*; 3) społeczno-kulturowy kontekst badań psychologicznych; 4) naruszanie reguł metodologicznych i etycznych przez badaczy-psychologów; 5) podejmowanie środków zaradczych i naprawczych.

[1] Adres do korespondencji: brzezuam@amu.edu.pl.

Słowa kluczowe: nauka, intersubiektywność, stabilność, racjonalność, replikacje, badanie psychologiczne, statystyka, test statystyczny, przedział ufności, „$p < 0{,}05$", moc testu statystycznego, wielkość efektu, *data fishing*, *p-hacking*, HARKing, oczekiwania interpersonalne, wskazówki sugerujące osobie badanej treść hipotezy badawczej (*demand characteristic*), efekt szuflady, wstępna rejestracja badań (*pre-registration research*)

Psychologia, zapewne jak i inne dyscypliny naukowe, też ulega „modowym" naciskom; bywa, że w dłuższym czy krótszym czasie obserwuje się wzmożone zainteresowanie jakąś problematyką **teoretyczną**. Ważne, gdy można o danym rozwiązaniu teoretycznym powiedzieć – z dużą dozą prawdopodobieństwa – że weszło do kanonu dokonań teoretycznych psychologii. Moim zdaniem np. tak właśnie można powiedzieć o dokonaniach na polu teorii inteligencji o, chyba powszechnie akceptowanej, teorii inteligencji: Cattella-Horna-Carrolla (znanej pod akronimem CHC[2]). Może warto zwrócić uwagę na fakt, że teoria CHC stała się teoretyczną podstawą dla najnowszych wydań *Skal inteligencji* sygnowanych nazwiskiem Davida Wechslera: dla dorosłych WAIS-IV (wydanie V w trakcie końcowych prac normalizacyjnych prowadzonych w USA – opóźnienie tych badań spowodowane było pandemią COVID) i dla dzieci WISC®-V[3]. Bywało jednak i tak, że (po latach) okazywało się, iż jakaś obiecująca „gwiazda" rozbłysła fałszywym światłem; tak było/jest np. z psychoanalizą czy – na znacznie mniejszą skalę – pseudonaukową koncepcją „ustawień rodzinnych" według Berta Hellingera.

Można też zauważyć w historii naszej dyscypliny występowanie mód **warsztatowych** (niektóre zostały przez psychologów zapożyczone z innych nauk empirycznych), dotyczących **metod zbierania danych** czy **metod analizy danych** (zwłaszcza statystycznych – ich dynamiczny rozkwit był wywołany postępem technologicznym w świecie komputerów i wręcz niesamowitym rozwojem bardzo wyrafinowanego oprogramowania statystycznego; pakiety typu SPSS czy SAS to już „zabawki"). Pojawiają się monografie, artykuły naukowe, opracowania techniczne[4] czy wystąpienia konferencyjne i warsztatowe skupiające się na jakimś temacie. Tak było – a nawet jeszcze jest – np. z metodami samoopisowymi (kwestionariusze osobowości) czy pseudonaukowymi metodami projekcyjnymi (np. *Test Rorschacha* czy *Test drzewa* Kocha.

Próbując „ogarnąć" dokonania jakiejś dyscypliny naukowej, a zwłaszcza tak młodej jak psychologia (mającej ok. 150 lat – co to jest wobec „długiego trwania"[5]

---

[2] Zob. dobre omówienie teorii CHC: Schneider i McGrew (2012).

[3] Ta ostatnia, piąta wersja WISC (*Wechsler Intelligence Scale for Children*® – 5th ed.) z 2014 r., jest od 2020 r. dostępna w polskiej adaptacji przeprowadzonej przez zespół Pracowni Testów Psychologicznych PTP (Joanna Stańczak, Anna Matczak, Aleksandra Jaworowska i Iwona Bac) – zob. https://www.practest.com.pl/wisc%C2%AE-v-skala-inteligencji-wechslera-dla-dzieci-%E2%80%93-wydanie-piate.

[4] Zob. rozwój języka „R" – też wykorzystanego do projektowania programów statystycznych, np. Schwarzer (2022).

[5] W rozumieniu historyka Fernanda Braudela (1902–1985).

fizyki, matematyki, biologii), trzeba nie tylko doszukiwać się tego, co łączy psychologię z innymi naukami empirycznymi, lecz także dostrzegać to, co stanowi o jej swoistości czy osobliwości. Tym, co łączy, jest, jak mniemam, struktura procesu badawczego. Przykładowo, psycholog, socjolog, biolog, psychiatra sięgają po te same narzędzia statystyczne, sprawdzając hipotezy. Zaś specyfika przedmiotu badań narzuca badaczowi ograniczenia metodologiczne (odnoszące się do metody) i etyczne (ograniczające psychologa w jego postępowaniu badawczym wobec osób badanych). To właśnie owa specyfika jest przede wszystkim „odpowiedzialna" za ujawniające się w badaniach prowadzonych przez psychologów niedoskonałości – np. zbyt niski poziom odtwarzalności wyników w przeprowadzanych replikacjach.

Problem podstawowy: jaką nauką jest (powinna być) psychologia?

Na intensywny rozwój **warsztatu eksperymentalnego** psychologów przed kilkudziesięciu laty wpływ miała, **zapożyczona od przyrodników**, a wymyślona przez wybitnego statystyka, badawczo działającego w obszarze doświadczalnictwa rolniczego Ronalda A. Fishera (1925/1938, 1935/1971)[6] **analiza wariancji** (ANOVA)[7], będącą (obok MANOVA) statystycznym modelem współczesnego eksperymentowania w psychologii (od lat 50. ubiegłego wieku) – co umożliwiło psychologom wyjście poza porównania tylko dwóch grup (eksperymentalnej i kontrolnej) i związane z nim dwie nowe możliwości badawcze: testowanie zależności krzywoliniowych i testowanie interakcji zachodzących między dwiema i większą liczbą zmiennych niezależnych. To oznaczało – może dziś zapomniany/niedoceniany – znaczący postęp w testowaniu nowych hipotez badawczych. Przez dziesiątki lat psychologia importowała nowości statystyczne (także pod tym względem zbliżając się do nauk przyrodniczych).

W ostatnich latach w polskiej psychologii pojawiły się publikacje poświęcone problematyce **mocy testu** (*statistical power*) w odniesieniu, zwłaszcza, do zastosowań testów do oceny istności różnic, np. *t*-Studenta czy testów stosowanych w obrębie modeli ANOVA i MANOVA. Problem nie jest nowy[8], ale został „dzięki"

[6] Podręcznik Fishera: *Statistical method for research workers* doczekał się aż 14 wydań, w tym 2 pośmiertne. Także popularny był drugi napisany przez niego podręcznik: *The design of experiments* (Fisher, 1935/1971 – 9 wydań, ostatnie ukazało się w 1978 r. po śmierci autora w 1960 r.). Fisher żył w latach 1890–1962. Po II wojnie światowej, moim zdaniem, przede wszystkim trzy podręczniki (znacząco uzupełniane w kolejnych wydaniach) napisane z myślą o psychologach kształtowały praktykę badawczą psychologów planujących eksperymenty według wymagań statystycznego modelu ANOVA: Edwards (1950/1960/1968/1972); Winer (1962/1971; ostatnie: Winer, Brown i Michels, 1991); Kirk (1968/1982/1995; ostatnie: Kirk, 2012).

[7] W polskiej psychologii: Brzeziński i Stachowski (1981/1984), Brzeziński (2000/2012).

[8] Wspomnę tylko, tytułem przykładu, dwie antologie starszych prac powstałych z udziałem psychologów (w Polsce przeszły bez echa): Henkel i Morrison (1970); Harlow,

pracom międzynarodowych zespołów nad wstydliwym dla środowiska psychologów zagadnieniem **niepełnej powtarzalności** (*reproducibility)* wyników psychologicznych badań empirycznych wydobyty na światło dzienne.

Chciałbym od razu zaznaczyć – a do tego wrócę w dalszej części artykułu – że **za niezadawalający poziom powtarzalności wyników badań empirycznych nie odpowiadają jedynie niedostatecznie finezyjnie stosowane przez psychologów metody statystyczne**.

Powstało już sporo prac, krytycznych wobec metodologicznej niedojrzałości psychologii – bo i tak mocne, jednak, jak sądzę, formułowane z nadmiernym krytycyzmem, zarzuty można wyczytać w tych pracach. Bodajże najszerszym echem odbiła się, opublikowana w jednym z dwóch najbardziej prestiżowych światowych czasopism naukowych: *Science* (drugie, a może pierwsze, to *Nature*) praca 125 autorów (zob. Open Science Collaboration, 2015)[9].

Warto, z konieczności skrótowo, przybliżyć pracę tego zespołu. Przeprowadzono **replikacje** badań empirycznych zaprezentowanych w 100 artykułach (na 488) opublikowanych w 2008 roku, w trzech prestiżowych psychologicznych czasopismach naukowych: *Psychological Science* (PSCI)*, Journal of Personality and Social Psychology* (JPSP)*, Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP: LMC). Aż 32 artykuły (z 55) pochodziły z JPSP, 28 artykułów (na 39) wybrano z JEP: LMC, a 39 (na 64) z PSCI. Jedynie 2 artykuły obejmowały po 2 replikacje. Jeśli chodzi o zakres tematyczny badań, to obejmował on 43 badania o profilu poznawczym (*cognitive*) i 57 o profilu społeczno-osobowościowym (*social-personality*). Wnioski, w wielkim skrócie (bardzo szczegółowe dane źródłowe dostępne są pod wskazanym w bibliografii linkiem), są następujące: o ile w 97% badań oryginalnych uzyskano wyniki istotne statystycznie ($p < 0{,}05$), o tyle w badaniach replikacyjnych ten procent był znacząco niższy i obejmował tylko 36%. Analiza wartości wskaźników *wielkości efektu* (*effect size*) pokazała, że tylko 47% wartości wskaźników uzyskanych w oryginalnych badaniach mieściło się w granicach 95% przedziału ufności dla wartości tych wskaźników z powtórzonych badań. Warto zauważyć, iż **zespół autorów ograniczył swoje analizy do badań przedstawionych tylko w trzech czasopismach, z jednego rocznika 2008, sprofilowanych na tematykę poznawczą (43 badania) i społeczno-osobowościową (57 badań)**. Psychologia jednak nie ogranicza się wyłącznie do tych obszarów badawczych. Niemniej dane te stały się podstawą do poważnych, teoretycznych i metodologicznych dyskusji, ale też i komentarzy o charakterze hejterskim kierowanych pod adresem **całej psychologii**.

Mulaik i Steiger (1997). Chciałbym też – przede wszystkim – przywołać dwie klasyczne (pojawiające się w różnych antologiach tekstów o profilu metodologicznym) prace Jacoba Cohena (1990/2006, 1994/2006), które, wspólnie z moim przyjacielem, prof. Jerzym Siutą (1943–2018) z Uniwersytetu Jagiellońskiego, postanowiliśmy udostępnić przed laty w tłumaczeniu na język polski: Brzeziński i Siuta (1991); zwracam uwagę na bardzo wówczas ograniczony, zwłaszcza dla studentów, dostęp do wersji oryginalnych. Moim zdaniem formułowane w nich oceny i wnioski jeszcze dziś **nic nie straciły na aktualności**; to powinna być **obowiązkowa lektura** dla magistrantów i doktorantów psychologii.

[9] Zob. 7773 cytowań wg Google Scholar na 14 listopada 2022 r.

„Jądrem" dyskusji nad wieszczonym przez krytyków (uwiedzionych tym artykułem) upadkiem psychologii było skupienie się na wskaźnikach statystycznych: **istotność statystyczna**: $p < 0{,}05$ *versus* $p > 0{,}05$, oraz niezadowalających wartościach wskaźników **wielkości efektu** (uzyskanych w replikowanych badaniach).

Moim zdaniem ten projekt to dobry punkt wyjścia do dalszych analiz replikacyjnych, które już zostały podjęte. Warto też wyjść w stronę innych, poznawczo i aplikacyjnie istotnych, o odmiennych podejściach metodologicznych badań empirycznych. Tak czy inaczej, replikacje (obok udostępniania danych surowych, publikowania prac, w których statystyki testowe, np. t czy F, nie uzyskały wartości spełniających „magiczne" kryterium $p < 0{,}05$ oraz metaanaliz) są poważnym środkiem „terapeutycznym" na wywołany chaos i naruszenia etyczne spowodowane przemożną i nieposkromioną chęcią drukowania czegokolwiek gdziekolwiek (zob. *casus* tzw. drapieżnych czasopism, *predatory journals*)[10] i niszczącym poddaniem się nakazowi „publikuj albo giń" (*publish or perish*). Miernoty naukowe też podlegają temu nakazowi – zwłaszcza aktualnie w Polsce (szał publikacyjny: „punktoza" i „slotoza"; co jeszcze naukowi biurokraci wymyślą?).

Czy nie należy jednak spojrzeć na problem ze znacznie szerszej – aniżeli podyktowana stosowanymi (wbrew zaleceniom zespołów ekspertów APA[11]) przez redakcje czasopism naukowych praktykami edytorskimi – perspektywy? Bynajmniej ich nie lekceważę, ale też nie traktuję wskaźnika istności statystycznej, identyfikowanej z wartością $p < 0{,}05$, jako nienaruszalnej „świętości" – zob. Skipper i ł Guenther (1967/1970). Też zdaję sobie sprawę z tych ograniczeń – zwłaszcza gdy są przykładane, bywa, że bezrefleksyjnie, do **słabych pomiarowo danych**. Zauważmy, że psychologia tylko w części swoich działów może przywoływać „twarde" wyniki pomiarowe. Nadal, niestety, są to – i to w nadmiarze – dane samoopisowe (różne kwestionariusze osobowości i skale szacunkowe). Przykładanie do takich wyników mocnych (według założeń modeli statystycznych) narzędzi statystycznych tylko stwarza **pozory precyzji i naukowości** – niestety. I być może nie należy zbytnio kwestionować przywołanych wyżej wyników zespołu Open Science Collaboration?

Oczywiście **psychologia – jak zresztą każda dyscyplina naukowa – nie jest uprawiana w izolacji społecznej (w swoistej wieży z kości słoniowej)**. Praktyka badawcza podlega nie tylko wewnętrznym, swoistym dla psychologii, zaplanowanym przez badacza i poddanym jego efektywnej kontroli **oddziaływaniom wewnętrznym**. Podlega też wpływom pochodzącym z szeroko pojmowanego **kontekstu zewnętrznego**. Powinniśmy zdawać sobie z tego sprawę.

---

[10] Podejrzane tytuły można sprawdzić na zestawieniu owych „drapieżnych" tytułów: *Cabell's Journalytics and Predatory Reports*; lista ma charakter komercyjny (trzeba wykupić dostęp) i jest dostępna pod adresem (tam też szczegółowe informacje): https://www2.cabells.com/predatory.

[11] Zob. Wilkinson i Task Force on Statistical Inference, American Psychological Association Science Directorate (1999); APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008); American Psychological Association (2020).

Jeżeli – z niewiedzy czy niedouczenia, albo, co gorsza, świadomego ignorowania owego kontekstu – tych oddziaływań nie uwzględnimy, to trudno spodziewać się satysfakcjonującej powtarzalności wyników naszych badań empirycznych. Chyba że dotyczą one trywialnych pytań o z góry przewidywalnych odpowiedziach. Ale wtedy po co angażować czas i pieniądze (często są to pieniądze podatnika) w takie – udające badanie naukowe – przedsięwzięcie?

Niestety, gdy przyjrzymy się zawartości polskich czasopism psychologicznych i tzw. zeszytów naukowych czy prac zbiorowych, to coraz więcej w nich artykułów (od strony statystycznej nawet „szkolnie" poprawnych – chociaż nieuwzględniających wszystkich standardowych procedur), ale merytorycznie miałkich. Dominują też artykuły prezentujące bądź nowe kwestionariusze osobowości, bądź ograniczone do omówienia wyników pozyskanych za pomocą metod samoopisowych. Wystarczy sięgnąć po kilka kwestionariuszy do odpowiedniej szuflady… To łatwe – od strony warsztatowej – ale to wcale nie oznacza, że inspirujące poznawczo. Ot, trzeba „coś" napisać, to piszę. Jakieś „drapieżne" czasopismo to, za stosowną opłatą, wydrukuje (oczywiście w języku angielskim!). Muszę (chcę) założyć, że np. Redakcja *Przeglądu Psychologicznego* takiego artykułu „naukowego" nie wydrukuje.

Samo cyzelowanie analiz statystycznych nie wystarczy, a i „wkładanie" osób badanych do – jeszcze nie tak dawno niedostępnego badaczom (w tym także psychologom) – tomografu czy skanowanie mózgu też nie wystarczy. Niestety, **psychologia (ta przez duże „P") jest trudną badawczo dyscypliną naukową**. O jej „trudności" nie stanowi zaś (chociaż go nie lekceważę) stopień komplikacji technologicznej, z którym mamy do czynienia np. w technicznie zaawansowanych badaniach fizyków atomowych prowadzonych w CERN w Genewie.

Jest też trudna dlatego, że to jedna osoba (badacz) poddaje badaniom drugą osobę (uczestniczącą w tych badaniach). Tak jest często w psychologii klinicznej. To bywa czasem naprawdę bardzo trudne (zwłaszcza gdy w badaniach biorą udział małe dzieci czy osoby z różnymi niepełnosprawnościami – nie tylko intelektualnymi).

Przed wielu laty psycholog amerykański Saul Rosenzweig (1907–2004) pisał o **trzech osobliwościach** (*peculiarities*) badań eksperymentalnych w psychologii (Rosenzweig, 1933; zob. Larsen, 2005)[12]: (1) badacz staje się elementem sytuacji badawczej, (2) wpływ na zachowanie się osoby badanej w sytuacji badawczej mają takie zmienne związane z osobą badaną ją charakteryzujące, jak: osobowość, motywacja itp., (3) zawiązuje się interakcja: badacz – osoba badana. Nawiasem mówiąc, ten, jak mniemam, bardzo ważny artykuł poprzedził prace następujących psychologów: Orne (1962/1991, 1973/1993) czy Rosenthal (1966/2009; też: Blanck, 1993; Trusz, 2013[13]).

[12] Cytowany Larsen zwrócił uwagę: „[…] in his first published article ("The experimental situation as a psychological problem" 1933), in which he explored the reciprocal interaction between experimenter and subject and laid the foundation for the work on experimenter expectancy effects that flourished a generation later." (s. 259)

[13] Zwracam szczególną uwagę psychologów na obszerny (liczący 817 stron) i reprezentatywny wybór tłumaczeń prac – tych klasycznych i tych współczesnych – poświęconych *efektowi Rosenthala* (Trusz, 2013).

Prace wyżej wymienionych psychologów zwróciły uwagę na to, że osoby badane są w stanie rozpoznać cel badania i, zgodnie z tym, modyfikować swoje zachowanie w eksperymencie. Orne mówił o **wskazówkach** (*demand characteristics*) **sugerujących osobie badanej treść hipotezy badawczej**. Jest ona w stanie rozpoznać, jakie jej zachowania są oczekiwane przez badacza i – w zależności od tego, jak postrzega badacza (przyjaznego czy zagrażającego) – będzie starała się zachować zgodnie z rozpoznaną treścią hipotezy albo niezgodnie z nią. Z kolei – jak pokazały badania (poddane wielu replikacjom) – przeprowadzone przez psychologa społecznego i metodologa Roberta Rosenthala (badacz, ale też: psychoterapeuta, nauczyciel, sędzia, trener sportowy), generuje **oczekiwania interpersonalne** (*interpersonal expectations*) wobec zachowania się osoby, uczestniczącej w badaniu naukowym, „modelując" w jej zachowaniu te jego atrybuty, które są zgodne z treścią hipotezy. Nawiązując do tytułu antologii tekstów sporządzonej przez Arthura G. Millera (1972: *The social psychology of psychological research*), badania nad psychologicznymi uwarunkowaniami procesu badawczego można nazwać *psychologią społeczną badania psychologicznego* albo *psychologią metodologii psychologii*.

Dziś – zwłaszcza w „miękkich" metodologicznie obszarach badawczych (np. psychologia kliniczna czy psychologia zdrowia) – niewielu psychologów postrzega ową interakcję: „badacz – osoba uczestnicząca w badaniu" jako odrębne, ważne źródło wariancji. Wielka szkoda.

Gdy mam odpowiedzieć na kluczowe pytania: „Jaką nauką jest psychologia?", „Jakie twierdzenia wygłaszane w jej imieniu mogą kandydować do miana **naukowych**?", to odpowiadam, z głębokim przekonaniem, następująco: **psychologia jest nauką empiryczną** i pod tym względem nie różni się od np. biologii. Każde jej nowe twierdzenie (a właściwie „kandydatka" na takowe: hipoteza) musi być skonfrontowane – *via* test empiryczny – z faktami empirycznymi. Niekoniecznie musi to być eksperyment laboratoryjny. Może to być eksperyment terenowy (*field experiment*) czy badanie typu klinicznego (to jednak powinno być poprowadzone według metodologicznych standardów postępowania profesjonalnego przyjętych w psychologii klinicznej, której praktyka (diagnostyczna i terapeutyczna) powinna być oparta na dowodach empirycznych (*evidence-based practice in psychology*; zob. American Psychological Association, 2006/2016; American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006; Brzeziński, 2016).

Wybitny przedstawiciel Szkoły Lwowsko-Warszawskiej, Kazimierz Ajdukiewicz (1949/2003, 1958) mówił o: (1) spełnianiu przez daną wypowiedź naukową **zasady intersubiektywnej komunikowalności i intersubiektywnej kontrolowalności** (sprawdzalności) oraz (2) o tym, że stopień przekonania, z jakim badacz będzie głosił dane twierdzenie, powinien być proporcjonalny do stopnia pewności jego uzasadnienia; ta zaś zależna jest od pewności metody (najlepiej, gdy jest to badanie eksperymentalne, a nie korelacyjne), którą badacz posłużył się w procedurze konfrontowania tego twierdzenia z faktami empirycznymi – jest to treść **zasady racjonalnego uznawania przekonań**. Obie zasady: (1) zasada intersubiektywności i (2) zasada racjonalnego uznawania przekonań składają się na **zasadę racjonalności** po prostu. Pisałem (Brzeziński, 2019):

> **Zasada racjonalności** powinna być bezwzględnie przestrzegana przez psychologów-badaczy. Specyfika psychologii jako nauki empirycznej nakazuje psychologom uznawać za naukowe jedynie takie wypowiedzi, które oparte zostały na rezultatach badawczych pozyskanych w dobrze **kontrolowanych** badaniach empirycznych. (s. 17)

Oczywiście takie pojmowanie psychologii wyklucza z pola nauki „psychologię różnorakich praktyk" (bywa, że bardzo „egzotycznych") diagnostycznych czy, niemalże cudownie, uzdrawiających. Oczywiście, samo opowiadanie czy pisanie głupot nie jest prawnie zabronione. Można też przejmować się tym, co piszą autorzy „harlekinów" psychologicznych. Można nawet, jak to czyni się na wydziałach teologii katolickiej lokowanych na świeckich uniwersytetach (nie udajmy jednak, że jest to nauka), zajmować się problematyką egzorcyzmów. Żyjemy w wolnym kraju. Natomiast trudno zaakceptować, że takie poglądy są wygłaszane na niektórych wyższych uczelniach. Przecież wszystkie one są nadzorowane przez Ministerstwo Edukacji i Nauki. Uzyskany dyplom magistra psychologii (a bywa, że tylko licencjata psychologii) pozwala owe „naukowe" poglądy przekuć na niezłe dochody z prywatnej praktyki sprofilowanej „psychologicznie". I to powinno być restryktywnie kontrolowane. Do tego jednak potrzebna jest działająca (a nie będąca w stanie uśpienia) ustawa o zawodzie psychologa; tej jednak cały czas nie mamy. I w ten sposób z „pomocą" ministerstwa oraz Polskiej Komisji Akredytacyjnej postępuje lawinowo proces psucia psychologii – i jako nauki, i jako praktyki.

Jak uniknąć trywialnych wyników? Czy $p < 0{,}05$ powinno przesądzać o wartości naukowej badania empirycznego?

Opracowania statystyczne danych empirycznych, przedstawiane w artykułach publikowanych dzisiaj w reprezentujących wysoki poziom naukowy czasopismach naukowych z zakresu psychologii, znacząco odbiegają od tych, które publikowano w tych samych czasopismach 50 lat temu. Jest to pochodna samego rozwoju teoretycznego i metodologicznego naszej dyscypliny naukowej, wzrostu społecznej świadomości metodologicznej, ale też postępu technologicznego: lepszej aparatury pomiarowej i testów psychologicznych, wydajnych komputerów osobistych i specjalistycznego oprogramowania (też statystycznego!). Jeżeli odwołać się do wyróżnionych przez filozofa Hansa Reichenbacha (1938/1989) **dwóch kontekstów**: **odkrycia** (*context of discovery*) i **uzasadniania** (*context of justification*), to ów pierwszy kontekst „odpowiada" za poziom naukowy psychologii, za tworzenie nowych teorii empirycznych, a drugi za poziom metodologiczny prowadzonych badań empirycznych, których celem jest sprawdzanie tych teorii – najlepiej, moim zdaniem, w duchu popperowskiego (Popper, 1974/1977) **falsyfikacjonizmu**: negatywnym sprawdzaniu, które polega na podejmowaniu przez badacza rygorystycznych prób obalenia, falsyfikacji teorii czy hipotezy, a nie na pozytywnym sprawdzaniu, **konfirmacji**: szukaniu faktów empirycznych tę teorię czy hipotezę potwierdzających.

W coraz większym zakresie poważne badania naukowe prowadzone są przez psychologów w zespołach interdyscyplinarnych. To na styku różnych dyscyplin powstają nowe, niestandardowe rozwiązania naukowe. Pozytywnym przykładem niech będzie dynamicznie rozwijająca się kognitywistyka (*cognitive science*). I to jest, jak sądzę, przyszłość empirycznej psychologii – nie izolowanie się od innych dyscyplin, zwłaszcza tych bardziej zaawansowanych metodologicznie (aparatura, pomiary, zaawansowane analizy ilościowe). Zatem korzystne będzie szukanie partnerów w obszarze nauk o mózgu czy biologii, a nie pedagogiki.

Nie jest odkrywcze stwierdzenie, że nie każda wypowiedź, nawet ta wydrukowana, wnosi „coś" nowego do psychologii. Co zatem można (trzeba) powiedzieć o wypowiedziach pretendujących do tego, aby uznać je za poważne, a nie trywialne (też te, formułowane – czy nie nazbyt często – przez psychologów)? I znowu odwołam się do wybitnego filozofa i metodologa Kazimierza Ajdukiewicza (1957/2020), który w eseju: *O wolności nauki* pisał, że jakaś tematyka podejmowana przez badacza zasługuje na miano naukowej, gdy spełnia cztery warunki. I tak, **po pierwsze**, gdy jej treść odnosi się do – jak pisze – „takich spraw, którymi jakaś nauka się zajmuje, a nie takich, które dla każdej nauki są obojętne" (s. 9). Ponadto to, co proponuje badacz, musi „wzbogacać naukę w sposób istotny". **Po drugie**, oczekuje się, aby propozycja badacza była „z należytą ścisłością sformułowana" (s. 9). **Po trzecie**, wnioski formułowane na podstawie przeprowadzonych badań powinny uwzględniać – tak bym to ujął – moc eksplanacyjną zastosowanej przez badacza metody: „[…] stanowczość, z jaką się wygłasza swe twierdzenia, jest dostosowana do stopnia ich uzasadnienia" (s. 9). Wreszcie, **po czwarte**, badacz nie może być ignorantem w badanej dziedzinie (s. 11).

Gdybyśmy poprzestali **tylko i wyłącznie** na kryterium istotności statystycznej, uznającym daną hipotezę badawczą za empirycznie pozytywnie sprawdzoną (osobiście wolę mówić o jej zadowalającym uprawdopodobnieniu), to musielibyśmy się wówczas zgodzić na „udowadnianie" jej prawdziwości za pomocą zabiegu zwiększania *ad absurdum* wielkości grup – co słusznie wydrwił Jacob Cohen (1994/2006):

> W niepublikowanej pracy Meehl i Lykken sporządzili tablice kontyngencji dla 15 zmiennych i próby liczącej 57 000 [sic! – J. M. B.] uczniów szkół ponadpodstawowych, uwzględniając zawód i wykształcenie ojca i matki, liczbę rodzeństwa, płeć, kolejność urodzeń, plany dotyczące dalszej edukacji, postawę rodziców wobec szkoły, lubienie szkoły, wybór szkoły, plany zawodowe na najbliższe dziesięć lat, preferencje religijne, sposób spędzania wolnego czasu i organizację szkoły. **Wszystkie ze 105 korelacji chi-kwadrat utworzonych przez skombinowanie zmiennych metodą „każda z każdą" okazały się istotne, przy tym 96% spośród nich było istotnych na poziomie $p < 0{,}00001$** [wyróż. – J. M. B.]. (s. 110)

I wyraźnie kpiąc, napisał: „wszystko jest powiązane ze wszystkim".

Magia wielkich prób (za każdą cenę – nawet tak dużą, że doprowadzi badacza do absurdalnych wniosków, ale „uświęconych" wartością $p < 0{,}0001$!), czasami kreuje artefakty (jak w przytoczonym cytacie z artykułu Cohena). Wiele szkód (nie tylko w psychologii) wprowadził upowszechniający się nowy **pseudostandard metodologiczny** nadający terminowi ***istotny*** (istotny związek korelacyjny,

istotna różnica między…, istotny wpływ *X* na *Y* itp.) wyłącznie sens statystyczny, odnoszący się do procedury statystycznej NHST[14] i wiążący decyzję zasadnego odrzucenia H_0 z prawdopodobieństwem $p < 0{,}05$. Dlaczego badacze tak postępują i dlaczego chowają swój raport, który nie spełnia tego kryterium do szuflady? Ano dlatego, że zostali skutecznie wytrenowani przez statystyczne autorytety w obszarze zastosowań statystyki do „obiektywnego" uznania swoich dokonań za godne upowszechnienia w społeczności badaczy. W tym treningu swoją znaczącą rolę odegrały praktyki selekcyjne stosowane przez redakcje czasopism naukowych i praktyki recenzenckie stosowane w procedurach awansowych i konkursowych. Zostały one „trafnie" odczytane przez szybko uczących się autorów artykułów (w miejsce milgramowskiego szoku elektrycznego potencjalnym autorom redakcje fundowały szok odrzucenia artykułu z powodu nieprzekroczenia kryterium $p < 0{,}05$). A to było niekiedy bardzo boleśnie odczuwane przez uczciwie postępujących badaczy. I tak pokolenie za pokoleniem utrwalały się te praktyki. Czyżby i tu zadziałała ekspresowo metoda „doboru naturalnego" Darwina?

Nie chcę tu powtarzać argumentów wysuwanych przez krytyków takiego „zero–jedynkowego" postępowania. Przecież czym innym jest wielkość ryzyka, które skłonny jest akceptować badacz odrzucający prawdziwą H_0 (możliwość popełnienia *błędu I rodzaju* o prawdopodobieństwie α), gdy jego koszt społeczny jest niewielki (np. student zaliczy pracę licencjacką, która trafi w otchłań cyfrowego archiwum), a znacząco czym innym jest uznanie tego wyniku za podstawę podjęcia prac nad nową metodą terapeutyczną (wysoki koszt społeczny błędu). W tym drugim przypadku społecznie uzasadnione jest przejście na bardziej rygorystyczny poziom tego prawdopodobieństwa, np. $p < 0{,}001$, co zwiększa **pewność decyzji** badacza, ale nie przesądza o trafności teoretycznej. Nie zapominajmy jednak o tym, że badacz może manipulować wielkością *N* tak, by za wszelką cenę przekroczyć próg istności statystycznej. Wybitny psycholog-statystyk William L. Hays (1925–1995; 1973, s. 422–424) zamieścił w swoim popularnym podręczniku *Statistics for the social sciences* pkt 10.22: „Can a sample size be too large?", a w nim takie zdanie: „trivial associations may well show up as significant results when the sample size is very large" (s. 424).

Badacz powinien też skupiać się na kontroli ryzyka popełnienia *błędu II rodzaju*, nieodrzucenia fałszywej H_0 o prawdopodobieństwie ß. I tu – ostatnio – także w literaturze psychologicznej (późno, ale dobrze, że został ten problem dostrzeżony – poświęcenie temu numeru specjalnego *Przeglądu Psychologicznego* jest też widomą oznaką dostrzeżenia rangi problemu) przywrócono należytą rangę pojęciu **mocy testu**. Powszechnie wiadomo, że moc testu wzrasta wraz ze wzrostem wielkości próby. Badacz powinien – planując badanie! – postępować elastycznie. Trudną do przecenienia zasługą Cohena (1988) jest opublikowanie fundamentalnej pracy *Statistical power analysis for the behavioral sciences*. Nie można jej pominąć w poważnym studiowaniu problematyki mocy testu.

Nie będą tu rozpisywał się na temat mocy testu, gdyż – także w polskiej literaturze psychologicznej – dostępne są opracowania tematyki **istotności**

[14] Akronim: *Null Hypothesis Significance Testing*.

statystycznej oraz **mocy testu**, że wymienię parę najnowszych: Piotra Wolskiego (2016a, 2016b, 2016c), Tytusa Sosnowskiego i Liliany Jarmakowskiej-Kostrzanowskiej (2020), tłumaczenie podręcznika statystyki dla psychologów i pedagogów Bruce'a M. Kinga i Edwarda W. Miniuma (2009), w którym bardzo dużo uwagi – jak na podręcznik o charakterze wprowadzającym w tematykę statystycznej analizy danych – poświęcono tej tematyce. Wspomnę też pierwsze podjęcie – w polskiej literaturze psychologicznej – tego zagadnienia w odniesieniu do planowania eksperymentów wedle modelu ANOVA (Brzeziński i Stachowski, 1981/1984).

Dopełnieniem **triady wskaźników** (dwa pierwsze elementy to: **poziom istności statystycznej** i **moc testu**), które muszą być każdorazowo brane pod uwagę w fazie planowania badania empirycznego, a nie tylko już po przeprowadzeniu badania (np. formułując hipotezy *ad hoc*), jest wskaźnik **wielkości efektu ES** (*effect size*) (zob. American Psychological Association, 2020; Grissom i Kim, 2005, 2011; King i Minium, 2003/2022; Rosenthal i in., 2000; Wilkinson i Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999). To stało się już od dłuższego czasu – ale jeszcze nie w Polsce! – **standardową procedurą** (m. in. zalecaną przez raporty APA – zob. przypis 10). W najnowszym, znacząco uzupełnionym, podręczniku metodologii badań psychologicznych (Brzeziński, 2019) zamieściłem i omówiłem dwie ryciny rekomendujące dla każdego testu istotności wskaźnik/wskaźniki wielkości efektu (Ryc. 10.4, s. 221 oraz Ryc. 10.5, s. 223).

Sanford Labowitz (1970) podał aż 11 kryteriów wyboru poziomu istotności. Zwłaszcza zaś – gdy **eksplorujemy** interesujący naukowo obszar problemowy – rozluźniłbym te kryteria i nie ignorował $p = 0{,}15$ czy $p = 0{,}20$. Badacz powinien – przed podjęciem brzemiennych we wnioski decyzji statystycznych – najpierw dokładnie się przyjrzeć wynikom, ich rozkładom, a dopiero potem wyciągać wnioski z pomocą odpowiednich narzędzi statystycznych (tego nauczyłem się – w mojej edukacji statystycznej – z jednej z ważniejszych książek: Johna B. Tukeya (1977): *Exploratory data analysis*.

Sumując, badacz powinien zerwać z niedobrą tradycją wiążącą się z poszukiwaniem „za wszelką cenę" (najczęściej, bez przemyślanego uzasadnienia, zwiększając wielkość próby) wyniku statystycznie istotnego (negatywne z etycznego punktu widzenia zjawisko *data fishing* czy *p-hacking*). Zdarza się to dość często w badaniach, prowadzonych – za pomocą amerykańskiego panelu *Amazon Mechanical Turk*, *AMT* (zob. Aguinis i in., 2021; Brzeziński, 2023; Buchanan i Scofield, 2018; Buhrmester i in., 2018; Keith i Harms, 2017; Saad, 2021; Webb i in., 2022) czy wzorowanych na nim polskich paneli – przez socjologów, psychologów społecznych, psychologów zdrowia czy pedagogów. Te badania nie są zbyt metodologicznie wyrafinowane. Wykorzystują metody samoopisowe: kwestionariusze osobowości, skale postaw, ankiety. Aby „przeskoczyć" poprzeczkę (jak w skoku wzwyż na igrzyskach olimpijskich): „$p = 0{,}05$" realizowane są na wielkich próbach. Z kolei nie zawsze (np. gdy mamy bardzo ograniczony dostęp do potencjalnych osób badanych – jak to dzieje się w badaniach klinicznych, np. osoby o rzadko występujących w populacji niepełnosprawnościach czy są bardzo wysokie koszty badania każdej osoby) możliwe jest dotarcie do dużych prób. Przecież dla tego typu „trudnych" badań, uwzględniających małe N, wymyślono testy istotności różnic, takie jak: *test t*, *test dokładny* Fishera (*Fisher Exact test for*

2x2 Tables), *test* χ^2, *analiza sekwencyjna Walda*, czy *testy nieparametryczne* (dla skali porządkowej, np. testy: Manna-Whitney'a-Wicoxona czy Kruskala-Wallisa i Friedmana. O nich piszą współczesne podręczniki statystyki i są też uwzględnione w „menu" pakietów statystycznych.

Wnioski

Podsumowanie moich rozważań (dziękuję Redakcji za zaproszenie do napisania artykułu i prof. Piotrowi Wolskiemu za konstruktywne uwagi do pierwotnej wersji tekstu) – w znaczącej mierze sprowokowanych opublikowanym w *Science* artykułem zespołu Open Science Collaboration (2015) oraz wywołaną nim falą krytyki skierowanej pod adresem praktyk badawczych stosowanych w psychologii (z wyraźną tendencją do naduogólnień) chciałbym ująć w dwóch częściach, odpowiednio zatytułowanych: (1) Dlaczego tak się stało i nadal się dzieje? (2) Jak można, bo nie wątpię, że można, minimalizować straty (też te wizerunkowe); inaczej: jakie należy podjąć środki zaradcze?

I jeszcze jedno: **nie tylko psychologowie grzeszą** (stosują *p-hacking* i HARKing[15], fałszują wyniki, zmyślają wyniki, dopisują się do prac, w których ich udział był minimalny, mało znaczący czy wręcz zerowy, a także plagiatują). I nie tylko ich wyniki badawcze w niezadowalającym stopniu poddają się satysfakcjonującej replikacji. Także w innych dyscyplinach naukowych – w nieznanym stopniu – dokonuje się transformacja „nauki akademickiej" w „naukę śmieciową" (*scientific junk*) (Grabski, 2015, s. 180; nie można się spodziewać, myśląc racjonalnie, że uprawia się w ten sposób prawdziwą naukę)[16]. Jak pisze Maciej W. Grabski:

---

[15] Akronim: *Hypothesizing After the Results are Known.*

[16] Żeby nie być gołosłownym, podam spektakularny przykład nadużycia w „twardej" nauce. W poczytnym dzienniku *Gazeta Wyborcza* (z 28 października 2022 r.) ukazał się artykuł Pauliny Mozolewskiej: *Skandal w przełomowych badaniach* (s. 16) z nadtytułem: „Co z leczeniem chorych na Alzheimera?". Poprzedziła go informacja autorki: „W badaniach nad chorobą Alzheimera mogło dojść do manipulacji [krytyka dotyczy artykułu S. Lesné'ego i K. H. Ashe, opublikowanego w 2006 r. w *Nature* – przyp. J. M. B.] – wskazuje dziennikarskie śledztwo. Lekarze i naukowcy zastanawiają się, jak wpłynie to na badania nad potencjalnymi lekami na Alzheimera" (s. 16). Przeprowadzone przez redakcję *Science* pół roku trwające śledztwo doprowadziło – jak na razie – do, jak pisze autorka, wniosku wydrukowanego przez *Science*, iż: „[…] kluczowe wyniki, na których opierają się liczne prowadzone przez wiele lat badania nad chorobą Alzheimera, mogły być zmanipulowane lub celowo sfałszowane" [chodzi konkretnie o manipulację dokumentacją zdjęciową – przyp. J. M. B.]. Artykuł uzupełnia obszerny wywiad, który autorka przeprowadziła z prof. Tomaszem Gabryelewiczem z Instytutu Medycyny Doświadczalnej i Klinicznej im. M. Mossakowskiego i prezesem Polskiego Towarzystwa Alzheimerowskiego. Gabryelewicz w tym wywiadzie powiedział: „Oskarżenia dotyczą zdjęć przedstawiających analizy eksperymentów oceniających poziomy białka. Wyniki z takich badań obrazuje się jako prążki i w uproszczeniu można powiedzieć, że ich wielkość i grubość określa poziom danego białka. Na razie nie można jednoznacznie ocenić, czy wątpliwości związane z tymi

> Wraz ze wzrostem „miękkości danych" i oddalaniem od głównych tematów nauki oraz słabnięciem związków z pierwszorzędnymi ośrodkami badawczymi możliwość pojawienia się nieuczciwości naukowych gwałtownie rośnie, a prawdopodobieństwo ich wykrycia maleje. W takich warunkach nauka akademicka łatwo przekształca się w coś, co coraz powszechniej nazywa się nauką śmieciową (*junk science*) […] Ilość nauki śmieciowej rośnie wykładniczo wraz z liczebnością instytucji naukowych oraz działających poza systemem *peer review* lokalnych czasopism, chociaż i renomowane periodyki nie mogą się przed nią uchronić. […]
>
> Co gorsza, nauka śmieciowa często stanowi użyteczny element manipulacji, gdyż celowo zafałszowując dane oraz naciągając ich interpretację, a także manipulując analizami naukowymi, jest wykorzystywana do wsparcia założonych z góry punktów widzenia, tworząc żerowisko dla działających bezkarnie i niejednokrotnie utytułowanych manipulatorów, hochsztaplerów i oszustów oraz dla żądnych sensacji mediów. (s. 180–181)

Przyglądając się krytycznie rozwojowi psychologii w Polsce, który jest (też) pochodną jakości kształcenia psychologów na studiach magisterskich (wszak, chcemy tego czy nie, w sposób naturalny – jesteśmy śmiertelni – następuje stopniowa, naturalna wymiana kadr), bardzo niepokoi mnie wzrastająca lawinowo liczba ośrodków kształcących przyszłych psychologów. Biorąc pod uwagę wielkość populacji kadry akademickiej (co najmniej ze stopniem doktora), nie jest możliwe porządne wykształcenie tylu psychologów, ilu kształci się obecnie w Polsce, szczególnie w uczelniach niepublicznych (moim zdaniem jest do dobry biznes – zwłaszcza gdy są to studia tanie w prowadzeniu: pedagogika, zarządzanie, nauki polityczne i … psychologia)[17].

Psucie psychologii rozpoczyna się od przygotowania przez studenta byle jakiej pracy magisterskiej (odpowiedzialni za to są, przede wszystkim, jej promotor i recenzent – skąd jednak wziąć tylu dobrze przygotowanych i zaangażowanych potencjalnych opiekunów prac magisterskich?).

Dlaczego?

Zwróćmy teraz uwagę na możliwe przyczyny mnożenia się bylejakości. Kilka czynników patologicznych, jak sądzę, ma wpływ na występowanie w społeczności

---

zdjęciami są wynikiem oszustwa z premedytacją, czy próbą «podrasowania» wyników. […] Przeważają opinie, że choć Lesné mógł podrasować zdjęcia, to nie są one najważniejszym elementem zakwestionowanego artykułu. Niektórzy eksperci analizujący fotografie sugerowali, że rzekome manipulacje mogą być artefaktami cyfrowymi, które występują przypadkowo podczas przetwarzania obrazu. **Nienależnie od tego, czy była to próba oszustwa, czy tylko podretuszowanie w Photoshopie fotografii ukazujących wyniki analizy białek, pozostaje uczucie zażenowania i duży niesmak**" [wyróż. – J. M. B.]. (s. 16–17)

[17] Aktualne dane można pozyskać z wyszukiwarki Ministerstwa Edukacji i Nauki: https://radon.nauka.gov.pl/dane/studia-prowadzone-na-okreslonym-kierunku.

badaczy niepożądanych i – co tu dużo mówić – wstydliwych czy nawet karygodnych zachowań.

Pierwszy – pycha, rywalizacja, chęć utrzymywania się w czołówce „najlepszych". Warunki pracy oraz zagrożenia typu materialnego (np. rozwiązanie umowy o pracę) nie są tu bez znaczenia. Chęć bycia wśród najlepszych to życie w nieustannym stresie. Jedyne, co powoduje takimi osobami jak Diederic Stapel z Uniwersytetu w Tilburgu (wszak należał do elity psychologów społecznych, a i troski materialne były mu obce), to nieustająca troska o to, aby nie odpaść z czołówki, aby być zawsze obecnym na prestiżowych konferencjach i drukować (i być cytowanym!) w najlepszych czasopismach w branży. A jak pomysłów już nie starcza, to słabnie odporność na pokusy i wchodzi się na równię pochyłą.

Drugi – presja pracodawcy: kierownika katedry, dyrektora instytutu, dziekana wydziału czy rektora. Ostatnimi laty w Polsce (podkreślmy, że to polska osobliwość!) nasilił się nacisk wywierany przez kierownictwo jednostek naukowych na pracowników, aby przynosili punkty za „tłuste" publikacje, aby przyspieszyli gromadzenie dorobku naukowego niezbędnego do uruchomienia postępowania habilitacyjnego. Wszak nagromadzenie przez jednostkę dużej liczby owych punktów umożliwi jej uzyskanie dobrej kategorii w stosowanej przez Ministerstwo Edukacji i Nauki oraz Komisję Ewaluacji Nauki co cztery lata ewaluacji uzyskanych wyników badań naukowych w danej dyscyplinie naukowej (tu: psychologii). Taka nadmierna, mechaniczna bibliometryzacja oceny publikowanego dorobku naukowego w skrajnych przypadkach owocuje wyborem „drogi na skróty": dopisywaniem się do publikacji innych osób, sztucznym rozdrabnianiem i powielaniem publikacji, plagiatowaniem, kupowaniem całych czy tylko części prac statystycznych (np. zaawansowanych analiz statystycznych przyprowadzonych przez wyspecjalizowane firmy czy specjalistów, też opłacanych „pod stołem"), podejmowaniem prób (niestety, na ogół zwieńczonych sukcesem) publikowania artykułów quasi-naukowych w *predatory journals* czy takichże książek, poprawianiem danych itp. (zob. Brzeziński i Oleś, 2021[18]).

Trzeci – społeczne przyzwolenie i nikłe konsekwencje czynu. Patologicznym zachowaniom sprzyja również brak jednoznacznie stanowczej reakcji społeczności akademickiej na naruszenia standardów akademickich, a zwłaszcza władz uczelni (na wszystkich jej poziomach! – też zamiatanie pod dywan patologii psujących wizerunek szkoły).

Czwarty – nadmiar słabych (zwłaszcza w sektorze niepublicznym) szkół wyższych. Mimo i tak już obniżonych wymagań kadrowych przy powoływaniu nowych kierunków studiów i podtrzymywaniu (tu czasami chciałoby się użyć terminu „reanimacji") tych, które, w imię przyzwoitości chociażby, powinny być zlikwidowane. O awans naukowy ubiegają się osoby, które ani nie odczuwają takiej potrzeby, ani nie są utalentowane, ani nie są w stanie napisać porządnego artykułu naukowego. Pamiętajmy, że i tu działa prawo rozkładu normalnego. Co zatem mogą zrobić? Albo odejść z uczelni (tylko dokąd?), albo próbować wyżej opisanej swoistej „drogi na skróty".

[18] Zob. rozdz. 10, pkt 10.2: „Etyczny kontekst badania naukowego", s. 411–475.

Piąty – praktyka publikacyjna wydawców czasopism psychologicznych. Psychologowie wiedzą, że aby opublikować artykuł, musi on zdawać sprawozdanie z badań, w których „coś" wyszło. Oznacza to, że dla wydawców liczą się tylko artykuły informujące o badaniach, w których badacz/autor uzyskał wynik statystycznie istotny na minimalnym wymaganym poziomie: $p < 0{,}05$. Robi się tedy wszystko (także manipuluje danymi!), aby jednak „wyszło". Zauważmy, że istotność statystyczną utożsamia się z rzeczywistą siłą oddziaływania zmiennej niezależnej na zmienną zależną. Dopiero od niedawna poważne czasopisma wymagają, aby autorzy podawali także wartości wskaźników *wielkości efektu*, które informują właśnie o sile wpływu jednej zmiennej (lub interakcji dwóch i większej liczby zmiennych) na zmienną zależną, a nie tylko *poziom istotności*.

Środki zaradcze

Jakie zatem można podjąć środki zaradcze? Myślę, że można mówić o **czterech, uzupełniających się środkach**.

Pierwszy – respektowanie zasady – nazwę ją ***zasadą jawności***. Ograniczone ramy artykułu empirycznego (a zwłaszcza, gdy jest to tzw. *short report*) uniemożliwiają drukowanie szczegółowych informacji dotyczących charakterystyk badanych grup czy bardziej szczegółowych danych z opisu statystycznego wyników. Badacz powinien być jednak gotowy do udostępniania takich danych. Może za pośrednictwem redakcji czasopisma, która będzie depozytariuszem (przez jakiś czas) tych danych. Formułowany już przez niektóre redakcje wymóg dostarczania na ich żądanie (bo zachodzi podejrzenie naruszenia etyki – aby nie pojawił się kolejny Stapel ze swoimi „genialnymi" publikacjami[19]) **danych surowych**, aby możliwe było przeprowadzenie ich reanalizy. Nie przyjmuję do wiadomości, że dane są własnością badacza i tylko on może się nimi posługiwać. Taka postawa jest nie do zaakceptowania zwłaszcza wówczas, gdy badania były finansowane ze środków publicznych (płaci za nie podatnik!), a tak jest w systemie grantowym Narodowego Centrum Nauki.

Drugi – wymaganie **replikowania** badań (Neuliep, 1991; Wolski, 2016b). Tylko wyniki, które innym badaczom uda się powtórzyć, mają wartość naukową.

Trzeci – **zmiana polityki wydawniczej prowadzonej przez wydawców czasopism naukowych**. Dziś czasopisma niechętnie publikują artykuły stanowiące replikacje wcześniej opublikowanych wyników badań. Redakcje zastrzegają, że publikują tylko wyniki oryginalne! Konsekwencje tego są takie, iż **nie wiemy, ile nieopublikowanych artykułów zalega w szufladach badaczy**, dlatego że ich autorzy nie uzyskali „uświęconej" wartości $p < 0{,}05$, a nie chcieli postępować wbrew etyce i „poprawiać" danych. Stąd mówi się o negatywnym *efekcie szuflady* (zob. np. Rosenthal, 1979). Zauważmy, że taka stronnicza polityka wydawnicza wpływa na stronniczość metaanaliz (niedopuszczone do druku negatywne wyniki nie są „widoczne" dla metaanalizy), zawyżając ich wyniki.

[19] Zob. opis tego przypadku: Budzicz (2015).

Aby temu zapobiec, trzeba dopuścić nowe podejście do akceptacji przez redakcje artykułów. Nadzieje budzi (choć nie u wszystkich badaczy) inicjatywa wydawnicza, do której zaczęły się przyłączać, ale w ograniczonym zakresie, znaczące pisma. Polega ona na tym, że recenzowaniu podlega nie tylko gotowy tekst, lecz także cała koncepcja badania empirycznego (przed jego przeprowadzeniem). Jeżeli spotka się ona z pozytywnymi opiniami recenzentów, a jego wykonanie będzie zaakceptowane przez recenzentów, to redakcja zapewnia jej autora, że wyniki (niezależnie od tego czy „coś" wyszło: $p < 0{,}05$ czy nie: $p > 0{,}05$) przeprowadzone zgodnie z zaopiniowaną koncepcją będą opublikowane. Ten nowy format publikacji nosi nazwę „**wstępnej rejestracji**" (*pre-registration research*).

Czwarty – będący w jakimś sensie konsekwencją trzeciego środka zapobiegawczego (o czym pisałem wyżej) – **elastyczne podejście do wartości *p*.** Dlaczego to właśnie wartość $p < 0{,}05$ ma być traktowana jako bezwzględna miara wartości naukowej uzyskanego wyniku? Przecież to jest kwestia li tylko **konwencji**. Tak też zostało to ujęte przez Fishera (1925/1938), autora tego progu akceptacji z odwołaniem do właściwości rozkładu normalnego[20]. Racjonalne natomiast jest przejście na wskaźniki **wielkości efektu** i na **przedziały ufności** (zob. King i Minium, 2003/2022; Loftus, 2008, 2012). Lista poważnych badaczy – metodologów i statystyków – takie podejście polecających jest długa. Zwłaszcza **podawanie wartości granic przedziału ufności jest bardzo naukowo wartościowe**. No cóż, czasami są one, jak pisał Cohen (1994/2006), „kłopotliwie szerokie":

> Rzadko jednak spotyka się artykuły, w których znajdowałyby się informacje o przedziałach ufności. Podejrzewam, że jednym z powodów tego jest fakt, że są one tak **kłopotliwie szerokie**! Jednak ten **ich rozmiar powinien motywować nas do ulepszania naszych metod pomiaru i do prób redukowania wariancji błędu w naszych narzędziach** (tak jak to prawie wiek temu rekomendował Student). Zresztą szerokość przedziałów ufności jest czymś podobnym do tego, czym jest analiza mocy względem testowania istności – szerokość ta zmniejsza się wraz ze wzrostem wielkości próby, podobnie jak wielkość zwiększa moc procedur testowania hipotez zerowych. (s. 114–115)

Chciałbym zauważyć, że po prawie 30 latach od daty druku artykułu Cohena niewiele się zmieniło w praktyce publikowania wyników badań psychologicznych.

* * *

Mam nadzieję, że zdołałem przekonać Czytelnika, iż psychologia jest – gdy ją potraktować poważnie, wnikliwie – trudną dyscypliną badawczą. Także

[20] Czytamy u Fishera: „[…] The value for which P = .05, or 1 in 20, is 1,96 or nearly 2; it is **convenient** to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviations are thus formally regarded as significant". (s. 46)

niełatwa jest aplikacja jej teorii czy pojedynczych twierdzeń, które wyszły „zwycięsko" z konfrontacji z testami empirycznymi, w sferze praktyki społecznej: diagnostycznej, orzeczniczej, penitencjarnej czy terapeutycznej. Owa aplikacja ma tylko wówczas sens (jest także etyczna), gdy spełnia coraz bardziej wyśrubowane wymagania metodologiczne. Zachodzi zatem ścisła zależność postępu w skuteczności stosowanych przez psychologów-profesjonalistów procedur pomocowych od postępu dokonującego się w badaniach naukowych (np. opracowanie nowych procedur pomiarowych, które pozwolą na formułowanie zarówno rzetelniejszych i trafniejszych diagnoz, jak i takichże terapii). Aby tak się stało, psychologowie nie powinni nadmiernie koncentrować swojej uwagi na specyficznych problemach technicznych odnoszących się do, np. przywołanego w tym artykule, zagadnienia mocy testu statystycznego zastosowanego do oceny hipotezy zerowej. Mówiąc inaczej – a nie lekceważę ważności tego zagadnienia – opracowanie w języku współczesnej statystyki poprawnie zebranych wyników to tylko jeden z elementów całego postępowania badawczego. Statystyka jest wyłącznie narzędziem, które będzie odpowiedzialnie użyte jedynie wówczas, gdy badacz-psycholog nie tylko je dobrze pozna (techniczna biegłość posłużenia się jakimś pakietem statystycznym typu SPSS to zbyt mało), lecz także gdy pozna granice jego uzasadnionego (określonego założeniami modelu: testu istotności różnic, wskaźników wielkości efektu, przedziałów ufności, miar korelacji). Statystyka, niestety, dostarczyła recenzentom i redaktorom czasopism naukowych pozornie tylko prostego, ilościowego kryterium ważności uzyskanego przez psychologa wyniku – współczynnika istotności statystycznej: $p < 0{,}05$. W artykule starałem się pokazać, że jest to złe podejście, że owa „istotność statystyczna na poziomie $p < 0{,}05$ nie może stanowić jedynego „zero–jedynkowego" kryterium uznawania wyniku za naukowo interesujący. Mechanicznie podejmowane decyzje wydawnicze w procedurze kwalifikowania artykułu do druku w jakimś czasopiśmie naukowym na podstawie kryterium „$p < 0{,}05$", deformują rozwój nauki (przykładem takiej deformacji jest przywołany przeze mnie tzw. *efekt szuflady*). Potrzebne jest – jak starałem się przekonać Czytelnika – elastyczne podejście w wyborze wartości p. Poza tym niezbędne jest też odwołanie się do wartości *wielkości efektu*. Trzeba też – przy ocenie istotności różnic – odwołać się do przedziałów ufności – jeżeli poziom pomiarowy danych na to pozwala.

I jeszcze jedno. Psychologowie zdają się zapominać o tym, że badanie psychologiczne ma charakter psychologiczny i że wykryte przed laty przez Rosenthala czy Orne'a efekty nie mogą być ignorowane. Cenę, którą przyjdzie zapłacić badaczowi, gdy o tym zapomni, będzie ustalanie nie faktów – jak pisał przywoływany tu kilkakrotnie Rosenthal – ale **artefaktów**.

Bibliografia

Aguinis, H., Villamor, I., Ramani, R. S. (2021). MTurk Research: Review and Recommendations. *Journal of Management*, *47*(4), 823–837.https://doi.org/10.1177/0149206320969787

Ajdukiewicz, K. (1949/2003). *Zagadnienia i kierunki filozofii. Teoria poznania. Metafizyka*. Czytelnik.

Ajdukiewicz, K. (1957/2020). O wolności nauki. *Nauka*, *2*, 7–24. https://doi.org/10.24425/nauka.2020.132629

Ajdukiewicz, K. (1958). Zagadnienie racjonalności zawodnych sposobów wnioskowania. *Studia Filozoficzne*, *4*, 14–29.

American Psychological Association. (2006/2016). Praktyka psychologiczna oparta na dowodach. Raport sporządzony przez Grupę Roboczą ds. praktyki opartej na dowodach, powołaną przez Zarząd Amerykańskiego Towarzystwa Psychologicznego, przeł. L. Kalita. W: L. Cierpiałkowska i H. Sęk (red.), *Psychologia kliniczna* (s. 739–758). Wydawnictwo Naukowe PWN.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (wyd. 7). Author.

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066X.61.4.271

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology. *Why do we need them? What might they be?*, *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Blanck, P. D. (red.). (1993). *Interpersonal expectations. Theory, research, and applications*. Cambridge University Press.

Brzeziński, J. (2012). *Badania eksperymentalne w psychologii i pedagogice* (wyd. popr.). Wydawnictwo Naukowe Scholar.

Brzeziński, J. (2016). Towards a comprehensive model of scientific research and professional practice in psychology. *Current Issues in Personality Psychology*, *4*(1), 2–10. https://doi.org/10.5114/cipp.2016.58442

Brzeziński, J. M. (2019). *Metodologia badań psychologicznych. Wydanie nowe*. Wydawnictwo Naukowe PWN.

Brzeziński, J. M. (2023). Pytania do psychologów prowadzących badania naukowe. W: A. Jonkisz, J. Poznański SJ i J. Koszteyn (red.), *Zrozumieć nasze postrzeganie i pojmowanie człowieka i świata. Profesorowi Józefowi Bremerowi SJ z okazji 70-lecia urodzin* (s. 289–311). Wydawnictwo Naukowe Akademii Ignatianum.

Brzeziński, J. M., Oleś, P. K. (2021). *O psychologii i psychologach. Między uniwersytetem a praktyką społeczną*. Wydawnictwo Naukowe PWN.

Brzeziński, J., Siuta, J. (red.). (2006). *Metodologiczne i statystyczne problemy psychologii*. Wydawnictwo Naukowe UAM.

Brzeziński, J., Stachowski, R. (1981/1984). *Zastosowanie analizy wariancji w eksperymentalnych badaniach psychologicznych*. Państwowe Wydawnictwo Naukowe.

Buchanan, E., Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*(3), 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

Budzicz, Ł. (2015). Dyskusja „po Stapelu". Wokół rzetelności badań i publikacji w psychologii. *Roczniki Psychologiczne*, *18*(1), 9–24.

Buhrmester, M. D., Talaifar, S., Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (wyd. 2). L. Erlbaum.

Cohen, J. (1990/2006). O tym, czego się nauczyłem (jak dotąd), tłum. R. Polczyk. W: J. Brzeziński i J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (s. 75–99). Zysk i S-ka Wydawnictwo.

Cohen, J. (1994/2006). Ziemia jest okrągła ($p < 0{,}05$), przeł. R. Polczyk. W: J. Brzeziński i J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (s. 100–118). Zysk i S-ka Wydawnictwo.

Edwards, A. L. (1950/1960/1968/1972). *Experimental design in psychological research*. Holy, Rinehart and Winston.

Fisher, R. A. (1925/1938). *Statistical methods for research workers* (wyd. 7. zm. i rozszerz.). Oliver & Boyd.

Fisher, R. A. (1935/1971). *The design of experiment* (wyd. 8). Oliver & Boyd.

Grissom, R. J., Kim, J. J. (2005). *Effect sizes for research. A broad practical approach*. The Psychology Press, Taylor and Francis Group.

Grissom, R. J., Kim, J. J. (2011). *Effect sizes for research. Univariate and multivariate applications* (wyd. 2). Routledge, Taylor and Francis Group.

Harlow, L. L., Mulaik, S. A., Steiger, J. H. (red.). (1997). *What if there were no significance tests?* L. Erlbaum.

Hays, W. L. (1973). *Statistics for the social sciences* (wyd. 2). Holt, Rinehart and Winston [wyd. 1, 1963: *Statistics for psychologists*; ostatnie, wyd. 5: *Statistics* ukazało się w 1994 r.].

Henkel, E., Morrison, D. E. (red.). (1970). *The significance test controversy: A reader*. Butterworths.

Keith, M. G., Tay, L., Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, *8*, 1359. https://doi.org/10.3389/fpsyg.2017.01359

King, B. M., Minium, E. W. (2003/2022). *Statystyka dla psychologów i pedagogów*, przeł. M. Zakrzewska. Wydawnictwo Naukowe PWN.

Kirk, R. E. (1968/1982/1995). *Experimental design: Procedures for the behavioral sciences* (wyd. 3). Brooks/Cole.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (wyd. 4). Sage.

Labowitz, S. (1970). Criteria for selecting a significance level: A note on the sacredness of .05. W: E. Henkel i D. E. Morrison, (red.), *The significance test controversy: A reader* (s. 166–171). Butterworths.

Larsen, R. J. (2005). Saul Rosenzweig (1907–2004). *American Psychologist*, 60(3), 259. https://doi.org/10.1037/0003-066X.60.3.259

Loftus, G. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. W: H. Pashler i J. Wixted (red.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (s. 339–390). John Wiley & Sons, Inc. https://doi.org/10.1002/0471214426.pas0409

Miller, A. G. (red.). (1972). *The social psychology of psychological research*. The Free Press.

Neuliep, J. W. (red.). (1991). *Replication research in the social sciences*. Sage.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://www.researchgate.net/publication/281286234_Estimating_the_reproducibility_of_psychological_science

Orne, M. T. (1962/1991). Eksperyment psychologiczny z punktu widzenia psychologii społecznej ze szczególnym uwzględnieniem wpływu zmiennych sugerujących hipotezę i ich implikacji, przeł. J. Siuta. W: J. Brzeziński i J. Siuta (red.), *Społeczny kontekst badań psychologicznych i pedagogicznych. Wybór tekstów* (s. 15–32). Wydawnictwo Naukowe UAM.

Orne, M. T. (1973/1993). Komunikowanie się w sytuacji eksperymentalnej: dlaczego jest ono istotne, jak jest oceniane i jakie ma znaczenie dla trafności ekologicznej, przeł. J. Siuta i K. Tatarczuch. W: J. Brzeziński (red.), *Psychologiczne i psychometryczne problemy diagnostyki psychologicznej* (s. 31–68). Wydawnictwo Naukowe UAM.

Popper, K. (1974/1977). *Logika odkrycia naukowego*, przeł. U. Niklas. Państwowe Wydawnictwo Naukowe.

Reichenbach, H. (1938/1989). Trzy zadania epistemologii [przeł. W. Sady: §1: *The three tasks of epistemology*. W: H. Reichenbach, *Experience and prediction* (s. 3–16). University of Chicago Press]. *Studia Filozoficzne*, *7–8*, 205–212.

Rosenthal, R. (1966/2009). Experimenter effects in behavioral research. Appleton-Century-Crofts. W: *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's classic books* (s. 287–666). Oxford University Press.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 838–641.

Rosenthal, R., Rosnow, R. L., Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge University Press.

Rosenzweig, S. (1933). The experimental situation as a psychological problem. *Psychological Review*, *40*, 337–354.

Saad, D. (2021). Nowe narzędzia i techniki zwiększające trafność badań internetowych. *com. press*, *4*(1), 106–121. https://journals.ptks.pl/compress/article/view/248/163; https://doi.org/10.51480/compress.2021.4-1.248

Schneider, W. J., McGrew, K. S. (2012). The Cattell-Horn-Carroll model of Intelligence. W: D. P. Flanagan i P. L. Harrison (red.), *Contemporary intellectual assessment: Theories, tests, and issues* (s. 99–144). The Guilford Press.

Schwarzer, G. (2022). *General Package for Meta-Analysis. Version 6.0-0*. https://cran.r-project.org/web/packages/meta/meta.pdf

Skipper, J. K. Jr., Guenther, A. L., Nass, G. (1967/1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. W: R. E. Henkel i D. E. Morrison (red.), *The significance test controversy. A reader* (s. 155–160). Butterworths.

Sosnowski, T., Jarmakowska-Kostrzanowska, L. (2020). Do czego potrzebna jest moc statystyczna? W: M. Trojan i M. Gut (red.), *Nowe technologie i metody w psychologii* (s. 449–470). Liberi Libri. https://doi.org/10.47943/lib.9788363487430.rozdzial21

Trusz, S. (2013). *Efekty oczekiwań interpersonalnych. Wybór tekstów*. Wydawnictwo Naukowe Scholar.

Tukey, J. B. (1977). *Exploratory data analysis*. Addison-Wesley.

Webb, M. A., Tangney, J. P. (2022). Too Good to Be True: Bots and Bad Data From Mechanical Turk. *Perspectives on Psychological Science*. https://www.gwern.net/docs/psychology/2022-webb.pdf. https://doi.org/10.1177/17456916221120027

Wilkinson, L., Task Force on Statistical Inference American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Winer, B. J. (1962/1971). *Statistical principles in experimental design*. McGraw-Hill.

Winer, B. J., Brown, D. R., Michels, K. M. (1991). *Statistical principles in experimental design* (wyd. 3). McGraw-Hill.

Wolski, P. (2016a). Istotność statystyczna I. Nieodrobiona lekcja. *Rocznik Kognitywistyczny*, *9*, 27–35. https://doi.org/10.4467/20843895RK.16.003.5471

Wolski, P. (2016b). Istotność statystyczna II. Pułapki interpretacyjne. *Rocznik Kognitywistyczny*, *9*, 59–70. https://doi.org/10.4467/20843895RK.16.006.6412

Wolski, P. (2016c). Istotność statystyczna III. Od rytuału do myślenia statystycznego. *Rocznik Kognitywistyczny*, *9*, 71–85. https://doi.org/10.4467/20843895RK.16.007.6413