PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 9–26
DOI: 10.31648/przegldpsychologiczny.9455

Dwa głosy o kryzysie wiarygodności w psychologii

Arkadiusz Białek[1]
Uniwersytet Jagielloński, Instytut Psychologii
https://orcid.org/0000-0002-9002-4764

Piotr Wolski[2,3]
Uniwersytet Jagielloński, Instytut Psychologii
https://orcid.org/0000-0002-7028-6142

Streszczenie

Choć różne niedociągnięcia i wady sposobu prowadzenia badań i analizowania wyników w psychologii oraz innych naukach społecznych dostrzegano już dawno, ostatnie lata wyróżnia zarówno powszechność, jak i zakres tej krytyki. Pojawia się też więcej propozycji naprawy. W artykule skupiamy się na wybranych, kluczowych naszym zdaniem, obszarach kryzysu wiarygodności w psychologii. Piotr Wolski omawia te, które wiążą się z niewłaściwym rozumieniem i stosowaniem testów istotności, Arkadiusz Białek charakteryzuje niektóre z obniżających wiarygodność badań psychologicznych niewłaściwych praktyk badawczych oraz pokazuje, jak można im przeciwdziałać. Choć stosowanie dobrych praktyk badawczych może poprawić reprodukowalność i replikowalność wyników badań, to postulowana reforma powinna objąć swoim zakresem także sposób tworzenia teorii. Omawiana propozycja zasad tworzenia teorii w psychologii prowadzi do serii praktycznych kroków. Inaczej niż w modelu hipotetyczno-dedukcyjnym, za punkt wyjścia przyjmuje się w niej identyfikację i opis fenomenu. Sformułowane poprzez abdukcję wyjaśnienie fenomenu jest następnie formalizowane w równaniach matematycznych lub symulacjach komputerowych i weryfikowane. Przestrzeganie dobrych praktyk badawczych oraz poprawne tworzenie teorii ma szansę dostarczyć psychologii solidniejszych podstaw i uczynić ją nauką o kumulatywnym charakterze.

Słowa kluczowe: kryzys wiarygodności, wnioskowanie statystyczne, wartość p, testy istotności, niewłaściwe praktyki badawcze, tworzenie teorii

[1] Adres do korespondencji: a.bialek@uj.edu.pl.
[2] Adres do korespondencji: piotr.wolski@uj.edu.pl.
[3] Obaj autorzy wnieśli równy wkład w ten manuskrypt.

Ostatnie lata to okres burzliwych dyskusji i zmian w psychologii. Trudności w replikowaniu wyników (Ioannidis, 2005; Open Science Collaboration, 2015), ograniczenia ich generalizowalności (Yorkoni, 2022) oraz niedostatki teorii psychologicznych (Eronen i Bringmann, 2021; Oberauer i Lewandowsky, 2019) doprowadziły do podważenia wiarygodności badań i ogłoszenia stanu kryzysu. Choć wiele z identyfikowanych współcześnie problemów nie jest niczym nowym – na podobieństwa pomiędzy obecnym kryzysem a dyskusjami toczonymi w psychologii społecznej w latach 60. i 70. wskazuje Lakens (2023), a testowanie istotności hipotez zerowych krytykowano już w latach 30. ubiegłego wieku, czyli zanim zadomowiło się ono w psychologii (Cohen, 1994/2006) – to obecny kryzys wyróżnia powszechność świadomości problemów oraz obecność procedur i praktyk naprawczych. Owa odmienność związana jest częściowo z rozwojem technologicznym, jak np. pojawienie się platform umożliwiających prejestrację badań, danych i kodów analiz (np. OSF, Zenodo, GitHub) czy formatów dokumentów, ułatwiających reprodukowalność rezultatów (np. R Markdown, Quarto).

Brak miejsca uniemożliwia nam kompleksowe omówienie tutaj przejawów, przyczyn i proponowanych rozwiązań kryzysu wiarygodności w psychologii (przystępne, aktualne omówienie można znaleźć w pracy Nosek i in., 2022). Poniżej skupiamy się tylko na wybranych, kluczowych naszym zdaniem, obszarach. Piotr Wolski pisze o wnioskowaniu statystycznym i o tym, jak jego niewłaściwe rozumienie i stosowanie przyczynia się do obniżenia wiarygodności badań. Arkadiusz Białek charakteryzuje obniżające wiarygodność badań dyskusyjne praktyki i omawia interesującą propozycję zasad tworzenia teorii w psychologii.

Wnioskowanie statystyczne

Metodologia rzadko jest ulubionym przedmiotem osób studiujących psychologię. W swoim sondażu Haller i Kraus (2002) zaobserwowali problemy z właściwą interpretacją typowego wyniku testu *t*-studenta u 100% przebadanych studentów, ok. 90% badaczy i aż ok. 80% osób uczących metodologii. Choć specyficzny charakter zadań oraz ograniczenia próby każą traktować uzyskane wartości raczej anegdotycznie, ich wyniki współbrzmią z powszechną wśród nauczycieli metodologii obserwacją, że w kwestiach dotyczących wnioskowania statystycznego większość studentów, ale też duża część badaczy i badaczek czuje się niezbyt pewnie i najchętniej opiera się na gotowych schematach.

Najważniejszy z takich schematów interpretacyjnych, powielany w wielu podręcznikach, wykładach, w niezliczonych materiałach dydaktycznych i poradnikach internetowych, wreszcie przekazywany bezpośrednio między samymi badaczami, dotyczy interpretacji prawdopodobieństwa testowego *p*. Ma on swoją genezę we wczesnych pracach Ronalda Fishera, rozwiniętych i zmodyfikowanych później przez Jerzego Neymana we współpracy z Egonem Pearsonem, ale dziś rozpowszechniany jest głównie przez głuchy telefon licznych wzajemnie inspirujących się i powtarzających – nie zawsze precyzyjnie – źródeł. Jego istotę można w uproszczeniu streścić następująco: jeśli wyliczona w teście statystycznym

wartość prawdopodobieństwa p jest mniejsza od progu 0,05, wówczas testowany efekt uznaje się za istotny statystycznie. Konstatacja istotności statystycznej oznacza pewnego rodzaju walidację testowanego efektu: potwierdzenie, że dane uzasadniają stwierdzenie – z wystarczająco małym ryzykiem błędu – iż zaobserwowany w próbie efekt nie jest przypadkową fluktuacją, ale odzwierciedla realnie istniejącą prawidłowość. Wartość p to warunkowe prawdopodobieństwo wystąpienia w próbie efektu takiego lub większego niż zaobserwowany, gdyby ów efekt w populacji nie był obecny, tzn. gdyby prawdziwa była tzw. hipoteza zerowa. Jeśli hipotezę zerową uda się odrzucić, efekt można uznać za istotny statystycznie. Według Fishera (1971) „O każdym eksperymencie można powiedzieć, że istnieje tylko po to, by dać faktom szansę obalenia hipotezy zerowej" (s. 16). W praktyce opisany schemat redukuje się zwykle do prostej zasady: jeśli efekt jest istotny statystycznie, tzn. $p \leq 0{,}05$, wyniki są rzetelne i można je uogólniać na populację – z 5-procentowym lub mniejszym, zależnie od wartości p, ryzykiem błędu. Niestety – jak pokazują m.in. wyniki wspomnianego sondażu Hallera i Krausa – opisany schemat interpretacyjny, a zwłaszcza jego uproszczona wersja, często prowadzi użytkowników do błędnych wniosków, m.in. przeceniania znaczenia istotności statystycznej, wyciągania za daleko idących wniosków o populacji, złego rozumienia natury błędu i nieadekwatnej oceny jego ryzyka.

Poważnym problemem jest niewłaściwe rozumienie wartości p. Przyzwyczailiśmy się traktować ją tak, jakby była obiektywnym, niezależnym kryterium wiarygodności (istotności) wyniku. Tymczasem p nie jest parametrem populacji, tylko statystyką z próby – obciążoną błędem losowym tak samo jak testowany wynik. Proste symulacje pokazują, że – zwłaszcza przy niewielkich próbach – kolejne powtórzenia tego samego eksperymentu prowadzą do zasadniczo różnych ocen wielkości efektu i w konsekwencji także diametralnie różniących się między sobą ocen prawdopodobieństwa p oraz decyzji co do istotności statystycznej (Cumming, 2008; Halsey i in., 2015).

Test istotności statystycznej jest użytecznym narzędziem, pomagającym oddzielić realne efekty od losowych fluktuacji, ale działa prawidłowo tylko wtedy, gdy zarówno prawdopodobieństwo błędu I rodzaju (fałszywego alarmu – uznania losowej fluktuacji za efekt obecny w populacji), jak i II rodzaju (pominięcia – uznania efektu istniejącego w populacji za losową fluktuację obecną tylko w próbie) są odpowiednio małe. Popularna konwencja pilnuje nas, żeby ryzyko fałszywego alarmu nie było większe niż 5%, nie ma jednak równie powszechnego zwyczaju, który by nas zabezpieczał przed błędem pominięcia i jednocześnie gwarantował wielkość próby ograniczającą błąd losowy do akceptowalnego poziomu. Moc statystyczna, czyli prawdopodobieństwo niepopełnienia błędu drugiego rodzaju, jest w przypadku typowych badań w psychologii bulwersująco mała – mniejsza niż 50% (Bakker i in., 2012; Cohen, 1962; Rossi, 1990; Sedlmeier i Gigerenzer, 1989). W życiu codziennym nie dałoby się funkcjonować, opierając się na tak mało rzetelnych sposobach określania stanu rzeczy. Wyobraźmy sobie, co byłoby, gdybyśmy w piekarni co drugi raz patrzyli na półki pełne pieczywa i wychodzili rozczarowani z przekonaniem, że właśnie się skończyło; gdybyśmy próbując się umyć, co drugi raz niepotrzebnie rozpakowywali nowe mydło, uznawszy niesłusznie, że poprzednie zostało zużyte; albo gdybyśmy otwierali drzwi na co

drugi dzwonek, a i wtedy co drugi raz zamykali je gościom przed nosem zaraz po otwarciu, konstatując, że pod drzwiami nie ma nikogo, a dzwonek nam się tylko przywidział… Szczęśliwie ewolucja zadbała, byśmy w codziennym funkcjonowaniu „odruchowo" opierali się na odpowiednio mocnych danych i tworzyli wystarczająco rzetelne modele rzeczywistości. Jednak w przypadku wnioskowania statystycznego, paradoksalnie, nasze decyzje bywają mniej rozsądne od tych codziennych. Intuicje statystyczne (także profesjonalistów, Bakker i in., 2016) są zawodne, podobnie jak złote reguły w rodzaju granicy $N = 30$, rzekomo oddzielającej próby duże od małych, albo zasady minimum 30 obserwacji na komórkę tabeli w eksperymentach wieloczynnikowych. Dlatego tak jak nie szacujemy wartości p na oko, również potrzebnej wielkości próby nie powinniśmy ustalać *ad hoc*, tylko wyliczyć ją – adekwatnie do oczekiwanej wielkości efektu i zakładanej mocy statystycznej. Podstawowe obliczenia nie są wymagające, a odpowiednie narzędzia programowe, także darmowe są łatwo dostępne (np. Faul i in., 2007). Badania są kosztowne i pracochłonne, więc wielkość próby wynika zwykle z kompromisu między chęcią maksymalizowania mocy statystycznej a względami praktycznymi. Wspominaliśmy, że wartość p bywa absolutyzowana – często użytkownicy testów statystycznych nie mają świadomości jej obciążenia błędem losowym i nawet wyniki badań o małej mocy uznają za wiarygodne, jeśli tylko te spełniają „magiczne" kryterium istotności statystycznej $p \leq 0{,}05$. Takie myślenie może zachęcać do badawczego „pokera": wprawdzie nie mamy czasu ani pieniędzy na duże badania, ale teoria uzasadniająca oczekiwanie istnienia efektu wydaje się przekonująca, więc może nam się poszczęści i uda się uzyskać potrzebną wartość $p \leq 0{,}05$ już w małej próbie. Prawa wielkich liczb jednak nie da się oszukać – każdy wniosek oparty na badaniu zbyt małej próby jest obciążony dużym ryzykiem pomyłki, tak samo pozytywny, jak negatywny. Wierzyć – wciąż z ograniczonym zaufaniem – możemy tylko tym wynikom, które pochodzą z badań o wystarczająco dużej mocy statystycznej. Jeśli rozumieć wiarygodność zgodnie z pierwotną ideą Fishera (1971), jako uzasadnione oczekiwanie, że kolejne badania tego samego efektu będą powtarzalnie przynosiły statystycznie istotne wyniki (s. 14), w psychologii wiarygodny jest tylko *mniej niż co drugi* z publikowanych, statystycznie istotnych wyników (Boyce i in., 2023; Open Science Collaboration, 2015). Możemy poprawić tę niezbyt chlubną statystykę, podchodząc bardziej rygorystycznie do planowania badań. Należy zwrócić na ten ważny element większą uwagę w dydaktyce, recenzenci powinni bardziej powszechnie oczekiwać od autorów uzasadnienia decyzji dotyczących próby, a nade wszystko sami powinniśmy planować swoje projekty tak, żeby dawały rozsądnie dużą pewność wykrycia efektów, których szukamy, jeśli te istnieją.

Innym ważnym problemem interpretacji testów statystycznych jest automatyczne uznawanie wyników statystycznie istotnych za znaczące w sensie potocznym. Tej nadinterpretacji sprzyja (a zarazem odzwierciedla ją) ważna zmiana, która zaszła w języku metodologicznym: choć Fisher (1971) pierwotnie odnosił przymiotnik „istotny" do sprzeczności wyników z hipotezą zerową, dziś odnosimy go raczej do samych wyników. Sugerujemy w ten sposób myląco, że to nie ich *niezgodność* z hipotezą przypadku jest znacząca, ale one same. Fisher przestrzegał też przed myleniem istotności statystycznej z praktyczną. W większości typowych

przypadków efekt statystycznie istotny to taki, co do którego tylko mamy rozsądną pewność, że nie wynosi 0. Nie znaczy to jednak, że ma znaczenie praktyczne. Istnieje wiele efektów, które choć różne od 0, są tak małe, że nie warto zaprzątać sobie nimi głowy. Dlatego metodolodzy zalecają obliczanie przedziałów ufności i uzupełnianie testów istotności szacunkami wielkości efektów (Wilkinson i in., 1999).

Z czysto językowego punku widzenia mogłoby się wydawać, że kiedy mówimy o efekcie statystycznie istotnym, mamy na myśli taki, który nie dość, że jest znaczący, to jeszcze znajduje potwierdzenie swojej wagi w procedurach statystycznych. Tymczasem w rzeczywistości chodzi tu o efekty, które są tylko sprzeczne z hipotezą zerową, czyli – w najbardziej typowym przypadku – różne od 0, ale nie wiadomo czy znacząco. Paradoksalnie zatem określenie „istotny statystycznie" nie znaczy więcej, ale mniej niż sam przymiotnik „istotny" w jego podstawowym sensie. Prowadzi to do konfuzji, ponieważ – czy chcemy, czy też nie – znaczenia słów przetwarzamy automatycznie i trudno nam zignorować skojarzenia, jakie owa fraza przywołuje. Pewnie dlatego istotność statystyczna jest tak często nadinterpretowana i przeceniana. Byłoby zapewne inaczej, gdybyśmy zamiast terminu „istotny" używali określenia „niezerowy". Konotacje tych terminów są zasadniczo różne – efekt *niezerowy* wprawdzie istnieje, ale ma nieznaną wielkość, może jest ważny, a może nie, trudno o nim cokolwiek powiedzieć bez dalszych, dokładniejszych badań lub analiz, do których przeprowadzenia zachęca; efekt *istotny* to raczej efekt ważny, duży, zasługujący na uwagę, stwierdzony z wystarczającą pewnością, skłaniający bardziej do uznania kwestii za rozstrzygniętą i zamknięcia badań. Trudno byłoby teraz zmieniać ugruntowaną przez dziesięciolecia tradycję językową, warto jednak podejmować działania edukacyjne, upowszechniające prawidłowe rozumienie istotności statystycznej.

Początkujący badacze i badaczki myślą czasem, że konstatacja istotności statystycznej waliduje zaobserwowany efekt i upoważnia do uznania, iż jego wielkość w populacji jest zbliżona do wartości w próbie. Można ich winić, że nie uważali na wykładzie ze statystyki, ale trzeba też przyznać, że taka życzeniowa interpretacja wyniku odpowiada zrozumiałemu psychologicznie oczekiwaniu: *chcielibyśmy* dysponować takim narzędziem statystycznej walidacji wyniku, które określałoby poziom ryzyka, że efekt w populacji odbiega od wartości zaobserwowanej w próbie o więcej niż pewien rozsądny margines błędu. Może dlatego część użytkowników testów zapomina, że prawdopodobieństwo p nie dotyczy interesującej ich populacji, ale próby, o którą – zdawałoby się – nie mają przecież powodu pytać, skoro właśnie ją zbadali. W ramach klasycznej statystyki frekwencyjnej nie da się jednak określać prawdopodobieństwa prawdziwości hipotez dotyczących populacji. Dlatego o istotności statystycznej z konieczności trzeba decydować na podstawie prawdopodobieństwa zaobserwowania określonych danych w przypadku prawdziwości hipotezy zerowej. To mało intuicyjne rozwiązanie sprzyja nieporozumieniom i życzeniowej interpretacji istotności. Ta zaś prowadzi do nieuprawnionych wniosków i zmniejsza skuteczność apeli o szersze stosowanie wskaźników wielkości efektu. Ktoś, kto fałszywie rozumie istotność statystyczną jako potwierdzenie, że efekt w populacji nie odbiega znacząco od zaobserwowanego w próbie, nie ma potrzeby dodatkowego sprawdzania wielkości efektu w populacji, bo myśli błędnie, że ją już zna.

Część problemów z interpretacją testów istotności można przypisać rozumiejącym je niewłaściwie użytkownikom, są jednak i takie, które wynikają z ograniczeń samej metody, a ściślej jej dyskusyjnej logiki (Westover i in., 2011). Interesujące badacza prawdopodobieństwo można oznaczyć P(H | D). Jest to warunkowe prawdopodobieństwo prawdziwości hipotezy H na temat populacji, w świetle zaobserwowanych w próbie danych D. Ponieważ dotyczy ono subiektywnej pewności w kwestii zachodzenia pewnego stanu rzeczy, a nie spodziewanej częstości występowania tego stanu w długiej serii powtórzeń, nie ma ono sensu w kategoriach tradycyjnej statystyki frekwencyjnej. Można je sensownie określić tylko w ramach statystyki bayesowskiej, popularnej dziś, ale przez Fishera uważanej za błędną. Dlatego w teście istotności zamiast P(H | D) obliczamy P(D | H), prawdopodobieństwo zaobserwowania w próbie danych D, w przypadku prawdziwości hipotezy H. Logika testu istotności opiera się na założeniu, że jeśli prawdopodobieństwo P(D | H), czyli *p*, jest małe, to także hipoteza zerowa jest mało prawdopodobna, czyli małe jest też P(H | D). To założenie na ogół nieźle się sprawdza, ale trzeba pamiętać, iż prawdopodobieństwa P(H | D) i P(D | H), choć powiązane, nie są tożsame. W skrajnym przypadku przy dużej różnicy między P(H | D) a P(D | H) test istotności może być nieakceptowalnie liberalny. Szczególnie ostrożnie należy podchodzić do wyników testu istotności w tych przypadkach, w których hipoteza zerowa jest wysoce prawdopodobna *a priori* (Wolski, 2016).

Część krytyków proponuje zastąpienie testów istotności przedziałami ufności (Cumming, 2014), część metodami bayesowskimi (Wagenmakers i in., 2018; Westover i in., 2011), radykałowie wręcz zakazują wnioskowania statystycznego (Woolston, 2015). Najwięcej zwolenników ma jednak chyba – mniej medialna, ale dla wielu bardziej przekonująca – opcja umiarkowana, wzywająca do bardziej wyważonych działań naprawczych – lepszego rozumienia tradycyjnych metod, uzupełnienia testów istotności szacowaniem wielkości efektów, staranniejszego projektowania eksperymentów i większej dbałości o ich replikowalność (Wasserstein, 2015; Wilkinson i in., 1999). Niezależnie od stanowiska wszyscy zgadzają się jednak, że osoby prowadzące badania naukowe powinny podejmować decyzje badawcze oraz interpretacyjne samodzielnie, nie cedując ich na żaden zrytualizowany schemat.

Niewłaściwe praktyki badawcze

Powyżej omówiono problemy towarzyszące wnioskowaniu statystycznemu oraz to, jak poprawnie stosować i interpretować wyniki testowania hipotez. Jednocześnie wiele z problemów współczesnej psychologii wynika ze stosowania metody testowania hipotez poza kontekstem, do którego została przeznaczona, czyli poza badaniami konfirmacyjnymi. Ci sami badacze, którzy dowodnie ukazali niemożność zreplikowania dużej części rezultatów badań psychologicznych (Open Science Collaboration, 2015), zwracają jednocześnie uwagę na konieczność wyraźnego oddzielenia generowania hipotez od ich testowania, kontekstu odkrycia od kontekstu uzasadnienia, badań eksploracyjnych od badań konfirmacyjnych czy w końcu predykcji od postdykcji (Nosek i in., 2018). Mieszanie tych kontekstów i prezentowanie badań eksploracyjnych jako konfirmacyjnych czy też postdykcji jako predykcji może, co prawda,

podnieść atrakcyjność zgromadzonych wyników i w konsekwencji przyczynić się do ich opublikowania, ale jednocześnie podnosi ryzyko popełnienia błędu I rodzaju i jest jedną z przyczyn niereplikowalności wyników badań w psychologii. Poszukiwanie w zgromadzonym zbiorze danych jakiejkolwiek statystycznie istotnej zależności jest formą manipulacji wynikami, określaną jako *p-hacking*, natomiast postępowanie polegające na prezentowaniu w raporcie z badań hipotez *post hoc* jakby były hipotezami *a priori* nazwano HARKing (*Hypothesizing After the Results are Known*; Kerr, 1998). Ze względu na to, że „hakowanie p" oraz formułowanie hipotez po poznaniu wyników są uznawane za najbardziej szkodliwe z „dyskusyjnych praktyk badawczych" (*Questionable Research Practices*, QRPs[4]), warto omówić je nieco szerzej.

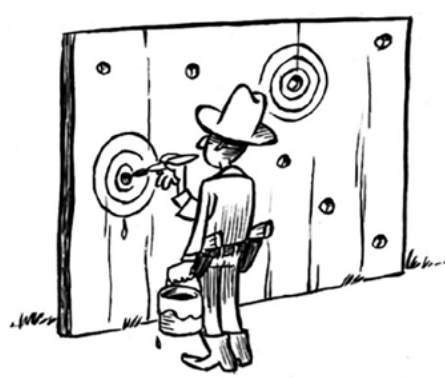
Jako pierwszy na problem mieszania generowania hipotez z ich testowaniem zwrócił uwagę De Groot (1956). Jednak ze względu na to, że jego publikacja ukazała się po holendersku, przez długi czas pozostała niezauważona (jej przekładu w 2014 roku dokonali Wagenmakers i in.). De Groot stwierdził, że choć w psychologii i szerzej w nauce oczywiście istnieje potrzeba badań eksploracyjnych, to używanie tego samego zbioru danych do zarówno generowanie hipotez jak i ich testowania jest niewłaściwą praktyką badawczą i prowadzi do wzrostu prawdopodobieństwa popełnienia błędu I rodzaju. Podkreślił on stanowczo, że przyjmowany w procedurze testowania istotności hipotezy zerowej poziom istotności alfa odnosi się do pojedynczej hipotezy. Przyjmując ten konwencjonalny próg – w psychologii najczęściej jest to 0,05 – badacze dopuszczają możliwość popełnienia pomyłki 1 na 20 razy, tj. dopuszczają, że testując hipotezę zerową (np. o zerowym związku między lękiem społecznym a osiągnięciami szkolnymi), zdecydują się ją błędnie odrzucić i uznać, że ten związek istnieje, choć w rzeczywistości go nie ma. W badaniach psychologicznych akceptuje się takie ryzyko popełnienia błędu, natomiast już w badaniach medycznych mogłoby ono okazać się zbyt duże, stąd przyjmowany tam próg istotności jest ustalony na poziomie 0,01 lub 0,005 (czyli dopuszczamy możliwość pomyłki 1 na 100 razy lub 1 na 200 razy). Sprawa wygląda jednak inaczej, gdy po zgromadzeniu danych zaczynamy w nich szukać wyniku istotnego statystycznie, czyli gdy postępujemy zgodnie z zasadą „torturowania danych": tak długo je „męczymy", aż znajdziemy jakąś, czasem nawet jakąkolwiek, istotną statystycznie zależność. W takiej sytuacji prawdopodobieństwo popełnienia błędu I rodzaju i bezzasadnego odrzucenia prawdziwej hipotezy zerowej znacząco wzrasta. Omówmy ten proces, posługując się przykładem Dorothy Bishop (2019; 2021). Przyjmijmy wspólnie z nią, że mamy duży zbiór danych i szukamy związku między ręcznością a ADHD. Po przeprowadzeniu analizy okazało się, że zależność pomiędzy tymi zmiennymi jest nieistotna statystycznie. Niezniechęceni decydujemy się podzielić zgromadzone dane ze względu na wiek osób badanych i poszukać owej zależności u młodszych i starszych dzieci. Po takim podziale wciąż żadna z relacji nie okazała się istotna statystycznie, więc w kolejnym kroku, ze względu na to, że w badaniu dokonaliśmy pomiaru zarówno umiejętności, jak i preferencji poszukujemy związku między ręcznością a ADHD w tak wydzielonych podgrupach i dla

[4] Omówienie innych dyskusyjnych praktyk badawczych znajduje się w publikacji Andrade (2021).

różnych metod pomiaru. Nadal nic! Zatrzymajmy się na chwilę w naszym kroczeniu po „ogrodzie o rozwidlających się ścieżkach" (Gelman i Loken, 2014) i rozważmy, czy prawdopodobieństwo popełnienia błędu I rodzaju nadal wynosi 0,05. Niestety, odpowiedź brzmi: „nie". Na tym etapie „męczenia danych", gdy poszukujemy zależności w czterech wydzielonych podgrupach i „zadowolimy się" znalezieniem zależności statystycznie istotnej w którejkolwiek z nich, wynosi ono 0,19, zgodnie ze wzorem $(1-(1-0{,}05)^4)$. Kontynuujemy jednak nasze poszukiwania i decydujemy się podzielić dotychczasowe zbiory ze względu na płeć i wyodrębnić w nich grupę dziewczynek i grupę chłopców. Wciąż brak wyniku istotnego statystycznie. Zdaliśmy sobie jednak sprawę, że nasi badani różnili się ze względu na miejsce zamieszkania, więc dzielimy ich na grupę z terenów miejskich i grupę z terenów wiejskich. I jest! Znaleźliśmy zależność, która jest istotna statystycznie! W grupie młodszych dziewczynek mieszkających w miastach, gdy weźmiemy pod uwagę pomiar umiejętności (a nie preferencji), związek pomiędzy ręcznością a ADHD jest statystycznie istotny. Jak zauważa Bishop, pozostaje nam teraz jedynie wymyślić uzasadnienie dla tego związku i przekonująco opisać w artykule. Jeśli dodatkowo przedstawimy go jako hipotezę i stwierdzimy, iż na etapie planowania badania oczekiwaliśmy, że ów związek będzie występował jedynie w grupie młodszych dziewczynek z dużych miast itd., będzie to HARKing. Oczywiście, taka prezentacja wyników nie jest jedynie opowiadaniem nieprawdziwej historii. Ważniejszym problemem jest to, że w toku naszego „męczenia danych" i poszukiwania „istotnego p" (*p-hacking*) w szesnastu z wyłonionych grup poziom alfa nie wynosi już przyjętego na wstępnie 0,05, ale 0,56 $((1-(1-0{,}05)^{16})$. Tym samym dopuszczamy możliwość pomylenia się nie w co 20 badaniach, ale w co 2 przypadku. Zatem prawdopodobieństwo „odkrycia" nieistniejącego związku jest bardzo duże i bliskie sytuacji, gdy rzucamy monetą, by stwierdzić, że dany wynik jest istotny statystycznie, choć w rzeczywistości on nie istnieje. Ten przerysowany przykład ukazuje realny problem generowania hipotez oraz ich testowania na tym samym zbiorze danych w sposób niewłaściwy.

Rysunek 1

Komiksowe zilustrowanie HARKingu (Dirk-Jan Hoek, CC-BY)[*]



[*] Badacz uprawiający HARKing jest niczym rewolwerowiec rysujący tarczę po oddaniu strzału, a nie przed nim.

Jak można zaradzić takim niewłaściwym praktykom? Pierwszym krokiem jest jasne odróżnienie badań eksploracyjnych od badań konfirmacyjnych i testowanie hipotez tylko w tych drugich (Wagenmakers i in., 2012). Jednak badacze, podobnie jak inni ludzie, podatni są na zniekształcenia poznawcze, w tym efekt „od zawsze to wiedziałem" (*hindsight bias*). W konsekwencji badacze mogą być przekonani, że „w sumie to spodziewaliśmy się tej zależności", bo „przecież jest ona sensowna". W celu uniknięcia i, de facto, ochronienia się przed takimi zniekształceniami konieczne jest – i to jest drugi krok przeciwdziałania powyżej opisanej sytuacji – wprowadzenie i przestrzeganie odpowiednich procedur polegających na prerejestrowaniu badań. W takim przypadku, zanim zbadamy pierwszą osobę, rejestrujemy nasze hipotezy, metodę i plan analizy danych na platformie typu *Open Science Framework* (https://osf.io/). Taka prywatna bądź publiczna rejestracja zostaje opatrzona datą (*timestamp*) i może nam służyć w wykazaniu, że nasze badania mają charakter konfirmacyjny i że wyboru analiz nie dokonaliśmy po przyjrzeniu się zgromadzonym danym. Możemy iść jeszcze krok dalej i zgłosić plan naszych badań do czasopisma, które akceptuje typ publikacji „zarejestrowany raport" (*registered report*)[5]. Szczegółowy opis pytania badawczego, hipotez, metod i planu analizy jest poddawany anonimowej recenzji. Jeśli na podstawie recenzji (lub po wprowadzeniu ewentualnych poprawek) zgłoszenie zostanie zaakceptowane, to raport z badań zostanie opublikowany bez względu na to, czy hipotezy uzyskają potwierdzenie i uzyskamy „wyniki istotne statystycznie". Oczywiście, badania muszą zostać przeprowadzone zgodnie z zaakceptowanym protokołem i artykuł poddawany jest ponownej recenzji, ale dotyczy ona głównie wyników oraz dyskusji i nie może kwestionować zaakceptowanych w pierwszym etapie rozstrzygnięć. Taki raport może zawierać dodatkowo, klarownie wyodrębnioną analizę *post hoc*, czyli może obejmować obok części konfirmacyjnej także część eksploracyjną. Zarejestrowany raport jest zatem formą publikacji empirycznych, w której oceniana jest poprawność propozycji badań, a nie istotność statystyczna wyników, co przeciwdziała „tendencyjności publikacyjnej" (*publication bias*), czyli skłonności czasopism i recenzentów do publikowania jedynie wyników istotnych statystycznie. Jednocześnie zarejestrowane raporty przeciwdziałają *p-hacking* i HARKing oraz umożliwiają klarowne oddzielenie zaplanowanych badań konfirmacyjnych od eksploracyjnej analizy danych[6].

Znaczenie teorii w psychologii

Choć postulowane powyżej procedury i standardy mają szansę przeciwdziałać wielu niewłaściwym praktykom badawczym (niska moc testu, *p-hacking*, HARKing) oraz instytucjonalnym (np. publikowanie jedynie wyników istotnych

[5] *Przegląd Psychologiczny* z początkiem 2024 r. wprowadzi możliwość zgłaszania zarejestrowanych raportów.

[6] Szersze i praktyczne omówienie dobrych praktyk badawczych można znaleźć w materiałach kursu *Best practices in statistical design and reporting* (Heyard, 2022).

statystycznie), i w konsekwencji uczynić wyniki badań psychologicznych reprodukowalnymi (tj. otrzymujemy takie same wyniki w ponownej analizie określonych danych) i replikowalnymi (tj. otrzymujemy takie same wyniki, przeprowadzając badanie z inną grupą osób), to są one współcześnie uznawane za niewystarczające. Nie usuwają one bowiem bardziej podstawowego niedostatku psychologii – „słabości" i niekumulatywnego charakteru jej teorii. Przestrzeganie zaproponowanych procedur i standardów może skorygować „maszynerię metody hipotetyczno-dedukcyjnej" (Scheel i in., 2021), ale nie podejmuje zagadnienia poprawności procesu powstawania teorii psychologicznych oraz tego, czy na podstawie ich treści można adekwatnie wyjaśniać i przewidywać rzeczywistość. Jest bowiem możliwe, że badacze będą z powodzeniem replikować wyniki badań, które opierają się na błędnej teorii (Szollosi i in., 2019) lub które zebrano za pomocą nietrafnych narzędzi pomiarowych. Zdanie sobie sprawy z takiej ewentualności spowodowało, że od 2019 roku postulaty zreformowania psychologii nie sprowadzają się jedynie do reformy praktyk badawczych, poprawności analiz statystycznych oraz wyciąganych na ich podstawie wniosków, ale obejmują swoim zakresem także reformę sposobów tworzenia teorii w psychologii (van Rooij i Baggio, 2020). Te podejmowane w ostatnich latach starania zmierzają w różnych kierunkach i dotyczą potrzeby doprecyzowania pojęć (Bringmann i in., 2022), sposobów pomiaru zmiennych psychologicznych (Flake i Fried, 2020) czy konceptualizacji np. zaburzeń psychicznych (Fried i in., 2022). Ze względu na ograniczenie objętości niniejszego artykułu nie jest możliwe nawet krótkie omówienie wszystkich z nich, więc ograniczymy się poniżej do dwóch kwestii, które uznajemy za szczególnie ważne, tj. do krótkiego omówienia znaczenia badań deskryptywnych oraz do nieco szerszej rekonstrukcji propozycji dotyczącej tego, jak tworzyć teorię w psychologii.

Choć ostrze krytyki Paula Rozina (2001) jest wymierzone przede wszystkim w psychologię społeczną, to wydaje się, że wykazane przez niego niedostatki dotyczą także innych obszarów „miękkiej" psychologii. W tej tworzonej początkowo wspólnie z Solomonem Aschem publikacji (choroba Ascha uniemożliwiła mu pełniejsze zaangażowanie się w prace nad tekstem) zwrócili uwagę, że psychologowie społeczni bardzo starają się postępować jak „dojrzali badacze", co ich zdaniem oznacza formułowanie hipotez i prowadzanie eksperymentów. Jednak to wyobrażenie jest błędne, gdyż w „dojrzałych naukach", jak np. w biologii, większy nacisk kładzie się na identyfikacje fenomenów (*phenomenon*[7]) i ich opis, a mniejszy na eksperymenty. Fenomen to stabilna, powtarzająca się i ogólna właściwość świata

[7] Angielski termin *phenomenon* przekłada się na język polski jako „zjawisko" lub „fenomen". Ze względu na to, że we współczesnej polszczyźnie „zjawisko" oznacza: 1 «to, co się wydarzyło», 2. «coś niezwykłego lub ktoś zadziwiający, wyjątkowy», 3. «nierealne, piękne widzenie senne lub urojenie», natomiast „fenomen" oznacza: 1. «rzadkie, niezwykłe zjawisko», 2. «osoba wyjątkowa, niezwykle uzdolniona», 3. «zjawisko fizyczne lub psychiczne będące przedmiotem poznania doświadczalnego», 4. «każdy fakt empiryczny będący punktem wyjścia badań naukowych» (SJP), to w niniejszym opracowaniu używamy terminu „fenomen" zgodnie z jego trzecim i czwartym znaczeniem.

(Haig, 2005). Zdaniem Rozina wiele badań prowadzonych w naukach przyrodniczych kierowanych jest „uzasadnioną, przygotowaną ciekawością" (*informed curiosity*), zaczyna się od identyfikacji fenomenu, jego opisu oraz ustalenia zakresu jego występowania. Często nie opierają się one na teorii, ale wynikają z potrzeby uchwycenia jakiegoś fenomenu w świecie, precyzyjnego opisania regularności i dopiero później stworzenia teorii, która ma go wyjaśnić. Wiele z przełomowych osiągnięć nauki, jak teoria ewolucji Darwina czy odkrycie DNA przez Watsona i Cricka, powstało w opisany powyżej sposób – badania nie były kierowane hipotezą lub modelem, ale ciekawością i miały charakter opisowy. Rozin stwierdza, że w uzasadnionej potrzebie stania się bardziej zaawansowaną nauką psychologia społeczna prześlizguje się po kluczowym etapie opisu badanego fenomenu. Postuluje on zatem wykonanie „kroku wstecz" i powrót do obserwacji oraz opisu zachowań społecznych. Podobny punkt wyjścia przyjmują metodolodzy dążący do zreformowania sposobów tworzenia teorii w psychologii, których propozycję omówiono poniżej.

Niemal pół wieku temu Paul E. Meehl (1978) w swojej przenikliwej, ale konstruktywnej krytyce naukowości psychologii uznał, iż „Teorie w «miękkich» obszarach psychologii pozbawione są kumulatywnego charakteru wiedzy naukowej. Zazwyczaj nie są one obalane ani potwierdzane, ale po prostu zanikają, gdy ludzie tracą zainteresowanie" (s. 806). Ta nieoptymalna sytuacja staje się jeszcze gorsza, gdy zdamy sobie sprawę, że w wielu obszarach psychologii lub w ogóle nie formułuje się teorii albo wyraża je jedynie werbalnie, a tym samym najczęściej nieprecyzyjnie (Robinaugh i in., 2021). Teorie psychologiczne są czasem tak nieprecyzyjne, że nie da się uznać, iż są błędne (Scheel, 2022). Uznając, że główną przyczyną tego stanu rzeczy jest to, iż „metodologiczny repertuar" posiadany przez większość psychologów obejmuje jedynie projektowanie badań w celu empirycznego przetestowania hipotez (najczęściej w ramach statystyki frekwencyjnej i procedury testowania istotności hipotezy zerowej), natomiast nie obejmuje poprawnego tworzenia teorii i dążąc do zmiany tego stanu rzeczy, poniżej zrekonstruowano procedurę budowania teorii w psychologii rozwijaną przez Borsbooma i współpracowników (Borsboom i in., 2021; Haslbeck i in., 2022; Van Dongen i in., 2022).

Borsboom i współpracownicy starając się ulepszyć proces tworzenia teorii w psychologii, proponują sekwencję praktycznych kroków pomocnych we właściwym konstruowaniu teorii. Omówienie owych kroków należy poprzedzić wyjaśnieniem używanych przez nich pojęć. Formułowane w psychologii teorie służą wyjaśnianiu fenomenów, których nie należy utożsamiać z danymi. Dane dostarczają świadectwa istnienia fenomenu, ale nie są z nim tożsame, gdyż są zawsze partykularne, tj. zgromadzone w danym miejscu i czasie; są efemeryczne i idiosynkratyczne (Haig, 2005). Natomiast relacje czy wzorce statystyczne zidentyfikowane w zgromadzonych danych wykraczają poza partykularność określonego zbioru danych i powinny wyłaniać się także w innych zbiorach. I to właśnie relacje statystyczne – identyfikowane w różnych zbiorach danych – reprezentują fenomeny. Przykładowo, dodatnia korelacja pomiędzy wynikami w skalach depresji i lęku jest identyfikowaną w różnych zbiorach danych relacją statystyczną i to ona reprezentuje stabilną i ogólną właściwość świata, czyli fenomen. Zatem

w metateorii Borsbooma i współpracowników fenomen jest utożsamiany z empirycznym uogólnieniem[8].

Ze względu na immanentną nieprecyzyjność teorii werbalnych badacze powinni starać się tworzyć teorie formalne. Borsboom i współpracownicy (Borsboom i in., 2021; Haslbeck i in., 2022; Van Dongen i in., 2022) proponują, aby proces konstruowania teorii podzielić na kroki[9]. W pierwszym kroku konieczne jest zidentyfikowanie fenomenu. Może nim być bądź empiryczne uogólnienie, np. niektórzy ludzie doświadczają napadów paniki oraz martwią się, że doświadczą ich w przyszłości, bądź zdolność (van Rooij i Baggio, 2021), np. do posługiwania się gestami wskazującymi, bądź zachowanie społeczne lub interakcja społeczna, np. droczenie się. W kolejnym kroku należy sformułować prototeorię. Ma ona charakter werbalnego i sformułowanego poprzez abdukcję wyjaśnienia danego fenomenu. Abdukcja[10] jest jednym z trybów (*mode*) – obok indukcji, bayesianizmu i metody hipotetyczno-dedukcyjnej – formułowania wyjaśnień w nauce (Fidler i in., 2018). Postulat posługiwania się nią w nauce sięga do prac Charlesa S. Peirce'a, który uznał, że „abdukcja polega na badaniu faktów i opracowaniu teorii wyjaśniającej je" („abduction consists in studying the facts and devising a theory to explain them") (za: Haig, 2005). Jeśli zatem w celu wyjaśnienia fenomenu badacz formułuje hipotezę lub prototeorię i uznaje, że jest ona warta dalszego dociekania, gdyż dostarcza lepszego wyjaśnienia fenomenu niż alternatywne hipotezy, to posługuje się abdukcją. Stąd abdukcja jest często, choć nie zawsze, utożsamiana z wnioskowaniem do najlepszego wyjaśnienia (*inference to the best explanation*; Haig, 2005). W kolejnym kroku tworzenia teorii sformułowane poprzez abdukcję wyjaśnienie fenomenu (werbalna prototeoria) powinno zostać sformalizowane i ujęte w postaci modelu formalnego. Model ów może zostać wyrażony na co najmniej dwa sposoby: w postaci równań matematycznych lub symulacji agensowych[11] (Borsboom i in., 2021). W tym pierwszym wypadku i korzystając z równań

[8] Nie jest to jedyne możliwe ujęcie, gdyż van Rooij i Baggio (2021) za fenomeny, których wyjaśnianiem powinni zajmować się psychologowie, uznają zdolności, np. zdolność do nabywania języka. Ze względu na to, że propozycja Borsboom i współpracowników wydaje się bardziej przystępna i potencjalnie bardziej użyteczna dla szerszego grona Czytelników, ograniczono się do przyjętych w jej ramach rozstrzygnięć.

[9] Niniejsze kroki nie ograniczają się do tych zawartych w najważniejszej publikacji Borsbooma i współpracowników (2021), ale zostały poszerzone i zmodyfikowane o rozstrzygnięcia zawarte w innych publikacjach członków jego zespołu (Haslbeck i in., 2022; Van Dongen i in., 2022).

[10] Choć termin ten może brzmieć dla części Czytelników obco, to opisywany przez niego tryb wnioskowania jest powszechny. Jeśli bowiem lekarz na podstawie obrazu klinicznego pacjenta uzna, że ma on zapalenie gardła i przepisze mu odpowiednie lekarstwa lub gdy ktoś dostrzegając osobę, która biegnie, rozglądając się i trzymając w ręku smycz, pomyśli, że uciekł jej pies i zapyta ją, czy nie potrzebuje pomocy, to oba te wnioskowania mają charakter abdukcji.

[11] Ze względu na to, że we współczesnym języku polskim termin „agent", na który zazwyczaj przekłada się anglojęzyczne *agent*, oznacza przedstawiciela firmy lub osoby lub pracownika wywiadu (SJP), proponujemy posługiwać się terminem „agens", oznaczającym wykonawcę czynności wskazanej czasownikiem. Przyjęcie takiej konwencji translatorskiej

różniczkowych, staramy się ująć najważniejsze komponenty fenomenu i relacje między nimi. Natomiast w tym drugim podejściu określamy właściwości agensów oraz reguły kierujące interakcjami pomiędzy nimi i otoczeniem, a następnie symulujemy proces ich rozwoju wykorzystując *Agent-based modelling* (ABM, Smaldino i in., 2015). Taki model, będący formalizacją prototeorii wyjaśniającej fenomen, ma charakter „modelu uproszczonego" (*toy model*), tj. reprezentuje on jedynie najważniejsze i wybrane właściwości fenomenu oraz części rzeczywistego świata, dającej mu początek (Beer, 2020; Haslbeck i in., 2022). Jest on „narzędziem myślenia" (*thinking tool*) – pozwala nam dociekać konsekwencji teoretycznych wyjaśnień (Borsboom i in., 2021). Taki model nie jest modelem danych czy modelem statystycznym, ale jest modelem teoretycznym. W odróżnieniu od sformułowanych jedynie werbalnie wyjaśnień wyjaśnienia ujęte w równaniach matematycznych lub kodach programistycznych są precyzyjne, umożliwiają ścisłe wydedukowanie przebiegu rozwoju systemu (Haslbeck i in., 2022) oraz są transparentne, co znacząco ułatwia komunikację między naukowcami i w konsekwencji kumulatywny przyrost wiedzy.

Podstawowej weryfikacji wartości eksplanacyjnej sformalizowanej w modelu prototeorii dostarcza fakt wywiedzenia z równań lub wyłonienia w symulacji badanego fenomenu (Borsboom i in., 2021). Natomiast bardziej szczegółowa weryfikacja poprawności modelu formalnego polega na wywiedzeniu z niego lub zasymulowanie danych (*theory-implied dataset*). Następnie dane te poddawane są takiej samej analizie statystycznej, jaką wykorzystano w analizie danych empirycznych. Ostatecznie porównuje się wyniki tych dwóch analiz (Haslbeck i in., 2022; Van Dongen i in., 2022). Zatem w tym kolejnym kroku tworzenia teorii oceniamy, w jakim stopniu model formalny dostarcza danych, których analiza daje wyniki podobne (podobny wzorzec relacji statystycznych) do uzyskanych w analizie danych empirycznych. Jeśli takie podobieństwo zostanie osiągnięte, to możemy uznać, że sformułowana teoria wyjaśnia fenomen (Van Dongen i in., 2022). Natomiast w wypadku rozbieżności pomiędzy wynikami tych dwóch analiz, staramy się wyjaśnić – ponownie poprzez abdukcję – ich przyczyny oraz odpowiednio zmodyfikować model formalny. Zatem opisywany tu proces konstruowania teorii jest iteracyjnym procesem jej poprawiania (Beer, 2020; Haslbeck i in., 2022). Jeśli ostatecznie osiągniemy oczekiwane podobieństwo[12], to w kolejnym kroku oceniamy wartość teorii ze względu na – przykładowo – sformułowane przez Kuhna właściwości dobrej teorii naukowej, tj. jej dokładność, spójność, zakres, prostotę i owocność (Borsboom i in., 2021), lub ze względu na eksplanacyjną dobroć teorii, tj. jej precyzję, solidność (*robustness*) i empiryczną relewancję (Van Dongen i in., 2022). Ostatecznie, w finalnym kroku konstruowania teorii wskazana jest ocena jej wartości predykcyjnej. Posługując się metodą hipotetyczno-dedukcyjną, wyprowadzamy z niej śmiałe predykcje, które narażają skonstruowaną teorię na możliwość odrzucenia (Borsboom i in., 2021; Haslbeck i in., 2022). Jest to

pozostaje także w zgodzie z łacińskim źródłosłowem angielskiego i polskiego terminu „agent", którym był „agens" – imiesłów czynny od „agere" «robić, czynić, wykonywać».

[12] Za niedostatek omawianej propozycji należy uznać fakt, że jej autorzy nie dookreślają, jaki stopień podobieństwa lub rozbieżności jest oczekiwany i dopuszczalny.

postępowanie ściśle konfirmacyjne, które powinno obejmować prerejestracje oraz symulacje danych i analiz związanych z określonym przewidywaniem (Haslbeck i in., 2022; Wagenmakers i in., 2012). Jeśli teoria przejdzie i ten test, a na jej podstawie możliwe będzie poprawne przewidzenie danych empirycznych, to można uznać ją za potwierdzoną i pomocną nie tylko w wyjaśnianiu, lecz także przywidywaniu i kontrolowaniu fenomenów psychologicznych (Haslbeck i in., 2022).

Zreferowany powyżej proces tworzenia teorii stawia przed badaczami duże wyzwania. Prowadzenie badań zgodnie z tymi wytycznymi wymaga większego nakładu czasu i pracy niż w tradycyjnym modelu hipotetyczno-dedukcyjnym. Wydaje się jednak, że takie postępowanie ma szansę dostarczyć psychologii bardziej solidnych podstaw, uczynić jej wyniki replikowalnymi, a nią samą – jak postulował Meehl (1978) – dziedziną wiedzy o kumulatywnym charakterze.

Tak jak stwierdzono na początku niniejszego artykułu, warto zaznaczyć, że wiele z omówionych powyżej niedostatków badań psychologicznych jest dostrzeganych od dawna[13]. Współcześnie coraz powszechniej akceptuje się konieczność bardziej przemyślanego planowania i przeprowadzania badań, a także rozważnego analizowania oraz interpretowania zgromadzonych danych i wyników. Mamy nadzieję, że również powyższe rozważania przyczynią się do intensyfikacji tych procesów.

Bibliografia

Andrade, C. (2021). HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *The Journal of Clinical Psychiatry*, *82*(1), 20f13804. https://doi.org/10.4088/JCP.20f13804

Bakker, M., van Dijk, A., Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., van der Maas, H. L. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Beer, R. D. (2020). Lost in words. *Adaptive Behavior*, *28*(1), 19–21. https://doi.org/10.1177/1059712319867907

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435. https://doi.org/10.1038/d41586-019-01307-2

Bishop, D. (2021). UBL & Elsevier seminars on Reproducible Research. YouTube https://www.youtube.com/watch?v=C-rk22as870&t=214s

[13] Warto przywołać także artykuł J. Cohena (1990), opublikowany po polsku w 2006 r. w tomie *Metodologiczne i statystyczne problemy psychologii* pod red. J. Brzezińskiego i J. Siuty, w przekł. R. Polczyka. Autor omawiając właściwości testowania hipotez, problem wielokrotnych porównań, moc testu czy wielkość efektu, dodaje jednocześnie, że o niektórych z tych spraw pisze od 20 czy nawet 30 lat, czyli od lat 60. XX w.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Boyce, V., Mathur, M. B., Frank, M. C. (2023, July 31). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. https://doi.org/10.31234/osf.io/dpyn6

Bringmann, L. F., Elmer, T., Eronen, M. I. (2022). Back to Basics: The Importance of Conceptual Clarification in Psychological Science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal Social Psychology*, *65*, 145–153.

Cohen, J. (1990/2006). Things I have learned (so far). *American Psychologist, 45*, 1304––1312. https://doi.org/10.1037/0003-066X.45.12.1304 (tłum. Cohen, J. (2006). O tym, czego się nauczyłem (jak dotąd). W: J. Brzeziński i J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (s. 75–99). Wydawnictwo Zysk i S-ka.

Cohen, J. (1994/2006). The Earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997 (tłum. Cohen, J. (2006). Ziemia jest okrągła ($p < 0{,}05$). W: J. Brzeziński i J. Siuta (red.), *Metodologiczne i statystyczne problemy psychologii* (s. 100–118). Wydawnictwo Zysk i S-ka.

Cumming, G. (2008). Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, *3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, *25*(1), 7–29.

de Groot, A. D. (1956/2014). The meaning of "significance" for different types of research (tłum. i adnotacje E.-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, H. L. J. van der Maas), *Acta Psychologica*, *148*, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Eronen, M. I., Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://link.springer.com/article/10.3758/BF03193146

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Fisher, R. A. (1971). *The design of experiments* (wyd. 9). Hafner Press.

Flake, J., Fried, E. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fried, E., Flake, J., Robinaugh, D. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*, 358–368. https://doi.org/10.1038/s44159-022-00050-2

Gelman, A., Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Haig, B. D. (2005). An Abductive Theory of Scientific Method. *Psychological Methods*, *10*(4), 371–388. https://doi.org/10.1037/1082-989X.10.4.371

Haller, H., Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online*, *7*(1).

Halsey, L. G., Curran-Everett, D., Vowler, S. L., Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), 179–185. https://doi.org/10.1038/nmeth.3288

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*(6), 930–957. https://doi.org/10.1037/met0000303

Heyard, R. (2022). *Best practices in statistical design and reporting*. University of Zurich. https://osf.io/t9rqm/

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Lakens, D. (2023, July 24). Concerns about Replicability, Theorizing, Applicability, Generalizability, and Methodology across Two Crises in Social Psychology. https://doi.org/10.31234/osf.io/dtvs7

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Mellor, D. T. (2018). The preregistration revolution. *PNAS* (*Proceedings of the National Academy of Sciences of the United States of America*), *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oberauer, K., Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years. *Journal of Consulting and Clinical Psychology*, *58*(5), 646.

Rozin, P. (2001). Social Psychology and Science: Some Lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Scheel, A. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1): e2295. https://doi.org/10.1002/icd.2295

Scheel, A. M., Tiokhin, L., Isager, P. M., Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744––755. https://doi.org/10.1177/1745691620966795

Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309.

SJP. (b.d.). *Słownik języka polskiego PWN online* (https://sjp.pwn.pl/).

Smaldino, P. E., Calanchini, J., Pickett, C. L. (2015). Theory development with agent-based models. *Organizational Psychology Review*, *5*(4), 300–317. https://doi.org/10.1177/2041386614546944

Szollosi, A., Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

van Dongen, N. N. N., van Bork, R., Finnemann, A., van der Maas, H., Robinaugh, D., Haslbeck, J. M. B., […] Borsboom, D. (2022, April 13). Productive Explanation: A Framework for Evaluating Explanations in Psychological Science. https://doi.org/10.31234/osf.io/qd69g

van Rooij, I., Baggio, G. (2020). Theory development requires an epistemological sea change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840X.2020.1853477

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J. Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wasserstein, R. (2015). ASA comment on a journal's ban on null hypothesis statistical testing. Retrieved 05 Aug 2015, Sente.

Westover, M. B., Westover, K. D., Bianchi, M. T. (2011). Significance testing as perverse probabilistic reasoning. *BMC Medicine*, *9*, 20. https://doi.org/10.1186/1741-7015-9-20

Wilkinson, L., APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

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

Woolston, C. (2015). Psychology journal bans P values. *Nature*, *519*(7541), 9.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, E1. https://doi.org/10.1017/S0140525X20001685

Podziękowania

Arkadiusz Białek dziękuje Róży Krycińskiej i Monice Szczygieł za uwagi dotyczące wcześniejszej wersji artykułu.