NAUKA 1/2017 • 143–156

KAROLINA KARPE *

O pogoni za wynikiem istotnym statystycznie. Konsekwencje rozpowszechnienia testowania istotności hipotezy zerowej w psychologii

Wprowadzenie

Sensem badań psychologicznych jest zdobycie wiedzy na temat ludzkiego zachowania. Aby móc wyprowadzić bardziej ogólne wnioski z obserwacji zebranych w czasie badań, konieczne jest prowadzenie wnioskowania statystycznego. Zasób obecnie wykorzystywanych w psychologii narzędzi statystycznych okrzepł. Analizy statystyczne przeprowadzane w nurcie testowania hipotez (NHST – *null hypothesis significance testing*) dominują w rachunkowej stronie opisu wyników badań psychologicznych. Przynajmniej pod względem ilościowego udziału w „rynku" statystyki w psychologii, alternatywne metody mają marginalne znaczenie. Wartości *p*, testy istotności różnic, korelacje – bez tych pojęć trudno wyobrazić sobie empiryczne (a zwłaszcza eksperymentalne) artykuły psychologiczne. NHST jest hybrydą dwóch podejść do testowania hipotez, które na początku dwudziestego wieku zostały zaproponowane przez Fishera oraz Neymana i Pearsona (Gigerenzer, 1993). Po wyczerpujący – historycznie i matematycznie – opis tych koncepcji, jak również różnic między nimi odsyłam do Jarmakowska-Kostrzanowska (2016). Wpływ NHST na psychologię sięga poza same techniki obliczeń: stanowi ona ramę projektową i interpretacyjną dla badań oraz ich wyników. Przede wszystkim jednak NHST przyczyniła się do ukształtowania rozpowszechnionych koncepcji wiążących naukowy sukces i porażkę z istotnością statystyczną otrzymanych wyników. Mają one znaczący wpływ na zachowania wszystkich uczestników „gry w naukę": badaczy, wydawców czasopism, recenzentów i pracowników administracji naukowej. Silne zakorzenienie tych koncepcji, w połączeniu z powszechnym niezrozumieniem ograniczeń i założeń tego nurtu analiz statystycznych mogą być widziane jako jedna z przyczyn obecnego kryzysu w psychologii. Objawia się on w nadmiarze wyników fałszywie pozytywnych w literaturze, trudnościach w replikacji opublikowanych wyników oraz coraz częstszych wykrywanych przypadkach występowania niewłaściwych praktyk badawczych. W niniejszym artykule omówione zostaną dwa najbardziej jaskrawe problemy

* Instytut Psychologii, Uniwersytet im. Adama Mickiewicza w Poznaniu, e-mail: kkarpe@amu.edu.pl

związane z niewłaściwym wykorzystywaniem analiz w nurcie NHST w psychologii: nadmierne przywiązanie do wyników istotnych statystycznie (i wartości $p < 0{,}05$) oraz niedoszacowanie mocy testów statystycznych wykorzystywanych w prowadzonych badaniach. Opisane zostaną również niewłaściwe praktyki badawcze, które mogą pojawiać się jako reakcja na te problemy.

Popularne błędy w stosowaniu NHST w psychologii

Ze względu na swoje znaczące negatywne konsekwencje, błędy powiązane ze stosowaniem NHST są tematem cieszącym się dużym zainteresowaniem metodologów i statystyków. Już prawie trzydzieści lat temu zauważano problemy metodologiczne takie jak zbytnie poleganie na dychotomicznych klasyfikacjach wyników (istotne/nieistotne), nieodpowiednia moc testów wykorzystywanych w badaniach czy niedocenianie wagi replikacji dla procesu badawczego (Rosnow i Rosenthal, 1989). Nickerson (2000) wymienia listę błędów popełnianych przez naukowców skupiającą się głównie na niewłaściwym rozumieniu założeń (np. przekonanie, że p to prawdopodobieństwo, że hipoteza zerowa jest prawdziwa, a 1-p to prawdopodobieństwo, że hipoteza alternatywna jest prawdziwa; przekonanie, że małe p jest dowodem na powtarzalność uzyskanych wyników; przekonanie, że poziom alfa to prawdopodobieństwo tego, że odrzucenie hipotezy zerowej okaże się błędem pierwszego rodzaju). Kirk (1996) z kolei wskazuje na trzy główne ograniczenia NHST, których niezrozumienie prowadzi do błędów w przeprowadzaniu i interpretacji analiz: NHST nie odpowiada na zadawane pytania badawcze (wyniki mówią o prawdopodobieństwie danych ze względu na hipotezę, nie o prawdopodobieństwie hipotezy ze względu na dane), średnie z dwóch prób zawsze są różne (przez co wymagana jest analiza mocy i refleksja nad sensownością próby), oraz arbitralność przyjmowanego poziomu alfa.

Te podsumowania pozwalają na zorientowanie się, jak podstawowych zagadnień dotyczy rozdźwięk pomiędzy popularnymi przekonaniami na temat NHST a faktycznym zrozumieniem jej zasad. Jak zauważa Garcia-Perez (2016), problemy, jakie obserwuje się w związku z NHST sprowadzają się do błędnego zrozumienia, błędnego wykorzystania i nierealistycznych oczekiwań względem możliwości metody. Pełna ocena wpływu tych błędów na stan nauki jest obecnie niemożliwa, ze względu na brak odpowiednich danych empirycznych. Biorąc pod uwagę to, że zagadnienia związane z NHST są powszechnie błędnie rozumiane przez naukowców i badaczy na wszystkich etapach kariery (Haller i Krauss, 2002) można jednak spodziewać się, że jest on znaczący. Wyróżnione zostaną – i dokładniej przybliżone – dwie kwestie związane z niepoprawnym wykorzystaniem i zrozumieniem NHST, które w oparciu o dotychczasowe analizy problemu zdają się mieć najbardziej niszczący wpływ na jakość wnioskowania statystycznego w psychologii.

Przywiązanie do wyników istotnych statystycznie i wartości $p < 0,05$

NHST, mimo możliwości popełnienia wielu błędów przy jej wykorzystywaniu, dalej ma nieodparty urok dla psychologów: naucza się jej na każdych studiach psychologicznych, jest prosta do zastosowania dzięki bogactwu programów statystycznych i przede wszystkim dostarcza prostego kryterium decyzyjnego do orzekania o znaczeniu otrzymywanych wyników. Prawdopodobnie najbardziej charakterystycznym dla psychologii uprzedzeniem związanym ze statystyką jest bezkrytyczne podejście do wartości $p = 0,05$ jako punktu demarkacyjnego pomiędzy „sukcesem" i „porażką naukową" (Bakker i Wicherts, 2012). Wartość p jest tym elementem, który w raporcie z badań budzi najwięcej emocji (Rosnow i Rosenthal, 1989) i może wydawać się wręcz, że głównym celem w prowadzeniu badań jest uzyskanie p mniejszego niż 0,05. Nie tylko w psychologii wartości p przypisuje się największą wartość informacyjną w treści artykułu (Goodman, 1999). Tymczasem zerojedynkowy podział wartości prawdopodobieństwa na istotne i nieistotne nie odzwierciedla ich prawdziwego rozkładu i znaczenia. Świadomość, że $p < 0,05$ to wartość wybrana arbitralnie zaciera się. Działa ona na wyobraźnię do tego stopnia, że stopień zaufania do danych stojących za jakąś hipotezą nie zmienia się proporcjonalnie do wartości prawdopodobieństwa, ale przynajmniej u części psychologów spada gwałtownie, kiedy wyniki stają się nieistotne statystycznie (Rosenthal i Gaito, 1964; Beauchamp i May, 1964, Pointevineau i Lecoutre, 2001).

Otrzymanie wyniku istotnego statystycznie, czyli o p mniejszym niż arbitralnie wybrany punkt 0,05, stało się wyznacznikiem odniesienia sukcesu w badaniu naukowym: grając w grę jaką jest psychologia, wygrywa się otrzymując istotne wyniki (Bakker, Wicherts, 2012). Takie wyniki są chętniej opisywane przez badaczy (Franco, Malhotra i Simonovits, 2014) i są oni przekonani, że łatwiej je opublikować (Ferguson i Heene, 2012). W efekcie psychologia i psychiatria są najbardziej „pozytywnymi" dziedzinami nauki: aż 91,5% wszystkich opublikowanych artykułów zawiera wynik istotny statystycznie (Fanelli, 2010). Trudno się dziwić, że w konsekwencji obcowania z taką literaturą, wyżej cenione są wyniki istotne statystycznie. W powszechnym przekonaniu, za istotnymi wynikami idą publikacje, za publikacjami granty, za grantami – dobre stanowiska pracy i prestiż. Cały ten proces może być zaburzony przez wynik o $p = 0,06$, które przecież mówi tylko tyle, że dane są o 1% mniej prawdopodobne ze względu na hipotezę alternatywną, niż byłoby to przy wyniku tuż poniżej granicy istotności (Rosnow i Rosenthal, 1989). Warto analizując tę kwestię pamiętać, że poziom alfa równy 0,05 jest arbitralnie przyjętą konwencją i nie dla każdego badania musi okazać się on odpowiedni.

Niedoszacowanie mocy

Tak ważny dla psychologów poziom istotności statystycznej powiązany jest z możliwością przyjęcia hipotezy alternatywnej, kiedy prawdziwa jest hipoteza zerowa (błąd

pierwszego rodzaju). Drugi błąd, który kładzie się cieniem na statystyce w badaniach psychologicznych dotyczy z kolei prawdopodobieństwa przyjęcia hipotezy zerowej, kiedy prawdziwa jest alternatywna (błąd drugiego rodzaju). Prawdopodobieństwo to informuje o mocy testu, opisując możliwość popełnienia tego błędu przy danej wielkości efektu, wielkości próby i poziomie alfa. Za odpowiedni poziom mocy uznaje się 80% (Cohen, 1992). Obliczenie mocy przed przeprowadzeniem badań wydaje się być rozsądnym działaniem. Przeprowadzenie takiego oszacowania i dobranie odpowiedniej wielkości badanej próby do spodziewanej siły efektu może uratować przed niepotrzebnymi kosztami, nieudanymi replikacjami czy porzuceniem obiecującej linii badań po otrzymaniu (fałszywie) negatywnego wyniku.

Tymczasem psychologowie jak ognia unikają szacowania mocy, w efekcie prowadząc badania o nieodpowiednim poziomie mocy – i to od lat. W latach sześćdziesiątych średnia moc testu w *Journal of Abnormal and Social Psychology* wynosiła 0,48, czyli prawdopodobieństwo wykrycia średniej wielkości efektu w tych badaniach było porównywalne do możliwości wyrzucenia reszki w rzucie monetą (Cohen, 1962). Sądzono, że może to wynikać z niedostatecznej edukacji i słabej dostępności informacji na temat analizy mocy, na co remedium miał być podręcznik „*Statistical power analysis for the behavioral sciences*" z 1969 roku oraz artykuł *Power primer* z 1992, w których Cohen dostarczył przystępnego i praktycznego opisu tego zagadnienia. Niestety, późniejsze analizy nie dostarczyły optymistycznych wniosków: ponowne zbadanie mocy w *Journal of Abnormal and Social Psychology* wykazało, że średnia moc w badaniach wręcz spadła do 0,25, a żaden z artykułów nie zawierał próby oceny mocy statystycznej (Seldmeier i Gigerenzer, 1989). Pomimo doskonałej dostępności informacji na temat oszacowywania mocy i wielokrotnego podkreślania wagi tego zagadnienia, sytuacja do dzisiaj nie uległa poprawie. Większość artykułów psychologicznych dalej nie zawiera informacji na temat analizy mocy (Bakker, van Dijk i Wicherts, 2012). Dodatkowo, intuicje badaczy dotyczące mocy przeprowadzanych badań są nadmiernie optymistyczne (Bakker, Hartgerink, Wicherts, van der Maas, 2016). Wiedza psychologiczna zagrożona jest więc nie tylko nadmiarem wyników fałszywie pozytywnych, ale również fałszywie negatywnych.

Strategie przyjmowane przez naukowców w celu poradzenia sobie z problemem nieistotnych wyników i niskiej mocy

Niedoszacowanie mocy w badaniach psychologicznych skutkuje problemami z wykryciem istniejących efektów. Jednocześnie, występuje znacząca presja na otrzymywanie i publikowanie wyników istotnych statystycznie. Taka sytuacja, w której jednocześnie ciężko jest otrzymać wynik istotny statystycznie i jest on uznawany za główny wyznacznik sukcesu badania może zachęcać do stosowania wybiegów zwiększających

prawdopodobieństwo otrzymania „publikowalnych" rezultatów. Zachowania mające na celu manipulację procesem badawczym nazywane są niewłaściwymi praktykami badawczymi (ang. *questionable research practices*) lub nierzetelnością badawczą (ang. *research misconduct*). Jednoznaczne zdefiniowanie tych praktyk jest trudne, lecz we wszystkich próbach ich opisu można dostrzec, że kluczową cechą jest odejście od dobrych praktyk badawczych (Smith, 2000).

Wśród tych praktyk najwyraźniej w świadomości badaczy istnieją trzy: fabrykowanie, fałszowanie i plagiat (często właśnie z nimi utożsamia się nierzetelność badawczą). Wyniki badań wskazują, że te trzy poważne przewinienia występują bardzo rzadko (Fanelli, 2009), jednak uznawane są za bardzo szkodliwe dla zaufania w środowisku i dla uzyskiwanej wiedzy (poza plagiatem) (Bouter i in., 2016). Wykryte przypadki fałszowania, fabrykowania i plagiatu spotykają się ze znaczącym potępieniem środowiska naukowego (patrz np. sprawa Dederika Stapela – Levelt, Drenth i Noort, 2012) i często prowadzą do wycofywania artykułów z czasopism naukowych oraz zakończenia kariery nieuczciwego badacza. Ich szkodliwość jest oczywista: do nauki wprowadzane są wyniki nieprawdziwe (fabrykowanie, fałszowanie danych) lub kradzione (plagiat); spada również zaufanie do procesu naukowego, zarówno wewnątrz akademii, jak i w oczach opinii publicznej. Te trzy wykroczenia naukowe popełniane są raczej z premedytacją: trudno wyobrazić sobie sfałszowanie danych czy sfabrykowanie ich wynikające z przypadkowego błędu bądź niedostatecznej świadomości metodologicznej.

Fałszowanie, fabrykowanie i plagiat są problemem istotnym, ale rzadko spotykanym. Tymczasem poza nimi występują inne, subtelniejsze, ale bardziej rozpowszechnione zakłócenia procesu badawczego. Jak znaczący jest zakres problemu, pokazują badania: ponad 30% ankietowanych badaczy przyznaje się do angażowania się w zachowania, które mogą być zakwalifikowane jako niewłaściwe praktyki badawcze inne niż fałszowanie, fabrykowanie i plagiat (Fanelli, 2009); ankietowani oceniali rozpowszechnienie tych praktyk wśród swoich współpracowników na 70%. Niewłaściwe praktyki badawcze mogą być wykorzystywane w celu zwiększenia publikowalności swoich rezultatów. Należy pamiętać, że „materiałem wyjściowym" w psychologii są najczęściej mało obiecujące dane. W sytuacji, kiedy opieranie się na poprzednich badaniach jest obciążone dużym ryzykiem polegania na przypadkowo istotnych rezultatach, a moc jest niewystarczająca, trudno o uzyskanie satysfakcjonujących wyników. To sprawia, że motywacja do sięgania do niewłaściwych praktyk badawczych może być bardzo silna: w obecnym systemie oceny pracy naukowej przeprowadzenie badania i nieopublikowanie jego wyników jest luksusem, na który nie każdy może sobie pozwolić. W tym artykule omówione zostaną trzy rozpowszechnione wątpliwe praktyki badawcze, ściśle związane z niewłaściwym wykorzystaniem NHST, które służą do „poprawiania" szans otrzymanych wyników na publikację.

Selektywne publikowanie

Szkodliwe dla rozwoju wiedzy działania mogą pojawiać się na różnych etapach procesu badawczego. Część z nich ma miejsce w czasie przeprowadzania badania i analizowania danych, kolejne – na etapie opisu wyników i publikacji. Ze strony wydawców czasopism naukowych występuje obciążenie publikacyjne (*publication bias*): wywierana jest presja w procesie akceptowania artykułów w czasopismach na istotne wyniki i innowacyjną treść. Na decyzję o publikacji wpływa nie tylko obiektywna wartość badania, ale też kierunek otrzymanych wyników (Chambers i in. 2014). Pożądane są prace przełomowe, inne, pokazujące interesujące zjawiska i przede wszystkim, potwierdzające postawione hipotezy. Mniejsze znaczenie mają prace, w których pomimo rzetelnych metod nie otrzymano istotnych statystycznie wyników czy artykuły dotyczących replikacji, zwłaszcza dokładnych. „Problemem" otrzymania wyników nieistotnych statystycznie często rozwiązywany jest przez nieopublikowanie takich „nieudanych" rezultatów: albo w całości, albo ukrywając te części badania (warunki badawcze, pojedyncze eksperymenty, grupy osób badanych) w których nie uzyskano oczekiwanych wyników. Zjawisko to nazywane jest efektem szuflady (*file drawer effect*). Przez badaczy nad rzetelnością naukową selektywne publikowanie jest uznawane za jeden z największych problemów współczesnej nauki (Bouter i in., 2016). Rosenthal (1979) zwraca uwagę, że powstały w ten sposób obraz wiedzy naukowej z danej dziedziny jest w oczywisty sposób zafałszowany. Być może wskutek tego zakłócenia znacząca część odkryć prezentowanych na łamach czasopism, jest tak naprawdę efektem błędów typu I. Jest to szczególnie prawdopodobne w świetle doniesień o niewystarczającej mocy w badaniach psychologicznych.

Takie praktyki nieuchronnie nasilają efekt szuflady, do tego stopnia, że często badacze nie podejmują nawet prób opisania uzyskanych negatywnych wyników, antycypując odrzucenie artykułu przez czasopisma (Franco, Malhotra i Simonovits, 2014). Niepublikowanie nieistotnych statystycznie wyników oraz selektywne raportowanie rezultatów badań może zostać uznane w szerszej perspektywie za formę fałszowania danych. Można zauważyć podobieństwa pomiędzy tymi zjawiskami a manipulacją danymi przez usuwanie z nich obserwacji, które zakłócają wyniki. W tym przypadku, zamiast pozbywać się na przykład pojedynczych odstających pomiarów, wygładza się uzyskany obraz przez całkowite pomijanie „nieistotnych" elementów badania. Chociaż niepublikowanie negatywnych wyników jest bardzo popularną praktyką, nie wystarcza ono jako jedyne wyjaśnienie występowania zafałszowań w obrazie, jaki wyłania się z analizy badań psychologicznych (Francis, Tanzmann i Matthews, 2014, Francis, 2014). Muszą towarzyszyć mu zatem inne zachowania ze spektrum niewłaściwych praktyk badawczych.

Manipulacje wartością *p* (*p*-hacking)

Pod pojęciem *p*-hackingu kryje się wiele manipulacji, które mają na celu osiągnięcie istotnych statystycznie wyników na podstawie uzyskanych w badaniu danych. Ze względu na to, że prace o pozytywnych wynikach są chętniej publikowane, naturalną strategią zwiększania swoich szans jest manipulowanie przebiegiem analiz tak, aby wyniki przekroczyły $p < 0{,}05$. Do osiągnięcia tego celu nie jest konieczne aktywne fabrykowanie wyników ani fałszowanie danych. Wystarczający może być sam wybór analiz statystycznych czy swoboda w porównywaniu warunków eksperymentalnych. *P*-hacking może pojawiać się w dwóch momentach w czasie trwania procesu badawczego (Chambers i in., 2014). Pierwszy z nich to etap zbierania danych, który może być kontynuowany tak długo, aż przeprowadzane analizy osiągną odpowiedni poziom istotności statystycznej (im większa próba, tym mniejsza musi być wielkość efektu, żeby okazał się on być istotny statystycznie). Porzucana jest w tym momencie pożądana praktyka określania odpowiedniej wielkości próby przez rozpoczęciem badania tak, by badanie dawało maksymalnie wiarygodne wyniki. Drugim momentem, w którym może pojawić się *p*-hacking, jest etap analizowania i opisywania otrzymanych wyników. Badacz może sięgnąć po bardziej liberalne testy statystyczne lub nie stosować poprawek na wielokrotne porównania. Istotne wartości *p* mogą być też uzyskiwane poprzez wyodrębnianie podgrup czy szukanie mediatorów i moderatorów relacji między zmiennymi, nieuzasadnionych przez teorię i nieprzewidywanych wcześniej. Natomiast fragmenty badania, które mimo tego nie „wyszły", mogą zostać pominięte w opisie. Ukrytym fragmentem może stać się na przykład warunek eksperymentalny, który nie sprawił, że badana zmienna zależna istotnie zmieniła wartości, lub grupa badanych, w której wyniki nie odpowiadały założonemu wzorcowi. Jak skuteczne są takie praktyki w uzyskiwaniu istotnych statystycznie wyników, pokazali Simmons, Nelson i Simonsohn (2011). Wykazali oni, że bez odnoszenia się do manipulacji uzyskanymi danymi możliwe jest uzyskanie istotnych statystycznie wyników z właściwie dowolnych danych. W połączeniu z selektywnym raportowaniem, *p*-hacking może dostarczać zupełnie pozbawionych sensu, ale istotnych statystycznie wyników.

Częścią winy za bezrefleksyjne stosowanie NHST i rozpowszechnienie manipulacji obliczeniami obciąża się popularne pakiety statystyczne, które pozwalają na proste i szybkie przeprowadzanie analiz (Brzeziński, 2012). Również kursy statystyki na studiach psychologicznych w dużej mierze skupiają się na praktycznym przeprowadzaniu typowych analiz za pomocą pakietów statystycznych, ze słabym uwzględnieniem nowych metod. Nauka odpowiedzialnego wykorzystania takich narzędzi i zrozumienia teorii statystycznych leżących u podstaw wykonywanych obliczeń to ważne zadanie w procesie kształcenia przyszłych psychologów-naukowców.

Przedstawienie hipotez *post hoc* jako hipotez *a priori* (HARKing)

Kolejne ze zjawisk pojawia się na etapie opisywania wyników – w przypadku, kiedy uzyskane rezultaty są istotne statystycznie (więc „publikowalne"), ale niezgodne ze wcześniejszymi przewidywaniami i początkowym celem badań. *HARKing* (hypothesizing afer the results are known) został zdefiniowany jako „prezentowanie hipotez *post hoc* (na przykład bazujących na lub potwierdzonych przez otrzymane wyniki) w raporcie z badania, jakby rzeczywiście były one hipotezami *a priori*" (Kerr, 1998). Badania poddane procedurze HARKingu to samospełniające się przepowiednie: każda hipoteza postawiona zostaje w toku badania potwierdzona. Kerr (1998) wyróżnia różne rodzaje HARKingu ze względu na prawdopodobieństwo hipotezy przed przeprowadzeniem analiz: czysty HARKing (raportowanie wyłącznie potwierdzonych hipotez), czysty HARKing z losowaniem (dodatkowe raportowanie dla większej wiarygodności kilku niepotwierdzonych hipotez), tłumienie „przegranych" hipotez (unikanie raportowania hipotez wiarygodnych i przewidzianych przed badaniem, ale niepotwierdzonych, w drugiej wersji publikowanie jedynie hipotez przewidywanych i potwierdzonych), raportowanie na podstawie prawdopodobieństwa post hoc oraz empiryczne inspiracje (dodanie do początkowych hipotez również tych sformułowanych już po uzyskaniu wyników). Różne rodzaje HARKingu mają zróżnicowaną szkodliwość.

Chociaż empiryczne badanie częstości występowania wszystkich wątpliwych praktyk badawczych stanowi poważne wyzwanie, HARKing szczególnie trudno wykryć. Nie zostawia on jednoznacznych śladów w ostatecznym tekście opublikowanego artykułu. W naukach społecznych nie ma zwyczaju prowadzenia dokładnych ewidencji przebiegu procesu badawczego, więc szansa, że pierwotne hipotezy zostaną zapisane, jest niewielka. Zmiana statusu prezentowanych hipotez z *a priori* na *post hoc* jest kwestią tylko wprowadzenia zmian w opisie. Kerr (1998) wskazuje kilka symptomów mogących wskazywać na obecność HARKingu: obecność podgrup, w których osiągnięcie danych wyników nie są bardziej prawdopodobne a priori od innych (na przykład, hipoteza mówi o tym, że dany efekt występuje jedynie w grupie mężczyzn powyżej 55 roku życia); „historie zbyt dobre, żeby były prawdziwe" – sytuację, w której na podstawie teorii stawia się hipotezy, które zostają idealnie potwierdzone, podczas gdy równie lub bardziej prawdopodobne a priori wydają się inne wnioski; słabe dopasowanie pomiędzy metodologią badania a testowanymi hipotezami. Jednak obecność takich elementów nie gwarantuje pewności, że wystąpił HARKing – mogą one być też po prostu oznakami nieidealnie zaprojektowanego, przeprowadzonego lub opisanego badania. Jakie są więc dowody na występowanie HARKingu w badaniach naukowych? Przykładowo, można powołać się na wyższy odsetek potwierdzonych hipotez w pracach, w których nie rejestruje się hipotez przed otrzymaniem wyników (opublikowane artykuły) niż w pracach doktorskich (Mazzola i Deuling, 2013). Potwierdzenia dla funkcjonowania zjawiska

HARKingu w psychologii, tym razem na polskim gruncie, dostarcza również sondażowe badanie przeprowadzone przez Zdybka, Walczaka i Zdybek (2012), gdzie niemal połowa studentów psychologii wskazała na występowanie którejś z form HARKingu w badaniach, w których uczestniczyli.

Konsekwencje błędów w NHST i niewłaściwych praktyk badawczych jako reakcji na nie

Wynik fałszywie pozytywny – niezależnie od tego, czy jest efektem celowej manipulacji danymi, czy bardziej „niewinnej" swobody w doborze i przeprowadzeniu analiz ma dokładnie taki sam wpływ na naukę. Każdy z nich jest cegiełką budującą obraz świata nieprzystający do rzeczywistości; na każdym z takich wyników mogą zostać oparte kolejne badania, eksplorujące nieistniejące problemy i przekłamane zależności. Fundusze na badania są marnowane na dociekania oparte na fałszywych wcześniejszych doniesieniach. Rozwój nauki jest spowalniany, a możliwe też, że w niektórych obszarach uniemożliwiany przez trudne do zweryfikowania, fałszywie pozytywne wyniki. Niewykluczone, że całe gałęzie badań są zapędzone w ślepą uliczkę – przykładem mogą być analizy wpływu tzw. póz energetycznych (ang. *power posing*) na samopoczucie i fizjologię.

Opublikowany w prestiżowym czasopiśmie *Psychological Science* artykuł (Carney, Cuddy i Yap, 2010) opisuje eksperymenty badające zależności pomiędzy przyjmowaną pozycją ciała a poziomem testosteronu, kortyzolu, poczuciem siły i skłonnością do podejmowania ryzyka. Autorzy stwierdzili, że przyjmowanie póz kojarzących się z „wysokim poziomem mocy" w kontraście do póz skojarzonych z „niską mocą" (trzymanie nóg na biurku vs. siedzenie z założonymi rękami) korzystnie wpływa na wymienione parametry. Badania te stały się podstawą popularnonaukowej książki oraz wykładu dla serwisu TED, w którym jedna z autorek oryginalnego badania w przystępny sposób zachęca do wykorzystywania póz energetycznych do „zmiany swojego życia". Interesujące i potencjalnie znaczące społecznie wyniki przełożyły się na duże zainteresowanie tematem również ze strony środowiska naukowego. Niestety, pięć lat później zarówno replikacja (Ranehil i in., 2015), jak i metaanalizy 33 artykułów prezentujących wyniki badań nad tym fenomenem wykazały, że efekt nie istnieje (Simonsohn i Simmons, 2015). Ilość czasu, funduszy i zaangażowania zainwestowanych w ponad 30 eksperymentalnych badań nad nieistniejącym efektem trudno jest oszacować. Artykuły dotyczące póz energetycznych dalej są dostępne i z dużym prawdopodobieństwem w dalszym ciągu będą wykorzystywane jako podstawa kolejnych badań na ten temat. Nie istnieje bowiem praktyka opatrywania artykułów komentarzami zawierającymi informacje o nieudanych replikacjach czy metaanalizach – mimo technicznej możliwości umieszczenia takich informacji na stronach internetowych czasopism (wykorzystywana

np. przy informowaniu o wycofaniu artykułu). Choć opis wykładu został uzupełniony, jest on lakoniczny i nie informuje, że efekt póz energetycznych został obalony. Gdyby wyniki replikacji rzeczywiście byłyby wykorzystywane do uaktualniania wiedzy – możliwe byłoby zaoszczędzenie znacznych ilości czasu i środków włożonych w badania nad efektami, które okazały się być jedynie wynikami fałszywie pozytywnymi.

Opisany przypadek przeprowadzenia kilkudziesięciu badań nad nieistniejącym zjawiskiem jest efektowny, ale nie mówi o zakresie, w jakim fałszywie pozytywne wyniki są problemem w skali całej psychologii. Próby odpowiedzi na to pytanie można podjąć na podstawie danych o odsetku, jaki w literaturze stanowią replikacje. Okazuje się, że w psychologii jest on dramatycznie niski (około 1% wszystkich opublikowanych artykułów: Makel, Plucker i Hegarthy, 2012). Większość badań nie zostaje więc nigdy zreplikowanych i zweryfikowanych. Dodatkowo, najwięcej udanych replikacji pochodzi od zespołów, które przeprowadzały pierwotne badanie. Z jednej strony można to przypisać temu, że oryginalni badacze lepiej znają procedurę, ale z drugiej możliwe jest, że nieudane replikacje przeprowadzone przez tych samych badaczy są bardziej zagrożone efektem szuflady (ponieważ publikacja wyników sprzecznych ze wcześniej uzyskanymi może być postrzegana jako niekorzystna dla autora). Wyniki badania powtarzalności badań psychologicznych przeprowadzone przez *Open Science Collaboration* pokazują, że nie tylko mała ilość replikacji jest problemem psychologii. Projekt, w którym niezależne zespoły powtórzyły sto znanych badań psychologicznych wykazał, że mniej niż połowę wyników oryginalnych badań udało się powtórzyć, a wielkości efektu w replikacjach były średnio o połowę mniejsze (Open Science Collaboration, 2015). Takie rezultaty mogą być przypisane masowej publikacji wyników fałszywie pozytywnych i niedostatecznej mocy badań. Kolejną negatywną konsekwencją nadreprezentacji wyników fałszywie pozytywnych jest przeszacowanie wielkości efektów w metaanalizach (Bakker, van Dijk i Wicherts, 2012). Niekorzystna sytuacja publikowania głównie wyników pozytywnych w psychologii dodatkowo kumuluje się z trudnościami w komunikowaniu społeczności naukowej przypadków wycofywania artykułów czy publikacji nieudanych replikacji. W systemie obciążonym tyloma błędami, nie można liczyć na to, że z biegiem czasu fałszywe odkrycia samoistnie ulegną korekcie.

Wnioski

Mimo dominującej pozycji NHST wśród metod statystycznych wykorzystywanych przez psychologów, jest ona powszechnie niezrozumiana i błędnie wykorzystywana. Tej sytuacji częściowo można przypisać winę za wykorzystywanie niewłaściwych praktyk badawczych, których występowanie negatywnie wpływa na jakość wiedzy w psychologii. Nadmierna swoboda w wykorzystywaniu narzędzi dostarczanych przez NHST doprowadza do problemów z powtarzalnością wyników. Z tego powodu możliwe jest, że nega-

tywne opinie i odczucia związane z tymi błędami i praktykami zostaną przeniesione na cały nurt statystyczny testowania istotności hipotezy zerowej.

Negatywna percepcja NHST niejednokrotnie skłaniała badaczy do szukania alternatyw. Zwrot w kierunku statystyk w nurcie bayesowskim był wielokrotnie obwoływany remedium na wszelkie bolączki związane ze stosowaniem NHST (Wagenmakers, 2007). Upatrywanie w Bayesie nadziei na poprawę jakości analiz może być jednak przedwczesne: stosowanie tego paradygmatu nie wyklucza nadużyć (np. *p*-hackingu) pojawiających się również przy wykorzystaniu NHST (Simonsohn, 2014). Jeszcze niedawno spodziewano się, że popularyzacja podejścia bayesowskiego zmniejszy swobodę i łatwość wielokrotnego analizowania danych ze względu na konieczność samodzielnego przeprowadzania obliczeń. Obecnie dostępne są już proste w obsłudze programy umożliwiające analizy statystyczne w nurcie bayesowskim (np. JASP, BUGS, BACC) i ta wątpliwa przewaga nad NHST odeszła do przeszłości.

Mimo zdecydowanej krytyki NHST, nawoływanie do odejścia od analiz w tym nurcie może okazać się wylewaniem dziecka z kąpielą. Użyteczność tego narzędzia została zmniejszona przez błędy w jego wykorzystaniu i interpretacji, ale w dalszym ciągu NHST może być bazą rzetelnego wnioskowania statystycznego o uzyskiwanych w psychologii danych. Odpowiednio wykorzystana, NHST jest narzędziem użytecznym i powszechnie znanym w środowisku psychologicznym. Wprowadzanie do praktyki innych metod, np. podawania przedziałów ufności, technik bayesowskich czy raportowania jedynie statystyk opisowych jest korzystne: w ten sposób zwiększa się wachlarz środków, które mogą być wykorzystane do opisania otrzymywanych danych. Należy bowiem pamiętać, że NHST nie jest narzędziem uniwersalnie pasującym do wszystkich danych, planów analiz i preferencji badacza. Mimo to, wyparcie NHST przez alternatywne metody doprowadziłoby do niepotrzebnego zubożenia „skrzynki narzędzi" psychologa-badacza.

Odpowiednia edukacja – nie tylko na poziomie studiów, ale również skierowana do praktykujących badaczy, wypływająca z towarzystw psychologicznych, instytucji oraz niejako wymuszana przez czasopisma psychologiczne (zarówno przez ich wymagania, jak i popularyzację przez publikowanie artykułów na ten temat) – powinna być centralnym punktem walki z rozpowszechnieniem negatywnych zjawisk, które można powiązać z wykorzystaniem NHST. Pomimo istnienia metod wyłapywania prostych błędów statystycznych w artykułach (Nuijten, Hartgerink, van Assen, Epskamp, Wicherts, 2016), podejmowanych prób wykrywania występowania *p*-hackingu w opublikowanych artykułach (Simonsohn, Nelson, Simmons, 2014), czy propozycji zapobiegania występowaniu błędów spowodowanych przez działania redakcji czasopism (Nuijten, 2016), moim zdaniem tylko poprawa świadomości statystycznej i metodologicznej może trwale zmniejszyć występowanie niewłaściwych praktyk badawczych w tym zakresie. W końcu,

jak trafnie zauważa Garcia-Perez (2016), NHST, podobnie jak noże czy siekiery, nie jest samo w sobie złe. Należy jedynie dbać o to, żeby takie narzędzia były wykorzystywane odpowiedzialnie i bez wywoływania szkód.

Bibliografia

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.

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.

Beauchamp K.L., May R.B. (1964) *Replication report: Interpretation of levels of significance by psychological researchers*. Psychological Reports, 14, 272.

Bouter L.M., Tijdink J., Axelsen N., Martinson B.C., Riet G. (2016). *Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity*. Research Integrity and Peer Review, 1(17).

Brzeziński J.M. (2012). *Co to znaczy, że wyniki przeprowadzonych przez psychologów badań naukowych poddawane są analizie statystycznej?* Roczniki Psychologiczne, 15(3), 7–40.

Carney D.R., Cuddy A.J.C., Yap A.J. (2010). *Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance*. Psychological Science, 21(10), 1363–1368.

Chambers C.D., Feredoes E., Muthukumaraswamy S.D., Etchells P.J. (2014). *Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond*. AIMS Neuroscience, 1(1), 4–17.

Cohen J. (1962). *The statistical power of abnormal-social psychological research: A review*. The Journal of Abnormal and Social Psychology, 65(3), 145–153.

Cohen J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cohen, J. (1992). *A power primer*. Psychological Bulletin, 112(1), 155–159.

Fanelli D. (2009). *How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data*. PLoS One, 4(5), e5738.

Fanelli D. (2010). *"Positive" results increase down the hierarchy of the sciences*. PLoS ONE 5(4), e10068.

Ferguson C.J., Heene M. (2012). *A vast graveyard of undead theories: publication bias and psychological science's aversion to the null.* Perspectives on Psychological Science, 7(6), 555–561.

Francis G. (2014). *The frequency of excess success for articles in Psychological Science*. Psychonomic Bulletin & Review, 21(5), 1180–1187.

Francis G., Tanzmann J., Matthews W.J. (2014). *Excess success for psychology articles in the journal Science*. PLoS ONE, 9(12), e114255.

Franco A., Malhotra N., Simonovits G. (2014) *Publication bias in the social sciences: Unlocking the file drawer*. Science, 345(6203), 1502–1505.

Garcia-Perez M.A. (2016). *Thou shalt not bear false witness against null hypothesis significance testing*. Educational and Psychological Measurement, 1–32.

Gigerenzer G. (1993). *The Superego, the Ego, and the Id in statistical reasoning.* W: Keren G., Lewis C. (red.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (s. 311–339). Hillsdale, NJ: Erlbaum.

Haller H., Krauss S. (2002). *Misinterpretations of significance: A problem students share with their teachers?* Methods of Psychological Research Online, 7(1).

Jarmakowska-Kostrzanowska L. (2016). *W statystycznym matriksie: kontrowersje wokół testowania istotności hipotezy zerowej (null hypothesis significance testing, NHST) oraz p-wartości. – z autorskimi komentarzami*. Pozyskano z: https://www.researchgate.net/publication/299567395_W_statystycznym_matriksie_kontrowersje_wokół_testowania_istotnosci_hipotezy_zerowej_null_hypothesis_significance_testing_NHST_oraz_p-wartosci_-_z_autorskimi_komentarzami

Kerr, N.L. (1998). *HARKing: hypothesizing after the results are known*. Personality and Social Psychology Review, 2, 196–217.

Kirk R. (1996). *Practical significance: A concept whose time has come*. Educational and Psychological Measurements, 56, 746–759.

Levelt W.J.M., Drenth P., Noort E. (red.). (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Commissioned by the Tilburg University, University of Amsterdam and the University of Groningen.

Makel M.C., Plucker J.A., Hegarty B. (2012). *Replications in psychology research: How often do they really occur?* Perspectives on Psychological Science, 7(6), 537–542.

Martinson B.C., Anderson M.S., de Vries R. (2005). *Scientists behaving badly*. Nature, 435(9), 737–738.

Mazzola J.J., Deuling J. K. (2013). *Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I-O journal articles*. Industrial and Organizational Psychology, 6(3), 279–284.

Nickerson R.S. (2000). *Null hypothesis significance testing: a review of an old and continuing controversy*. Psychological Methods, 5(2), 241–301.

Nuijten M.B. (2016) *Preventing statistical errors in scientific journals*. European Science Editing, 42(1), 8-10.

Nuijten M.B., Hartgerink C.H.J., van Assen M.A.L.M., Epskamp S., Wicherts J.M. (2016). *The prevalence of statistical errors in psychology (1985–2013).* Behavior Research Methods, 48(4), 1205–1226.

Open Science Collaboration (2015). *Estimating the reproducibility of psychological science.* Science, 349(6251), aac4716.

Poitevineau J., Lecoutre B. (2001). *Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated*. Psychonomic Bulletin & Review, 8(4), 847–850.

Ranehill E., Dreber A., Johannesson M., Leiberg S., Sul S., Weber R.A. (2015). *Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women.* Psychological Science, 26(5), 653–656.

Rękowski W., Przyłuska-Fiszer A., Różyńska J., Fijałkowska B. (2012). *Środowiskowe przyzwolenie na łamanie zasad dobrej praktyki badawczej w opinii społeczności akademickiej*. Zagadnienia Naukoznawstwa, 4(194), 251–268.

Rosenthal R. (1979). *The file drawer problem and tolerance for null results*. Psychological Bulletin, 86, 638–641.

Rosenthal R., Gaito, J. (1964). *Further evidence for the cliff effect in the interpretation of levels of significance*. Psychological Reports, 15, 570.

Rosnow R., Rosenthal R. (1989). *Statistical procedures and the justification of knowledge in psychological science.* American Psychologist, 44, 1276–1284.

Sedlmeier P., Gigerenzer G. (1989). *Do studies of statistical power have an effect on the power of studies?* Psychological Bulletin, 105, 309–316.

Simonsohn U. (2014). *Posterior-hacking: selective reporting invalidates Bayesian results also.* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2374040

Simonsohn U., Nelson L.D., Simmons J.P. (2014). *P-curve: A key to the file-drawer*. Journal of Experimental Psychology: General, 143(2), 534–547.

Simonsohn U., Simmons J. (2015). *Power Posing: reassessing the evidence behind the most popular TED talk*. http://datacolada.org/37

Simmons J.P., Nelson L.D., Simonsohn U. (2011). *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.* Psychological Science, 22(11), 1359–1366.

Smith R. (2000). *What is research misconduct?* [W:] C. White (red.) *The COPE Report 2000: the Committee on Publication Ethics.* (s. 7–11). Londyn: BJM Books.

Wagenmakers E-J. (2007). *A practical solution to the pervasive problems of p-values*. Psychonomic Bulletin & Review, 14, 779–804.

Zdybek P., Walczak R., Zdybek M. (2012). *Historia zwykłego oszustwa. Nieuczciwość akademicka widziana oczami studentów psychologii*. Psychologia Społeczna, 73(22), 234–244.

In chase of statistically significant result. Consequences of widespread use of NHST (null hypothesis significance testing) in psychology

NHST (null hypothesis significance testing) is the most popular statistical paradigm in psychology. Mistakes in interpretation of its assumptions and their consequences are topic for methodological and statistical discussion for over fifty years. Article presents two problems associated with NHST that are particularly prevalent in psychology: identifying non-significant results with research failure and conducting underpowered research. They can contribute to increase in exploiting questionable research practices in order to obtain desirable, significant outcomes. Three practices: *p*-hacking, HARKing and selective publishing are described, along with analysis of their impact on replication crisis in psychological science.

Key words: NHST, *p*-hacking, HARKing, selective publishing, replication crisis, false positives