Psychologia Społeczna
2015 tom 10 4 (35) strona 419–434

ISSN 1896-1800
doi: 10.7366/1896180020153505

Postscriptum do oszustwa Stapela: niepokojące dane, początek zmian?

Łukasz Budzicz

Instytut Psychologii, Uniwersytet im. Adama Mickiewicza

Artykuł stanowi uzupełnienie dyskusji dotyczącej rzetelności w nauce, jaka odbyła się na łamach *Psychologii Społecznej* (nr 3/2012). W ostatnich latach pojawiły się liczne nowe dane wskazujące na to, że Stapel mógł nie być odosobnionym przypadkiem, lecz symptomem większego kryzysu. Nawet jeśli większość badaczy nie fabrykuje danych, subtelne zafałszowania polegające na arbitralnej obróbce danych i wybiórczych prezentacjach wyników mogą być relatywnie częste i prowadzić do poważnego zniekształcenia obrazu rzeczywistości. Na wybiórcze publikowanie i zniekształcanie danych szczególnie wskazują analizy rozkładu wartości *p* w literaturze (*p-curve analysis*), analizy prawdziwości raportowania wartości *p*, oraz analizy skumulowanej mocy w artykułach z wieloma badaniami. Artykuł przedstawia także najważniejszy postulaty z toczonej obecnie dyskusji nad gruntownymi zmianami w praktykach badawczych i publikacyjnych. Wskazane są przykłady, gdzie takie zmiany zostały już zainicjowane.

Słowa kluczowe: *oszustwo Stapela*, *rzetelność danych w psychologii*, *psychologia fałszywie pozytywna*, *moc badań*

Oszustwo Stapela było największym wykrytym fałszerstwem w psychologii i jednym z większych w nauce w ogóle (Stroebe, Postmes, Spears, 2012). Praktycznie wszystkie ważne czasopisma z zakresu psychologii społecznej padły jego ofiarą. Historia ta jest dobrze znana w środowisku i nie będę tu przypominał szczegółów, które są szeroko opisane w innych miejscach (Bhattacharjee, 2013; Klebaniuk, 2012; Levelt Committee, Noort Committee, Drenth Committee, 2012; zob. też: Levelt, 2012). W tekście chciałbym natomiast zwrócić uwagę na wiele pojawiających się w ostatnich latach systematycznych analiz wskazujących na nieprawidłowości we wzorze niektórych danych obecnych w czasopismach psychologicznych. Mogą one wskazywać na to, że Stapel nie był odosobnionym przypadkiem, lecz symptomem szerszego kryzysu.

Prawdopodobnie większość badaczy nie fabrykuje danych w potocznym sensie, tj. poprzez raportowanie wyników badań, które nigdy się nie odbyły. W badaniu Johna, Loewensteina i Preleca (2012), 1,7% aktywnych badaczy (psychologów) przyznało się do fabrykacji danych, równolegle ankietowani szacowali, że w ich dyscyplinach średnio 10% badaczy fałszuje dane. Jednak wątpliwe praktyki badawcze (*questionable research practice* – QRP), polegające na arbitralnych analizach i wybiórczych prezentacjach wyników, mogą poważnie zniekształcać obraz rzeczywistości w nauce (Simmons, Nelson, Simonsohn, 2011). Przykładem takiej praktyki może być zbieranie wielkiej liczby zmiennych i raportowanie tylko tych danych, które przyniosły wyniki istotne statystycznie. Według wspomnianego badania Johna i in. 20–70% badaczy przyznaje się do stosowania tego rodzaju praktyk i, co gorsza, zasadniczo są one uważane za dopuszczalne. Tabela 1 przedstawia rozpowszechnienie QRP w tymże badaniu i ocenę ich dopuszczalności (w badaniu wzięło udział ponad 2 tys. badaczy spośród 6 tys., do których wysłano zaproszenia).

Tego typu praktyki zniekształcają obraz rzeczywistości w stopniu niewiele mniejszym niż zwykłe fałszerstwo. Społeczność badaczy uzyskuje informacje o efektach o zawyżonej sile, ma nierealistyczne oczekiwania co do możliwości replikacji określonych efektów, a nieopublikowane (nieistotne) dane mogą zawierać informacje na temat ważnych moderatorów niwelujących efekt. Najgorszym skutkiem jest zapewne presja na innych badaczy, aby

Łukasz Budzicz, Instytut Psychologii, Uniwersytet Adama Mickiewicza, ul. A. Szamarzewskiego 89/AB, 60-568 Poznań,
e-mail: lukasz.budzicz@gmail.com

Tabela 1
Odsetek badaczy przyznających się do stosowania wybranych wątpliwych praktyk badawczych w badaniu Johna i in., 2012

Wątpliwa praktyka badawcza	Procent badanych, którzy przyznali się do stosowania danej praktyki (%)	Średnia ocena dopuszczalności danej praktyki (w nawiasie odchylenia standardowe)
Nieraportowanie wszystkich wykorzystanych zmiennych	66,5	1,84 (0,39)
Zbieranie dodatkowych danych po sprawdzeniu, czy już posiadane dane są istotne	58,0	1,79 (0,44)
Wybiórcze raportowanie tylko tych warunków eksperymentalnych, pomiędzy którymi zanotowano istotne różnice	27,4	1,77 (0,49)
Nieuprawnione zaokrąglanie wartości p (np. raportowanie wartości $p = 0{,}054$ jako $p < 0{,}05$)	23,3	1,68 (0,57)
Selektywne raportowanie tylko tych badań, które wyszły	50,0	1,66 (0,53)
Decydowanie o tym, czy wykluczyć określone dane po sprawdzeniu wpływu takiej operacji na rezultaty	43,4	1,61 (0,59)
Opisywanie nieoczekiwanego wcześniej odkrycia jako przewidzianego od samego początku	27,0	1,50 (0,60)
Fabrykowanie (tj. raportowanie nieistniejących) danych	1,7	0,16 (0,38)

Adnotacja. W kolumnie 2 przedstawiono wyniki grupy, która była dodatkowo motywowana do mówienia prawdy. Wyniki grupy kontrolnej były kilka procent niższe (tj. 0–7%).

Wyniki w kolumnie 4 na podstawie średnich ocen w skali: 0 – *całkowicie niedopuszczalne*; 1 – *w pewnym stopniu dopuszczalne*; 2 – *dopuszczalne*.

pokazywać w artykułach badania jako serię sukcesów i pomijać wyniki niejednoznaczne, a więc przedkładać atrakcyjność tekstu nad jego rzetelność[1].

Tego typu praktyki są potencjalnie możliwe do wykrycia poprzez systematyczne analizy literatury za pomocą odpowiednich technik matematyczno-statystycznych. Wiele takich analiz przeprowadzono w ostatnim czasie i warto je tutaj przedstawić, gdyż dają lepszy obraz stanu nauki niż pojedyncze artykuły empiryczne.

Skumulowana moc badań i wybiórcze publikowanie

Czasopisma empiryczne z dziedziny psychologii składają się w olbrzymiej większości z „udanych" badań, przy czym kryterium sukcesu jest przede wszystkim istotność statystyczna wyniku, która niekoniecznie musi iść w parze z jego prawdziwością i znaczeniem teoretycznym albo praktycznym (por. np. Brzeziński, 2012; Wagenmakers, 2007). W jakiejś mierze trudno się dziwić redakcjom. Gdyby miano publikować każde badanie o sensownych hipotezach i rozsądnej metodologii niezależnie od ostatecznego wyniku, wówczas czasopisma byłyby zarzucone raportami o nieistotnych wynikach, co prawdopodobnie jeszcze bardziej utrudniłoby poruszanie się w monstrualnej liczbie artykułów, jakie już się ukazują. Z nieskończonej liczby możliwych kombinacji zmiennych najbardziej przyciągają umysł te, które rzeczywiście na siebie wpływają, a z niedziałających na ogół mniej wynika dla naszego rozumienia świata. We wszystkich więc naukach to wyniki pozytywne stanowią większość zawartości czasopism. Okazuje się jednak, że w bardzo różnym stopniu. Fanneli (2010) przeprowadził analizę bibliometryczną na dużej próbie artykułów naukowych i pokazał, że w psychologii (wraz z psychiatrią) największy odsetek artykułów zawiera dane pozytywne (91,5%). Najmniej artykułów potwierdzających wyjściową hipotezę publikuje się w astrofizyce (ok. 70%). Biorąc jednak poprawkę na to, że w artykułach psychologicznych testuje się dużo więcej hipotez, w psychologii jest w przeliczeniu na artykuł ponad pięć razy więcej wyników pozytywnych niż w naukach o kosmosie. Różnice te są szczególnie wyraźne w badaniach podstawowych.

[1] Również w raporcie poświęconym oszustwu Stapela zwrócono uwagę na to, że same redakcje czasami zachęcały Stapela i jego współautorów do przekłamań, szczególnie pomijania informacji o niewychodzących badaniach (por. Levelt, 2012, s. 53).

Łatwość uzyskiwania wyników pozytywnych pokazywana w literaturze kontrastuje z tym, co sami badacze mówią o eksperymentowaniu. „Nie wiem, jak to wygląda u was, ale większość moich danych nie zostało nigdy opublikowanych w dobrych czasopismach ani nawet w słabych czasopismach. Większość moich danych leży sobie w szufladzie (w ostatnim czasie na dysku komputera)" – to słowa redaktor naczelnej *Perspectives on Psychological Science* Barbary Spellman (2012, s. 58). „Jako czynny psycholog społeczny jestem przekonany, że co roku tysiące tak zwanych nieudanych badań ląduje w szufladach bez żadnej szansy na publikację (sam oczywiście mam takie we własnym biurku – mnóstwo pracy, żadnych efektów)" – to z kolei polski psycholog społeczny Jarosław Klebaniuk (2012, s. 216). Mógłbym dodać do tego podobne obserwacje. Prowadząc przedmiot poświęcony badaniom empirycznym, nadzorowałem studenckie replikacje m.in. badań nad wpływem prymowania nietypowych ról płciowych na poczucie własnej wartości w domenie zawodowej (Rudman, Phelan, 2010), nad wpływem aktywacji konceptu śmiertelności na poziom religijności (Norenzayan, Hansen, 2006) czy zagrożenia patogenami na postrzeganie atrakcyjności fizycznej (Little, DeBruine, Jones, 2006). Praktycznie nigdy replikacje te nie dawały wyników zbliżonych do oryginalnych. W przypadku względnie prostych, „klasycznych" efektów nigdy nie miałem problemów z ich replikacją (np. efektu Stroopa; efektu rotacji mentalnych; efektu pamiętania w zależności od poziomów przetwarzania itd.).

Artykuły w takich czasopismach, jak *Journal of Personality and Social Psychology* (dalej: JPSP) składają się z kilku badań. Przejrzałem streszczenia wszystkich artykułów opublikowanych w JPSP w roku 2013. W 98 tekstach przedstawiono wyniki więcej niż jednego oryginalnego badania empirycznego, łącznie 435 badań (68 tekstów zawierało 4 lub więcej badań, maksymalnie 10). Wszystkie badania w obrębie jednego artykułu potwierdzały wyjściową hipotezę. Autorzy w żadnym z abstraktów nie wspominają o znalezieniu jakichś danych, które nie potwierdzałyby ich hipotezy. (W olbrzymiej większości były to wyniki na rzecz jakiejś oryginalnej hipotezy, choć zdarzały się wyjątkowo teksty, których celem była krytyka określonej teorii). Wyniki były zwykle przedstawione jako przewidziane od samego początku przez założenia teoretyczne (por. też: Kerr, 1998), sporadycznie badanie było przedstawiane jako eksploracyjne w swej naturze. Brian Nosek (Bones, 2012) ironizował, że Bem (2011) w swoim kontrowersyjnym badaniu o prekognicji nie pokazał niczego nowego, gdyż każdy numer JPSP dostarcza mnóstwa dowodów na prekognicję.

Kilka udanych pod rząd badań kłóci się nie tylko z codziennym laboratoryjnym doświadczeniem, ale również z twardymi prawami rachunku prawdopodobieństwa. Pomocne będzie tu przypomnienie pojęcia mocy badania[2]. Moc jest to prawdopodobieństwo, że przy założonej sile efektu, liczebności grupy badanej oraz poziomie istotności statystycznej[3] określone badanie da wynik istotny statycznie (Cohen, 1992). Hipotetyczne efekty niekoniecznie występują w naturze, ale nawet gdy występują, nie zawsze pokażą się w badaniu w postaci istotnych statystycznie wyników. To wina szczególnie tzw. błędu próby (*sampling error)*, czyli różnych przypadkowych czynników, które wpływają na to, że wynik jest odległy od reprezentatywności. Im większa próba, tym przypadek gra mniejszą rolę (moc jest większa). Znaczenie ma też siła efektu. Nie trzeba zbadać wielu osób, żeby wykazać bardzo silny efekt np. wyższego wzrostu u mężczyzn niż kobiet (np. Lippa, 2009). Wykazanie, że 14-latki są wyższe od 13-latków wymaga już zebrania dużo większej próby, gdyż efekt ten jest znacznie słabszy. Na rysunku 1 przedstawiłem przykładowe wyliczenia mocy w zależności od wielkości próby i siły efektu. Zasadniczo nie można manipulować wielkością efektu (chyba że uciekamy się do QRP), dlatego jedynym sposobem zwiększania mocy jest zwiększanie próby badanej.

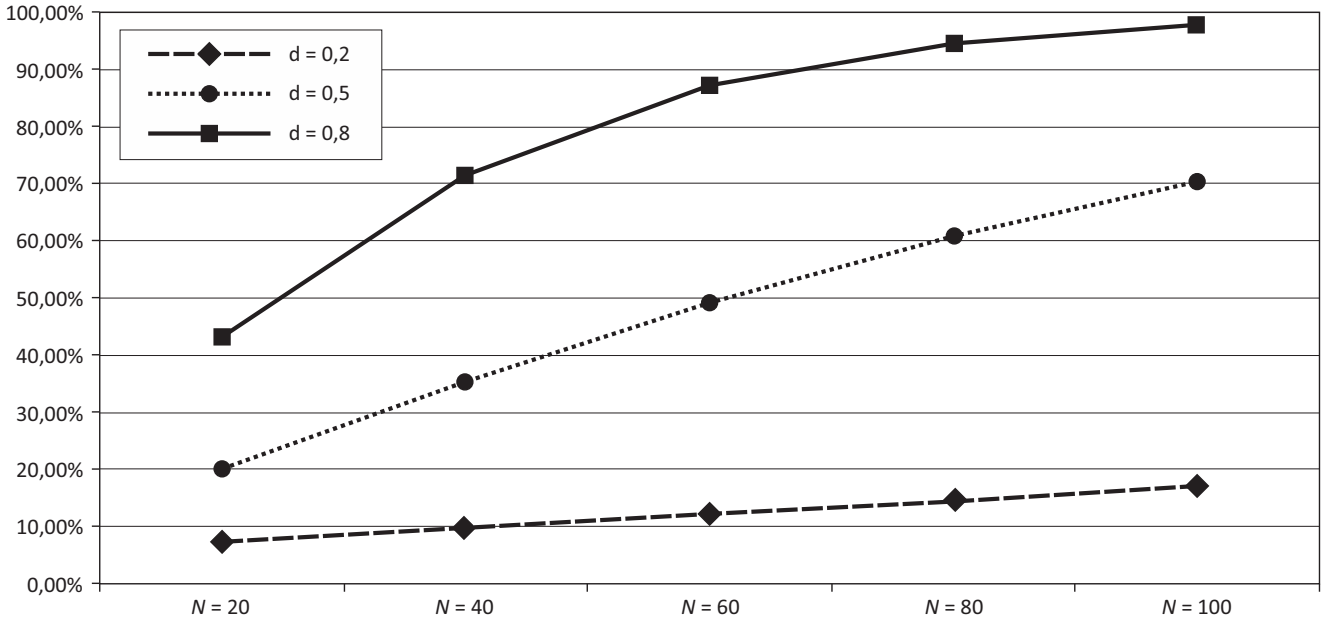
Problem z artykułami w czasopismach psychologicznych polega na tym, że siła efektów na ogół nie jest duża, a próby niezbyt liczne. Richard, Bond i Stokes-Zoota (2003) zestawili wielkości efektów z 322 metaanaliz w psychologii społecznej i otrzymali średni efekt $r = 0{,}21$ (czyli około $d = 0{,}4$). Takie wyliczenie jest jednak mylące, gdyż rozkład efektów jest wysoce prawoskośny i najczęstsza wartość to $0 < r < 0{,}1$ (co odpowiada $d < 0{,}2$; por. rysunek 1). Tak jak typowe zarobki są niższe niż średnia krajowa, tak samo typowy efekt jest sporo niższy niż owo $r = 0{,}21$. Bakker, Dijk i Wicherts (2012) oszacowali średnią wielkość efektu w publikowanych raportach psychologicznych na około $d = 0{,}50$.

Średnia wielkość prób w badaniach to około 40 osób (Marszalek, Barber, Kohlhart, Holes, 2011), 20–30 osób (Tressoldi, 2012) lub 24 osoby w jednej celce (Wetzels i in., 2011). Moc typowych badań psychologicznych wynosi więc około 35%. Za przyzwoity poziom mocy uważa się 90%.

Tak niska moc jest wymowna i zastanawiająca. Dlaczego badacze tak dobierają wielkości prób, skoro wiedzą, że większość da nienadające się do publikacji wyniki? Jedną z odpowiedzi może być po prostu brak odpowiedniej świadomości metodologicznej (Maxwell, 2004; Sedlmeier,

[2] Moc wydaje się pojęciem dużo słabiej znanym w społeczności psychologów niż pojęcie istotności statystycznej. Przykładowo, choć olbrzymia większość artykułów psychologicznych zawiera analizy istotności statystycznej, to analiza spodziewanej mocy pojawia się w 3% przypadków (Fritz, Scherndl, Kühberger, 2013).

[3] Poziom istotności statystycznej praktycznie zawsze przyjmuje się jako 0,05, więc ta akurat zmienna nie ma znaczenia w kalkulacjach.



Rysunek 1. Przykładowe wyliczenia mocy.

Adnotacja. Wykres przedstawia moc prostego badania polegającego na porównaniu średnich dwóch grup przy założonej wielkości efektu i liczbie osób badanych (liczebność każdej grupy: *N*/2).

Gigerenzer, 1989; Vankov, Bowers, Munafò, 2014). W czasopismach niezwykle rzadko spotyka się analizy oczekiwanej mocy badania (Fritz, Scherndl, Kühberger, 2013). Możliwe jednak, że czasami niska moc jest elementem przemyślanej strategii. Pierwszym kryterium publikacyjnym jest istotność statystyczna wyników, dlatego zamiast przeprowadzać jedno bardzo duże badanie lepszą pod kątem publikacji strategią może być przeprowadzenie wielu małych badań i uwzględnienie w raporcie tylko istotnych efektów (por. symulacje metaanaliz; Bakker i in., 2012)). Część da wyniki istotne z racji czystego przypadku, dodatkowo przy małej próbie istnieje szansa uzyskania efektów o większej sile niż rzeczywiste (Ioannidis, 2008). Dodajmy do tego, że przygniatająca większość badań w psychologii społecznej opiera się na zmiennych samoopisowych (względnie takich, które mierzy komputer; Baumeister, Vohs, Funder, 2007), a więc dla kogoś, kto dysponuje dostępem do licznych badanych (studentów), przeprowadzenie wielu małych badań nie musi stanowić dużego problemu logistycznego.

Zestawienie bardzo niskiej mocy z bardzo wysokimi wskaźnikami pozytywnych wyników wskazuje na to, że publikowane raporty nie zawierają całej prawdy i mamy do czynienia z publikowaniem wybiórczym[4] (*publication bias*). Można to próbować wykazać zarówno na poziomie pojedynczych artykułów, jak i na poziomie metaanaliz.

W pokazywaniu możliwego wybiórczego publikowania na poziomie pojedynczych artykułów empirycznych poprzez obliczanie skumulowanej mocy[5] wyspecjalizował się w ostatnich latach szczególnie Gregory Francis z Purdue University. W tabeli 2 przedstawiłem zakwestionowane przez niego teksty.

Francis został skrytykowany przez Simonsohna (2012), który zwrócił uwagę, że taka metoda jest czymś w rodzaju wybierania wisienek z tortu. W uproszczeniu jego krytyka opiera się na założeniu, że niektóre mało prawdopodobne wyniki są takie z racji czystego przypadku. Dostatecznie długo przeglądając zupełnie rzetelne teksty, prędzej czy później natrafi się na takie dane. Wybierając badania według niesprecyzowanych kryteriów, Francis nie jest w stanie wskazać, czy nietypowy wynik to kwestia przypadku, czy też celowe zniekształcenia. Odpowiedzią Francisa (2014) na ten zarzut jest systematyczna analiza, w której uwzględnił wszystkie artykuły zawierające przynajmniej

[4] Zwane też w literaturze polskiej *zniekształceniem publikacyjnym*.

[5] Obliczając skumulowaną moc badań, należy podjąć decyzję, czy wyliczymy ją na podstawie średniej wielkości efektu, czy na podstawie wielkości efektów obserwowanych w pojedynczych badaniach. Decyzje te mają matematyczne wady i zalety. Po szczegóły odsyłam do Schimmack (2012, s. 555). Niezależnie od decyzji skumulowana moc jest jednak porównywalna.

Tabela 2
Teksty wskazujące na zbyt niską skumulowaną moc badań i prawdopodobne wybiórcze publikowanie lub QRP

Źródło	Zakwestionowany efekt	Korelacja pomiędzy wielkością próby a wielkością efektu
Francis (2012a)	Przyszłe, nieznane jeszcze wydarzenia wywierają niewielki, ale eksperymentalnie dostrzegalny wpływ na teraźniejsze decyzje (Bem, 2011)	$r = -0{,}89$
Francis (2012b)	Obiekty „pożądane" są widziane jako bliższe (ang. *wishful seeing*; Balcetis, Dunning, 2010)	$r = -0{,}75$
Francis (2012c)	Osoby z niższych klas społecznych mają większą tendencję do pomagania (Piff, Stancato, Côté, Mendoza-Denton, Keltner, 2012)	$r = -0{,}38$
Francis (2012d)	Niemoralne zachowanie wzbudza potrzebę fizycznego oczyszczenia (Zhong, Liljenquist, 2006)	$r = -0{,}92$
Francis (2012e)	Antycypacja nieprzyjemnej sytuacji powoduje bardziej negatywną ocenę podobnych sytuacji z przeszłości (Galak, Meyvis, 2011)	$r = -0{,}88$
Francis (2013)	Mężczyzna fotografowany na czerwonym tle jest postrzegany przez kobiety jako atrakcyjniejszy i mający wyższą pozycję społeczną (Elliot i in., 2010)	$r = -0{,}80$

Opracowanie własne na podstawie danych zawartych w cytowanych artykułach.

cztery badania opublikowane w *Psychological Science* w latach 2009–2012. Takich artykułów było 44, spośród których w 36 (82%) wzór danych wskazywał na małe prawdopodobieństwo sukcesu, jaki miał rzekomo miejsce, relatywnie do niskiej skumulowanej mocy badań[6]. Francis nazwał swój wskaźnik „testem zbyt dużego sukcesu" (*test for excessive success*). Konceptualnie bardzo podobny wskaźnik zaproponował Schimmack (2012) i nazwał go indeksem nieprawdopodobieństwa (*incredibility index)*. Schimmack ograniczył się do wskazania wysokiego nieprawdopodobieństwa dwóch serii badań: kontrowersyjnego efektu prekognicji Bema (2011) oraz niekontrowersyjnego, ale bardzo często cytowanego efektu zwiększania siły woli pod wpływem glukozy (Gaillot i in., 2007).

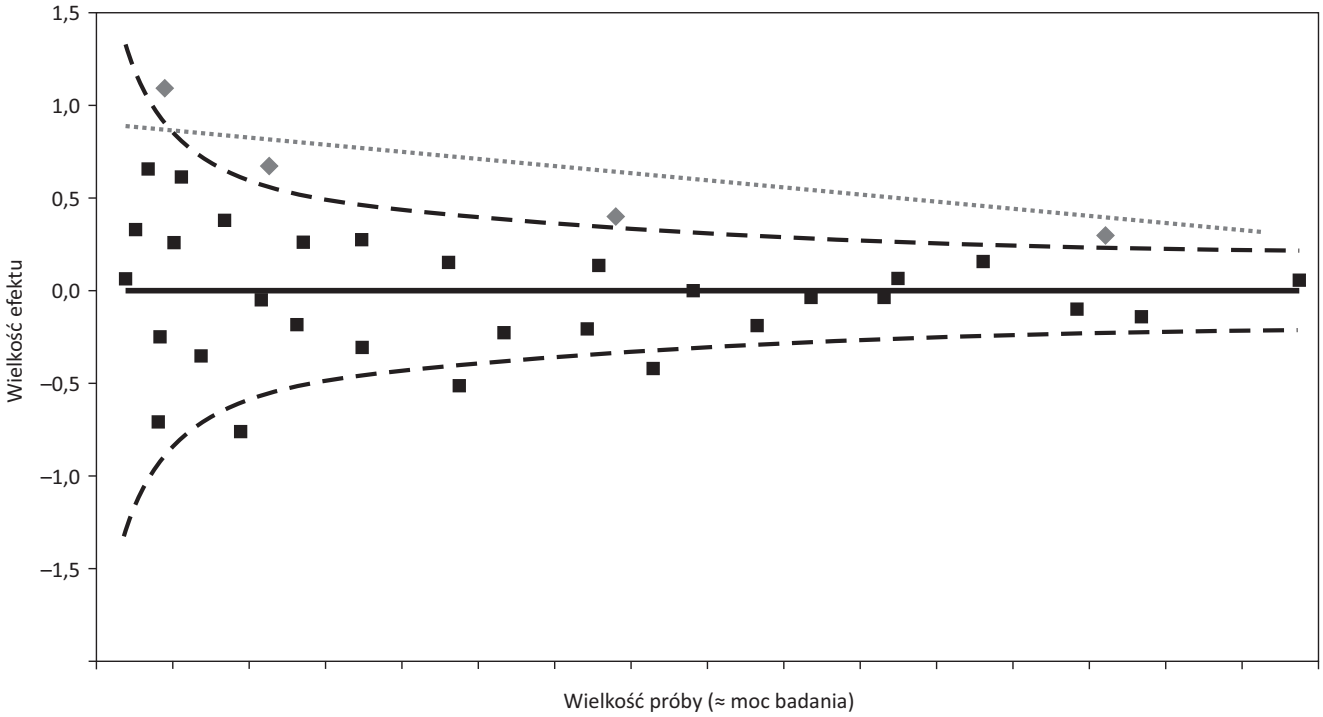
Zbliżonym, ale technicznie znacznie prostszym testem wybiórczego publikowania jest sprawdzenie korelacji między wielkością próby a wielkością efektu w serii badań. Matematycznie rzecz biorąc, średnia wielkość efektu nie będzie skorelowana z wielkością próby (w odpowiednio dużej serii badań, por. rysunek 2; empiryczny przykład: Lippa, 2009, rys. 6). Jednak w przypadku mniejszych prób wariancja wielkości efektów jest większa, a tylko silniejsze efekty będzie można uznać za istotne. Tak więc w przypadku wybiórczego publikowania może występować negatywna korelacja między wielkością próby a siłą efektu (*negative n-r correlation*), co zostało schematycznie przedstawione na rysunku 2. W zasadzie we wszystkich zakwestionowanych przez Francisa tekstach (tabela 2) występuje bardzo silna negatywna korelacja między wielkością próby a siłą efektu.

Przykład tego zjawiska pokazuje metaanaliza badań nad skutecznością programów nauczania matematyki (Slavin, Smith, 2009), w której wielkość efektu systematycznie malała wraz ze wzrostem wielkości grupy badanej. Przykładowo, dla badań z najmniej licznymi grupami ($N < 50$) średnia wielkość efektu wynosiła 0,44 (d Glassa), ale w badaniach z najliczniejszymi grupami ($N > 2000$) średnia wielkość efektu wynosiła już tylko 0,09.

Levine, Asada i Carpenter (2009) zbadali 51 metaanaliz ze wszystkich subdyscyplin psychologii, obejmujących łącznie 75 efektów. W 59 z nich (79%) stwierdzono negatywną korelację między wielkością próby a siłą efektu, z czego w przypadku 21 efektów była to dosyć wyraźna zależność ($r < -0{,}3$). Warto też zauważyć, że w koncepcyjnie zbliżonym badaniu pokazano, iż metaanaliza badań nad trafnością „kultowego" narzędzia w psychologii społecznej, czyli testu utajonych skojarzeń (*implicit association test*), zawiera nadmiar efektów pozytywnych relatywnie do niskiej mocy badań (Bakker i in., 2012).

Systematyczną analizę zależności między siłą efektu a wielkością próby w psychologii ogólnej przeprowadzili Kuhberger, Fritz i Scherndl (2014). Wylosowali oni 1000 tekstów z bazy PsycINFO z roku 2007. Odrzucono teksty przeglądowe, teoretyczne, metaanalizy, badania empiryczne opisowe i eksploracyjne oraz teksty niezawierające wszyst-

[6] Jako „nadmierny sukces" przyjęto mniej niż 0,1 (10%). Przykładowo, jeśli przyjmiemy, że w hipotetycznej serii 11 badań moc pojedynczego badania wynosi 80%, to prawdopodobieństwo sukcesu wszystkich będzie wynosiło około 8,5%. W przypadku badań o bardziej realistycznej mocy, dajmy na to 40%, prawdopodobieństwo sukcesu czterech badań wynosi tylko 2,56%.



Rysunek 2. Schematyczne przedstawienie przyczyn negatywnej korelacji między wielkością próby a wielkością efektu przy wybiórczym publikowaniu.

Adnotacja. Przerywane linie przedstawiają granicę, w której obrębie wyniki są nieistotne. Z uniwersum wszystkich badań (punkty szare i czarne) publikowane są tylko istotne, wskazujące na efekt w jednym kierunku (punkty szare). Szara kropkowana linia przedstawia negatywną zależność między wielkością próby a siłą efektu. Na wykresie założono brak efektu (d = 0,0), jednak taka praktyka publikacyjna może występować też w przypadku niezerowych.

kich danych (szczególnie wielkości efektów). Ostatecznie uwzględniono 395 badań. Występowała w nich wyraźna negatywna zależność między wielkością próby a siłą efektu r = –0,54 (korelacja była nieco mniejsza r = –0,45, jeśli odrzucono dane ze skrajnymi wielkościami prób, tj. mniej niż 10 i więcej niż 1000).

Istnieje możliwość, że negatywna korelacja między wielkością próby a siłą efektu wynika z wcześniejszej analizy mocy i doboru próby na tej podstawie (przykładowo, w przypadku spodziewanych słabych efektów badacze dobierają większą próbę, żeby zmaksymalizować moc). Jest to jednak wątpliwe – w tylko 3% artykułów empirycznych znajdują się formalne analizy mocy (Fritz i in., 2013), niewielki odsetek badaczy przyznaje się do ich stosowania (Vankov, Bowers, Munafò, 2014); ponadto w obrębie jednego artykułu badania często różnią się bardzo subtelnie, trudno więc przypuszczać, że badacze byli w stanie z góry przewidzieć, które procedury, bodźce, sposoby operacjonalizacji zmiennych itd. przyniosą silniejsze efekty. Również we wspomnianej analizie Kuhbergera i in. (2014) występowanie formalnej analizy mocy nie miało wpływu na siłę korelacji między wielkością próby a siłą efektu.

Pamiętając o tym, że małe próby nie będą się średnio wiązać z wyższymi efektami (a tylko większą wariancją siły efektów), interpretacja negatywnej zależności próba – siła efektu wzmacnia tezę, że badacze często decydują się na relatywnie małe badania, które nie wiążą się z dużym ryzykiem (mniejsze koszty finansowe i czasowe), a następnie udane badania są publikowane, natomiast nieudane nie są zgłaszane do publikacji lub są odrzucane przez redaktorów. W przypadku małych prób zapewne redaktorów musi dodatkowo przekonać wyraźny efekt, a taki jest łatwiejszy do uzyskania w przypadku niewielkiej mocy (Ioannidis, 2008).

Na koniec tej sekcji chciałbym przypomnieć starsze badanie Coopera, DeNevea i Charltona (1997), które prawdopodobnie ciągle zachowuje aktualność. Sprawdzano w nim, jaki był los badań zgłoszonych do komisji etycznej jednego z amerykańskich uniwersytetów. Przygotowanie wniosku do takiej komisji jest oznaką, że badanie wyszło poza etap mglistych planów i z dużym prawdopodobieństwem zostanie przeprowadzone. Spośród badań, które ostatecznie zrealizowano i których wyniki były istotne, opublikowano (lub przynajmniej wysłano do czasopisma) 74%, dodatkowe 14% zaprezentowano na konferencjach. Natomiast spośród

badań z nieistotnymi wynikami aż 91% nie doczekało się ani publikacji, ani choćby referatu. Brak „interesujących wyników" lub nieistotne wyniki były jedną z częściej podawanych przyczyn zarzucenia planów publikacyjnych (artykuł nie precyzuje, niestety, na czym polega to, że wyniki są nieinteresujące). Jako przyczyny zaniechania często podawane są też problemy w procedurze (*design problems*), trudno jednak rozstrzygnąć, na ile problemy te są rzeczywiste, a na ile są rodzajem racjonalizacji. Badacze mogą mieć tendencję do przeceniania proceduralnej wartości badań z istotnymi wynikami i nadmiernego doszukiwania się błędów w danych z wynikami nieistotnymi. Być może dla rzetelności postępu naukowego lepiej by było, gdyby oceny wartości danych (też nieistotnych) mogła dokonywać cała społeczność badaczy, a nie tylko sam zainteresowany. Badacze nie zawsze muszą być najlepszymi sędziami we własnej sprawie.

Publikacja tylko istotnych wyników, co oczywiste, zniekształca obraz rzeczywistości, ale też ustanawia nadmiernie wysokie wymagania wobec badaczy, nierealistyczne w takiej nauce, jak psychologia, i może niestety sprzyjać mniejszym i większym przekłamaniom. Liczba stanowisk na uczelniach jest ograniczona w stosunku do liczby zainteresowanych, podobnie ilość miejsca na łamach czasopism jest dużo mniejsza niż liczba manuskryptów (w najlepszych czasopismach psychologicznych odsetek odrzuceń to 60–80%; Suls, Martin, 2009). Osoby najbardziej zdeterminowane mogą zechcieć „pomóc szczęściu", a przypadek Stapela pokazał, że dysponując inteligencją, wprawnym piórem i znajomością standardów można to robić przez kilkanaście lat. Ostatecznie Stapel został zdemaskowany przez donos współpracowników, a nie przez nieudane replikacje jego badań czy zakwestionowanie ich na etapie *peer-review*.

Rozkład wartości *p* w literaturze

Inny matematyczny sposób na wykazanie, że literatura może zawierać nadmiar wątpliwych danych polega na analizie rozkładu wartości *p* w artykułach empirycznych (*p-curve analysis*). Założenia takiej analizy przedstawili Simonsohn, Nelson i Simmons (2014). Podstawowe założenie, potwierdzone w matematycznych symulacjach, jest takie, że rozkład wartości *p* dla niezerowych efektów jest prawoskośny. Im mocniejszy efekt, tym będzie bardziej prawoskośny. Mówiąc najprościej, jeśli dany efekt rzeczywiście występuje, to częściej będziemy spotykali badania z wartościami $0 < p < 0{,}01$ niż $0{,}04 < p < 0{,}05$, a te częściej niż $0{,}16 < p < 0{,}17$. Jeśli żaden efekt nie występuje (hipoteza zerowa jest prawdziwa), wówczas jest tak samo prawdopodobne, że uzyskamy $0 < p < 0{,}01$, jak i $0{,}04 < p < 0{,}05$ (i prawdopodobieństwo to wynosi w każdym wypadku 1%).

Autorzy przetestowali swoją hipotezę empirycznie, losując dwie grupy badań z JPSP. Pierwsza grupa, zdaniem autorów „prawdopodobnych", to badania, w których testowano proste efekty beż żadnych dodatkowych współzmiennych i efekty te okazały się istotne. Druga wylosowana grupa to badania, w których były istotne tylko w interakcji ze współzmiennymi (np. płcią). Wyodrębnianie współzmiennych samo w sobie nie jest błędem, szczególnie jeśli ma to dobre uzasadnienie teoretyczne. Jednak testując dowolną zależność, mamy większą szansę, że znajdziemy jakieś istotne efekty, jeśli w analizach uwzględnimy współzmienną (Simmons i in., 2011). Dla danych ze współzmiennymi i bez nich zaobserwowano istotnie różny rozkład (tabela 3). W pierwszym przypadku statystyka *p* najczęściej przyjmowała wartość od 0,04 do 0,05, w przypadku danych bez współzmiennych najwięcej danych notowano dla wartości $p < 0{,}01$. Rozkład wartości *p* dla prostych efektów był bardzo zbliżony do rozkładu wartości *p* dla hipotetycznego rozkładu danych o mocy $d = 0{,}33$ (por. Simonsohn i in., 2014, rysunek 3). Dla danych potencjalnie rzetelnych wzór rozkładu *p* jest prawoskośny, oprócz wyników tuż poniżej wartości *p*. Może to wskazywać na takie stosowanie QRP, żeby sprowadzić dane do pożądanej wartości $p < 0{,}05$.

Podobną koncepcyjnie analizę przeprowadzili Masicampo i Lalande (2012). Wyodrębnili wartości *p* w przedziale $0{,}01 < p < 0{,}1$ w artykułach empirycznych z trzech znaczących czasopism[7] z 12 numerów z roczników 2008 i 2007.

Tabela 3
Rozkład częstości wartości p *w badaniach potencjalnie nastawionych na łowienie* p *(tj. ze współzmiennymi) i danych bez współzmiennych*

	$0{,}0 < p < 0{,}01$	$0{,}01 < p < 0{,}02$	$0{,}02 < p < 0{,}03$	$0{,}03 < p < 0{,}04$	$0{,}04 < p < 0{,}05$
Dane potencjalnie wątpliwe (ze współzmiennymi)	5%	5%	15%	35%	40%
Dane potencjalnie rzetelne (bez współzmiennych)	45%	23%	14%	5%	14%

Adnotacja. Dane w Simonsohn i in. (2014)

[7] *Journal of Experimental Psychology*: *General, JPSP, Psychological Science.*

Ogółem wyodrębniono ponad 3500 wartości *p*. Analiza ich rozkładu wykazała, że miał on kształt wyraźnie prawoskośny [o kształcie zbliżonym do funkcji wykładniczej y = 1/(2^x)], jednakże rozkład nieco poniżej wartości 0,05 wyraźnie odstawał od reszty. Takich wartości *p* było nieproporcjonalnie dużo niezależnie, jakie czasopismo analizowano. Autorzy przyznają, że wadą ich analizy jest brak kontroli zależności danych – wiele wartości pochodzi od tych samych autorów i z tych samych artykułów. Największą wadą tej analizy wydaje się jednak wrzucenie do jednego worka wszystkich możliwych wartości *p* znalezionych w artykułach. Wiele z nich trafia do tekstów w wyniku konwencji (żeby np. zaraportować wszystkie korelacje między skalami kwestionariuszy), dlatego nie mają większego teoretycznego znaczenia. Trudno spodziewać się tu jakichś przekłamań. Można natomiast oczekiwać, że tylko bardzo niewielka część wartości *p* dotyczy kluczowych hipotez i byłoby interesujące otrzymać zbliżoną analizę, ale tylko dla wartości *p* z testowania hipotez, które są kluczowe z teoretycznego punktu widzenia.

Wyniki Masicampo i Lalande (2012) zostały powtórzone w badaniu Leggetta, Thomas, Loetschera i Nichollsa (2013), wykonanym na tych samych czasopismach. Sprawdzali oni również zmianę na przestrzeni czasu, porównywali roczniki czasopism z 2005 i 1965 roku. W nowszych artykułach efekt nadreprezentacji *p* tuż poniżej 0,05 był wyraźniejszy. Szczególnie w JPSP zanotowano wzrost takich danych na przestrzeni dekad. Podobnie Kuhberger i in. (2014) zaobserwowali, że w analizowanej grupie 395 badań występował gwałtowny wzrost badań z wynikami nieznacznie poniżej $p = 0,05$ w porównaniu do wyników nieznacznie powyżej tej wartości.

Systematyczna analiza rozkładu wartości *p* w literaturze jest stosunkowo nową techniką i oprócz wymienionych artykułów nie zidentyfikowałem większej liczby tekstów jej używających. Niemniej jednak wszystkie 4 artykuły wskazują na statystycznie nieprawdopodobną anomalię w takich rozkładach, a więc pośrednio na manipulacje danymi i/lub wybiórcze publikowanie. Jest przy tym możliwe, że często przyczyna jest dużo prostsza: autorzy zwyczajnie podmieniają w tekście wartość *p* (bez zmiany innych statystyk). Tej praktyce jest poświęcona następna sekcja.

Nieprawidłowości w raportowaniu wartości *p* w literaturze

O nierzetelne raportowanie wartości *p* pytali John i in. (2012) i dla przypomnienia prawie co czwarty badacz przyznał się do takich praktyk. Przekłamania polegające na niewłaściwym podawaniu wartości *p* (np. $p < 0,05$, gdy w rzeczywistości $p = 0,057$) dokładniej przeanalizowała grupa holenderskich badaczy. Bakker i Wicherts (2011; badanie 1) wylosowali sześć czasopism psychologicznych, trzy o wysokiej cytowalności (IF > 4) i trzy o relatywnie niskiej[8] (IF < 1,5), a następnie przeanalizowali wszystkie artykuły empiryczne z rocznika 2008. Z nich wyodrębniono wszystkie testy istotności statystycznej, w których używano statystyk chi kwadrat, *F* i *t*. Następnie, uwzględniając podane wielkości tych testów oraz liczbę stopni swobody, wyliczono, czy są one zgodne z wielkościami *p*, jakie podano w artykułach. Okazało się, że blisko połowa artykułów zawierała przynajmniej jedną statystykę z błędami. W większości przypadków były to błędy, które nie zmieniały zasadniczo istoty wyniku (np. $p < 0,01$, gdy w rzeczywistości $p = 0,01$[9]), ale sporadycznie zdarzały się poważniejsze błędy, polegające na niewłaściwym podawaniu wartości *p*, gdyż prawdziwa wartość powodowała zmianę interpretacji wyniku z istotnego na nieistotny lub odwrotnie. Jak można było przewidzieć, w blisko 80% przypadków błąd taki polegał na zaklasyfikowaniu wyniku nieistotnego jako istotnego. Przynajmniej jeden błąd takiego rodzaju miało 17% artykułów z czasopism o wysokiej cytowalności i 19% artykułów z czasopism o niskiej cytowalności.

Biorąc pod uwagę, że w pierwszym badaniu uwzględniono tylko sześć czasopism, autorzy postanowili przeanalizować 300 losowych artykułów ze wszystkich czasopism psychologicznych uwzględnionych w bazie PsycINFO (również rocznik 2008; Bakker, Wicherts, 2011, badanie 2). Wyniki były względnie zbliżone do pierwszego badania, tj. około 35% artykułów zawierało przynajmniej jeden błąd, a w 7% artykułów błąd zmieniał interpretację wyniku.

Wicherts, Bakker i Molenaar (2011) w podobny sposób przeanalizowali artykuły z dwóch uznanych czasopism (konkretnie JPSP oraz *Journal of Experimental Psychology*: *Learning, Memory and Cognition*). Skontaktowali się następnie z autorami z prośbą o podzielenie się surowymi danymi. Tylko 42% autorów przychyliło się do prośby. Pozostali nie odpowiedzieli na zapytania (26%), obiecali to zrobić, ale nie zrobili (24%) lub wprost odmówili podzielenia się danymi ze względu na brak czasu (6%).

W artykułach, których autorzy podzielili się danymi, nie znaleziono błędów polegających na zamianie wartości *p* w taki sposób, że zmieniało to decyzje o przyjęciu lub odrzuceniu hipotezy zerowej (choć znajdowano inne

[8] Te czasopisma to*: Journal of Child Psychology and Psychiatry, Development and Psychopathology, Journal of Personality and Social Psychology* (wysoki IF) oraz *Journal of Black Psychology, Journal of Applied Developmental Psychology, Journal of Research in Reading* (niski IF).

[9] Uwzględniano również to, że podana wartość statystyki *p* może być zaokrągleniem. Jeśli więc autorzy podawali $p = 0,03$, nie traktowano tego jako błąd, jeśli wartość *p* mieściła się w przedziale $< 0,025; 0,035 >$.

drobne błędy). Natomiast w tekstach, w których znaleziono takie błędy, w żadnym przypadku autor nie podzielił się surowymi danymi. Ogółem, gdy autor nie podzielił się danymi, w 25% były to artykuły z poważnymi błędami i w 50% z innymi rodzajami błędów. Ciekawym odkryciem jest też wyliczenie, że autorzy nieudostępniający danych mieli ogólnie „słabsze" dane, tj. wyższe wartości *p*.

Laggett i in. (2013) we wspomnianej już analizie rozkładu wartości *p* dokonywali samodzielnych wyliczeń ich wartości, niezależnie od danych pochodzących od autorów. Odkryli, że w 36 przypadkach na 93 (38%), mimo że raportowano $p = 0{,}05$, w rzeczywistości wartość ta była wyższa. Wszystkie takie wypadki dotyczyły JPSP, przy czym było ich znacznie więcej w 2005 roku niż w 1965 (42 vs. 19).

Wyniki tych analiz są z oczywistych względów niepokojące. Błędne wartości statystyki *p* zostały wykryte na podstawie danych opublikowanych w artykułach przez samych autorów. Jest to wyjątkowo proste, a mimo to, jak wynika z opisanych analiz, powszechne fałszerstwo polegające na zmianie wartości *p* przy braku zmiany wielkości statystyk i liczby stopni swobody. Błędy te raczej nie brały się tylko z nieuwagi lub niedbalstwa. Zdecydowana większość błędów, choć nie zmieniała decyzji o istotności wyniku, polegała na zaniżeniu wartości *p* w stosunku do rzeczywistej (por. Bakker, Wicherts, 2011; rysunek 3). Autorzy artykułów z dużymi błędami, zmieniającymi interpretację wyniku, nie dzielili się swoimi surowymi danymi, co może wskazywać na to, że zdawali sobie sprawę, iż ich dane są naciągane. Bakker i in. zauważyli też, że dużej liczby wyników nie można było skontrolować, gdyż nie zawierały kompletu danych (często pomijano liczbę stopni swobody, choć sporadycznie zdarzały się artykuły z samymi wartościami *p*). Ponadto, co zaskakujące, nie znaleziono większych różnic w liczbie niedokładnych wartości *p* między czasopismami o wysokim i niskim prestiżu (mierzonym cytowalnością). W artykułach z czasopism o wysokim prestiżu było dużo mniej błędów polegających na niepełnym raportowaniu danych, co może wskazywać na to, że redaktorzy i recenzenci znają i potrafią wymusić odpowiednie standardy. Jednak wysokie standardy nie zabezpieczają przed – relatywnie prostymi do wykrycia – błędami w danych statystycznych. Liczba artykułów z tego typu błędami wskazuje też na ograniczoną kontrolę danych przez redaktorów i recenzentów. Tego typu analizy oczywiście nie udzielają przy tym odpowiedzi na pytanie, w jakiej części artykułów wartość *p* została sprytniej zmieniona przez równoległe sfałszowanie wartości statystyk.

Replikacja danych w psychologii

Uzupełnieniem tych rozważań powinna być próba oszacowania częstości replikacji w psychologii. Wszystkie problemy „psychologii fałszywej pozytywnej" nie miałyby żadnego praktycznego znaczenia, gdyby w psychologii regularnie powtarzano badania. Tak się raczej nie dzieje. Przykładowo, JPSP nie publikował do tej pory replikacji (słów *replication* lub *replicate* w tytule artykułu w tym czasopiśmie użyto zaledwie 28 razy, z czego 22 przypadki pochodzą z lat 1965–1990). W ostatnich miesiącach jednak zmienia politykę, więcej o tym w następnej sekcji.

Interesującą systematyczną próbę zbadania częstości replikacji w psychologii podjęli Makel, Plucker i Hegarty (2012). Automatycznie przeanalizowano tekst wszystkich artykułów ze 100 czasopism o najwyższym IF od roku 1900 lub od pierwszego numeru. Szukano słów o rdzeniu *replicat*. Takie słowa pojawiły się w 1,6% artykułów. Samo użycie słowa znaczy jeszcze niewiele, i może ono pojawiać się w różnych kontekstach (np. „przyszłe pokolenia badaczy powinny spróbować zreplikować nasze wyniki…"), dlatego autorzy przeczytali losowe 500 artykułów. Stwierdzono, że 68% z nich to rzeczywiście replikacje, ostateczny wskaźnik replikacji skorygowano więc do 1,07%. Jakkolwiek należy dodać kilka zastrzeżeń. Około 30% replikacji były to replikacje w tym samym artykule, a 20% były to replikacje wykonane przez autora/-ów oryginalnego badania. Co oczywiste, nie są to replikacje uznawane przez społeczność badaczy za najbardziej wartościowe, trudno bowiem oczekiwać, żeby autorzy wysyłali do czasopism sprzeczne dane w obrębie jednego manuskryptu albo żeby podkopywali później swoją pracę, publikując nieudane replikacje. Istotnie, jak się okazało spośród replikacji wykonanych przez tego samego autora odsetek udanych wynosił przeszło 90%, natomiast jeśli wykonywał ją niezależny badacz, wówczas odsetek spadał do 64%.

Autorzy przeanalizowali niewątpliwie potężny korpus wiedzy, choć można się zastanawiać, na ile przeszukiwanie tekstu pod kątem jednego słowa kluczowego jest trafne. Być może część badaczy nie używa tego słowa albo nieświadomie powtarza badania o zbliżonych hipotezach, które zostały już przeprowadzone. Nie można też wykluczyć umyślnego niewspominania o wcześniej przeprowadzonych podobnych badaniach, żeby dodać swojemu odkryciu nimbu nowości. W przypadku doboru słów jedyna alternatywa, jaka przychodzi na myśl to *repetition* (czasami też niektóre badania określano jako *repeat study*, ale to raczej w starszej literaturze). Sytuacje nieświadomej replikacji mogą występować, nikt nie zna całej literatury, a prowokacja Petersa i Ceci (1982) pokazała, że redaktorzy nie potrafili nawet rozpoznać artykułów sprzed kilku lat z własnych czasopism. Tak czy owak, nawet gdyby replikacji było pięć razy więcej niż wynika ze wskazań zespołu Makel, byłoby ich ciągle relatywnie niewiele na tle wszystkich badań.

Autorzy (Makel i in., 2012) nie rozróżnili niestety w swojej analizie replikacji dokładnych i konceptualnych. Choć omawiają we wprowadzeniu to rozróżnienie, to jednak w sekcji *Metoda* ich artykułu nie ma informacji o tym, jakie były kryteria uznania artykułu za replikację. Przypuszczam, że liczba replikacji konceptualnych jest dużo wyższa niż ów 1% (por. np. Wojciszke, 2006), nawet jeśli badacze nie opisują swoich badań tymi słowami. Interesujące byłoby też zbadanie tekstów spoza setki najbardziej prestiżowych czasopism, choć moje osobiste wrażenie jest takie, że tzw. słabe czasopisma nie mają większej liczby replikacji.

W każdym razie wydaje się, że na każdy opublikowany artykuł, będący replikacją, przypada kilkadziesiąt tekstów, które replikacjami nie są. Lub, mówiąc inaczej, każde pojedyncze odkrycie ma znikome szanse na opublikowane replikacje. Co więcej, nawet gdy sporadycznie replikacje się ukazują, są to częściej replikacje udane. Prawdopodobnie redaktorzy niechętnie publikują nieudane replikacje, a łaskawszym okiem patrzą na udane replikacje konceptualne.

Niedawno ukazał się w *Science* raport z badania polegającego na dokładnej replikacji stu efektów psychologicznych (i jest to bodaj pierwsze w historii nauki tego rodzaju przedsięwzięcie: Open Science Collaboration, 2015). Replikowane efekty były pierwotnie opisane w rocznikach 2008 trzech prestiżowych czasopism (*Psychological Science, Journal of Personality and Social Psychology* oraz *Journal of Experimental Psychology: Learning, Memory and Cognition).* Dbano o to, żeby replikacje miały wysoką moc (zwykle wyższą niż oryginalne badanie). Międzynarodowy zespół kilkuset badaczy pod kierunkiem Briana Noseka ostatecznie zreplikował z sukcesem około jednej trzeciej badań. Część środowiska psychologów społecznych sceptycznie przyjęła ten raport (por. np. Wójcik, Cisłak, Doliński, 2015), jednak w moim przekonaniu nie można go zignorować. W powszechnym przekonaniu opublikowany raport z badania w dobrym czasopiśmie to gwarancja, że opisywany efekt jest wysoce prawdopodobny. Tymczasem okazuje się, że to prawdopodobieństwo w psychologii jest jednak dosyć niskie. Jest wręcz większe prawdopodobieństwo, że efektu nie da się powtórzyć! Raport ten sam w sobie byłby wysoce niepokojący a w kontekście opisywanych w literaturze szeregu nieprawidłowości jego wymowa jest jeszcze mocniejsza. Zaryzykowałbym tezę, że stanowi on ostateczny dowód na głęboki kryzys typowych praktyk badawczo-publikacyjnych w psychologii.

Dyskusja nad zmianami w psychologii

Oszustwo Stapela oraz omawiane we wcześniejszych sekcjach niepokojące wskaźniki zainicjowały szeroką dyskusję o potrzebie znaczącej zmiany praktyk badawczych i publikacyjnych. Omówię teraz najważniejsze wnioski z tych dyskusji oraz wskażę interesujące sygnały gotowości do zmian.

Nie sądzę, żeby istniały absolutnie skuteczne sposoby wykrywania fałszerstwa i przekłamań na etapie recenzji artykułów. U Stapela znaleziono post factum wiele nieprawidłowości, ale można sobie wyobrazić badacza, który ma większą matematyczną świadomość, więc do tworzenia danych używa np. programów generujących liczby pseudolosowe o zadanych wartościach. Dlatego w ostatecznym rozrachunku mechanizmem kontroli nierzetelności w nauce (i nieświadomego błądzenia) jest mechanizm replikacji dokonywanych przez niezależnych badaczy. Duża część głosów dotyczy więc zwiększenia roli replikacji w psychologii (Asendorpf i in., 2013; Francis, 2014; Nosek, Spies, Motyl, 2012), oczywiście replikacji dokładnych, gdyż replikacji konceptualnych, wydaje się, jest niemało. Nie są to nowe postulaty. W historii psychologii istniały przynajmniej dwa czasopisma poświęcone publikacjom replikacji (*Replications in Social Psychology*, *Representative Research in Social Psychology*). Żadne z nich nie odniosło większego sukcesu i dzisiaj już nie istnieją. Inne czasopismo, *Journal of Articles in Support of the Null Hypothesis,* publikuje rocznie 1–6 artykułów (Giner-Sorolla, 2012). Łatwiej więc powiedzieć niż zrobić. Jak powszechnie wiadomo, „system" nie zachęca badaczy do systematycznych replikacji (mniejsza szansa na publikacje, wymóg nowatorstwa ze strony instytucji przyznających granty itp.), z drugiej wydaje się, że same czasopisma nie patrzą na nie przychylnie, zakładając, że artykuły replikacyjne będą miały średnio niższą cytowalność niż artykuły przedstawiające „odkrycie".

Pewne wyłomy zostały jednak już zrobione. Czasopismo *Perspectives on Psychological Science* (piąte najczęściej cytowane czasopismo ogólne) ogłosiło, że otwiera sekcję poświęconą rejestrowanym replikacjom (Association for Psychological Science, 2013). Opublikowane wcześniej badania, które zostaną zakwalifikowane do programu, mają podlegać specjalnej procedurze. Najpierw ma powstać dokładny protokół badania, który zostanie udostępniony w internecie. Każdy badacz będzie mógł zgłosić akces do projektu, w zamian za co będzie mógł liczyć na współautorstwo publikacji. Wyniki niezależnych zespołów zostaną zestawione za pomocą metod metaanalitycznych, a artykuł ma zostać opublikowany niezależnie od wyników (autorzy oryginalnego badania będą zaproszeni do współuczestnictwa). Jak podkreśla redakcja, nacisk zostanie położony na siłę efektu, a nie na dychotomiczne decyzje „replikacja udana–nieudana", gdy kryterium sukcesu jest tylko istotność statystyczna wyników.

Największą rewolucją jest bodaj ogłoszona w połowie 2014 roku decyzja JPSP o publikacji replikacji. Specjalne artykuły replikacyjne mają być publikowane tylko w wersji elektronicznej, co jednak wydaje się nie mieć większego znaczenia, biorąc pod uwagę malejące znaczenie wydań drukowanych. Podkreślono, że preferowane są replikacje dokładne, a nie konceptualne, replikacje kilku badań, a nie jednego oraz replikacje przeprowadzone przez niezależnych badaczy (nie są to warunki niezbędne). Głównym kryterium decydującym o publikacji ma być znaczenie replikowanego badania oraz odpowiednio wysoka moc. Analogiczne zasady wprowadziło kolejne ważne czasopismo APA *Journal of Experimental Psychology: General.* Inne znaczące czasopismo z dziedziny psychologii społecznej, *Journal of Experimental Social Psychology,* w nowej polityce publikacyjnej również zapowiedziało gotowość publikacji replikacji.

Pojawiają się też sugestie, że czasopisma powinny otworzyć swoje łamy nie tylko na replikacje, lecz także na wyniki negatywne (nieistotne statystycznie; np. Nosek, Bar-Anan, 2012; Schimmack, 2012). Dosłowne spełnienie tego postulatu budzi we mnie pewne obawy. Już w tej chwili liczba nowych danych w dowolnej subdyscyplinie psychologii jest monstrualna. Dołożenie do tego bliżej nieokreślonej liczby badań bez żadnych istotnych efektów z pewnością nie ułatwi poruszania się po literaturze. Uważam, że bardzo rozsądny kompromis zaproponowały dwa czasopisma: *Attention, Perception and Psychophysics* (Wolfe, 2013) oraz *Cortex* (Chambers, 2013). Zadeklarowały one utworzenie sekcji z rejestrowanymi badaniami (w tym replikacjami). Zasadnicza ocena manuskryptu odbywać się będzie przed zebraniem danych empirycznych. Recenzenci mają otrzymywać artykuły zawierające tylko wstęp teoretyczny i metodę. Mają oni ocenić, na ile badania są sensownie zaprojektowane, teoretycznie uzasadnione i poznawczo interesujące (w przypadku replikacji oceniają ponadto, czy istnieją mocne przesłanki, żeby wątpić w rzetelność oryginalnego badania). Decyzja będzie też uwarunkowana spodziewaną mocą badania, która powinna wynosić minimum 90%. Jeśli rozstrzygnięcie będzie pozytywne, autorzy mają rok na przeprowadzenie badań i przesłanie kompletnego manuskryptu, a wspomniane czasopisma mają zamiar je opublikować niezależnie od wyników. Jak wspomniałem, jest to rozsądny kompromis między publikowaniem wszystkiego a rozpowszechnianiem tylko istotnych wyników (występującym obecnie). Z nieskończonej liczby możliwych kombinacji zmiennych, jakie można przetestować, skończona ich liczba jest faktycznie interesująca teoretycznie i praktycznie. Czasopisma więc będą mogły wybrać to, co jest dla nich ciekawe. Mają gwarancję raczej rzetelnych wyników, bo badacz, po wstępnej akceptacji, nie ma już żadnego interesu w ich koloryzowaniu. Dla badacza również jest to korzystna sytuacja, gdyż może się skupić na zaprojektowaniu jednego dużego badania i sensownym jego uzasadnieniu. Jeśli zostanie ono zaakceptowane przez redakcję, ma gwarancję, że nie zmarnuje wielu miesięcy/lat na zbieranie danych, których nigdzie nie będzie mógł opublikować. Mam nadzieję, że ta praktyka zostanie zaadaptowana powszechnie, gdyż, niestety, te dwa czasopisma obejmują tylko dwie dosyć wąskie subdyscypliny psychologiczne.

Alternatywnie, w przypadku już ukończonych badań zaproponowano procedurę *result blind review* (Greve, Bröder, Erdfelder, 2013). Recenzenci (niekoniecznie wszyscy) otrzymywaliby artykuły bez sekcji wyników i tylko na tej podstawie dokonywali oceny. Zakładając, że istotność wyników jest ważnym kryterium atrakcyjności tekstu, procedura ta powinna zmniejszyć wagę tego kryterium przy decydowaniu o dopuszczeniu do druku. Z tego, co wiem, jednak żadne czasopismo nie zdecydowało się na razie na takie rozwiązanie. Co więcej, koncept istotności statystycznej, krytykowany przez statystyków od bardzo dawna (Huberty, 2002), zdaje się powoli odchodzić do lamusa. W zaleceniach sformułowanych przez *Society for Personality and Social Psychology* (Funder i in., 2014), redaktorów *Psychological Science* (Cumming, 2014), a wcześniej przez APA (Wilkinson,1999) wyraźnie się postuluje, żeby większy nacisk kłaść na referowanie przedziałów ufności i wielkości efektów (uzasadnienie matematyczno-metodologiczne tego problemu wykracza poza łamy tego artykułu; odsyłam szczególnie do artykułu Cohena, 1994/2006).

Wysuwane są też postulaty, które mają potencjalnie zminimalizować arbitralność analiz prowadzącą do „psychologii fałszywej pozytywnej". Szczególnie sugeruje się pójście śladem medycyny i szerokie prerejestrowanie badań przed ich wykonaniem (Aveyard i in., 2013; Miguel i in., 2014). Jeśli badacz określałby zawczasu, jakie zmienne uwzględni w modelu, jakie hipotezy go interesują, jakimi zasadami będzie się kierował przy analizie (np. przy wyłączaniu danych), wówczas możliwości twórczego odnajdywania istotnych prawidłowości zostałyby ograniczone. Utworzono specjalny internetowy serwis *Open Science Framework* (osf.io), który umożliwia badaczom prerejestrowanie badań. Zwraca się uwagę na to, że znaczna część badań jest w pewien sposób uprzednio rejestrowana, gdyż badacze, zgłaszając wnioski do komisji etycznych, przedstawiają zwykle istotne informacje też o planowanych analizach statystycznych. Dodatkowa rejestracja tych danych w odpowiednich serwisach nie wydaje się dużym obciążeniem.

Chcąc ograniczyć arbitralność decyzji i wybiórczość prezentacji danych, Simmons i in. (2012) postulują, żeby

czasopisma wprowadziły wymóg deklarowania przez autorów, czy przedstawili w raporcie wszystkie zmienne, warunki eksperymentalne, sposoby ustalania wielkości próby oraz sytuacje usuwania przypadków odstających. W każdej sekcji poświęconej metodzie, autorzy mogliby dodawać standardową formułę: „Opisaliśmy tutaj, w jaki sposób ustaliliśmy wielkość próby, w jaki sposób wyłączaliśmy dane, wszystkie grupy oraz wszystkie zmienne wykorzystane w badaniu". W ten sposób, jeśli badacze stosowaliby arbitralne metody zbierania i analizy danych, musieliby skłamać, żeby temu zaprzeczyć (a jest spora psychologiczna różnica między kłamstwem a niemówieniem całej prawdy). Inna grupa (LaBel i in., 2013) utworzyła stronę internetową (psychDisclosure.org; ang. *disclosure* = ujawnianie), która daje autorom tekstów empirycznych możliwość ujawnienia wyżej wspomnianych dodatkowych informacji o ich badaniach, normalnie niewymaganych przez redakcje. Od 2014 roku autorzy chcący publikować artykuły empiryczne w *Psychological Science* muszą dostarczyć powyższe informacje do redakcji. Żeby umożliwić autorom przedstawienie maksymalnie pełnego obrazu badania, sekcje *Metoda* i *Wyniki* nie są zaliczane do limitu znaków w tekście (Association for Psychological Science, 2014).

Pojawia się też regularny postulat znaczącego zwiększenia dostępu do danych surowych, co może sprzyjać wykrywaniu nieprawidłowości i fałszerstw (Miguel i in., 2014; Simonsohn, 2013; Wicherts, Bakker, 2012). Choć formalnie badacze mają obowiązek[10] udostępniać takie dane, okazuje się, że gdy przychodzi co do czego, większość tego nie robi (Wicherts, Borsboom, Kats, Molenaar, 2006). Trudno się temu dziwić. Analiza danych dokonana przez niezależnego badacza może w najgorszym wypadku zakwestionować rzetelność badacza. Można też twórczo odczytać z nich coś nowego, w takim wypadku jednak splendory spadną na kogo innego. Badacze niestety rozumieją tę logikę. Simonsohn (2013) zauważył: „Uszkodzenia dysków twardych, skradzione laptopy, uszkodzone pliki, zepsucie serwera, niekompatybilność oprogramowania, niechlujne kartoteki zdarzają się nazbyt często, czasopisma powinny więc dbać o dodatkowe kopie danych" (s. 1876). Postuluje się, żeby surowe dane były wysyłane razem z manuskryptem, a dane te powinny być później ogólnie dostępne (z wyjątkiem sytuacji szczególnych). Praktycznie wszystkie najważniejsze czasopisma mają wersje elektroniczne, więc jak się wydaje, istnieją wszelkie możliwości techniczne, żeby czasopisma udostępniały dane w internecie. Istnieje wiele repozytoriów, w których można już teraz zdeponować i/lub udostępniać dane (np. datadryad.org; easy.dans.knaw.nl; zenodo.org).

Omówię na koniec krótko zmiany zadeklarowane przez niektóre inne czasopisma. *The Journal of Social Psychology* (Grahe, 2014) i *Psychological Science* (Association for Psychological Science, 2014) zaczęły zachęcać autorów do udostępniania nie tylko danych, ale też materiałów niezbędnych do przeprowadzania badań oraz rejestrowania badań przed ich wykonaniem. Na pierwszej stronie artykułów pojawiły się kolorowe znaczki (*badges*) informujące o tym, czy badacz wykonał te zalecenia. *Archives of General Psychology*, nowe czasopismo APA typu *open access* (Cooper, VandenBos, 2013) wymaga, aby badacz udostępniający artykuł zdeponował dane w otwartym repozytorium (podobne wymogi ma *PLoS One*, czołowe czasopismo *open access*). Wiele zaleceń dotyczących zmian w praktykach publikacyjnych wydała organizacja *Society for Personality and Social Psychology* (Funder i in., 2014). Wydawane przez nią czasopismo *Personality and Social Psychology Bulletin* zaczęło wymagać m.in. udostępnienia przez autorów wszystkich materiałów wykorzystanych w badaniu. Czasopismo *Social Psychology* wydało specjalny numer poświęcony tylko replikacjom badań. Zestaw zaleceń dotyczących publikacji wydał *Journal of Experimental Social Psychology* (JESP, Editorial Guidelines, 2014). Autorzy są m.in. zachęcani do szczerego opisywania procesu generowania hipotez, unikania „łowienia *p*" oraz udostępniania wszystkich istotnych materiałów i informacji. Powstało też czasopismo w zupełnie nowej formule *Journal of Open Psychology Data* (Wicherts, 2013). Jego zasadniczym celem jest publikacja danych surowych, tekst artykułu jest tylko dodatkiem do nich i ma przede wszystkich służyć ich objaśnieniu. Możliwe jest w nim publikowanie danych z artykułów wcześniej opublikowanych w innych czasopismach.

Podsumowanie

Około rok po sprawie Stapela wykryto kolejne dwa przestępstwa przeciwko rzetelności w psychologii (niestety, też dotyczyły psychologów społecznych). Uri Simonsohn (2013) z Uniwersytetu Pensylwanii zwrócił uwagę na anomalie w danych opublikowanych przez Smeestersa (Smeesters, Liu, 2011) oraz Sannę (Sanna, Chang, Miceli, Lundberg, 2011) na łamach *Journal of Experimental Social Psychology* (pismo podobnej klasy, jak JPSP). Dane charakteryzowały się nadzwyczajną „gładkością", i to mimo relatywnie niewielkich prób. U Smeestersa wszystkich sześć porównań grup eksperymentalnych wykazywało istotne różnice, a u Sanny średnie w grupach różniły się znacząco, miały jednak niezwykle podobne odchylenia standardowe. Simonsohn za pomocą serii symulacji matematycznych

[10] Ten obowiązek określają wymogi redakcyjne czasopism oraz punkt 8.14 kodeksu etycznego APA.

wykazał skrajne nieprawdopodobieństwo wystąpienia takiego wzoru danych. Odbiegały one także od typowych danych uzyskiwanych w podobnych badaniach. Po śledztwie zarówno Smeesters, jak i Sanna zrezygnowali ze stanowisk na uniwersytetach (Yong, 2012). Jest to skandal zupełnie innej miary niż Stapela (któremu wycofano przeszło 60 tekstów), jednak kolejny raz wskazuje na to, że istnieją problemy w wykrywaniu oszustwa, które potencjalnie jest wykrywalne już na poziomie *peer-review*.

W badaniu ludzkiego zachowania prawdopodobnie nie można uzyskiwać tak „czystych" i teoretycznie jednoznacznych wyników, jak w dojrzalszych naukach przyrodniczych (choć i tam nie zawsze występują). Badane efekty są złożone, na ogół słabe, wchodzą w rozmaite interakcje wyższego stopnia (Meehl, 1978), nie wspominając o niskiej precyzji narzędzi pomiarowych i całej gamie potencjalnych artefaktów. Mimo to trudno oprzeć się wrażeniu, że w psychologii w ostatnich kilkunastu latach nasila się pościg za „nowatorstwem". Czołowe czasopisma jeszcze do niedawna wykazywały gotowość do publikowania tylko „badań wykraczających poza współczesny stan wiedzy", podobnie instytucje przyznające granty. Równolegle oszustwo Stapela, Sanny, Smeestersa i przywołane tutaj analizy pokazują, że recenzenci przepuszczają regularnie teksty z nieprawdopodobnym wzorem danych. Tekstom tym nie można jednak odmówić jednego: są „nowatorskie", atrakcyjne i zasadniczo przedstawiają tylko dowody na wsparcie wyjściowych hipotez.

Można wyobrazić sobie zmiany czysto techniczne, polegające na tym, że manuskrypty analizują zawodowi statystycy. W takim wypadku jednak „wyścig zbrojeń" doprowadziłby prawdopodobnie tylko do tego, że „łowiący *p*" zainwestowaliby w swoją edukację statystyczno-matematyczną (np. tworzyli dane przy użyciu generatorów liczb pseudolosowych o zadanych wartościach). Zasadnicza zmiana musi polegać na przewartościowaniu tego, co jest naprawdę ważne w naszej nauce. W moim przekonaniu ważniejszy jest wysiłek badawczy od istotności statystycznej wyników (oczywiście przy założeniu, że pytanie badawcze jest znaczące i metoda sensowna), ważniejsza jest prawdziwość od atrakcyjności manuskryptu, ważniejsze są niezależne replikacje niż rzekome nowatorstwo. Jesteśmy dopiero na początku drogi, ale w ciągu ostatniego roku dokonała się mała rewolucja, wiele czasopism, w tym kilka bardzo ważnych, zadeklarowało daleko idące zmiany w polityce publikacyjnej. Jest to najlepszy dowód na to, że dotychczasowy model badawczo-publikacyjny był w dużej części wadliwy. Szkoda, że potrzeba było Stapela, żebyśmy to w pełni sobie uświadomili. Z drugiej jednak strony być może bez szoku, jakim było jego oszustwo, nie udałoby się zainicjować tak szerokich zmian.

Gdy kończyłem pierwszą wersję tego manuskryptu, konkluzja była negatywna. Sytuacja jednak rozwija się tak dynamicznie w ostatnich miesiącach, że wierzę, iż jeśli zmiany się utrzymają, jesteśmy na drodze ku lepszej nauce.

LITERATURA CYTOWANA

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K. i in. (2013). Recommendations for increasing replicability in psychology: *European Journal of Personality, 27* (2), 108–119.

Association for Psychological Science (2013). *Registered replication reports.* Pobrano z: https://www.psychologicalscience.org/index.php/replication (1.10.2015).

Association for Psychological Science (2014). *2014 Submission Guidelines.* Pobrano z: https://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions (1.10.2015).

Aveyard i in. (2013). *Trust in science would be improved by study pre-registration*. Pobrano z: http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration.

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., Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678.

Balcetis, E., Dunning, D. (2010). Wishful seeing more desired objects are seen as closer. *Psychological Science*, *21* (1), 147–152.

Baumeister, R., Vohs, K., Funder, D. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*, 396–403.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times.* Pobrano z http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?_r=0.

Bones, A. K. (2012). We knew the future all along scientific hypothesizing is much more accurate than other forms of precognition. A satire in one part. *Perspectives on Psychological Science*, *7* (3), 307–309.

Brzeziński, J. (2012). Co to znaczy, że wyniki przeprowadzonych przez psychologów badań naukowych poddawane są analizie statystycznej? *Roczniki Psychologiczne*, *15* (3), 7–39.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

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

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

Cooper, H., DeNeve, K., Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2* (4), 447–452.

Cooper, H., VandenBos, G. R. (2013). Archives of scientific psychology: A new journal for a new era. *Archives of Scientific Psychology*, *1* (1), 1–6.

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

Elliot, A. J., Niesta Kayser, D., Greitemeyer, T., Lichtenfeld, S., Gramzow, R. H. i in. (2010). Red, rank, and romance in women viewing men. *Journal of Experimental Psychology: General*, *139* (3), 399–417.

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

Francis, G. (2012a). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review*, *19*, 151–156.

Francis, G. (2012b). The same old New Look: Publication bias in a study of wishful seeing. *I-Perception*, *3* (3), 176–178.

Francis, G. (2012c). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences*, *109*, 1587.

Francis, G. (2012d). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review*, *19* (6), 975–991.

Francis, G. (2012e). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7* (6), 580–589.

Francis, G. (2013). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*, 292–296.

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

Fritz, A., Scherndl, T., Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory and Psychology*, *23* (1), 98–122.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., West, S. G. (2014). Improving the dependability of research in personality and social psychology recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*, 3–12.

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice i in. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology*, *92*, 325–336.

Galak, J., Meyvis, T. (2011). The pain was greater if it will happen again: The effect of anticipated continuation on retrospective discomfort. *Journal of Experimental Psychology: General*, *140* (1), 63–75.

Giner-Sorolla, R. (2012). Will we march to utopia, or be dragged there? Past failures and future hopes for publishing our science. *Psychological Inquiry*, *23* (3), 263–266.

Grahe, J. E. (2014). Announcing open science badges and reaching for the sky. *The Journal of Social Psychology*, *154* (1), 1–3.

Greve, W., Bröder, A., Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*, *18* (4), 286–294.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62* (2), 227–240.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19* (5), 640–648.

JESP Editorial Guidelines (2014). Pobrano z: www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-editorial-guidelines (1.10.2015).

John, L. K., Loewenstein, G., Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23* (5), 524–532.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2* (3), 196–217.

Klebaniuk, J. (2012). Profesor Stapel na dopingu. O upiększaniu psychologii społecznej. *Psychologia Społeczna*, *7*, 213–217.

Kuhberger, A., Fritz, A., Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*, e105825.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A. i in. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, *8* (4), 424–432.

Leggett, N. C., Thomas, N. A., Loetscher, T., Nicholls, M. R. (2013). The life of *p*: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66* (12), 2303–2309.

Levelt Committee, Noort Committee, Drenth Committee (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Pobrano z: https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf (1.10.2015).

Levine, T. R., Asada, K. J., Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76* (3), 286–302.

Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior*, *38* (5), 631–651.

Little, A. C., DeBruine, L. M., Jones, B. C. (2011). Exposure to visual cues of pathogen contagion changes preferences for masculinity and symmetry in opposite-sex faces. *Proceedings of the Royal Society B: Biological Sciences*, *278* (1714), 2032–2039.

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.

Marszalek, J. M., Barber, C., Kohlhart, J., Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331–348.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Masicampo, E. J., Lalande, D. R. (2012). A peculiar prevalence of *p* values just below.05. *The Quarterly Journal of Experimental Psychology*, *65* (11), 2271–2279.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9* (2), 147–163.

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.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A. i in. (2014). Promoting transparency in social science research. *Science*, *343* (6166), 30–31.

Murayama, K., Pekrun, R., Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, *18,* 107–118.

Nosek, B. A., Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Nosek, B. A., Spies, J. R., Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Norenzayan, A., Hansen, I. G. (2006). Belief in supernatural agents in the face of death. *Personality and Social Psychology Bulletin*, *32* (2), 174–187.

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

Peters, D. P., Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*, 187–255.

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences*, *109* (11), 4086–4091.

Richard, F. D., Bond Jr., C. F., Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7* (4), 331–363.

Rudman, L. A., Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology*, *41* (3), 192–202.

Sanna, L. J., Chang, E. C., Miceli, P. M., Lundberg, K. B. (2011). Rising up to higher virtues: Experiencing elevated physical height uplifts prosocial actions [retracted article]. *Journal of Experimental Social Psychology*, *47*, 472–476.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17 (*4), 551–566.

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

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.

Simmons, J. P., Nelson, L. D., Simonsohn, U. (2012). A 21-word solution. *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology*, *26*, 4–7.

Simonsohn, U. (2012). It does not follow evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in Press). *Perspectives on Psychological Science*, *7* (6), 597–599.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24* (10), 1875–1888.

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.

Slavin, R., Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31* (4), 500–506.

Smeesters, D., Liu, J. E. (2011). The effect of color (red versus blue) on assimilation versus contrast in prime-to-behavior effects [wycofany artykuł]. *Journal of Experimental Social Psychology*, *47*, 653–656.

Spellman, B. (2012). Introduction to the special section: Data, data, everywhere… especially in my file drawer. *Perspectives on Psychological Science*, *7* (1), 58–59.

Stroebe, W., Postmes, T., Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7* (6), 670–688.

Suls, J., Martin, R. (2009). The air we breathe: A critical look at practices and alternatives in the peer-review process. *Perspectives on Psychological Science*, *4*, 40–50.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*, *3*.

Vankov, I., Bowers, J., Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67* (5), 1037–1040.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, *14* (5), 779–804.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6* (3), 291–298.

Wicherts, J. M. (2013). Science revolves around the data. *Journal of Open Psychology Data*, *1* (1), 1–4.

Wicherts, J. M., Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence*, *40* (2), 73–76.

Wicherts, J. M., Bakker, M., Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6* (11), e26828.

Wicherts, J. M., Borsboom, D., Kats, J., Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61* (7), 726–728.

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

Wojciszke, B. (2006). Systematycznie modyfikowane autoreplikacje. Logika programu badań empirycznych w psychologii. W: J. M. Brzeziński (red.), *Metodologia badań psychologicznych:*

wybór tekstów (s. 44–68). Warszawa: Wydawnictwo Naukowe PWN.

Wójcik, A., Cisłak, A., Doliński, D. (2015). Psychologia to nie ściema. *Gazeta Wyborcza*, *219,* 43.

Wolfe, J. M. (2013). Registered reports and replications in attention, perception, & psychophysics. *Attention, Perception, and Psychophysics*, *75*, 781–783.

Yong, E. (2012). Uncertainty shrouds psychologist's resignation. Pobrano z: http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968 (1.10.2015).

Zhong, C. B., Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313* (5792), 1451–1452.

Aftermath of the Stapel case: More alarming data, the beggining of change?

Łukasz Budzicz

Institute of Psychology, Adam Mickiewicz University in Poznan

Abstract

This article adds to the discussion that took place in *Psychologia Społeczna* (no. 3/2012) concerning the reliability of research findings in social psychology. In recent years a number of new data indicated that Stapel's fraud might not have been an isolated case, but a symptom of a larger crisis. Even if most researchers do not fabricate data, subtle falsification involving arbitrary data processing and selective presentation of results may be relatively frequent and lead to a distorted picture of reality. Especially telling in this context are analyses, that show improbable distributions of *p*-values (*p*-curve analysis), distortions in reporting *p*-values, and very low cumulative statistical power of research studies. This article presents the most important voices about how to change research and publication practices. Described were examples where such changes have already been initiated.

Keywords: *Stapel's fraud*, *reliability of data in psychology*, *false-positive psychology*, *statistical power*

Złożono tekst: 13.04.2014
Złożono poprawiony tekst: 2.08.2014/ 19.10.2014
Zaakceptowano do druku: 23.11.2014