

Uniwersytet im. Adama Mickiewicza w Poznaniu

Instytut Psychologii

Łukasz Budzicz

**Wartość poznawcza badań empirycznych
w psychologii społecznej**

Pracę przyjmuję:

Rozprawa doktorska napisana pod kierunkiem
prof. dr. hab. Jerzego Brzezińskiego
w Instytucie Psychologii UAM

.....
Podpis promotora, data

Poznań, 2015



SPIS TREŚCI

ABSTRAKT	4
WPROWADZENIE	6
UŻYWANE W TEKŚCIE SKRÓTY.....	10
 ROZDZIAŁ 1. RZETELNOŚĆ I REPLIKOWALNOŚĆ BADAŃ EMPIRYCZNYCH W PSYCHOLOGII	 11
1.1. Afera Stapela i fabrykowanie danych w psychologii.....	11
1.2. Psychologia „fałszywie pozytywna”	21
1.3. Psychologia „fałszywie pozytywna” – próba empirycznego zbadania rozpowszechnienia zjawiska.....	28
1.4. Moc statystyczna badań i wybiórcze publikowanie	34
1.5. Rozkład wartości p w literaturze	50
1.6. Nieprawidłowości w raportowaniu wartości p w literaturze	53
1.7. Dzielenie się danymi	56
1.8. Replikacja danych w psychologii.....	60
1.9. Podsumowanie najważniejszych wniosków o rzetelności danych w psychologii	68
1.10. Skutki kryzysu „poststapelowskiego”	71
1.11. Nierzetelność niezależna od badacza	78
 ROZDZIAŁ 2. PROBLEMY TRAFNOŚCI ZEWNĘTRZNEJ BADAŃ W PSYCHOLOGII SPOŁECZNEJ	 84
2.1. Wprowadzenie	84
2.2. Przykłady alternatywnego podejścia	89
2.3. Problem uteoretyzowania obserwacji.....	93
2.4. Ograniczenia trafności metod samoopisowych	101
2.5. Ograniczenia trafności metod mierzących zmienne utajone	111
2.6. Studenci jako osoby badane – potencjalne ograniczenia trafności	123
2.7. Badanie osób z cywilizacji zachodniej – potencjalne ograniczenia trafności	131
2.8. Hipotetyczne przykłady zwiększania trafności zewnętrznej badań.....	136
 ROZDZIAŁ 3. KUMULACJA WIEDZY W PSYCHOLOGII.....	 142
3.1. „Metakryzys” psychologii – wprowadzenie	142
3.2. Konkretnie przykłady „metakryzysu”	144
3.3. Porównanie dojrzałości psychologii na tle innych nauk (szczególnie przyrodniczych)	155
3.4. Dojrzałość nauki a metoda – argumentacja Rozina	160
3.5. Kumulacja wiedzy w psychologii – wprowadzenie do modelu.....	167
3.6. Model rozwoju psychologii #1 – prawidłowości niskiego i wysokiego poziomu	168
3.7. Model rozwoju psychologii #2 – źródła postępu.....	175

3.8. Model rozwoju psychologii #3 – przedmiot psychologii	180
3.9. Model rozwoju psychologii #4 – „nowatorskie” odkrycia vs. lepsze rozumienie efektu.....	185
3.10. Model rozwoju psychologii – operacjonalizacja modelu	187
3.11. Analiza operacjonalizacji na przykładach	192
ROZDZIAŁ 4. METODOLOGIA BADANIA WŁASNEGO.....	197
4.1. Wprowadzenie	197
4.2. Wybór próby.....	199
4.3. Właściwa analiza.....	201
4.4. Dane surowe badania	206
4.5. Podsumowanie kodowanych zmiennych.....	206
ROZDZIAŁ 5. ANALIZA ILOŚCIOWA WYNIKÓW	208
5.1. Pozyskana próba artykułów do analizy	208
5.2. Opis artykułów zawierających oryginalne badania	209
5.2.1. Liczba cytowań.....	209
5.2.2. Podstawowe charakterystyki cykli badań w zależności od czasopisma	214
5.2.3. Klasyfikacja badania, rodzaj badań nawiązujących oraz rodzaj zmiennej zależnej w zależności od czasopisma.....	218
5.2.4. Czasopismo a potwierdzenie oryginalnej hipotezy i niezależność zespołu badawczego	220
5.3. Opis cykli badań – podstawowe charakterystyki	221
5.3.1. Niezależny zespół	221
5.3.2. Potwierdzenie wyjściowej hipotezy	222
5.3.3. Rok badania	222
5.3.4. Liczba osób badanych	225
5.3.5. Grupa badana	229
5.3.6. Narodowość badanych.....	229
5.3.7. Cel badania, rodzaj badania, sposób manipulacji zmiennymi oraz rodzaj zmiennej zależnej	232
5.4. Związki pomiędzy poszczególnymi zmiennymi	234
5.4.1. Co wpływa na potwierdzenie oryginalnej hipotezy badawczej?	234
5.4.2. Co wpływa na liczbę osób badanych?	237
5.4.3. Co wpływa na wybór grupy badanej (student vs. nie-student)?	240
5.4.4. Co wpływa na wielkość kratek w badaniach eksperymentalnych?	241
5.4.5. Co wpływa na cel badania?	243
5.4.6. Co wpływa na rodzaj badania?	244
5.4.7. Co wpływa na rodzaj manipulacji?	246
5.4.8. Co wpływa na wybór zmiennej zależnej?.....	246
5.5. Zróżnicowanie metod badawczych	248

ROZDZIAŁ 6. ANALIZA JAKOŚCIOWA POSZCZEGÓLNYCH CYKLI BADAŃ.....	253
6.1. Uwagi wstępne.....	253
6.2. Szczegółowa analiza jakościowa poszczególnych cykli badań.....	254
6.3. Podsumowanie.....	320
ROZDZIAŁ 7. PODSUMOWANIE ANALIZY ORAZ WNIOSKI NA PRZYSZŁOŚĆ	323
7.1. Wprowadzenie	323
7.2. Ograniczenia metody	327
7.3. Wnioski – częstość replikacji.....	329
7.4. Wnioski – trafność zewnętrzna (grupa badana)	331
7.5. Wnioski - trafność zewnętrzna (metody badawcze)	333
7.6. Wnioski – ocena zdolności programów badawczych do generowania wiedzy	338
7.7. Wskazania na przyszłość.....	348
Załącznik 1. Wykaz badań uwzględnionych i uzasadnienie nieuwzględnienia niektórych badań w analizie	355
Załącznik 2. Dane surowe	359
LITERATURA	360

ABSTRAKT

W ostatnich latach pojawia się coraz więcej wątpliwości dotyczących rzetelności i trafności badań empirycznych w psychologii, szczególnie psychologii społecznej. W 2011 wyszło na jaw wieloletnie, szeroko zakrojone oszustwo Stapela (później zakwestionowano też rzetelność kilku innych psychologów społecznych). Pojawiły się systematyczne analizy wskazujące na to, że w psychologii rzadko pojawiają się replikacje badań (szczególnie dokładne), badania mają niską moc, literatura jest zdominowana przez wyniki pozytywne (statystycznie istotne). Badacze przyznają się do stosowania metod, które zwiększają prawdopodobieństwo błędu fałszywie pozytywnego (np. wybieranie do publikacji tylko istotnych porównań między grupami eksperymentalnymi). Dodatkowo kwestionowana jest trafność badań, wskazuje się, że choć psychologia jest nauką o zachowaniu, to zachowanie pojawia się niezwykle rzadko, a badanymi są najczęściej studenci z krajów zachodnich.

W rozprawie poprzez systematyczną analizę praktyk badawczych autor chce odpowiedzieć na pytanie na ile te zarzuty są uzasadnione. Postawione są trzy zasadnicze pytania: a) czy istnieje mechanizm intersubiektywnej sprawdzalności odkryć poprzez stosowanie replikacji dokładnych, b) czy badania mają dostateczną trafność zewnętrzną, szczególnie czy efekty psychologiczne są sprawdzane na grupach innych niż studenckie i osobach z krajów innych niż zachodnie, oraz czy zmienne zależne mają naturę behawioralną, c) czy ma miejsce kumulatywny przyrost wiedzy, czy raczej badania są ze sobą luźno powiązane i nie tworzą zintegrowanego systemu wiedzy.

W celu przeprowadzenia analizy wyodrębniono z 4 czołowych czasopism psychologii społecznej (*Journal of Personality and Social Psychology*, *European Journal of Social Psychology*, *Journal of Experimental Social Psychology* i *Personality and Social Psychology Bulletin*) 40 oryginalnych efektów i śledzono badania nad nimi w okresie prawie 13 lat. Dało to łącznie próbę 472 badań.

Analiza pokazała, że w psychologii replikacje dokładne mają miejsce niezwykle rzadko (około 3 % wszystkich badań, z czego większość to replikacje wcześniejszych badań kwestionariuszowych). Badanymi są w zdecydowanej mierze studenci (prawie 80 % prób) lub inne grupy względnie łatwo dostępne (np. studenci MBA, uczniowie

szkół, badani w internecie). Występuje również zdecydowana nadreprezentacja badanych z pewnych krajów. I tak badani z USA stanowili ponad 52 % prób. Badani z 9 najczęściej badanych krajów stanowili ponad 90 % prób.

Dodatkowo rzadko stosuje się zmienne behawioralne jako zmienne wyjaśniane. Najczęściej (76 %) natura zmiennej wyjaśnianej ma charakter samoopisowy. Niezwykle rzadko stosuje się eksperymenty terenowe, lub reanalizuje zarejestrowane ludzkie zachowania lub wytwory. Bardzo rzadko sięga po takie rodzaje badań, gdzie badani mogą w jakiś sposób spontanicznie zareagować. Wydaje się, że nierzadko badacze ograniczają swoje ryzyko, często przeprowadzając stosunkowo niewielkie badania, z których do publikacji są wybierane tylko te z pozytywnymi wynikami.

Trudno również mówić o systematycznym przyroście zintegrowanej wiedzy. W przypadku wielu efektów występują liczne badania nawiązujące, ale są one najczęściej bardzo luźno powiązane z wyjściowym efektem. Nie wyjaśnia się systematycznie rozbieżności między badaniem wyjściowym, a nawiązującymi. Nie ma prób zwiększania precyzji przewidywania zachowania. Nie zanotowano także poprzedzania badań systematycznymi obserwacjami.

Autor w podsumowaniu stawia tezę, że przyczyną obecnego kryzysu jest kultura „sukcesu” (wszystkie wyniki muszą istotnie statystyczne) i „nowatorstwa” (artykuły muszą przedstawiać „odkrycia”). Prowadzi ona do tego, że poświęca się mało uwagi systematycznemu zbadaniu jednego efektu, ale przeskakuje się od jednej do drugiej hipotezy bez gromadzenia największej możliwie wiedzy. Autor przedstawia możliwe przyczyny takiego stanu rzeczy wynikające ze złożonych interakcji między badaczami, redaktorami czasopism i opinią publiczną.

WPROWADZENIE

Ostatnie lata nie były łatwe dla psychologii, szczególnie dla psychologii społecznej. Środowiskiem wstrząsnęła afera Dederika Stapela, który przez lata fabrykował dane i następnie publikował je w najlepszych periodykach (z samego tylko *Journal of Personality and Social Psychology* wycofano 14 artykułów). Afera pokazała dużo więcej niż nierzetelność jednego badacza, ale powiedziała też coś znaczącego o stanie dyscypliny. Okazało się, że możliwe jest systematyczne publikowanie sfabrykowanych danych w najbardziej prestiżowych czasopismach, a społeczność nie jest w stanie wykryć tego, ani poprzez *peer-review*, ani w szczególności poprzez nieudane replikacje takich badań. Równolegle Bem (2011) opublikował w najbardziej prestiżowym czasopiśmie psychologii społecznej serie badań wskazujących na występowanie prekognicji – zjawiska całkowicie kłócącego się z elementarną wiedzą o tym, jak działają zależności przyczynowo-skutkowe w świecie. Zespół niezależnych badaczy nie mógł następnie opublikować w tym samym czasopiśmie nieudanych replikacji tych badań.

Rozpoczęła się szeroka dyskusja w literaturze. Pojawiły się, niestety, głosy, że psychologia jako nauka w dużej mierze nie jest procesem rzetelnego zbierania i publikowania solidnych danych o rzeczywistości, ale jest raczej swego rodzaju „konkurem piękności”, gdzie premiowani są Ci, którzy zaprezentują nieskazitelne wyniki, podparte atrakcyjną narracją. Wskazywano, że społeczność psychologów mogła zapomnieć o tym, że biorąc pod uwagę złożoność przedmiotu badania i stan rozwoju nauki, takie wyniki są na ogół niemożliwe. Redaktorzy i recenzenci jednak do tej pory preferowali przede wszystkim jednak „atrakcyjność” przekazu, nic więc dziwnego, że psychologia wespół z psychiatrią okazały się być „liderem” wśród wszystkich nauk w odsetku opublikowanych „udanych” badań tj. takich gdzie potwierdzono wyjściową hipotezę (Fannelli, 2010). Nie wydaje się, żeby wynikało to z tego, że teorie psychologów są bardziej wyrafinowane i precyzyjne niż teorie fizyków czy chemików.

Powodów do niepokoju jest znacznie więcej. Symulacje matematyczne pokazały, że przy dużej liczbie testowanych zmiennych i odpowiednich zabiegach statystycznych możliwa jest „psychologia fałszywie pozytywna”: z praktycznie każdego zbioru

danych da się „wyciągnąć” jakieś istotne zależności (Simmons, Nelson, Simonsohn, 2011). Równolegle badania ankietowe psychologów pracujących na amerykańskich uczelniach, pokazały, że większość badaczy przyznaje się do stosowania tych sztuczek i nie widzi w nich nic nagannego (John, Loewenstein, Prelec, 2012). Gdy przyjrano się szczegółowo wielu danym okazywało się, że wartości p często są zaokrąglane w niewłaściwy sposób (Wicherts, Bakker, Molenaar, 2011), ich rozkład odbiega od oczekiwanego (Masicambo i Lalande, 2012), moc statystyczna badań jest niska (Bakker, Dijk i Wicherts, 2012), a wśród opublikowanych badań replikacje należą do rzadkości (Makel, Plucker i Hegarty, 2012). Społeczność badaczy ludzkiego zachowania musiała zmierzyć się z pytaniem, na ile korpus wiedzy w ich dziedzinie jest rzetelny.

Głosy podważające rzetelność dołączyły do wcześniejszych wypowiedzi kwestionujących trafność metod używanych przez psychologów (oraz bardzo ograniczony dobór prób). Zdaniem Baumeistera, Vohs i Fundera (2007) psychologia stała się „nauką o samoopisach i ruchach palców” – badane są pośrednie wskaźniki zachowania, szczególnie przy pomocy różnego rodzaju kwestionariuszy i technik opartych na czasach reakcji. Zachowania jako takiego jest mało, i mniej niż było kiedyś. Rozin (2001), idąc za Ashem, twierdzi, że psycholodzy chcą „przyspieszyć czas przesuwając wskazówki zegara”: nadmiernie polegają na eksperymencie, a zupełnie pomijają obserwacyjną i opisową część procesu badawczego, która jest podstawą sukcesu nauk przyrodniczych. Dodatkowo w każdej dekadzie pojawia się kilka analiz wskazujących na ograniczenia związane z wykorzystywaniem głównie studentów do badań, i, co gorsza, wszystko odbywa się ciągle po staremu. Nawet, jeśli sporadycznie sięga się po innych badanych niż studenci, tak czy owak są to osoby z krajów zachodnich, a więc tracimy możliwość opisu całej różnorodności ludzkiego zachowania.

Pierwotnie ta praca miała być poświęcona problemowi trafności badań w psychologii społecznej, jednak ze względu na wydarzenia ostatnich 3 lat, postanowiłem rozszerzyć ją o bardziej elementarne pytanie: na ile publikowane dane są rzetelne, czyli rzeczywiście prawdziwe (na tyle na ile prawda jest osiągalna przy ograniczeniach metod badawczych i ludzkiego umysłu w ogóle). Dlatego w części teoretycznej chciałbym zaproponować model wartości poznawczej danych psychologicznych, składający się z trzech warstw.

Podstawą jest rzetelność danych, a więc to, czy badacze przedstawiają rzeczywiste wyniki, nieoparte na fabrykacji danych, lub ich wygładzaniu i wybiórczym publikowaniu. Probierzem rzetelności jest przede wszystkim możliwość replikacji (rozdział 1).

Nawet, jeśli dane są względnie prawdziwe, i tak mogą niewiele mówić o tym, jak ludzie zachowują się w codziennych kontekstach. Czy krótki eksperyment w którym młody człowiek zagra w brutalną grę komputerową, a następnie wypełni kwestionariusz z pytaniami o odczuwaną złość, będzie rzeczywiście wskazywał na efekt potencjalnego wzrostu agresji (np. bójek), wskutek wielu lat grania (rozdział 2)?

Wreszcie, może się zdarzyć, że dane, które są w miarę rzetelne i w miarę trafne, nie przyczyniają się jednak do systematycznej kumulacji wiedzy, przede wszystkim poprzez rozwój teorii. Mogą być tylko luźno powiązanymi odkryciami, czy mówiąc dosadniej „zbieractwem faktów”, w niewielkim stopniu integrującymi się w większą całość. Postaram się zarysować model, po czym poznamy, że następuje efektywna kumulacja wiedzy, a nie tylko przyrost danych (rozdział 3).

Czytelnikowi należą się też pewne wyjaśnienia terminologiczne. Używam w tytule tej pracy zwrotu „psychologia społeczna”, choć tak naprawdę wolałbym określenie „nauka o zachowaniach ludzi w codziennych kontekstach”. Nie stosuję go tylko dlatego, że jest językowo dosyć niezgrabny. Badania robione w ramach tzw. psychologii społecznej ze wszystkich psychologicznych subdyscyplin są najbliższe codziennemu ludzkiemu zachowaniu. Jednak na pewno nie tylko „psychologia społeczna” wchodzi w skład tej definicji (podobnie jak nie wszystko, co ukazuje się w czasopiśmie ze słowami „psychologia społeczna” w tytule, jest choćby w najbardziej odległy sposób związane z badaniem zachowania ludzi w takich kontekstach). „Psychologia społeczna” jest więc w tym kontekście pewnym skrótem myślowym.

Jestem przy tym mocno przekonany, że to, jakie badania ukazują się pod „szyldem” poszczególnych subdyscyplin jest w dużej mierze płynne i arbitralne. Badania, które ukazują się w czasopiśmie z psychologii społecznej często mogłyby być równie dobrze zaliczone do ekonomii behawioralnej, psychologii osobowości, psychologii ewolucyjnej, psychologii motywacji, rozmaitych obszarów psychologii stosowanej i tak dalej (i vice versa). Dlatego, choć w analizach empirycznych zawartości

czasopism wychodzę od tego, co jest publikowane w czasopismach, które mają „psychologię społeczną” w tytule, to jednak rozwinięcia badań szukam już we wszystkich czasopismach. Nie ograniczam się do jakiejś sztywnej definicji „psychologii społecznej” i w pierwszych rozdziałach gdzieś tam podaję też przykłady z innych subdyscyplin.

Choć część empiryczna tej pracy jest poświęcona psychologii społecznej, to jednak zaznaczam, że w części teoretycznej często cytuję artykuły, które dotyczą psychologii jako całości. Na podstawie własnego rozeznania nie sądzę, żeby inne subdyscypliny psychologii były znacząco różne pod kątem rzetelności i dojrzałości. Jedyne w czym psychologia społeczna może szczególnie się wyróżniać na tle innych subdyscyplin jest że bardzo mocny nacisk na problem trafności zewnętrznej. Na ogół przyjmuje się tutaj, że wyniki odnoszą się do tego, co ludzie robią w swoim codziennym życiu, nawet jeśli nie jest to wprost wyartykułowane i nawet jeśli procedura badania jest odległa od jakichkolwiek codziennych doświadczeń. W innych obszarach dociekań np. badaniu elementarnych procesów poznawczych, trafność zewnętrzna może nie mieć kompletnie żadnego znaczenia. Zwykle pytając o zasadność badań w psychologii społecznej, musimy pytać o przekładalność badań na codzienne sytuacje.

Praca ta nie mogłaby powstać bez licznych wskazówek i inspiracji promotora profesora Jerzego Brzezińskiego. Panu profesorowi jestem też szczególnie winny podziękowania za to, że liberalnie podchodził do moich zwykle zbyt optymistycznych planów dotyczących kończenia kolejnych etapów pracy (nie analizowałem tutaj na ile efekt „planning fallacy” jest rzetelnie udokumentowany w literaturze, tym niemniej na podstawie swoich doświadczeń wierzę, że ten punkt systemu wiedzy psychologicznej jest akurat mocny).

Podziękowania kieruję do koleżanek i kolegów ze studium doktoranckiego za inspirujące dyskusje w trakcie dyżurów (szczególnie Piotrowi Haładzińskiemu i Aleksandrze Bujacz).

Dziękuję rodzicom za wsparcie.

Wreszcie szczególne podziękowania kieruję do mojej żony Małgorzaty, która w tej ciężkiej pracy zawsze miała dla mnie niewyczerpane pokłady ciepła i czułości.

UŻYWANE W TEKŚCIE SKRÓTY

APA – *American Psychological Association*

APS – *Association for Psychological Science*

EJSP – *European Journal of Social Psychology*

IAT - *Implicit Association Test* (test utajonych skojarzeń)

IF – *Impact Factor* (wskaźnik wpływu)

JEP – *Journal of Experimental Psychology*

JEP:G - *Journal of Experimental Psychology: General*

JESP – *Journal of Experimental Social Psychology*

JPSP – *Journal of Personality and Social Psychology*

PB – *Psychological Bulletin*

PS – *Psychological Science*

PSPB – *Personality and Social Psychology Bulletin*

PoPS – *Perspectives on Psychological Science*

QRP – *Questionable Research Practice* (ang. wątpliwe praktyki badawcze)

W części teoretycznej (rozdziały 1-3) liczba cytowań na podstawie Google Scholar, stan na czerwiec 2015 roku.

ROZDZIAŁ 1

RZETELNOŚĆ¹ I REPLIKOWALNOŚĆ BADAŃ EMPIRYCZNYCH W PSYCHOLOGII

1.1. Afera Stapela i fabrykowanie danych w psychologii

Choć dekretowanie „kryzysu” psychologii towarzyszy nam od bardzo dawna (Asch, 1952/1987; Ring, 1967; Elms, 1975; Bevan, 1991; Staats, 1999; Rozin, 2001; Rand i Illardi, 2005), wydaje się jednak, że ostatnie lata są pod tym względem szczególne. W relatywnie krótkim czasie opublikowano szereg tekstów poddających w poważną wątpliwość rzetelność typowych danych w psychologii (cały numer PoPS² vol. 7(6), 2012; Asendorf i in., 2013; Ferguson, 2013; Fiedler, 2011; John, Loewenstein, Prelec., 2012; Kepes i McDaniel, 2013; LeBel i Peters, 2011; Masicambo i Lalande, 2012; Mitchell, 2012; Molenar, 2011; Murayama, Pekrun, Fiedler., 2014; Simmons, Nelson, Simonsohn, 2011; Simonsohn, 2013; Simonsohn, Nelson, Simmons, 2014; Tressoldi, 2012; Yong, 2012a, Wicherts, Bakker i Fiedler, 2011; w literaturze polskiej: Brzeziński, 2012, Klebaniuk, 2012). W czasopiśmie PoPS określono obecną sytuację dyscypliny mianem *crisis of confidence* (Pahsler i Wagenmakers, 2012), co wymownie można przetłumaczyć na język polski zarówno jako „kryzys zaufania”, jak i „kryzys pewności”.

Nie będzie przesadą powiedzieć, że być może najnowszego kryzysu nie byłoby, gdyby nie afera Dederika Stapela, określana nawet przez nie-psychologów „oszustwem na niebywałą skalę” (np. Vogel, 2011). Holenderski „badacz”, uznawany wcześniej za genialne dziecko psychologii społecznej, wyróżniany branżowymi nagrodami, oraz piastujący wysokie stanowiska uniwersyteckie, jak się okazało, publikował przez prawie 15 lat artykuły w najlepszych czasopismach oparte na sfabrykowanych lub mocno wybiórczo przedstawionych danych. Odkrycie przestępczego procederu było – jak na skalę oszustwa – bardzo przyziemne. Kilku współpracowników nabrało wątpliwości, co do dziwnego zachowania Stapela. Miał m.in. chwalić się udanym eksperymentem, z istotnymi wynikami, ale zapytany o

¹ Słowo „rzetelność” jest używane w tym rozdziale w sensie „prawdziwość, wiarygodność”, a nie jako rzetelność w sensie psychometrycznym.

² Skróty nazw czasopism w wykazie skrótów.

różnice międzypłciowe nie był w stanie pokazać danych surowych (por. Bhattacharjee, 2013). Gdyby nie te w gruncie rzeczy mało fortunate „wpadki”, Stapel prawdopodobnie dalej publikowałby 5-10 artykułów rocznie w najlepszych czasopismach psychologicznych.

Komitet powołany do zbadania praktyk Stapela ostatecznie zakwestionował 55 spośród jego 137 artykułów (w niezakwestionowanych był na ogół współautorem nie przeprowadzającym bezpośrednio badań) (Levelt Committee, Noort Committee, Drenth Committee, 2012; dalej: Levelt, 2012). W następnych 10 tekstach, choć ponad wszelką wątpliwość nie udowodniono fałszerstwa, uznano niektóre dane statystyczne za wysoce nieprawdopodobne (Enserink, 2012). Praktycznie wszystkie znaczące czasopisma psychologii społecznej padły ofiarą oszustwa. Po śledztwie komitetu wycofano 14 artykułów z JPSP, 5 z JESP, 7 z EJSP, 5 z PSPB, 3 z *British Journal of Social Psychology*, 2 z PS, jedno z *Science*³, oraz z szeregu innych, mniej znanych czasopism. Stapel nie zaczynał od pełnego oszustwa, polegającego na wyjęciu danych „z kapelusza”. Jak wynika z raportu (Levelt, 2012, rdz. 4.2) na początku rzeczywiście przeprowadzał eksperymenty na studentach (w niektórych „tylko” dokonywał przekłamań, żeby „pomóc” danym). Splendory, jakie na niego spadały, były zbyt przyjemne, by łatwo je porzucić i niczym nałogowiec, coraz głębiej wchodził w przestępczy proceder. W późniejszym okresie swojej kariery już nie dbał o to, żeby jakiegokolwiek badania w ogóle przeprowadzać. Sam Stapel, w książce wydanej po odkryciu oszustwa (o wymownym tytule „Wykolejenie”), pisał „uzależniłem się od cyklu dociekania, badań, odkryć, publikacji i aplauzu.” (Stapel, 2012, za: Enserink, 2012).

Waga afery nie polega rzecz jasna na tym, że w psychologii znalazł się oszust. Skoro są w innych naukach, trudno się spodziewać, żeby psychologia była w tym względzie jakaś szczególna. Doniosłość oszustwa wynika z czegoś innego. Jak już wspominałem, Stapel został nakryty z powodów zasadniczo niezwiązanych z samą nauką, skutek nieporozumień ze współpracownikami. Wydaje się to dobitną oznaką tego, że w psychologii nie wszystko działa tak jak powinno. Ponad 60 artykułów z około dwoma setkami badań z najpoczytniejszych czasopism nie zostało nigdy zakwestionowane dzięki działaniu mechanizmu intersubiektywnej sprawdzalności (replikacji). To znacząco odróżnia aferę Stapela od innych oszustw naukowych,

³ Wyliczenie własne na podstawie danych z PsycINFO.

szczególnie w naukach przyrodniczych (Stroebe, Postmes, Spears., 2012), gdzie wykrycie oszustwa było spowodowane niemożliwością powtórzeń eksperymentów⁴. W fizyce miało miejsce oszustwo podobnego kalibru. Jan Hendrik Schön, swego czasu młoda gwiazda fizyki materiałowej, publikował jeden artykuł co parę tygodni w czasopismach pokroju *Science i Nature*. Jak się okazało, większość danych sfałszował. Wycofano w konsekwencji m.in. 8 artykułów z *Science*, 7 z *Nature* i 6 z *Physical Review*. Pod względem skali jest to więc afery porównywalna do stapelowskiej. Jedno ją jednak odróżnia: oszukańcze praktyki Schöna trwały kilkanaście miesięcy, a nie 15 lat jak u Stapela. Żadne laboratorium nie potrafiło uzyskać materiałów podobnej jakości jak rzekomo przez niego wytwarzane, przez co zaczęto uważnie przyglądać się jego artykułom. Od pierwszych publikacji eksperci potrzebowali niecałe 2 lata, żeby wykryć anomalie w danych i zakwestionować rzetelność jego pracy (Reich, 2009). Natomiast przed wyjściem na jaw oszustwa Stapela nie pojawiła się żadna poważna systematyczna krytyka jego odkryć, oraz nie ukazała się żadna nieudana replikacja jego badań⁵.

Afera Stapela może wskazywać pośrednio na dużo bardziej fundamentalne problemy, a mianowicie na to, że w psychologii odkrycia są ze sobą słabo powiązane teoretycznie, tj. ma miejsce bardziej „zbieractwo” faktów, a nie budowanie dobrze zintegrowanego systemu wiedzy. Nowe odkrycia trudno umieścić na tle szerszego „krajobrazu” teoretycznego, gdyż albo go nie ma, albo jest zbyt ogólny, więc „odkrycia” Stapela nie zaprzeczały żadnym paradygmatom, a wyjaśniające je teorie można było utworzyć ad hoc. Więcej uwagi poświęcę temu problemowi w rozdziale 3.

Warto spojrzeć na to co konkretnie badał, gdyż mówi to wiele o samej dyscyplinie. „Odkrył” przykładowo m.in. takie prawidłowości psychologiczne:

- w bardziej chaotycznym środowisku (np. zanieczyszczonym śmieciami) ludzie w większym stopniu dyskryminują mniejszości (np. siadają dalej takich osób na publicznej ławce). Badanie zostało opublikowane w *Science* (Stapel i Lindenberg, 2011)

⁴ Być może, gdyby Stapel publikował odkrycia znacząco kwestionujące uznane paradygmaty, wówczas jego dane byłyby częściej replikowane i oszustwo łatwiej wykryte. Prekognicja Bema (2011) w relatywnie krótkim czasie doczekała się wielu nieudanych replikacji.

⁵ Wzmianek o takich nieudanych replikacjach stapelowskich badań nie ma w licznych artykułach poświęconych aferze. Podobnie, ja nie mogłem znaleźć w bazie PsycINFO żadnych nieudanych replikacji prac Stapela (ściślej: żadnych dokładnych replikacji).

- po bardzo krótkiej ekspozycji (120 ms) obrazków przedstawiających bodźce wzbudzające strach lub wstręt badani następnie częściej uzupełniali brakujące litery w niepełnych słowach tak, żeby powstałe słowo było zgodne z eksponowanymi bodźcami emocjogennymi oraz raportowali korespondujące emocje. Efekty te nie występowały lub były znacznie słabsze dla ultrakrótkiego czasu ekspozycji (40 ms) (Ruys i Stapel, 2008)
- eksponowanie „zhumanizowanego” środowiska aktywizuje normy z nim związane, ale samo środowisko niekoniecznie. Przykładowo osoby, którym eksponowano zdjęcia restauracji z ludźmi, zgadzały się następnie w wyższym stopniu z normami dotyczącymi kulturalnego jedzenia, ale efekt ten nie występował, gdy pokazywano zdjęcia pustych restauracji oraz w przypadku norm niezwiązanych z jedzeniem. Analogicznie, biblioteka z ludźmi aktywizowała normy związane z zachowaniem ciszy i spokoju (Joly, Stapel, Lindenberg, 2008)
- u badanych studentów eksponowanie informacji o studencie odnoszącym sukcesy („zagrożającym”), choć obniża poczucie własnej wartości, podnosi zarazem poziom wykonania testu mierzącego kreatywność. Samoafirmacja paradoksalnie znosi ten efekt (Johnson i Stapel, 2007)

Wszystkie opisane badania okazały się sfalszowane i zostały wycofane. Ich lektura jest wysoce doświadczeniem dającym wiele do myślenia. Badania te nie są w żaden sposób szczególne. Nie odstają od typowego artykułu z solidnego czasopisma. Nie opisują jakichś przełomowych lub sensacyjnych odkryć, szczególnie takich, które głęboko kłóć się z dotychczasową wiedzą (tak jak Bem, 2011). Są napisane w poprawny, dosyć atrakcyjny sposób, hipotezy są przyzwoicie uzasadnione, oczywiście z odwołaniem do wcześniejszych badań (nie tylko Stapela). Opisane procedury są zgodne z zasadami sztuki, używane są odpowiednio wyrafinowane metody statystyczne. Efekty, choć przekonujące, nie są jakieś szczególnie silne i nie wszystkie badane prawidłowości są istotne (sam Stapel zwracał uwagę na to, żeby sfabrykowane efekty nie były bardzo mocne, gdyż będą mało przekonujące dla redaktorów, por. Bhattacharjee, 2013). Ewidentnie, Stapel wiedział jak pisać, żeby wyglądało to wiarygodnie i żeby jego teksty publikowano.

Efekty te może i są jakoś interesujące, choć wydają się, że nie ma w nich nic sensacyjnego. Ot, badania jak tysiące innych w psychologii społecznej. Ta

„normalność” jest w tym wypadku czymś szczególnie niepokojącym. Nie mając dzisiejszej wiedzy, trudno się domyślić oszustwa, gdyż są to badania zbyt podobne do niezliczonej ilości innych w dyscyplinie. To z kolei musi nasuwać pytanie, czy spośród tysięcy innych nieodbiegających od standardów, przyzwoicie napisanych i nieźle uzasadnionych artykułów wiele też nie jest sfalszowanych? Ile jeszcze w psychologii mamy takich „Stapelów”, którzy chcieli pomóc swojej karierze, robiąc mniejsze i większe przekłamania, równocześnie (jak to zwykle u naukowych oszustów bywa) głęboko wierząc, że ogólnie opisywane prawidłowości są jak najbardziej prawdziwe?

Afera Stapela pokazało, że osoba inteligentna i potrafiąca się wpisać w standardy jest w stanie coś takiego robić. Raport zakwestionował nie tylko moralne standardy jednego badacza, ale też praktyki wydawnicze całej dyscypliny. Prawdopodobnie łatwiej publikować i robić karierę przy pomocy „gładkich” wyników, podpartych atrakcyjną narracją, niż danych zebranych rzetelnie, ale bardziej niejednoznacznych i mniej efektownych. Jeśli badacz chce być stuprocentowo rzetelny, zadanie mogą mu utrudniać same czasopisma. Komitety powołane do zbadania afery odkryły, że redaktorzy i recenzenci nie tylko chętnie wierzyli w (prawie) zawsze udane badania Stapela, ale również zachęcali go i jego współautorów do przekłamań. Cytując raport: „[...] współautorzy artykułów Stapela przyznawali, że czasami redaktorzy i recenzenci prosili, żeby pewne zmienne były usunięte, ponieważ dzięki temu wyniki byłyby bardziej zgodne z rozumowaniem i tokiem narracji. Skutkowało to usuwaniem niepożądanych rezultatów. Recenzenci również wymagali, aby nie wszystko przeprowadzone analizy były opisywane, przykładowo poprzez nie wspomnianie warunków dla których nie znaleziono efektów, które były początkowo oczekiwane. Czasami recenzenci nalegali na przeprowadzenie badań „pilotażowych” już po badaniu, które były potem raportowane jako przeprowadzone przed właściwym badaniem. W ten sposób procedura badań oraz wybór bodźców były uzasadnione rzekomą wcześniejszą wiedzą. [...]” (Levelt, 2012, s. 53).

Sama lektura artykułów Stapela nie pozwala, w moim przekonaniu, wykryć oszustwa przez niespecjalistów. Analiza danych w artykułach przeprowadzona przez zawodowych statystyków wykazała jednak szereg nieprawidłowości. Eksperti często zwracali uwagę na takie nieprawdopodobne sytuacje jak identyczne dane w niezależnych eksperymentach (np. średnie i odchylenia standardowe),

niewystępowanie brakujących danych, bardzo niskie statystyki F wszędzie tam, gdzie nie spodziewano się istotnych efektów, zbyt silne efekty jak na skale o niskich rzetelnościach lub skale z jedną pozycją itp. Gdy analizowano dane surowo znajdowano mechanicznie przeklejone kolumny zmiennych (opis nieprawidłowości w poszczególnych artykułach: Levelt, 2012, s. 69-100). Tworzenie całkiem fałszywych danych nie jest, wbrew pozorom, łatwe. Jednak fakt, że dane te nie wzbudzały przez szereg lat wątpliwości redaktorów i recenzentów, musi prowadzić do pytania, na ile mechanizmy peer-review są rzeczywiście skuteczne w wykrywaniu oszustwa. Być może, jak sugerują niektórzy (Nosek, Bar-Anan, 2012) peer-review służy przede wszystkim sprawdzeniu czy proponowany tekst jest dostatecznie „atrakcyjny” i nie odstaje zasadniczo od przyjętych standardów. I niewiele ponad to.

Oczywiście trudno oczekiwać od recenzentów umiejętności wykrywania każdego możliwego oszustwa, ale z perspektywy czasu interesujące jest pytanie ilu z recenzentów i redaktorów prac Stapela poprosiło o dane surowe lub materiały wykorzystane w eksperymentach. APA w ogłoszeniach o poszukiwaniu recenzentów⁶ twierdzi, że „recenzowanie jest czasochłonne; potrzeba około 1-4 godzin na jeden manuskrypt”. Trudno w tym kontekście nie postawić pytania, czy rzeczywiście przez 4 godziny można zrobić wiele, np. ile razy można przeczytać z dużą uwagą 20 stronicowy artykuł z JPSP zawierający 5-8 badań, na ile można go przemyśleć, przeanalizować podstawy teoretyczne, sprawdzić cytowane źródła, zastanowić się nad sensownością procedury i wykorzystanych narzędzi, nie mówiąc o przeanalizowaniu danych surowych czy sprawdzeniu poprawności obliczeń, a o napisaniu samej recenzji nie wspominając? Pytanie jest retoryczne, zakładałbym raczej, że przez 4 godziny można zrobić kontrolę podstawowych standardów, szczególnie najłatwiejszych do weryfikacji, ewentualnie przedstawić własny pogląd na sprawę i wypomnieć najbardziej oczywistych błędów. I jest to w dużym kontraście do tego, jak praktyka peer-review jest przedstawiana w mediach (por. np. tekst Wojciecha Orlińskiego (2013a) z *Gazety Wyborczej* mocno gloryfikujący peer-review).

Oszustwo na taką skalę jest czymś niezwykle rzadkim nawet w naukach przyrodniczych (por. Stroebe, Postmes, Smers, 2012). W psychologii jest tym

⁶ Standardowe ogłoszenie APA o poszukiwaniu recenzentów można przeczytać wyszukując w bazie PsycARTICLES teksty o tytule „Reviewers wanted”

bardziej szczególne, bo oszustwa są tu wykrywane rzadko. Powstaje oczywiście zasadnicze pytanie, czy wynika to z tego, że ich nie ma, czy nie są po prostu wykrywane. Omówię dane, którą mogą dać wskazówki, co do odpowiedzi na to pytanie.

Wspomniani Stroebe i in. (2012) zebrali 40 przypadków dobrze udokumentowanych oszustw w nauce odkrytych w latach 1974-2012. Z psychologów oprócz Stapela wymieniają oszustwa dwóch osób o raczej mało znanych nazwiskach (Roxana Gonzales; Karen Rugiero) oraz Marca Hausera, psychologa i biologa, badacza bardziej znanego z badań naczelnych. W przypadku tych trzech badaczy zakwestionowano ich jeden lub dwa artykuły (Hausera z *Cognition*, Gonzales z PSPB oraz JESP, a Rugiero z PSPB oraz JPSP). Olbrzymia większość listy to oszustwa w medycynie. Na pozór są więc powody do optymizmu. Gdy ukazał się raport poświęcony oszustwu Stapela, w którym postawiono pod znakiem zapytania integralność i standardy w psychologii społecznej, organizacja *European Association of Social Psychology* (2012) powołała się na ten artykuł, który miał być jakoby dowodem na to, że psychologia wypada nad wyraz korzystnie na tle innych nauk, a przypadek Stapela to odosobniona „czarna owca” (innych merytorycznych argumentów trzystronicowe oświadczenie nie zawierało).

I faktycznie, trudno zaprzeczyć, że relatywnie niewiele oszustw jest wykrywanych. Nie musi to oznaczać jednak, że ich nie ma. Być może większa liczba udokumentowanych oszustw w medycynie wynika nie tyle z niższych standardów etycznych, ile z wyższej weryfikowalności odkryć, szczególnie w obszarze badań stosowanych nad skutecznością leków i terapii. Niewykrywanie oszustw psychologii na tle medycyny może wynikać z niskiej kultury replikacji badań, niewielkiej teoretycznej integracji (teorie nie przewidują precyzyjnie co powinno, a co nie powinno występować), niewielkiej przekładalności odkryć podstawowych na badania aplikacyjne (z wielu odkryć nic konkretnego dla praktyki może nie wynikać, więc trudniej wykryć ich błędność w codziennej obserwacji) oraz tego, że badania aplikacyjne nie dotyczą zwykle, jak w medycynie, kwestii życia i śmierci (nawet jeśli terapie lub inne techniki oddziaływania są nieskuteczne jest to trudniej wykrywalne, a subiektywne poczucie poprawy może być związane z działaniem efektu placebo, por. Prioleau, Murdock i Brody, 1983).

Tezę o niewykrywaniu oszustw zdaje się potwierdzać badanie Johna, Loewensteina i Preleca (2012). Wysłali oni do 6 tys. psychologów, pracowników naukowych amerykańskich uczelni, zaproszenie do wzięcia udziału w anonimowej ankiecie dotyczącej stosowania wątpliwych praktyk badawczych w psychologii. Wzięło udział nieco ponad 2100 osób. Spośród uczestników 1,7 % przyznał się do tego, że chociaż raz w karierze fabrykował dane. W liczbach bezwzględnych daje to 36 osób. Swobodnie można jednak przyjąć, że nie wszyscy się przyznali, oraz że wśród osób, które nie wzięły udziału w badaniu odsetek ten może być wyższy. Używając odpowiednich algorytmów (zwanym bayesowskim serum prawdy; ang. bayesian truth serum) John z zespołem szacują, że w psychologii nawet 10 % badaczy może fabrykować dane.

Nawet jednak jeśli weźmiemy pod uwagę tylko te kilkadziesiąt osób, które wprost przyznało się do oszustw, jest to liczba mocno kontrastująca z 4 opisanymi przypadkami udokumentowanych oszustw w psychologii w ostatnich 12 latach (Stroebe i in., 2012).

Po pierwszej redakcji tego podrozdziału, życie niestety dopisało dalszy rozdział historii oszustw w psychologii. Niechlubne grono oszustów w psychologii zwiększyło się o następnych dwóch badaczy publikujących w czasopismach o wysokiej renomie. Uri Simonsohn (2013) wykrył podejrzany wzór danych w dwóch artykułach z psychologii społecznej. Omówię je nieco szerzej, gdyż same badania również są wymowne.

Pierwszy zakwestionowany artykuł opisywał serię badań pod kierunkiem Lawrence Sanna (Sanna, Chang, Miceli i Lundberg, 2011) na łamach JESP. Odkryto ciekawy efekt polegający na tym, że zwiększanie swojego położenia w pionie (np. poprzez wjeżdżanie w górę ruchomymi schodami) podnosi u ludzi standardy moralne np. zwiększa skłonność do pomagania innym albo poziom empatii (np. objawiający się niepodawaniem innym bardzo ostrego sosu do jedzenia). Inaczej mówiąc wznoszenie się w górę, „uwzniośla” nasze myślenie i działanie. Dla niespecjalisty efekt ten może wydawać się wydumany i efekciarski, jednak gdyby rzeczywiście był prawdziwy, byłby to ciekawy wkład w teorię poznania ucieleśnionego (ang. *embodied cognition*, np. Anderson, 2003), i mówiłby coś ważnego o związku między ciałem a procesami psychologicznymi. Simonsohn dostrzegł jednak w danych pewną anomalię. W trzech eksperymentach średnie wyniki osób w trzech grupach (osoby

„wznoszące”, „opadające” i grupa kontrolna) były znacząco różne (różniły się 1,5-2 odchyleniami standardowymi, albo o około 100 %), jednak same wielkość odchyłeń standardowych były każdorazowo niezwykle podobne (różnice rzędu 2-3 %). Simonsohn stworzył „indeks podobieństwa” odchyłeń standardowych (w niezależnych grupach eksperymentalnych), a następnie przeprowadził 100 tys. symulacji, zadając wielkość średnich w poszczególnych i uogólnione odchylenie standardowe takie jak w oryginalnym artykule. Indeks podobieństwa uzyskany przez Sanna był skrajnie mało prawdopodobny (tj. zbliżony rozkład uzyskano w 15 ze 100,000 symulacji)⁷. Jako dodatkowy mechanizm kontrolny Simonsohn sprawdził wyniki uzyskiwane przez innych badaczy w tym obszarze. I znowu wyniki Sanna wyróżniały się. W następstwie artykułu uczelnia Sanny podjęła śledztwo. Stwierdzono m.in. że współautorzy nie widzieli eksperymentów. Ostatecznie z JESP wycofano 3 artykuły Sanny, a on sam zrezygnował ze stanowiska na uniwersytecie (Yong, 2012b).

Kolejny artykuł zakwestionowany przez Simonsohna pochodził od holenderskiego badacza Dirka Smeestersa (Smeesters i Liu, 2011). Badanie dotyczyło wpływu koloru na prymowanie (nie)inteligentnego zachowania pod wpływem egzemplarzy obiektów, które stereotypowo cechują się inteligencją (np. profesorowie) lub nie cechują się (np. stadionowi chuligani, modelki). Na podstawie (rzekomego) pilotażu stwierdzono, że kolor niebieski powoduje przyciąganie, a kolor czerwony odpychanie. We właściwym badaniu osobom więc prezentowano materiały w teczce w jednym z tych kolorów, następnie osoby opisywały cechy profesora (lub modelki, jako rzekomej antytezy inteligencji), a następnie rozwiązywały test wiedzy ogólnej (zgodnie z dosyć znanym badaniem Dijksterhuisa (1998) prymowanie konceptem profesora lub chuligana odpowiednio zwiększało lub obniżało wyniki takich testów). Tak więc osoby opisujące profesora (czyli prymowane schematem poznawczym profesora) pod wpływem koloru czerwonego mniej podświadomie identyfikowały się z „konceptem” profesora i uzyskiwały niższe wyniki w teście wiedzy, natomiast w przypadku „konceptu” modelki uzyskiwały wyniki wyższe. W przypadku koloru niebieskiego było dokładnie na odwrót. Tak czy owak teoretycznie 6 kombinacji powinno przynieść wysokie wyniki, a 6 niskie wyniki testów wiedzy. I tak się też stało, wszystkie 12 kombinacji przyniosło wyniki zgodne z oczekiwaniami. Simonsohn

⁷ Opisuje tutaj proces analiz matematycznych w dużym uproszczeniu, po szczegóły odsyłam do tekstu.

założył, że nawet gdyby efekt taki występował to taka liczba trafień jest nadzwyczajnie mało prawdopodobna (badanie też nie miało wielkiej mocy, por. 1.3, w jednej komórce występowało około 16 osób). Tak jak w poprzednim przypadku przyjęto założenia dotyczące wielkości efektów i wykonano 100 tys. symulacji. Rozkład taki, jaki uzyskał Smeesters okazał się skrajnie nieprawdopodobny (zbliżony rozkład wypadł w 21 symulacjach). Podobnie jak w przypadku Sanny, również macierzysta uczelnia Smeestersa podjęła śledztwo. Zakwestionowano kilka dalszych jego artykułów (Report of the Smeesters Follow-Up Investigation Comitee, 2014). Smeesters przyznał się do fabrykacji danych. Ostatecznie został on usunięty z uczelni, a jego artykuły zostały wycofane (Retraction Watch, 2012, Yong, 2012c).

Wydaje się, że w miarę pewna odpowiedź na pytanie, ile danych w psychologii jest sfałszowanych jest zasadniczo niemożliwa. Prawdopodobnie znacznie więcej niż wskazują na to oficjalne wskaźniki wykrywalności oszustw. Wspomniany wcześniej optymizm *European Association of Social Psychology* (2012) wydaje się nieuzasadniony (warto też dodatkowo zauważyć, że w obrębie samej psychologii olbrzymia większość wycofanych artykułów należy do psychologii społecznej; jednak nie musi to świadczyć o braku oszustw w innych subdyscyplinach, ale potencjalnie o braku kontroli). Warto też zauważyć, że zarówno w przypadku Stapela, Sanny, Smeestersa, Gonzales i Rugiera zakwestionowane dane zostały opublikowane w czasopismach z „najwyższej półki” (Stapelowi hurtowo wycofano artykuły też z wielu mniej znanych artykułów). Zastanawia brak czasopism o niskim wpływie. Być może wysokie wymagania takich czasopism są trudniejsze do przeskoczenia, a więc badacze mają większą motywację by fałszować dane. Niewykluczone jednak, że nadreprezentacja renomowanych czasopism wynika z tego, że artykuły w nich są czytane dużo częściej i jest dużo większe prawdopodobieństwo, że ktoś im się przyjrzy dokładnie. Osobiście skłaniam się do tej drugiej hipotezy.

Fabrykowanie nieistniejących danych jest niewątpliwie „najniższym kręgiem piekła” w hierarchii naukowych przestępstw (por. Neuroskeptic, 2012). Jednak, nie trzeba wymyślać nieistniejących danych, żeby poważnie zniekształcać obraz rzeczywistości. Można oszukiwać dużo subtelniej poprzez arbitralną analizę i prezentację wyników rzeczywiście przeprowadzonych badań. O ile nikt nie zakłada, że większość danych w psychologii jest zwykłym fałszerstwem, o tyle istnieją pewna przesłanki, żeby

sądzić, że znaczna ich część może być przekłamana w sposób bardziej wyrafinowany.

1.2. Psychologia „falszywie pozytywna”⁸

W mniej więcej tym samym czasie, gdy wybuchła afera Stapela ukazał się artykuł Simmonsa, Nelsona i Simonsona (2011) o „psychologii fałszywie pozytywnej”. Tekst zdobył dużą popularność i spekulowałbym, że afera Stapela bardzo się do tego przyczyniła. O wrażeniu, jakie zrobił na społeczności badaczy, świadczy to, że przez 3 lata był cytowany ponad 900 razy⁹. Simmons i jego zespół, przeprowadzając tysiące symulacji na losowych danych wykazali, że nie trzeba wymyślać zupełnie fałszywych danych, żeby dowolny zbiór danych odpowiednio uatrakcyjnić tj. „podkreślić” do poziomu istotności statystycznej przy rzeczywistym braku jakichkolwiek efektów. Fakt ten jest oczywisty dla każdej w miarę zorientowanej matematycznie osoby i w literaturze był już wcześniej analizowany (Ioannidis, 2005; Maxwell, 2004). Być może odświeżeniu tego problemu w wykonaniu Simmonsa i współpracowników pomógł zabieg polegający na tym, że przeprowadzono również „badania”, w którym wybiórczo raportując dane „dowodzono” występowanie absurdalnych efektów. Przykładowo, „okazało” się, że manipulacja eksperymentalna polegająca na słuchaniu piosenki *When I’m 64* zespołu The Beatles *zmniejsza* wiek osób badanych.

W zasadniczej części tekstu Simmons i in. opisali kilka wątpliwych praktyk badawczych (ang. *questionable reaserch practice* – QRP), konkretnie:

- a) używanie wielu zmiennych zależnych i raportowanie tylko tej, która „wyszła”;
- b) dodawanie kolejnej grupy badanych tak długo, aż nie osiągnie się istotności statystycznej (i, oczywiście, zaprzestawanie dalszych badań, gdy tylko istotność zostanie osiągnięta);
- c) uwzględnianie płci jako dodatkowej współzmiennnej (należy wyjaśnić, że jest to metoda czasami używana jak najbardziej zasadnie, szczególnie jeśli oczekujemy,

⁸ Gra słów polegająca na połączeniu zwrotów „psychologia pozytywna” i „błąd fałszywie pozytywny”

⁹ Wszystkie liczby cytowań wymieniane w rozdziałach 1-3 na podstawie Google Scholar, stan na czerwiec 2015. Google Scholar ma dosyć „liberalne” kryteria kwalifikowania tekstów jako manuskrypty naukowe, więc liczby cytowań w bazach zbierających same czasopisma są zwykle niższe (np. około 20-40 % w SCOPUS)

że płęć wchodzi w interakcje z manipulacją eksperymentalną – tym niemniej porównanie różnic dwóch grup oraz porównanie interakcji różnic z płcią jako współzmienną zwiększa prawdopodobieństwo, że któryś z tych efektów będzie istotny).

d) wybiórcze raportowanie grup eksperymentalnych analogicznie jak w praktyce pierwszej – tylko tych między którymi wykryto istotne różnice

Statystycznie patrząc, powyższe zabiegi opierają się na fałszowaniu liczby stopni swobody. Analogiczna sytuacja ma miejsce wtedy, gdy chcemy wyrzucić kostką sześciocenną jedną liczbę (dajmy na to szóstkę), ale rzucamy pięcioma kostkami. Jeśli wypadnie gdzieś szóstka, tą kostkę zostawiamy, a pozostałe chowamy. W „raporcie” z rzutu przedstawiamy sytuację tak, jakbyśmy od początku rzucali jedną kostką. Prawdopodobieństwo wyrzucenia szóstki jest znacznie wyższe niż $1/6$ (wynosi ponad 60 %).

Simmons i in. przeprowadzili kilkanaście tysięcy symulacji na zupełnie losowych danych i wykazali, że stosowanie wymienionych 4 praktyk zwiększa prawdopodobieństwo popełnienia błędu fałszywego pozytywnego z tradycyjnych 5 % do, odpowiednio, 9,5 %, 7,7 %, 11,7 % i 12,6 %¹⁰. Stosowanie wszystkich 4 metod powoduje, że istnieje aż 61 % szans na znalezienie jakiegoś istotnego efektu w danych.

Autorzy w dyskusji przedstawili szereg zaleceń dla badaczy oraz recenzentów manuskryptów, służących unikaniu tych zniekształceń. Zalecają m.in. raportowania danych o wszystkich grupach eksperymentalnych i zmiennych zależnych, jasnych i określanych przed badaniem reguł dobierania badanych, odpowiedniej liczebności badanych grup, w przypadku obróbki danych (np. eliminacji skrajnych wyników) podawania również wyników bez takiej obróbki. Recenzenci mają być strażnikami tego, że dane nie są oparte na arbitralnych decyzjach analitycznych. Powyższe zalecenia wydają się być niemożliwe do wprowadzenia bez najważniejszej wskazówki, a mianowicie takiej, żeby recenzenci i redaktorzy wykazywali dużą większą tolerancję niż obecnie wobec wyników, które nie są doskonałe. Do problemu zmian w praktykach publikacyjnych powrócę w rozdziale 1.10.

¹⁰ Wielkość tych liczb zależy od dodatkowych założeń np. ile jest zmiennych zależnych, z których wybieramy tylko działające. Dla naszych potrzeb szczegóły nie są tu istotne, zainteresowanego Czytelnika odsyłam do oryginalnej publikacji.

Jeśli opisane przez Simmonsa sztuczki rzeczywiście są stosowane może to tłumaczyć, dlaczego podczas lektury naukowych artykułów empirycznych mam, wcale nierzadko, poczucie, że wiele decyzji jest wysoce arbitralnych. Trudno oczywiście oczekiwać od badaczy, że wytłumaczą każdą decyzję stojącą za wyborem każdego elementu procedury, trudno też oczekiwać, że dane badanie musi wyjaśnić wszystko np. wylistować wszystkie możliwe moderatory danego efektu, albo elementy procedury niezbędne do tego, żeby efekt wystąpił. Badacz nie musi wiedzieć z góry, jakie modyfikacje sprawią, że dany efekt nie zadziałała. Rzetelnie opisane odkrycie nie musi implikować automatycznie tego, że wszystko już na ten temat wiemy itd. Tym niemniej, nawet pomimo tych zastrzeżeń, niektóre decyzje badaczy nasuwają wrażenie, stosowania opisywanych wcześniej QRP. Aby lepiej zobrazować idee Simmonsa przedstawię przykład badania, które uważam osobiście za podejrzone.

Gervais i Norenzayan (2012) w artykule w *Science* przedstawili dane empiryczne wskazujące na to, że analityczne myślenie jest predykatorem braku wiary religijnej. Jedno badanie miało charakter korelacyjny (mierzono związek między zdolnością do rozwiązywania problemów, w których narzucało się intuicyjne, ale nieprawdziwe rozwiązanie¹¹, a deklarowaną wiarą w Boga), pozostałe 4 eksperymentalny (manipulacja polegała m.in. na oglądaniu rzeźby myślącego filozofa lub w grupie kontrolnej – rzeźby atlety; prymowaniu słowami związanymi z analitycznym myśleniem lub neutralnymi itd.). W każdym z badań uzyskano wyniki wskazujące na związek analitycznego myślenia z brakiem wiary. Efekty były umiarkowanie silne i kształtowały się w zakresie $0,3 < d < 0,6$. Wątpliwości budzi we mnie następująca rzecz: w poszczególnych badaniach eksperymentalnych używano innych zmiennych zależnych. I tak w jednym badani określali poziom wiary w Boga na skali od 0 do 100, w drugim odpowiadali na pytania dotyczące ich wiary w Boga, anioły i diabła, każde na skali od 1 do 7, a w jeszcze innym odpowiadali na 10 pytań na temat ich religijności (na każde w skali od 1 do 7; np. „Moje przekonania religijne są tym, co naprawdę leży u podstaw mojej filozofii życiowej”). Jednak nawet w tej niekonsekwencji badacze byli *niekonsekwentni*, gdyż w ostatnim badaniu wrócono do zmiennej zależnej „wiera w Boga w skali 0-100”. Autorzy nie uzasadnili, dlaczego używali różnych zmiennych zależnych w każdym badaniu, choć w gruncie rzeczy

¹¹ Przykładowo: „za arbuza i jabłko zapłaciłeś 1,10 dolara. Arbuza kosztował dolara więcej niż jabłko. Ile kosztował arbuza?”. Intuicyjnie narzuca się odpowiedź „1 dolar za arbuza”, jednakże prawidłowa odpowiedź to 1,05 dolara.

każdorazowo testowali tą samą hipotezę. O ile różnorodne procedury eksperymentalne są zrozumiałe, o tyle różnorodne zmienne zależne utrudniają rzetelną interpretację wyników, ponieważ można się zastanawiać czy poszczególne zmienne zależne mierzą dokładnie to samo i czy tak naprawdę w każdym eksperymencie nie jest testowana inna hipoteza. Przypuszczam intuicyjnie, że miary te prawdopodobnie mierzą coś bardzo zbliżonego, tym niemniej ich arbitralne stosowanie budzi uzasadniony niepokój, może bowiem wskazywać, że każdorazowo używano kilku zmiennych zależnych i zaraportowano tylko te, które wyszły (lub były różne kombinacje manipulacji i zmiennej zależnej i przedstawiono tylko działające). Na wybiórcze raportowanie badań wskazuje też, obliczona przeze mnie, bardzo wysoka negatywna korelacja pomiędzy siłą efektu, a wielkością próby (w 4 badaniach eksperymentalnych: $r = -0,97$; dlaczego może to wskazywać na nierzetelność w opisie badań patrz: rdz. 1.4).

Aby dalej zilustrować problem arbitralności, podam przykład hipotetycznego badania zaczerpnięty od Gelmana i Lokena (2013) i twórczo przeze mnie rozwinięty. Badanie to dotyczy różnic w rozwiązywaniu problemów matematycznych pomiędzy zwolennikami demokratów i republikanów w USA. Badacz zakłada, że różnice występują w zależności od kontekstu, w jaki ubierze problem: demokraci lepiej poradzą sobie z problemem, który dotyczy opieki zdrowotnej, a republikanie lepiej z matematycznie analogicznym problemem, ale ubranym w szaty działań militarnych. Dla pewności badacz dodaje jeszcze każdej grupie po dwa problemy z innych ideologicznie nasyconych tematów (dajmy na to, dostęp do broni i aborcja). Zbiera też szereg dodatkowych danych demograficznych.

Założmy, że po zebraniu danych nie ma zakładanego efektu dla całej próby, ale okazuje się, że występuje on tylko u badanej grupy mężczyzn? Można to bardzo łatwo wytłumaczyć, wszak mężczyźni mają silniejsze przekonania ideologiczne (tu hipotetyczny badacz może przywołać szereg danych). A co jeśli różnice dotyczą tylko kobiet? Badacz też stworzy do tego bogatą historię wskazując na to, że kobiety są bardziej wrażliwe na kontekst (i wzmocni to szeregiem źródeł). Może efekt nie występuje dla żadnej płci analizowanej osobno (wewnątrzgrupowo), ale pojawia się istotna interakcja głównego efektu z płcią?

Dalej, założmy, że badacz pytał o sympatię partyjne na skali 1-7 (skrajni Demokraci to 1, a skrajni republikanie 7). Powstaje kolejne pytanie, kogo ma porównywać?

Osoby, które zaznaczyły 1-3 z osobami 5-7? A może ideologiczne „jedyńki” i „siódemki”? Co zrobić z osobami neutralnymi? Wyłączyć je, a może kluczowe porównania wskażą na różnice pomiędzy osobami zaangażowanymi politycznie a neutralnymi? Może żadna analiza nie da istotnego efektu, ale jeśli wrzucimy wiek do modelu, to okaże się, że występuje istotna interakcja, do której można dorobić dalszą historię (np. młodszy są bardziej skrajni w swoich postawach politycznych vs. młodszy mają mniej skryzalizowane i bardziej płynne poglądy – do każdego twierdzenia też znajdzie się bibliografia).

„Hipotezy badawcze” można dalej mnożyć w nieskończoność. Przy każdym odpowiednio rozbudowanym badaniu istnieje mnóstwo możliwych porównań i arbitralnych decyzji, z których kilka może okazać się istotne statystycznie z racji czystego przypadku. Potem, biorąc pod uwagę obfitość literatury, można wyniki lepiej lub gorzej uzasadnić wcześniejszymi danymi.

Wspomniani Gelman i Loken nie ograniczyli się tylko do hipotetycznego przypadku badań nad demokratami i republikanami. Wskazali też kilka przykładów, gdzie badania wykazują ponadprzeciętnie wysoki poziom arbitralności.

I tak w badaniu Petersena, Szyncera, Sella, Cosmides i Tooby’ego (2013) wskazano, że mężczyźni o większej sile fizycznej są bardziej przeciwni redystrybucji dochodów. Podtytuł artykułu brzmi „siła fizyczna reguluje postawę mężczyzn względem przekładania interesu własnego ponad redystrybucję dochodów¹²”. Założono, kierując się pewną adaptacyjną logiką, że mężczyźni mający większą „siłę przetargową” (w formie siły fizycznej) będą mniej skłonni domagać się społecznych programów zmniejszania nierówności. Gelman i Loken wskazują jednak, że w dwóch z trzech badań nie mierzono rzeczywistej siły, ale jej przybliżony wskaźnik tj. obwód bicepsa (i trudno się nie zgodzić, że tytuł „wielkość bicepsa wpływa na postawy względem redystrybucji dochodów” jest bardziej poprawny naukowo, choć mniej atrakcyjny niż tytuł oryginalnie użyty). Dalej nie dowiedziono w żadnym z badań występowania dosłownie takiego efektu, ale za to wskazano na występowanie efektu interakcji z dodatkową zmienną, statusem społeczno-ekonomicznym (im wyższy status tym negatywna korelacja siła-poparcie dla redystrybucji dochodu była silniejsza). Zdaniem Gelmana i Lokena, gdyby zaobserwowano interakcję o przeciwnym kierunku, rzeczony efekt można by równie efektownie wyjaśnić (nie

¹² Upper-body strength regulates men’s assertion of self-interest over economic redistribution.

wspominając o możliwościach tłumaczenia innych potencjalnych interakcji, np. z liczbą rodzeństwa). Również sam pomiar zmiennej zależnej wydaje się być cokolwiek wątpliwy. W materiałach uzupełniających sami Peterson i in. wskazują, że ostateczne skale poparcia dla redystrybucji były złożone z różnych pozycji testowych w zależności od kraju (np. „w argentyńskiej próbie występuje niski poziom alfy Cronbacha [...], skala utworzona z pozycji 1 i 4 osiąga zadawalającą wielkość $\alpha = .65$ ”), co rzecz jasna tłumaczy się specyfiką myślenia politycznego w różnych rejonach świata.

Inne analizowane przez Gelmana i Lokena badanie ma równie „medialną” wymowę. Wg Durante, Arsen i Griskeviciusa (2013) faza cyklu miesięcznego wywiera wpływ na intencje wyborcze kobiet. Przykładowo, samotne owulujące kobiety stają się bardziej liberalne, a kobiety posiadające partnera w czasie owulacji stają się bardziej konserwatywne. Uzasadnienie teoretyczne w dużym uproszczeniu skupia się na korzyściach, które płyną z przedstawiania się jako osoba o konserwatywnych poglądach (a więc też „prorodzinnych”) w zależności od tego, czy zabezpieczyło się dostęp do długotrwale inwestującego partnera. Gelman i Loken pokazują, że wbrew wymowie abstraktu przeprowadzono tylko porównania międzygrupowe (nie porównywano preferencji politycznych tych samych kobiet w różnych fazach cyklu). Samo uzasadnienie teoretyczne jest cokolwiek arbitralne (np. gdyby się okazało, że owulujące kobiety w związkach stają się bardziej liberalne, można by to tłumaczyć ich gotowością do „łowienia” genów poza stałym związkiem i tym samym różnicowaniem puli genowej swojego potomstwa, co, jak wykazano w biologii ewolucyjnej, występuje u niektórych gatunków). Dalej, wielkość efektu (ponad 20 % różnicy w preferencjach kandydata w wyborach prezydenckich USA) jest nieproporcjonalnie duża do znanych innych badań nad zmianą preferencji politycznych. Sama moc badania też jest słaba. Tym bardziej ten efekt jest wątpliwy, że autorzy mieli bardzo wiele zmiennych, które mogli analizować (wiek, pochodzenie etniczne, rasa, posiadanie potomstwa itd.) i z których mogli sobie „wybrać” moderatory. Sam status związku też został utworzony w dyskusyjny sposób (jako „związane” uznano kobiety, które były w małżeństwie, zaręczone, lub mieszkające z partnerem, ale już nie „chodzące na randki”, które z kolei wrzucono do kategorii „wolne”). Decyzja o tym, które kobiety były w fazie płodnej również też jest po części arbitralna (u autorów to kobiety w dniach 6.-14. cyklu, choć Gelman i Loken

przedstawiają alternatywne dane, np. organizacje działające w obszarze planowania rodziny wskazują raczej przedziały 12.-17. lub 10.-17. itd.).

Warto wspomnieć, że obydwie analizowane tutaj badania nie ukazały się w jakichś podrzędnych czasopiśmie przez nikogo nie czytanych, ale w samym *Psychological Science*, a więc czasopiśmie niewątpliwie w ścisłej czołówce czasopism empirycznych (Gelman i Loken analizują też mnóstwo arbitralnych decyzji w słynnym badaniu Bema, 2011, opublikowanym w JPSP).

Od siebie dodam znany mi osobiście przykład z pracy magisterskiej, którą konsultowałem (Jamrużka, 2013). Jej celem była replikacja interesującego efektu poznawczego polegającego na tym, że zmiana na obrazku polegająca na dodaniu obiektu ożywionego jest łatwiej wykrywana przez umysł niż zmiana polegająca na dodaniu obiektu nieożywionego. Oryginalne badanie zostało opublikowane w wysoce prestiżowym *Proceedings of the National Academies of Sciences* (New, Cosmides, Tooby, 2007). W replikacji udało się ten efekt powtórzyć, ale był on znacznie słabszy niż w oryginalnym badaniu. Co jednak najważniejsze, okazało się, że sam dobór bodźców w zupełności wystarcza do tego, żeby wykazać efekt taki jak w oryginale, lub efekt zupełnie przeciwny. Dużo większe znaczenie niż sama „ożywioność” bodźca miały jego cechy czysto fizyczne np. relatywna jasność na tle innych obiektów.

Warto przytoczyć na koniec tego podrozdziału opinię Gelmana i Lokena:

Zbiór danych może być analizowany na tak wiele sposobów [...], że bardzo niewiele wynika z informacji, że w badaniu osiągnięto wynik $p < .05$. W skrócie, jest łatwo znaleźć jakieś porównania $p < .05$, jeśli nic w rzeczywistości się nie dzieje, ale szuka się dostatecznie długo. A dobry naukowiec ma wiedzę, jak szukać długo, i jak potem do wyniku dodać przekonującą historię. (Gelman, Loken, 2014, s. 1).

W podanych tutaj przykładach istnieją przesłanki, żeby podejrzewać, czy postulowany efekt rzeczywiście występuje, czy może jest artefaktem wynikającym z niekompetencji (celowej lub niezamierzonej) statystyczno-matematycznej badacza. Jednak absolutnej pewności nie ma i tego rodzaju dane są co najwyżej pośrednim dowodem na występowanie psychologii fałszywie pozytywnej. Można jednak takich dowodów szukać bardziej bezpośrednio, pytając badaczy o stosowanie wątpliwych praktyk.

1.3. Psychologia „falszywie pozytywna” – próba empirycznego zbadania rozpowszechnienia zjawiska

Fakt, że teoretycznie możliwe jest naginanie danych, oraz fakt, że konkretne badania wydają się cokolwiek arbitralne, nie mówi nam nic na temat tego, jak bardzo jest to rozpowszechnione. Najbardziej kompleksową i chyba jedyną próbę empirycznego oszacowania częstości wątpliwych praktyk badawczych przeprowadzili John, Loewenstein i Prelec (2012). Zaproszenie do udziału wysłano prawie 6 tys. psychologów akademickich w USA. Ostatecznie w badaniu wzięło udział nieco ponad 2 tys. Badanym zapewniono anonimowość i motywowano do mówienia prawdy (ankietowani wskazywali organizację pożytku publicznego, której autorzy badania wysyłali datki, co rzeczywiście miało miejsce).

Pytano przede wszystkim o to, czy osoby kiedykolwiek realizowała jedną z wymienionych wątpliwych praktyk badawczych oraz jakie jest – zdaniem ankietowanego – rozpowszechnienie tychże praktyk wśród badaczy. Dodatkowo badani szacowali, jaka część badaczy przyznałaby się do wątpliwych praktyk, i m.in. na tej podstawie starano się oszacować rzeczywiste rozpowszechnienie wątpliwych praktyk badawczych (omówienie logiczno-matematyczne tej metodologii: Prelec, 2004). Pytano również o dopuszczalność poszczególnych praktyk w 3 stopniowej skali. Tabela 1.1. zawiera rozpowszechnienie wątpliwych praktyk na podstawie samopisów, szacunków dotyczących innych badaczy, oraz oszacowanie rzeczywistego rozpowszechnienia przez autorów artykułu. Ostatnia kolumna zawiera ocenę dopuszczalności danej praktyki.

TABELA 1.1. Rozpowszechnienie wątpliwych praktyk badawczych

Wątpliwa praktyka badawcza	Procent badanych, którzy przyznali się do stosowania danej praktyki¹³	Szacowane przez ankietowanych badaczy rozpowszechnienie danej praktyki w społeczności naukowej¹⁴	Oszacowanie rozpowszechnienia praktyki na podstawie bayesowskiego serum prawdy	Średnia ocena dopuszczalności danej praktyki¹⁵ (w nawiasie odchylenia standardowe).
1. Nie raportowanie wszystkich wykorzystanych zmiennych zależnych	66,5 %	60 %	78 %	1,84 (0,39)
2. Zbieranie dodatkowych danych po sprawdzeniu czy już posiadane dane są istotne	58,0 %	62 %	72 %	1,79 (0,44)
3. Wybiórcze raportowanie tylko tych warunków eksperymentalnych, pomiędzy którymi zanotowano istotne różnice	27,4 %	38 %	42 %	1,77 (0,49)
4. Rezygnacja ze zbierania danych wcześniej niż zaplanowano, ze względu na znalezienie oczekiwanych wyników	22,5 %	41 %	36 %	1,76 (0,48)
5. Nieuprawnione zaokrąglanie wartości p (np. raportowanie wartości $p=0,054$ jako $p<0,05$)	23,3 %	41 %	39 %	1,68 (0,57)
6. Selektywne raportowanie tylko tych badań, które „wyszły”	50,0 %	61 %	67 %	1,66 (0,53)
7. Decydowanie o tym, czy wykluczyć określone dane po sprawdzeniu wpływu takiej operacji na rezultaty	43,4 %	45 %	62 %	1,61 (0,59)
8. Opisywanie nieoczekiwanego wcześniej odkrycia jako przewidzianego od samego początku	27,0 %	50 %	54 %	1,50 (0,60)
9. Twierdzenie, że na wyniki nie mają wpływu zmienne demograficzne, podczas gdy w rzeczywistości nie wiadomo lub mają wpływ	4,5 %	22 %	13 %	1,32 (0,60)
10. Falszowanie (fabrykowanie) danych	1,7 %	10 %	9 %	0,16 (0,38)

Źródło: dane zawarte w John i in. (2012)

¹³ Ścisłej: zaraportowano tutaj wyniki grupy, która była dodatkowo motywowana do mówienia prawdy. Wyniki grupy kontrolnej były zwykle kilka procent niższe (tj. niższe o 0-7 %).

¹⁴ Ta wielkość jest oszacowana przeze mnie na podstawie wysokości słupka wykresu, gdyż autorzy nie przedstawili w artykule tabeli z dokładnymi danymi.

¹⁵ Skala: 0 – całkowicie niedopuszczalne; 1 – w pewnym stopniu dopuszczalne; 2 – dopuszczalne

Z badania wynikają w większości niewesołe wnioski. Zanim je omówię, wspomnę jednak najpierw o kilku wątpliwościach, co do samej metodologii i prezentowanych wyników. Główna wątpliwość nasuwa się taka: zadawano pytanie o to, czy komuś zdarzyło się przynajmniej raz popełnić określoną wątpliwą praktykę. Odpowiedź pozytywna jest sama w sobie bardzo pojemna i może obejmować zarówno osoby, które systematycznie „podkręcają” dane wypaczając zupełnie wynik, jak również osoby, którym zdarza się to niezwykle rzadko lub w taki sposób, że dotyczy to kwestii marginalnych i nie ma praktycznie żadnego znaczenia dla rzetelności danych. Samo zestawienie nie uwzględnia też sytuacji, gdy nieuprawnione działania zostały niejako wymuszone, bo redaktorzy i recenzenci wymagali „wygładzenia” danych (a afera Stapela pokazała, że i tak się zdarza np. Levelt, 2012 s. 53). W artykule zaraportowano co prawda dodatkowe badanie, w którym przebadano grupę ponad 130 badaczy – uczestników „konferencji poświęconej badaniom behawioralnym”. Wśród osób, które twierdziły, że zdarzały się im wątpliwe praktyki 64 % twierdziło, że zrobiło to „raz lub dwa razy”, 26 % twierdziło, że robi to sporadycznie, a 10 %, że często. Wydawać by się mogło więc, że nawet jeśli są stosowane, to wątpliwe praktyki są raczej rzadkie. Niemniej trzeba odnotować, że była to grupa dużo mniejsza, mniej reprezentatywna i prawdopodobnie anonimowość była gorzej zapewniona (niestety, to badanie to jest bardzo pobieżnie opisane).

John i in. mogli też zaznaczyć, jaki procent badaczy angażował się w przynajmniej jedną praktykę (lub więcej), a jaki odsetek był całkowicie „czysty”. Warto to wiedzieć, gdyż robi pewną jakościową różnicę, jeśli – przykładowo – badacze, którzy stosują praktykę 6 i 7 są to te same osoby, albo te dwa zbiory pokrywają się w niewielkim stopniu.

Tak czy owak, niektóre pesymistyczne wnioski są niezaprzeczalne. W badaniu wzięło udział nieco ponad 30 % wszystkich osób, do których wysłano zaproszenie (kilkakrotnie je ponawiano, jeśli ktoś nie odpowiedział od razu). Można założyć, że duża część osób systematycznie fałszujących dane nie podejdzie w ogóle do takiej ankiety. Odsetek przyznających się jest tak czy owak niemały, a trudno przypuszczać, żeby ktoś się przyznawał do nieuczciwych praktyk nie stosując ich, wydaje się, że odwrotna sytuacja jest dużo bardziej prawdopodobna (na co wskazuje też Fanelli, 2009). W przypadku niektórych kategorii ponad 50 % badanych przyznaje się do praktyk, które, jak wiadomo, prowadzą do zawyżonej liczby wyników

falszywych pozytywnych. Jeśli założyć, że metodologia „bayesowskiego serum prawdy” jest trafna¹⁶ odsetek badaczy stosujących analizowane przez Simmons i in. (2011) praktyki takie jak nie raportowanie zmiennych zależnych, dokładanie osób do krutek aż do skutku i wybieranie „działających” porównań między grupami wynosi odpowiednio 78 %, 72 %, i 42 %. Jednym z najbardziej pesymistycznych wniosków jest, że wszystkie wątpliwe praktyki badawcze (poza fałszowaniem danych) nie są postrzegane jako coś złego, ale są oceniane jako „raczej dopuszczalne” lub wręcz „dopuszczalne”. W przyszłych badaniach byłoby interesującym rozstrzygnąć czy wynika to z niskiej świadomości metodologicznej, czy też raczej z cynicznego pragmatyzmu. John i in. (2012) podają też rezultaty z rozbiem na poszczególne subdyscypliny. Psychologia społeczna, która jest w szczególny sposób analizowana w tej dysertacji, miała badaczy najczęściej przyznających się do QRP.

Chciałbym poświęcić tu uwagę praktyce polegającej na opisywaniu nieprzewidzianego odkrycia jako hipotetycznie założonego od początku¹⁷ (nr 8 na liście w Tabeli 1.1.). Jest to inny zabieg niż pozostałe, gdyż sam w sobie nie ma na celu zwiększenia szans na wynik istotny statystycznie. Na pierwszy rzut oka może wydawać się przede wszystkim zabiegiem stylistyczno-retorycznym, nie do końca uczciwym, ale na pozór nieszkodliwym. Tak jednak nie jest, na co zwraca uwagę Kerr (1998) w artykule szeroko omawiającym tę niestety częstą praktykę. Jeśli badacz odkryje jakieś zjawisko, a następnie dorobi do niego odpowiednią „ideologię” teoretyczną (co w nauce takiej jak psychologia nie jest trudne, jak starałem się pokazać wcześniej), i przedstawia swoje odkrycie jako przewidziane od początku, wówczas tworzy w innych badaczach nieuprawnione poczucie siły przewidywania teorii. A skoro teoria jest w stanie „wygenerować” takie odkrycia, to coś w niej musi być na rzeczy. Jednak, jak dobrze wiemy, teoria niczego nie wygenerowała.

Problem jest niemały, gdyż lektura czasopism takich jak JPSP czy PSPB pokazuje, że praktycznie wszystkie odkrycia były „przewidziane”. Bones (2012) ironizowała, że Bem (2011) w swoim badaniu o prekognicji nie pokazał niczego nowego, gdyż każdy numer JPSP dostarcza mnóstwo dowodów na prekognicję. Do problemu roli teorii w generowaniu hipotez powrócę w rozdziale 3.

¹⁶ Artykuł teoretyczno-matematyczny, który ją opisuje został opublikowany przez Prelec-a (2004) w *Science*, niestety ze względu na skomplikowany aparat matematyczny, brakuje mi kompetencji, żeby ocenić tą metodę.

¹⁷ W języku angielskim opisywana jako „hypothesizing after the results are known” w skrócie HARKing

Falszerstwo – najgorszy rodzaj występku przeciwko rzetelności badań naukowych – szczęśliwie nie jest oceniane jako dopuszczalne, tym niemniej prawie 2 % badaczy przyznało się do jego stosowania, a odsetek ten może być dużo wyższy (wg autorów nawet 10 %). Przyjmując optymistyczne założenie, że odsetek 2 % jest reprezentatywny dla całej populacji badaczy, oznacza to, że tylko na uczelniach amerykańskich funkcjonują setki osób, które aktywnie przekłamują dane. Trudno jednak zaprzeczyć, że pozostałe praktyki też są rodzajem fałszerstwa¹⁸, nie polegają co prawda na tworzeniu nieistniejących danych, ale fałszują innym badaczom obraz rzeczywistości, nie dają bowiem możliwości oceny rzeczywistej siły efektu, prawdopodobieństwa replikacji oraz nie dają wglądu w opis całego procesu badawczego. Osobną kwestią jest to, o czym już wcześniej wspominałem, że redaktorzy i recenzenci najprawdopodobniej aktywnie promują niektóre rodzaje zafałszowań, gdyż nie jest możliwe – ze względu choćby na czysty rachunek prawdopodobieństwa – żeby kilka kolejnych badań o niskiej mocy i umiarkowanie silnych efektach dawało zawsze tak gładkie rezultaty, jak to na ogół spotykamy choćby w JPSP.

Wyniki badania zespołu Johna są dosyć zbliżone do wyników uzyskanych w meta-analizie badań nad częstością deklarowanych fałszerstw przeprowadzonej przez Fanelliego (2009), obejmującej wszystkie nauki (większość badań w niej uwzględnionych pochodziła z medycyny). Ważony odsetek badaczy przyznających się do fałszerstw wynosił 1,97 %, a 33 % przyznawało się do popełniania innych wątpliwych praktyk badawczych (niestety, Fanelli nie precyzuje dokładnie jakich). Ankietowani badacze szacując rozpowszechnienie fałszerstwa i wątpliwych praktyk wśród *innych* badaczy wskazywali średnio wartości 14 % i 72 %.

Wyniki Johna i in. potwierdzają więc, że psychologia fałszywa pozytywna przed którą ostrzegają Simmons i in. (2011) może mieć miejsce, a nie jest tylko matematyczną ciekawostką.

Idąc krok dalej Bakker, van Dijk i Wicherts (2012) postanowili sprawdzić jakie strategie badawcze są optymalne z punktu widzenia znajdowania potencjalnie publikowalnych (czyli istotnych) wyników. Użyli metafory gry: naukowcy mogą przyjmować różne „strategie” badawcze, a celem gry jest wynik istotny statycznie.

¹⁸ Praktyki 5. i 9. to w moim przekonaniu nie rodzaj, ale zwykłe fałszerstwo.

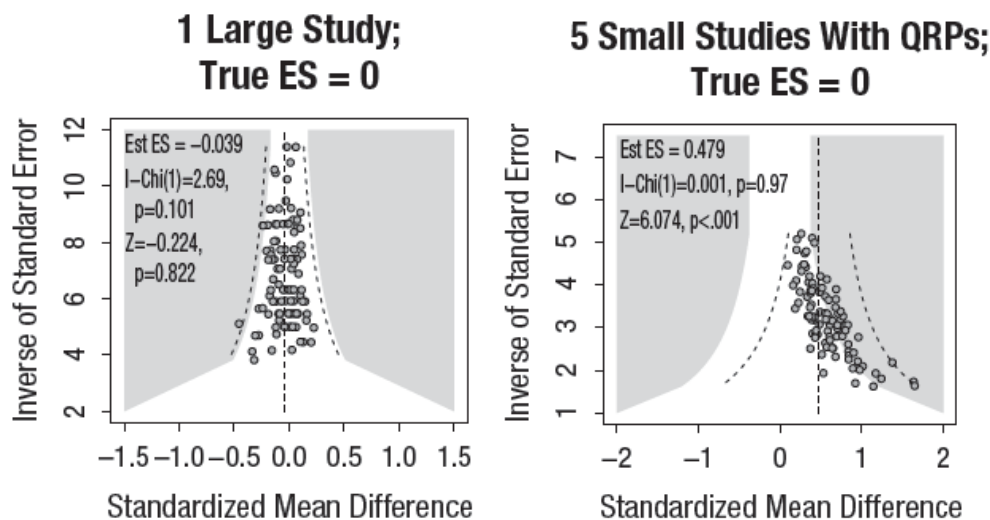
Baker i in. wygenerowali symulacje wielu badań i zestawiali je w „meta-analizy”.

Założono, że „gracze” mogą stosować jedną z 4 strategii:

- a) jedno duże badanie o próbie N
- b) jedno duże badanie o próbie N z zastosowaniem trzech QRP (tj. dwie zmienne zależne, raportowanie tylko istotnej; dodawanie badanych do czasu znalezienia istotności; usuwanie skrajnych wyników, jeśli przynosi to istotne wyniki)
- c) 5 małych badań o próbie $N/5$ z zaprzestaniem dalszych badań, gdy tylko istotny efekt zostanie znaleziony
- d) 5 małych badań o próbie $N/5$ z zaprzestaniem dalszych badań, gdy tylko istotny efekt zostanie znaleziony z zastosowaniem trzech QRP opisanych w punkcie b

Zauważmy, że wszystkie strategie wymagają zbliżonego wysiłku organizacyjnego, gdyż za każdym razem badamy tak samo dużą grupę. W przypadku strategii c i d wysiłek teoretycznie powinien być większy, gdyż w zamyśle przeprowadzamy 5 niezależnych badań. Jednak zmiany te mogą być symboliczne.

Wyniki nie były zaskakujące: największe szanse na znalezienie „jakichś” istotnych zależności miały strategię b. i d.. Były one tym większe, im słabszy był efekt i mniejsza próba N. Analogicznie, największe odchylenia siły znalezionej efektu od rzeczywistego występowały przy słabych efektach i małych próbach. Na rycinie 1.1 zaprezentowano wynik dwóch „meta-analiz” przeprowadzonych na wynikach symulowanych badań zgodnie ze strategią pierwszą i ostatnią, przy założeniu braku rzeczywistego efektu ($d = 0$). Ta ostatnia strategia bardzo wyraźnie wypacza jego wielkość, wskazuje bowiem, że wynosi ona około $d = 0,5$.



Źródło: przedruk z Bakker, van Dijk, Wicherts (2012)

W przypadku bardzo silnych efektów, wątpliwe strategie aż tak bardzo nie wypaczały rzeczywistego wyniku, bo istotne wyniki były znajdowane tak czy owak, a więc wybiórcze publikowanie nie było konieczne.

O ile przeprowadzenie jednego dużego badania wydaje się być strategią najlepszą w dążeniu do prawdy, o tyle niekoniecznie musi to być strategia najlepsza pod kątem potencjalnej publikacji.

Polski badacz o znaczącym dorobku podzielił się ze mną obserwacją ze swojego pobytu w zagranicznym ośrodku w połowie pierwszej dekady XXI wieku. Stwierdził, że standardową praktyką tam, było przeprowadzanie jednego nowego badania w ciągu tygodnia, z których potem zwykle ostatecznie jedno na dziesięć było wysyłane do czasopisma. Niestety, istnieją bardziej „twarde” dane wskazujące na to, że strategia badawcza polegająca na przeprowadzaniu wielu badań o względnie niskiej mocy, po to, żeby z ogromu danych dało się wyłowić istotne wyniki, jest wcale nierzadka. Szczególnie istotne są tu systematyczne analizy nad mocą statystyczną badań.

1.4. Moc statystyczna badań i wybiórcze publikowanie

Czasopisma publikujące oryginalne artykuły empiryczne w psychologii składają się w olbrzymiej większości z „udanych” badań. Cudzystów bierze się z tego, że kryterium

sukcesu jest przede wszystkim istotność statystyczna wyniku, niekoniecznie jego rzetelność (prawdziwość) i/lub znaczenie (teoretyczne czy praktyczne). W jakiejś mierze trudno się dziwić redaktorom. Gdyby miano publikować każde badanie o rozsądnej metodologii, niezależnie od wyniku, wówczas czasopisma byłyby zalane raportami o nieistotnych wynikach, co prawdopodobnie jeszcze bardziej utrudniałoby poruszanie się w monstrualnej liczbie artykułów, jakie już się ukazują. Z nieskończonej liczby możliwych kombinacji zmiennych najbardziej przyciągają umysł te, które rzeczywiście „działają”, a z „niedziałających” na ogół mniej wynika dla naszego rozumienia świata (wyjątki od tej reguły omówię na końcu tego podrozdziału). We wszystkich więc naukach wyniki pozytywne stanowią większość. Jednak w różnym stopniu. W dużej analizie bibliometrycznej Fanelli (2010a) pokazał, że psychologia wespół z psychiatrią znajduje się na czele listy wszystkich nauk z wynikami pozytywnymi. Najmniej artykułów potwierdzających wyjściową hipotezę publikuje się w astrofizyce (ok. 70 %). W psychologii odsetek ten wynosi 91.5 %. Biorąc jednak poprawkę na to, że w artykułach psychologicznych testuje się dużą więcej hipotez, Fanelli wyliczył, że w psychologii jest ponad 5 razy więcej wyników „pozytywnych” w przeliczeniu na artykuł niż w naukach o kosmosie. Różnice te są szczególnie wyraźne w badaniach podstawowych (w badaniach aplikacyjnych nie było dużych różnic między tj. w „twardych” naukach negatywne wyniki w czasopismach aplikacyjnych też nie były mile widziane).

Postanowiłem dodatkowo przetestować odkryte przez Fanelliego prawidłowości. Przejrzałem abstrakty wszystkich artykułów opublikowanych w JPSP w roku 2013. W 98 tekstach przedstawiono wyniki więcej niż jednego oryginalnego badania empiryczne, 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 zdarzyły się dwa teksty, których celem była krytyka innych teorii i przedstawiono tutaj dane im zaprzeczające). Jeśli abstrakt wspomina o tym, skąd autorzy wzięli określone hipotezy empiryczne, to były zwykle przedstawione jako „przewidziane” od samego początku przez założenia teoretyczne (por. też Kerr, 1998), sporadycznie badanie było przedstawiane jako eksploracyjne w swej naturze. Jak już wspominałem Bones (2012) ironizowała, że

Bem (2011) w swoim kontrowersyjnym badaniu o prekognicji nie pokazał niczego nowego, gdyż każdy numer JPSP dostarcza mnóstwo dowodów na prekognicję. Niestety, trudno się nie zgodzić.

Psycholodzy byliby więc nieformalnymi „mistrzami świata” w wyciąganiu trafnych wniosków o rzeczywistości (na podstawie precyzyjnych teorii?). Rzeczywistość najprawdopodobniej nie rysuje się tak różowo. Choć w normalnych empirycznych czasopismach nie spotyka się prawie nigdy wzmianek o „porażkach”, to czasami niektórzy badacze przyznają, że większość ich danych nie została nigdy opublikowana, gdyż zwyczajnie nie przyniosły istotnych wyników (i robią to raczej w esejach niż w artykułach empirycznych). „Nie wiem jak to wygląda u was, ale większość moich danych nie zostało opublikowanych w dobrych czasopismach, ani nawet słabych czasopismach, większość moich danych leży sobie w szufladzie (ok, większość niedawnych danych jest rozrzuconych po dysku komputera)” to słowa redaktor naczelnej PoPS Barbary Spellman (2012, s. 58). „Jako czynny psycholog społeczny jestem przekonany, że co roku tysiące tak zwanych nieudanych badań w szufladach bez żadnej szansy na publikację (sam oczywiście mam takie w własnym biurku – mnóstwo pracy, żadnych efektów)”, to z kolei cytata z polskiego psychologa społecznego Jarosława Klebaniuka (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 takich jak oryginalne. Dodam, że były to na ogół tzw. replikacje konceptualne, a nie dokładne, uczestnikami byli ochotnicy, którzy nie otrzymywali gratyfikacji w postaci wynagrodzenia lub punktów do zaliczenia, przeprowadzającymi badanie byli studenci, a nie zawodowi badacze, zwykle też badani byli testowani w masowych sesjach, a nie w indywidualnych kabinach. Mimo tego, trudność w replikacji względnie nowych, mało zbadanych efektów jest niepokojąca, tym bardziej, że 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. W pracy dyplomowej próbowałem replikować sensacyjny (i silny) efekt nieświadomego myślenia (Dijksterhuis, Bos, Nordgren, van Baaren, 2006). Jedno z badań ukazało się m.in. w *Science*. Również nie udało mi się powtórzyć wyników (Budzicz, 2008).

Gdy zacząłem czytać artykuły w JPSP, odczuwałem rodzaj podziwu i respektu dla badaczy, którzy w kolejnych sześciu czy ośmiu badaniach wykazywali prawdziwość wyjściowej hipotezy. Wszystko przebiegało bez jakichkolwiek „zgrzytów”, choćby w postaci nieudanych eksperymentów. Wraz ze zdobyciem szerszych doświadczeń eksperymentalnych, uznałem za podejrzane, że badacze od razu „utrafili” z procedurą, manipulacją eksperymentalną, sposobem operacjonalizacji zmiennych, a gdy już utrafili, to potem każde z kolejnych badanie potwierdzało wyjściową hipotezę. Czasopisma psychologiczne zdają się wskazywać, że coś takiego jest „normą”. W często cytowanej systematycznej analizie Sterlinga, Rosenbauma i Weinkama (1995) sprawdzono, że średnio 95 % artykułów w psychologii odrzuca hipotezę zerową, a różnice między poszczególnymi subdyscyplinami są minimalne (od 93 % w psychologii eksperymentalnej do 97 % w psychologii klinicznej).

Osiem wychodzących pod rząd badań w psychologii klóci się nie tylko z codziennym laboratoryjnym doświadczeniem, ale również z twardymi prawami rachunku prawdopodobieństwa. Zrozumieniu, dlaczego tak się dzieje, pomocne będzie przypomnienie pojęcie mocy badania.

Moc badania to prawdopodobieństwo, że – przy założonej sile efektu oraz liczebności grupy badanej – określone badanie da wynik istotny statycznie (Cohen, 1992). Badacze dążą do maksymalizacji mocy, gdyż, co oczywiste, nie chcą marnować czasu i środków na badania, które nie dadzą wyników potencjalnie publikowalnych w dobrych periodykach. Hipotetyczne efekty niekoniecznie występują w naturze, ale nawet, gdy występują nie zawsze pokażą się w badaniu w postaci istotnych statystycznie zależności, szczególnie z powodu tzw. błędu próby (ang. *sampling error*), czyli różnych przypadkowych czynników, które wpływają na to, że próba jest odległa od reprezentatywności. Im większa próba, tym przypadek gra mniejszą rolę, podobnie silniejsze efekty są łatwiej powtarzalne. W próbie liczącej 6 kobiet i 6 mężczyzn może się zdarzyć, że kobiety będą średnio wyższe, ale w próbie 40 kobiet i 40 mężczyzn jest to bardzo nieprawdopodobne (tym bardziej, że efekt różnicy wzrostu w zależności od płci jest bardzo silny, por. Lippa, 2009). Gdybyśmy

porównywali wzrost 13- i 14-latków, porównanie średnich wyników dwóch grup może nie wykazać występowania istotnego efektu, o ile próba nie będzie naprawdę bardzo duża.

W badaniach o niewielkiej próbie nie tylko trudniej uzyskać istotny wynik, ale też uzyskuje się zróżnicowane siły efektu często mocno odbiegające od prawdziwej wartości (Ioannidis, 2008). Przy czym odchylenie to może być w dowolnym kierunku, a więc jeśli efekt wynosi rzeczywiście $d = 0,5$, w badaniu o małej próbie jest spore prawdopodobieństwo, że efekt będzie miał wartość zarówno $d = 1,0$ jak i $d = 0,0$ (im większa próba tym częściej wielkość efektu będzie się oscylować wokół rzeczywistego $d = 0,5$). Przykładowo, Lippa (2009) opisał dane z dużego badania internetowego dotyczącego różnic międzypłciowych, w którym wzięło udział ponad 200 tysięcy osób z 53 krajów. Możemy umownie potraktować badanych z każdego kraju jako osobną próbę. Rycina 1.2 prezentuje wielkości efektu różnic międzypłciowych dla jednej z badanych zmiennych (konkretnie: siły popędu seksualnego) w zależności od logarytmu z wielkości próby. Punkty tworzą charakterystyczny „lejek” (ang. *funnel*), który zwęża się wraz ze zwiększaniem wielkości próby. Im większa próba, tym przypadek traci na znaczeniu, a siła efektu coraz bardziej zbliża się do rzeczywistej¹⁹. Im próby mniejsze, tym siły efektów są coraz bardziej rozrzucone, a więc odnajdujemy zarówno efekty dużo słabsze i silniejsze niż rzeczywiste.

W badaniu nie można modyfikować siły efektu (o ile nie stosuje się wątpliwych praktyk badawczych), dlatego zasadniczo jedynym sposobem zwiększenia mocy jest zwiększenie liczebności grupy badanej²⁰. Idealne byłyby grupy bardzo liczne, ale ze względu na ograniczone środki, badacze muszą znajdować kompromisy pomiędzy dużą mocą, a rozsądną liczbą osób badanych²¹. Choć wielkość efektu jest nieznana, można kierować się siłami efektów w podobnych badaniach. Znając hipotetyczną wielkość efektu i przyjmując określony poziom pożądanej mocy, można precyzyjnie wyliczyć liczbę badanych. Szczególnie popularny program do takich obliczeń to G-

¹⁹ Oczywiście, rzeczywista siła efektu to nie tylko kwestia dużej próby, ale też trafności narzędzia pomiarowego. W opisanym wypadku można się też zastanawiać czy przypadkiem różnice międzykulturowe nie grają jakiejś roli, choć kształt rozkładu wskazywałby na to, że nie jest to duża rola.

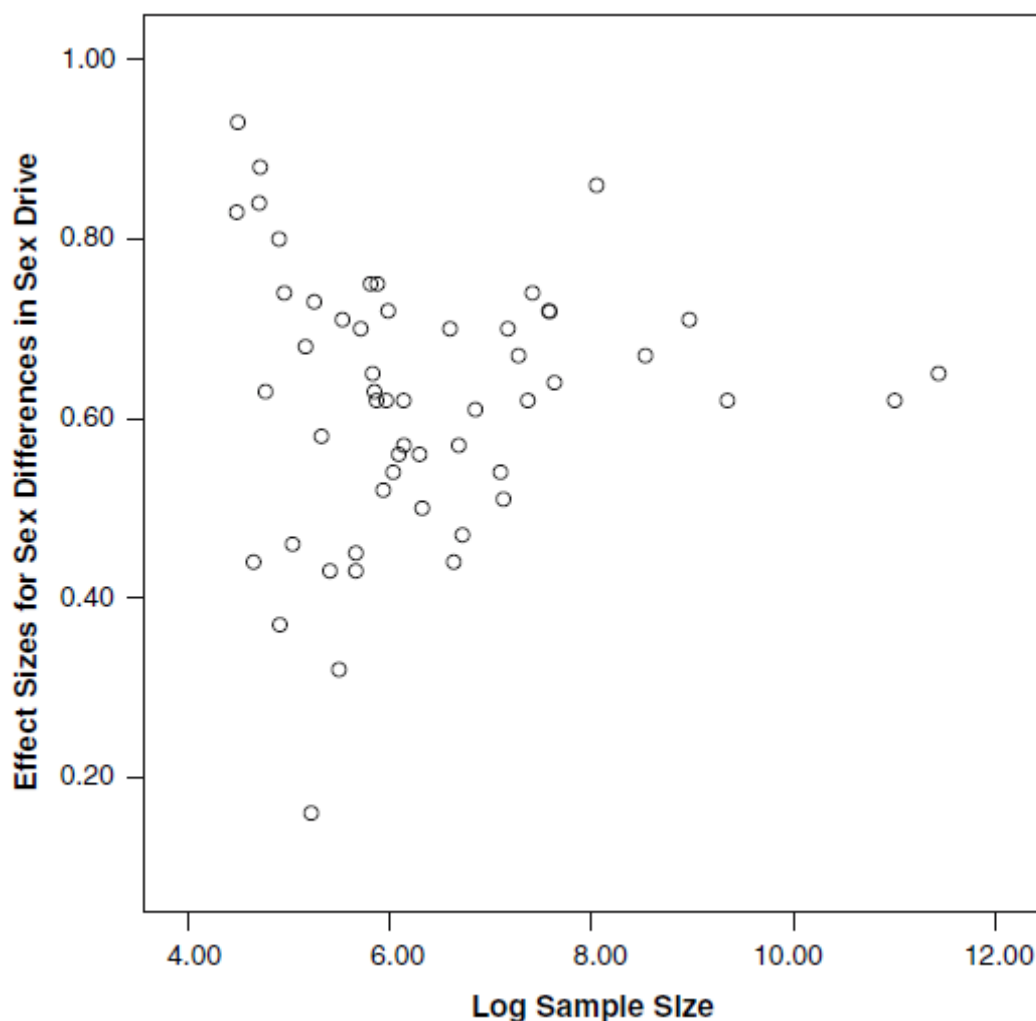
²⁰ Ale też bardziej rzetelne narzędzia zwiększają moc, por. LeBel, Paunonen (2011)

²¹ Na moc wpływa też założony poziom α . Przy bardziej wyśrubowanych poziomach (np. $p=0,001$) moc jest niższa niż przy „standardowym” $p=0,05$. Jako, że w psychologii prawie zawsze przyjmujemy poziom α równy 0,05 lub mniej, nie ma to zasadniczo znaczenia dla wyliczania mocy badania.

Power, wiele symulacji oczekiwanej mocy przedstawił też Cohen (1992). Zdaniem niektórych w psychologii istnieje niska statystyczna świadomość znaczenia mocy i wielkość grup badanych ustala się przede wszystkim na podstawie intuicji, a nie formalnej analizy mocy (Maxwell, 2004, Sedlmeier, Gigerenzer, 1989; Vankov, Bowers, Munafo, 2014).

Tak więc raport prezentujący 8 udanych badań o umiarkowanej sile efektów i niezbyt licznych próbach raczej nie zawiera całej prawdy. Muszą istnieć jakieś badania, których badacz nie opisał, lub, w najgorszym przypadku, opisane badania są sfałszowane lub zniekształcone. Wybiórcze publikowanie można analizować na poziomie zarówno serii badań opublikowanych w jednym artykule, meta-analiz, jak i całej nauki.

RYCINA 1.2. Rozkład wielkości efektu siły popędu seksualnego w zależności od płci w zestawieniu z logarytmem wielkości próby (w badaniu Lipa, 2009)



Źródło: przedruk z Lipa, 2009

W pokazywaniu możliwego wybiórczego publikowania na poziomie pojedynczych artykułów empirycznych poprzez obliczanie skumulowanej mocy²² wyspecjalizował się w ostatnich latach szczególnie Gregory Francis z Purdue University. W tabeli 1.2 przedstawiłem zakwestionowane przez niego teksty.

Przykładowo w jednym z artykułów (Francis, 2012e) przyjrzał się on dokładniej badaniu Galaka i Meyvis (2011) opisującym efekt polegający na tym, że osoby wspominały nieprzyjemne doświadczenie jako bardziej awersyjne, jeśli spodziewały się go powtórnie doświadczyć. Spośród ośmiu badań w siedmiu miały miejsce istotne efekty (ale i to nieistotne było bardzo blisko granicy 0,05). Analiza siły efektów oraz wielkości prób wskazywała jednak na to, że prawdopodobieństwo uzyskania takich wyników było skrajnie niskie, gdyż efekty były zawsze dosyć silne, a grupy mało liczne. Tak więc powinny były istnieć dodatkowe badania, nie opisane w tekście, bez istotnych efektów (ewentualnie dane w cytowanych badaniach zostały jakoś „podkreścone”, lub w najgorszym wypadku sfalszowane).

Pośrednią wskazówką wybiórczej publikacji jest też negatywna korelacja pomiędzy wielkością próby, a siłą efektu. W dużej próbie trudno przypadkowo uzyskać duży efekt, gdyż błąd próby jest mały. W małej próbie łatwiej uzyskać silny efekt, choć można też uzyskać efekt słabszy niż rzeczywisty. Statystycznie patrząc, nie ma żadnego związku między *średnią* siłą efektu, a wielkością próby. Choć średnie nie będą się na długą metę różnić, to mniejsze próby będą miały większą wariancję sił efektów (por. ryc. 1.2). Publikując tylko badania pozytywne, badania z mniejszymi próbami muszą siłą rzeczy zawierać silniejsze efekty, a badania z licznymi próbami słabsze efekty (co nie występowałoby, gdyby badacz publikował wszystkie raporty). Francis (2012e) zademonstrował, że w omawianych danych występuje silny ujemny związek pomiędzy wielkością próby a siłą efektu ($r = -0,86$), a prawdopodobieństwo uzyskania takiego rozkładu jest bardzo małe (w tym przykładzie wynosi mniej niż 1 %: Francis, 2012e, str. 588). Wybiórcze publikowanie samo w sobie nie jest oczywiście dowodem tego, że efekt nie występuje (jest zerowy), tym niemniej może zniekształcać obraz rzeczywistej siły efektu, bowiem niepublikowane badania z nieistotnymi efektami są to rzecz jasna, dane o słabszych efektach.

²² 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 swoje matematyczne wady i zalety. Po szczegóły odsyłam do Schimmack (2012, s. 555). Niezależnie od decyzji skumulowana moc jest jednak porównywalna.

TABELA 1.2. 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 laboratoryjnie mierzalny wpływ na teraźniejsze decyzje (Bem, 2011)	$r = -0,89$
Francis (2012b)	Obiekty "pożądane" są widziane jako bliższe (ang. <i>wishful seeing</i> ; 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 powodują 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)	$r = -0,80$
Schimmack (2012)	Podanie glukozy zwiększa siłę woli (Gaiot i in., 2007)	$r = -0,80$

Źródło: opracowanie własne na podstawie danych zawartych w cytowanych artykułach.

Francis został skrytykowany przez Simonsohna (2012), który zwrócił uwagę, że taka metoda też jest swego rodzaju techniką 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 znajdzie się na takie dane (analogicznie ktoś, kto długo gra w kasynie i w końcu wygrywa, niekoniecznie jest oszustem, może po prostu mieć wielkie szczęście). Zdaniem Simonsohna wybierając badania wg niesprecyzowanych kryteriów Francis nie jest w stanie wskazać, czy w nietypowym

wyniku zadziałał przypadek czy też celowe zniekształcenia. Odpowiedzią Francisa (2014) na ten zarzut jest systematyczna analiza, w której uwzględnił wszystkie artykuły zawierające przynajmniej 4 badania opublikowane w PS w latach 2009-2012. Ogółem 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 badań²³. Francis nazwał swój wskaźnik „testem zbyt dużego sukcesu” (ang. *test for excessive success*), konceptualnie bardzo podobny wskaźnik zaproponował Schimmack (2012) i nazwał go indeksem nieprawdopodobieństwa (ang. *incredibility index*). Schmick ograniczył się tylko do wskazania wysokiego nieprawdopodobieństwa dwóch serii badań: kontrowersyjnego efektu prekognicji Bema (2011) oraz niekontrowersyjnego, ale bardzo często cytowanego efektu zwiększania „siły woli” pod wpływem spożycia glukozy (Gaillot i in., 2007).

Opisywane prawidłowości mogą być też widoczne przy analizie efektów z dużo więcej niż jednego artykułu. Najprostszym sposobem sprawdzenia możliwości wybiórczej publikacji jest zmierzenie zależności pomiędzy wielkością efektów, a wielkością grupy badanej. Jedną z takich analiz w obszarze badań nad skutecznością programów nauczania matematyki przeprowadzili Slavin i Smith (2009). Wielkość efektu systematycznie malała wraz ze wzrostem grupy badanej. Dla badań z najmniej licznymi grupami ($N < 50$) średnia wielkość efektu wynosiła 0,44 (d Glassa), w przypadku badań tylko nieco większych ($51 < N < 100$) wielkość ta spadała do 0,29. W badaniach z najliczniejszymi grupami ($N > 2000$) średnia wielkość efektu wynosiła już tylko 0,09. Najwięcej badań obejmowało niewielkie grupy (46 badań z próbami $N < 100$; 49 z próbami $100 < N < 250$; ale tylko 23 z próbami $N > 2000$). Pamiętając o tym, że małe próby nie będą się *średnio* wiązać z wyższymi efektami (a tylko większą wariancją siły efektów), interpretacja tych danych może być taka, że badacze decydują się często na relatywnie małe badania, które nie wiążą się z dużym ryzykiem (mniejsze koszty finansowe i czasowe), a następnie „wychodzące” badania są publikowane, natomiast „niewychodzące” są chowane do szuflady, lub odrzucane przez redaktorów (w przypadku małych prób zapewne redaktorów może dodatkowo „przekonać” efekt o dużej sile). Badania na bardzo

²³ 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 4 badań wynosi tylko 2,56 %.

dużych próbach są zwykle finansowane grantami i wykonawcy częściej są zobowiązani do wykonania przynajmniej technicznego raportu.

Levine, Asada i Carpenter (2009) zbadali 51 meta-analiz ze wszystkich subdyscyplin psychologii obejmujących łącznie 75 efektów. W 59 z nich (79 %) stwierdzono negatywną korelację pomiędzy 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, że meta-analiza badań nad trafnością „kultowego” narzędzia w psychologii społecznej czyli testu utajonych skojarzeń (ang. *implicit association test*) zawiera nadmiar efektów pozytywnych relatywnie do niskiej mocy badań (Bakker i in., 2012).

Systematyczną analizę zależności pomiędzy siłą efektu, a wielkością próby w psychologii ogólnej przeprowadzili Kuhberger, Fritz i Scherndl (2014). Wylosowali oni początkowo 1000 tekstów z bazy PsycINFO z roku 2007. Odrzucono teksty przeglądowe, teoretyczne, meta-analizy, badania empiryczne opisowe i eksploracyjne, oraz teksty nie zawierające wszystkich danych (szczególnie wielkości efektów). Ostatecznie uwzględniono 395 badań. Gdy zestawiono wielkość próby i siłę efektu, okazało się, że występowała wyraźna negatywna zależność $r = -0,54$ między tymi dwiema zmiennymi (korelacja ta była nieco mniejsza $r = -0,45$, gdy odrzucono dane ze skrajnymi wielkościami prób tj. mniej niż 10 i więcej niż 1000). Gdybym miał wskazać tylko jeden artykuł najdobitniej pokazujący, że badania w czasopismach zawierają zniekształcony obraz rzeczywistości, wskazałbym właśnie tę analizę ze wszystkich cytowanych w tym rozdziale.

Istnieje teoretycznie możliwość, że negatywna korelacja pomiędzy 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 i in., 2014); dodatkowo, w obrębie jednego artykułu badania często różnią się bardzo subtelnie, trudno więc przypuszczać, żeby 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 pomiędzy wielkością próby,

a siłą efektu. Osobiście byłem świadkiem około 10 procedur wszczęcia przewodu doktorskiego i na żadnej nie analizowano mocy badania.

Na ogół jednak siłę efektów nie zestawia się z wielkością próby, ale ze skalkulowaną mocą badania. Zrobili tak między innymi Bakker, van Dijk, Wicherts (2012).

Wylosowali 11 meta-analiz opublikowanych w 2011 roku, oraz dodatkowo przeanalizowali popularną metaanalizę Greenwalda, Poehlmana, Uhlmana, Banaji'ego (2009) (ponad 1100 cytowań) dotyczącą trafności testu utajonych skojarzeń, a więc „kultowego” w ostatnich latach narzędzia w psychologii społecznej (por. rdz. 2.5). Zawarte w nich dane o wielkościach efektów zestawiono z mocą i sprawdzano, czy rozkład wielkości efektów jest dostatecznie symetryczny (po obydwu stronach średniej siły efektu) oraz czy dane nie zawierają nadmiaru efektów wyraźnie silniejszych od średniej wielkości efektu. W przypadku 4 meta-analiz znaleziono istotne odchylenie od spodziewanej symetryczności (polegające na nadmiarze silnych efektów w badaniach o relatywnie słabej mocy) oraz w przypadku 3 meta-analiz nadmiar efektów pozytywnych relatywnie do mocy i wielkości efektu. Dotyczy to też, meta-analizy Greenwalda i in. (2009), którą dodatkowo szerzej omówię w rdz. 2.5.

Tressoldi (2012) zebrał meta-analizy dotyczące efektów: mało kontrowersyjnych (prymowanie semantyczne), umiarkowanie kontrowersyjnych (efekt nieświadomego myślenia oraz efekt inkubacji), oraz skrajnie kontrowersyjnych (pozazmysłowa percepcja). Wielkość grup w tych badaniach kształtowała się standardowo na średnim poziomie 25-35 osób, natomiast same efekty (poza prymowaniem semantycznym) były słabe ($d < 0,3$). Oznacza to, że typowe opublikowane w czasopiśmie badania dotyczące tych zjawisk mają moc na poziomie 0,2-0,3 i grupy wielokrotnie mniejsze niż oczekiwane, żeby uzyskać przyzwoitą moc na poziomie 0,9. Tressoldi wyliczył, że badania nad efektem nieświadomego myślenia wymagałyby standardowo prób na poziomie 400 osób, żeby osiągnąć zadowalającą moc 90 %.

Analiz można też dokonywać na poziomie całej nauki. Najprostszy sposób polega na oszacowaniu średniej wielkości efektu występującego w danej dziedzinie i zestawienie tej wartości ze średnimi wielkościami prób. Bakker, Dijk i Wicherts (2012) analizując szerokie przeglądy literatury oszacowali średnią wielkość efektu w publikowanych raportach psychologicznych na około $d = 0,50$. Średnia całkowita

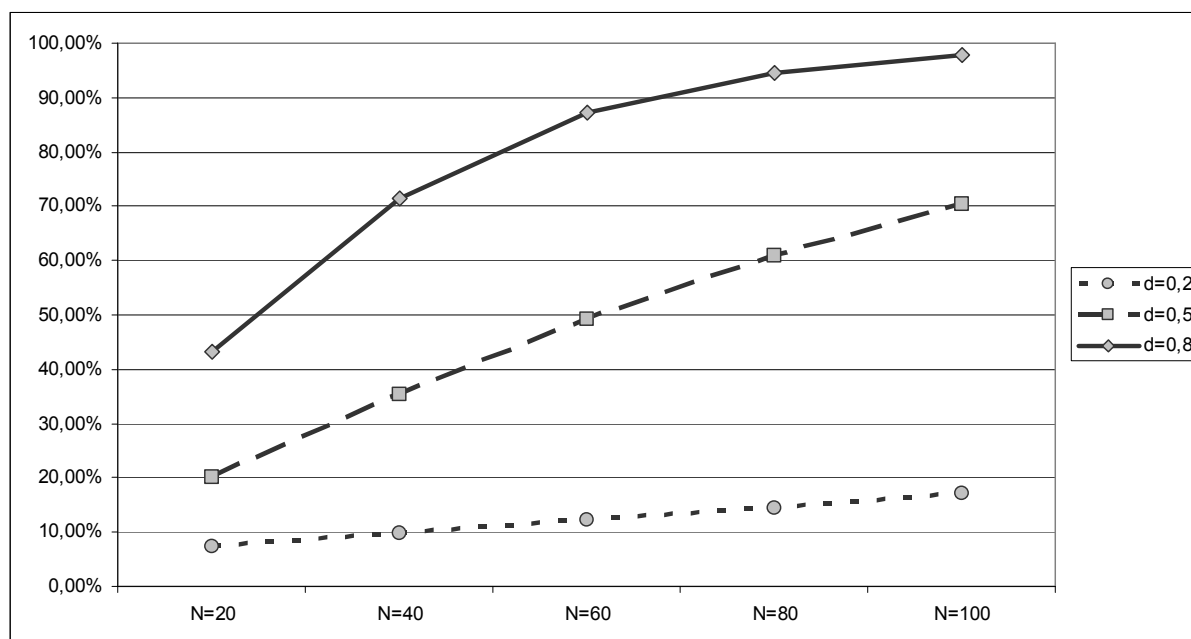
wielkość prób w badaniach to około 40 osób (Marszalek i in.²⁴) lub 24 osoby w jednej kratce (Wetzells i in., 2011²⁵). Bakker i in. zestawiając te dwie wartości wskazują na to, że moc typowych badań psychologicznych wynosi 0,35, a więc średnio istnieje tylko 35 % szans na znalezienie efektu, który rzeczywiście występuje. Za przyzwoity poziom mocy uważa się 90 % (Wolfe, 2013).

W psychologii społecznej najszerza meta-meta-analiza wielkości efektów została wykonana przez Richarda, Bonda i Stokes-Zoota (2003). Zestawiono wielkości efektów z 322 metaanaliz i otrzymano średni efekt $r = 0,21$, co w przybliżeniu odpowiada $d = 0,4$. Takie wyliczenie jest jednak mylące, gdyż prosta analiza wizualna częstości efektów wskazuje na to, że jest to rozkład wysoce prawoskośny i najczęstsza wartość to $r < 0,1$. Typowy efekt jest więc z pewnością dużo słabszy niż $r = 0,21$ (podobnie jak typowe zarobki są znacznie niższe niż średnia krajowa). Zakładając jednak nawet, że efekt wynosi faktycznie około $d = 0,4$, a wspomniane szacunki liczebności grup badanych są zaniżone i standardowa próba wynosi 50 osób w jednej grupie (np. w porównaniach dwóch grup w planie eksperymentalnym), moc wynosi w takim przypadku w przybliżeniu tylko 0,50. Na wykresie 1.1 przedstawiłem przykładowe wyliczenia mocy w zależności od wielkości próby i siły efektu przy założeniu standardowego poziomu $\alpha = 0,05$. Należy zwrócić uwagę, że o ile w przypadku bardzo silnych efektów, istnieje spora moc przy relatywnie niskich próbach, o tyle w przypadku słabych efektów (z jakimi na ogół spotykamy się w psychologii) moc utrzymuje się ciągle na niskim poziomie. W przypadku badania efektu o słabej sile $d = 0,2$, potrzeba aż 530 osób w jednej grupie, aby osiągnąć przyzwoitą moc 90%.

²⁴ Na podstawie zawartości czasopism: Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance, and Developmental Psychology

²⁵ Na podstawie zawartości czasopism: Psychonomic Bulletin & Review, Journal of Experimental Psychology: Learning, Memory and Cognition

WYKRES 1.1. Moc statystyczna badania w zależności od wielkości próby i siły efektu



Uwagi: dotyczy prostego porównania dwóch grup przy założonej sile efektu. Wielkość N dla jednej grupy (a więc całkowita próba badania = 2 x N)

Niska moc publikowanych badań jest pośrednim dowodem na wybiórcze publikowanie. Dowody takie można spróbować znaleźć w sposób bardziej bezpośredni tj. analizując, co dzieje się z danymi z przeprowadzonych badań, niezależnie od tego czy są istotne czy nie. Znany jest mi tylko jeden przykład takiej analizy. Cooper, DeNeve i Charlton (1997) sprawdzili, 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. Istotnie, ze 159 badań zgłoszonych i zaaprobowanych przez komisje etyczne, nie odbyły się tylko 4, a następne 4 zostały przerwane w trakcie. Zebrano więc 151 zbiorów danych, spośród, których przeanalizowano 121 zbiorów²⁶. Sporządzono pisemne raporty ze 105 zbiorów. Jako przyczyny braku raportu podawano najczęściej problemy z operacjonalizacją i procedurą, oraz to, że wyniki były „nieinteresujące” (50 %) lub nieistotne (31 %). Spośród pisemnych raportów tylko nieco więcej niż połowę wysłano do czasopisma (55). 50 raportów porzuconych na tym etapie

²⁶ Wśród 30 badań, gdzie zebrano dane, a które nie zostały przeanalizowali, badacze, na pytanie, dlaczego nie zrobili analiz, w olbrzymiej większości przypadków (90 %) zasłaniali się niepamięcią.

usprawiedliwiano brakiem zainteresowania publikacją (48 %), tym, że był to w zamyśle tylko projekt studencki (30 %), utratą zainteresowania tematem (26 %) oraz nieistotnymi (22 %) lub „nieinteresującymi” (20 %) wynikami. Ostatecznie 38 badań zostało opublikowanych w czasopismach i dalsze 3 w książkach. Badacze zapytali też o to, czy uzyskane wyniki były istotne czy nie i zapytali o los badań. 74 % badań z wynikami istotnymi zostało opublikowane lub chociaż wysłane do czasopisma, dodatkowe 14 % zostało zaprezentowane na konferencjach. Spośród nieistotnych danych aż 91 % nie doczekało się ani publikacji, ani choćby referatu. Tak więc kluczowa jest tu liczba 90 %. Taki odsetek danych jest przekazany społeczności badaczy, lub trafia do szuflady w zależności od samej tylko istotności statystycznej wyników.

Badanie Coopera i in. ma oczywiste ograniczenia wynikające z jego samoopisowego charakteru. Badacze, podobnie jak reszta populacji, mogą nie pamiętać dobrze przyczyn porzucenia projektów, a nawet zniekształcać te przyczyny, szczególnie po to, żeby racjonalizować swoje porażki i brak konsekwencji. Tym niemniej 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 (ang. *design problems*), trudno jednak rozstrzygnąć na ile problemy te są rzeczywiste, a na ile są rodzajem racjonalizacji. Tak czy owak, główny wniosek badania Coopera i in. jest niezaprzeczalny: niewielka część danych empirycznych ostatecznie ma szansę znaleźć się w publikacji (nieco poniżej 30 % wszystkich), ale szanse te radykalnie zwiększa istotność wyników. Nieistotność wyników powoduje praktycznie pewne zarzucenie planów publikacyjnych.

Warto jeszcze zauważyć, że wybiórcze publikowanie samo w sobie nie jest dowodem na brak efektu. Jeśli kilka badań (rzetelnie przeprowadzonych i opisanych) daje rzeczywiście istotne efekty, musiałoby istnieć kilkanaście, kilkadziesiąt badań z nieistotnymi wynikami, żeby te dane zniwelować (por. Ioannidis, 2008). Przykładowo, jeśli badacz przeprowadzi 20 badań, w około jednym, dwóch uzyska istotne wyniki przez sam przypadek ($20 \times 5\% = 1$). Ale już dziesięć istotnych wyników na dwadzieścia raczej nie jest przypadkiem. Wybiórcze publikowanie nie musi więc zasadniczo przekłamywać faktu występowania danego efektu, ale

najprawdopodobniej będzie przekłamywać informacje o jego sile. Niepublikowane badania siłą rzeczy będą charakteryzować się słabszymi efektami, a więc w zbiorczym zestawieniu metaanalitycznym siła efektu będzie mniej imponująca niż w opublikowanych danych. Wybiórcza publikacja ogranicza też możliwości systematycznej analizy efektu, czyli poznania jego natury. Badacze często nie przeprowadzają kilku identycznych badań, ale nieznacznie zmieniają elementy procedury, bodźce, sposoby operacjonalizacji itd. Niewychodzące badania mogą dawać nieistotne wyniki z powodu określonych moderatorów znoszących efekt. Taka informacja pozwala pełniej zrozumieć określone zjawisko. Oczywiście przy pojedynczym nieudanym badaniu, trudno rozstrzygnąć, czy jest to wpływ przypadku, czy raczej wpływ określonych moderatorów. Niemniej w dobrych meta-analizach w sposób systematyczny można przeanalizować wpływ określonych moderatorów na siłę efektu. Brak informacji o badaniach bez istotnych efektów będzie zniekształcał obraz wpływu moderatorów na efekt.

W badaniu Coopera i in. (1997) brak decyzji o wysłaniu do publikacji (albo chociaż sporządzenia raportu) usprawiedliwiano problemami z procedurą (i/lub operacjonalizacją), przy czym problemy te uświadamiano sobie już po zebraniu danych. Niewykluczone, że jest to rodzaj racjonalizacji. Choć nie znam danych empirycznych, które by to potwierdzały, spekulowałbym, że badacze mają tendencję do przeceniania proceduralnej wartości badań z istotnymi wynikami i nadmiernego doszukiwania się błędów w danych z wynikami nieistotnymi (recenzenci i redaktorzy zapewne podobnie). 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.

Na początku tego rozdziału zauważyłem, że poznawcza najbardziej interesujące są dane wskazujące na niezerowe efekty (im silniejsze tym lepiej). W dalszej części jednak zasadniczo krytykowałem wybiórczą publikację. Można zadać pytanie, kiedy wybiórcze publikowanie jest czymś szczególnie niepożądanym, a kiedy niewiele tracimy z niepublikowania danych negatywnych.

W moim przekonaniu, niezależnie od istotności, powinny być publikowane te dane, które są ważne, a więc takie, które odpowiadają na istotne pytania i w dużym stopniu przyczyniają się do rozwoju teorii oraz zrozumienia natury człowieka. Po drugie

powinny być publikowane te nieistotne wyniki, które przeczą wcześniejszym danym wskazującym na występowanie określonych efektów, tak, żeby nauka mogła realizować ważną funkcję samokorygowania. Po trzecie, uważam za głęboko niewłaściwe publikowanie w obrębie jednego artykułu tylko tych danych, które potwierdziły wyjściową hipotezę. Jeśli badacz przedstawia jakąś hipotezę powinien „wyspowiadać” się ze wszystkiego, co wie na ten temat, a nie tylko tego, co jest dla niego korzystne (na podobnej zasadzie, jak w wielu krajach prokurator musi pokazać sądowi też dowody świadczące potencjalnie o niewinności oskarżonego). W takich wypadkach niepublikowanie powinno mieć miejsce tylko w wyjątkowych sytuacjach, gdy wiadomo ponad wszelką wątpliwość, że dane są bezwartościowe (np. odkryto błąd w aparaturze wypaczający wynik pomiaru itp.).

W przypadku sytuacji, gdy seria badań nie przyniosła potwierdzenia jakiejś hipotezy (lub bardzo słabe potwierdzenie) publikowanie danych jest faktycznie dyskusyjne. Niektórzy (np. Nosek, Bar-Anan, 2012) uważają, że powinno się publikować wszystkie dane, które spełniają powszechnie przyjęte standardy metodologiczne i które nie zawierają oczywistych błędów w sztuce (oraz, rzecz jasna, są opisane w zgodzie ze standardami APA). Jednak w przypadku danych, które są głęboko wtórne, testują wątpliwe hipotezy, trudno mi się zgodzić z tą tezą. Sytuacja natomiast wygląda inaczej, gdy dane odpowiadają na „ponadczasowo” interesujące pytania, zarówno ze względu na zrozumienie natury ludzkiej, jak i ze względów praktycznych (pytania np. czy religijność, małżeństwo, uprawianie sportu (itd.) wpływa na poziom dobrostanu psychicznego; czy atrakcyjność fizyczna ułatwia manipulowanie innymi; czy oglądanie przemocy w telewizji lub grach komputerowych zwiększa poziom rzeczywistej agresji itd.) oraz używają nowatorskich niebanalnych metod (szczególnie danych behawioralnych – więcej o tym w rozdziale 2). Bardzo duży wysiłek organizacyjny (który będzie się wyrażał w bardzo wysokiej mocy) powinien być dodatkowym mocnym argumentem za publikacją, oczywiście przy założeniu, że jest spełniony wcześniejszy warunek dużej poznawczej istotności badania.

W literaturze zwraca się szczególnie uwagę na to, jak wybiórcze publikowanie jest niekorzystne poznawczo dla rozwoju bazy rzetelnej wiedzy. Uważam jednak, że publikacja tylko istotnych wyników jest też wysoce szkodliwa *psychologicznie*. Ustanawia nadmiernie wysokie wymagania wobec badaczy, które są nierealistyczne w nauce takiej jak psychologia. Brak negatywnych wyników w najlepszych

czasopismach może sprzyjać kulturze mniejszych i większych przekłamań. Liczba stanowisk na uczelniach jest ograniczona relatywnie do liczby zainteresowanych, podobnie ilość miejsca na łamach czasopism jest dużo mniejsza niż liczba chętnych (w najlepszych czasopismach psychologicznych odsetek odrzuceń to 80%-90%: Suls i Martin, 2009). Osoby najbardziej zdeterminowane mogą zechcieć „pomóc” szczęściu, a jak już wiemy, takich możliwości jest potencjalnie wiele. Przypadek Stapela pokazał, że dysponując inteligencją, wprawnym piórem i znajomością standardów można to robić bardzo długo przy niewielkim ryzyku wpadki. Z kolei osoby, które chcą zachować pełną rzetelność, mogą być zmuszone do wybiórczego publikowania przez występujące praktyki redakcyjne.

Badania nad mocą badań w psychologii nie są jedynym niepokojącym sygnałem, że z danymi nie wszystko jest w porządku. Inne takie sygnały to nieprawidłowościami przy raportowaniu wartości p , rozkład wartości p oraz nieudostępnianie danych surowych do analiz. Ostatecznym probierzem rzetelności nauki i jej zdolności do samokorekty jest praktyka systematycznych replikacji. Omówię teraz te obszary.

1.5. Rozkład wartości p w literaturze

Oprócz analizy mocy badań, istnieje jeszcze jeden matematyczny sposób na wykazanie, że literatura może zawierać nadmiar naciąganych danych. Tym sposobem jest analiza rozkładu wartości p w artykułach empirycznych (ang. *p-curve analysis*). Założenia takiej analizy przedstawili niedawno Simonsohn, Nelson, i Simmons (2013). Podstawowe założenie, sprawdzone w matematycznych symulacjach, jest takie, że rozkład wartości p dla niezerowych efektów powinien być prawoskośny, im mocniejszy efekt, tym bardziej prawoskośny (ale nawet dla słabych efektów systematyczne badania powinny generować rozkład prawoskośny). Mówiąc prościej, jeśli rzeczywiście występuje istotny efekt to dużo częściej będziemy spotykali badania z wartościami $0 < p < 0,01$ niż, dajmy na to $0,04 < p < 0,05$. 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 rzecz jasna 1 %), a więc rozkład

jest płaski. Jednak w przypadku praktyki „hakowania”²⁷ wartości p jej rozkład będzie inny niż prawoskośny, a więc przykładowo może być więcej wyników $0,04 < p < 0,05$ niż $0 < p < 0,01$.

Autorzy przetestowali swoją hipotezę empirycznie losując dwie grupy 20 badań z JPSP. Pierwsza grupa, zdaniem autorów „prawdopodobnych”, to badania, w których testowano jakieś proste efekty bez żadnych dodatkowych współzmiennych i efekty te okazały się być istotne. Druga wylosowana grupa to badania, w których występowały współzmiennne (np. płeć). Wyodrębnianie współzmiennych samo w sobie nie jest niczym nagannym, szczególnie jeśli ma to dobre teoretyczne uzasadnienie. Jednak testując dowolną zależność (dajmy na to jakiś prosty efekt manipulacji eksperymentalnej) mamy większą szansę, że znajdziemy jakieś istotne efekty, jeśli wrzucimy współzmienną do analiz. Analiza rozkładu p rzeczywiście pokazała, że istotne zależności częściej znajdowano ze współzmiennymi niż bez nich. Najwięcej statystyk p przyjmowało w pierwszym przypadku wartości od 0,04 do 0,05, w przypadku danych bez współzmiennych najczęściej danych notowano dla wartości $p < 0,1$. Przedstawia to Tabela 1.3. Rozkład wartości p dla prostych efektów był bardzo zbliżony do rozkładu wartości p dla hipotetycznego rozkładu danych o mocy $d = 0,33$ (por. Simmonsohn i in. 2013, ryc. 3). Zwróćmy uwagę, że dla danych potencjalnie rzetelnych, wzór rozkładu p jest prawoskośny poza wynikami tuż poniżej wartości p . Może to potencjalnie wskazywać na „naciąganie” danych o wyższej niż 0,05 wielkości p do wartości poniżej p .

TABELA 1.3. Rozkład częstości wartości p w badaniach potencjalnie nastawionych na „łowienie p ” (tj. ze współzmiennymi) i badania potencjalnie rzetelnych (bez współzmiennych)

Rodzaj danych	$0,0 < p < 0,01$	$0,01 < p < 0,02$	$0,02 < p < 0,03$	$0,03 < p < 0,04$	$0,04 < p < 0,05$
Dane potencjalnie „zhakowane”	5 %	5 %	15 %	35 %	40 %
Dane potencjalnie rzetelne	45 %	23 %	14 %	5 %	14 %

Źródło: na podstawie danych zawartych w Simmonsohn i in. (2013)

²⁷ Tak można przetłumaczyć to, literacko atrakcyjne, określenie, użyte w tytule artykułu (ang. *p-hacking*). Inna nazwa to „łowienie p ” (ang. *p-fishing*)

Konceptualnie zbliżoną analizę przeprowadzili Mascampo i Lalande (2012). Wyodrębnili wartości p w przedziale $0,01 < p < 0,1$ w artykułach empirycznych z trzech znaczących czasopism²⁸ z 12 numerów z roczników 2008 i 2007. Ogółem wyodrębniono ponad 3500 wartości p . Następnie ich rozkład poddano analizie (m.in. uwzględniając różną „grubość” analizowanych przedziałów). Miały one kształt wyraźnie prawoskośny (o kształcie zbliżonym do funkcji wykładniczej $y=1/(2^x)$), jednakże rozkład nieco poniżej wartości 0,05 wyraźnie odbiegał od reszty (miała miejsce na wykresie „górkę” tuż przed tą wartością). Takich wartości p było nieproporcjonalnie dużo, niezależnie od tego, jaki przyjęto przedział i niezależnie od tego, jakie czasopismo analizowano.

Charakterystyczny wzrost częstości tuż poniżej $p = 0,05$ może świadczyć o wybiórczym publikowaniu i o stosowaniu metod podkręcania danych.

Najprawdopodobniej wielu autorów, którzy otrzymują zbliżone do pożądanej wartości p (np. $p = 0,07$) podejmuje opisane wcześniej zabiegi mogące doprowadzić do niższych wartości. 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), i jako taka, nie ma większego teoretycznego znaczenia. Trudno spodziewać się tu jakichś przekłamań. Można natomiast oczekiwać, że 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 Mascampo i Lalande (2012) zostały powtórzone w konceptualnie analogicznym badaniu Leggetta i in. (2013), wykonanym na tych samych czasopismach. Sprawdzali oni również zmianę na przestrzeni czasu, mianowicie porównywali roczniki czasopism z 2005 i 1965 roku. W nowszych artykułach efekt nadprezentacji p tuż poniżej 0,05 był wyraźniejszy. Szczególnie w JPSP zanotowano wzrost takich danych na przestrzeni dekad (efekt był minimalny w JEP:G).

²⁸ Journal of Experimental Psychology: General, JPSP, Psychological Science

Podobnie Kuhberger i in. (2014) we wspomnianej już analizie 395 artykułów empirycznych, stwierdzili, że 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 poza wymienionymi artykułami nie mogłem zidentyfikować większej liczby tekstów jej używających. Tym niemniej wszystkie 4 artykuły wyraźnie wskazują na statystycznie nieprawdopodobne anomalie w takich rozkładach, potencjalnie wskazujące na manipulacje danymi i wybiórcze publikowanie. Nie można też wykluczyć, że często przyczyna jest dużo prostsza: autorzy zmieniają wartości p bez zmiany innych statystyk, po to tylko, żeby zejść poniżej „magicznego” progu 0,05. Ten wyjątkowo nieudolny sposób fałszerstwa jest na szczęście wykrywalny. Dzięki temu okazuje się, że występuje wcale nierzadko.

1.6. 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 $p = 0,057$), zostały dokładniej przeanalizowana przez holenderskich badaczy. Bakker i Wicherts (2011; badanie 1) wylosowali 6 czasopism psychologicznych, trzy o wysokiej cytowalności ($IF > 4$) i trzy o relatywnie niskiej cytowalności²⁹ ($IF < 1,5$), a następnie przeanalizowali wszystkie artykuły empiryczne z rocznika 2008. Z nich wyodrębniono testy istotności statystycznej, w których używano statystyk χ^2 , 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, który nie zmieniały zasadniczo istoty wyniku (np. $p < 0,01$, gdy w rzeczywistości $p = 0,01^{30}$), ale sporadycznie zdarzały się poważniejsze błędy polegające na takim niewłaściwym podaniu wartości p , że

²⁹ Te czasopisma to w grupie wysoko cytowanych: *JPSP*, *Journal of Child Psychology and Psychiatry* i *Development and Psychopathology*, a w grupie nisko cytowanych: *Journal of Black Psychology*, *Journal of Applied Developmental Psychology* i *Journal of Research in Reading*.

³⁰ Uwzględniano również to, że podana wartość statystyki p może być zaokrągleniem. Jeśli więc autorzy artykuły podawali $p = 0,03$, nie liczono tego jako błąd, jeśli wartość p mieściła się w przedziale $<0,025; 0,035>$.

prawdziwa wartość powodowała zmianę interpretacji wyniku z istotnego na nieistotny, lub odwrotnie. Jak można się było zawczasu domyślić, 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 6 czasopism, autorzy postanowili przeanalizować blisko 300 losowych artykułów ze wszystkich czasopism psychologicznych uwzględnionych w bazie PsycINFO (również rocznik 2008) (Bakker i Wicherts, 2011, badanie 2). Wyniki były w miarę zbliżone do pierwszego badania, tj. około 35 % artykułów zawierało przynajmniej jeden błąd, a w 7 % artykułów błąd zmieniał interpretację wyniku.

Koncepcyjnym rozwinięciem jest badanie Wicherts, Bakker i Molenaar (2011). W podobny sposób przeanalizowali oni artykuły z drugiej połowy roku 2004 z dwóch wpływowych czasopism³¹. Skontaktowali się następnie z autorami z prośbą o podzielenie się danymi surowymi. Badanie odbywało się w 2005 roku, a więc autorzy, zgodnie z obowiązującymi zasadami powinni posiadać jeszcze odpowiednie dane i udostępniać je na życzenie. Jednakże tylko 42 % autorów podzieliło się danymi. Pozostali nie odpowiadali 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 „grubych” błędów polegających na zamianie wartości p , w taki sposób, że zmieniało to decyzje o przyjęciu lub odrzuceniu hipotezy zerowej (choć znajdowano inne 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 nie udostępniający 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 podanych przez autorów. Odkryli, że w 36 przypadkach na 93 (38 %) pomimo tego, że raportowano p

³¹ JPSP, oraz JEP: *Learning, Memory and Cognition*

= 0,05, w rzeczywistości wartość ta wynosiła $p > 0,05$. Wszystkie takie wypadki dotyczyły JPSP, oraz 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 były wykrywalne na podstawie danych opublikowanych w artykułach przez *samych* autorów. Jest to wyjątkowo prosty, a mimo to, jak wynika z opisanych analiz, powszechny sposób fałszerstwa polegający 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 i Wicherts, 2011; rycina 3). Autorzy artykułów z grubymi błędami, zmieniającymi interpretację wyniku, nie dzielili się swoimi danymi surowymi, co może wskazywać na to, że zdawali sobie sprawę z tego, że 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). Dodatkowo, nieco zaskakująco, ale 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 nie udzielają przy tym odpowiedzi na pytanie, w jakiej części artykułów wartość p została zmieniona bardziej „inteligentnie”, poprzez równoległe sfalszowanie wartości statystyk.

Patrząc bardziej krytycznie na dwie wspomniane analizy, można zauważyć, że podobnie jak w przypadku analizowania rozkładu wartości przez Masicampo i Lalande (2012) wszystkie przeliczalne wartości p wrzucono do jednego worka, bez analizy tego, które dotyczą kluczowych z punktu widzenia artykułu hipotez, a które dotyczą hipotez marginalnych lub kompletnie nieistotnych.

1.7. Dzielenie się danymi

Wicherts i in. (2011) zauważyli, że autorzy, którzy nie chcą dzielić się danymi surowymi, mogą mieć na sumieniu mniejsze i większe zafałszowania, znajdowano w ich artykułach więcej błędów związanych z wartością p i ogólnie mieli słabsze dane w swoich badaniach. Dlatego warto przyrzeć się bliżej praktyce dzielenia się danymi surowymi. John i in. (2012) nie uwzględniają kategorii „Nie podzieliłem się danymi surowymi na żądanie, pomimo tego, że wcześniej się do tego zobowiązałem”, a szkoda, bo nie tylko Wicherts i in. (2011) zwracali uwagę na niechęć autorów do wysyłania swoich danych surowych.

Dzielenie się danymi jest obowiązkiem badacza, który biorą na siebie w momencie wysyłania artykułu do publikacji (APA, 2010). Badacze mylą się lub brakuje im odpowiednich kompetencji, więc reanaliza danych przez innych badaczy pozwala nieraz zakwestionować oryginalne wnioski (lub sposób pokazać coś nowego w sposób pozytywny).

Żeby to zilustrować omówię szerzej artykuł Blanton i in. (2009), gdzie poddano ponownej analizie wnioski z dwóch badań z wykorzystaniem testu utajonych skojarzeń (IAT – patrz rdz. 2.5) (w przypadku kilku innych badań, autorzy nie mogli zrobić analiz ze względu na nieudostępnienie danych przez autorów). Pierwsze badanie wskazywało na związek pomiędzy postawą względem osób czarnoskórych mierzoną testem IAT, a zachowaniem względem czarnoskórego eksperymentatora ocenianą przez sędziów kompetentnych (McConnell, Leibold, 2001). Autorzy analizy przedstawili szereg wątpliwości po ponownej analizie danych. Przykładowo zmienne zależne były arbitralnie wyznaczane bez jakiegokolwiek uzasadnienia, czemu takie a nie inne (np. zmienna jako oceny zachowania względem białego minus ocena zachowania względem czarnoskórego), gdy inne rodzaje zmiennych nie przynosiły analogicznych efektów (np. ocena zachowania względem osoby czarnoskórej). Istotne efekty wynikały z oceny jednego z sędziów, ale u drugiego sędziego takich efektów nie zanotowano. Poza tym, jak się okazało, usunięcie tylko jednej skrajnej obserwacji likwidowało też niektóre istotne efekty (np. zamiast $p < 0,05$ osiągnęto $p < 0,13$).

W drugim badaniu poddanym reanalizie (Ziegert i Hanges, 2005) osoby wcielały się w rolę menedżera oceniającego CV potencjalnych pracowników. Dodatkowo,

poprzez instrukcje, manipulowano też „klimatem”, podkreślając znaczenie równości i braku dyskryminacji lub znaczenie tego, żeby nowy pracownik był tej samej rasy, jak osoba oceniająca. Oceny były bardziej negatywne względem osób czarnych im bardziej negatywna była postawa względem osób czarnych w teście IAT, ale tylko w „klimacie” dyskryminacji. Autorzy przeanalizowali ponownie te dane, i nie byli w stanie odtworzyć istotności wyników niektórych złożonych interakcji (np. uzyskiwali $p < 0,07$, zamiast $p < 0,05$). Wizualna analiza rozrzutu danych ujawniła, że kluczowa istotna korelacja opiera się na 3 obserwacjach mocno odstających od reszty. Usunięcie tych obserwacji z analizy likwidowało istotność efektu.

Dodatkowo Blanton i Mitchell (2011) analizując inny artykuł dotyczący trafności IAT (Heider, Skowronski, 2007) odkryli m.in. selektywne raportowanie zmiennych zależnych, które zasadniczo wypaczało wnioski z badania.

Warto dodać, że dwa analizowane artykuły nie pochodziły z jakichś marginalnych czasopism, ale z JESP i *Journal of Applied Psychology*, a więc czasopism ewidentnie elitarnych w dziedzinie psychologii społecznej.

Powyższe analizy byłyby niemożliwe, gdyby autorzy nie podzielili się danymi. Nic więc dziwnego, że punkt 8.14 kodeksu etycznego APA (2010) głosi: „Po publikacji wyników badań psychologowie nie odmawiają dostępu do danych, na których oparli swoje wnioski, innym kompetentnym ekspertom, którzy chcieliby zweryfikować zasadnicze twierdzenia poprzez reanalizę i którzy chcieliby użyć danych tylko do tego celu [...]”. Oprócz pożytków z reanalizy wskazuje się też, że posiadanie indywidualnych danych umożliwia przeprowadzanie bardziej szczegółowych meta-analiz niż posiadanie tylko danych zbiorczych (Cooper, Patall, 2009).

Przypomnę, że 58 % badaczy nie podzieliło się danymi w badaniu Wicherts i in. (2011). Kilku innych autorów również podjęło systematyczne próby analizy skłonności badaczy do dzielenia się danymi. Wolins (1962) próbował uzyskać dane od 37 badaczy z artykułów³² z czołowych czasopism APA. Choć większość odpowiedziała, ostatecznie dane otrzymał od 9 badaczy (25 %), z czego kilku wysłało je wiele miesięcy po pierwotnym zapytaniu (przeanalizował 7 zbiorów danych, a w 3 znalazł poważne błędy w procedurach statystycznych). Nieco lepsze

³² Z lat 1959-1961.

rezultaty uzyskali Craig i Reese (1973), którzy na 53 zapytania otrzymali 20 zestawów danych (37%).

Powyższe zapytania o dane pochodzą jednak z zupełnie innej epoki, w której dane były w większości zgromadzone w formie papierowej, i ich sama wysyłka mogła czasami być sporym logistycznym problemem (w badaniu Craiga i Reese'a jeden z zapytanych prosił o czek na pokrycie kosztów obsługi biurowej). Zupełnie inaczej sprawa powinna przedstawiać się dzisiaj, gdy praktycznie wszystkie dane surowe mają postać cyfrową (wcześniej czy później). Dobrze zorganizowanemu badaczowi powinno zająć kilkanaście sekund dotarcie do odpowiedniego folderu, dodanie załącznika i odesłanie maila (przy optymistycznym założeniu, że dane są sensownie opisane). Przyjrzyjmy się zatem jak sprawy mają się w XXI wieku. Wicherts i in. (2006) wysłali prośby o dane do autorów 141 artykułów empirycznych, które ukazały się około pół roku wcześniej w 4 czołowych czasopismach APA³³. Cytując autorów:

Jako, że w skład 141 artykułów wchodziło łącznie 249 badań, spodziewaliśmy się realistycznie otrzymać 90 do 100 zestawów danych [...]. Niestety, pół roku później, po wymianie ponad 400 maili – i wysłaniu niektórym autorom szczegółowych opisów naszych celów badawczych, zgody komisji etycznych, podpisanych zobowiązań do niedzielenia się danymi, a nawet naszych CV – doszliśmy do marnych 38 pozytywnych reakcji i mieliśmy dane z 64 badań. Oznacza to, że 73 % autorów nie podzieliło się swoimi danymi (Wicherts i in., 2006, s. 727)

Spośród badanych czasopism najlepiej wypadli autorzy *Journal of Experimental Psychology* (41 % zgód), a najgorzej piszący w *Developmental Psychology* (15 %). JPSP uzyskał wynik zbliżony do średniej (22 %).

Mitchell (2012) usiłował uzyskać dane z badań z użyciem wspomnianego już IAT z 42 artykułów. Ostatecznie pozyskał 17 zestawów. W 9 przypadkach autorzy zasłaniali się utratą danych (np. wskutek przeniesienia na inną uczelnię lub awarii komputera), w 4 obiecali wysyłkę, ale tego nie zrobili, w kilku wypadkach maile nie spotkały się z żadną odpowiedzią. Tylko w 5 przypadków autorzy jawnie odmówili, co świadczy o tym, że autorzy zdają sobie sprawę ze swojego obowiązku, ale tak czy owak go nie wypełniają. Mitchell opisuje wymowny przypadek:

Wdałem się w długie negocjacje z jednym z autorów. [...] Ostatecznie nie udostępniono mi danych, pomimo moich powtarzanych oświadczeń dotyczących tego, jakie dokładnie analizy przeprowadzę, pomimo mojej deklaracji, że wyniki analiz wyślę najpierw autorowi, zanim

³³ Te czasopisma to: JPSP, *Developmental Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Experimental Psychology: Learning, Memory, and Cognition*

zostaną przekazane do czasopisma, i pomimo propozycji, że w przypadku braku zgody, co do wyników analizy, dane zostaną przekazane jakimś neutralnym ekspertom, aby rozstrzygnęli spór (Mitchell, 2012, s. 13).

Dane więc nie tylko są rzadko udostępniane, ale i badacz musi się bardzo napracować, żeby w ogóle je zdobyć. W związku z powyższym nie dziwi, że pojawiły się postulaty radykalnej zmiany modelu udostępniania danych (Simonsohn, 2013; Wicherts i Bakker, 2012; Klebaniuk, 2012). Zdaniem niektórych obecna praktyka „podziel się danymi surowymi na żądanie” jest niewystarczająca. Dane są, jeśli wierzyć samym deklaracjom, gubione, tracone wskutek awarii sprzętu lub kradzieży tegoż, albo po prostu autorzy deklarują gotowość odesłania danych, która nigdy nie następuje. Postulowanym rozwiązaniem byłaby równoległa i obligatoryjna publikacja danych surowych jako materiałów uzupełniających do artykułu w Internecie. Trudno oprzeć się wrażeniu, że obecny model udostępniania danych zupełnie nie przystoi do współczesnych czasów, ale bliżej mu do okresu, gdy jedyną formą komunikacji naukowej były papierowe czasopisma. Technicznie nic nie stoi obecnie na przeszkodzie, żeby autorzy dobrowolnie udostępniali dane w Internecie, lub żeby czasopisma same je udostępniały jako dane uzupełniające. Istnieje już do tego stosowna infrastruktura informatyczna (zob. rdz. 1.10). Oczywiście w szczególnych przypadkach, gdy dane są wrażliwej natury, lub są tego rodzaju, że możliwa jest imienna identyfikacja badanych, należałoby zastosować dodatkowe obostrzenia. Być może wartym rozważenia byłby model, jaki stosuje wydawnictwo Oxford University Press, które udostępnia na platformie internetowej ryciny i wykresy ze swoich podręczników akademickich, ale tylko dla osób, które przejdą weryfikację jako nauczyciele akademicy³⁴. Dane skrajnie wrażliwe mogłyby być udostępniane po dodatkowej „ręcznej” weryfikacji i zgodzie badacza. Sądzę jednak, że problem ten dotyczy naprawdę niewielkiego ułamka danych w psychologii społecznej (por. też Mitchell, 2012).

Obowiązek składowania i udostępniania danych z pewnością nie jest absolutną gwarancją braku oszustw, bowiem nawet upublicznione dane mogą być zwyczajnie sfabrykowane. Przypadek Stapela pokazuje jednak, że sfabrykowanie danych nie budzących wątpliwości nie jest, wbrew pozorom, łatwe. Redaktorów nic by nie kosztowała zmiana polegająca na wymaganiu wysłania danych surowych równolegle

³⁴ Na własnej skórze ten model pozytywnie przetestowałem.

z wysyłką manuskryptu, a tylko nieduży dodatkowy wysiłek byłby wymagany do szerokiego udostępniania danych na platformach internetowych. Póki co jednak nie słyhać o wielu planach takich zmian. Przysłowiową „jaskółką” w tym względzie wydaje się być nowe czasopismo APA w formule Open Access – *Archives of Scientific Psychology*. W pierwszym redakcyjnym artykule zaznaczono, że jednym z warunków publikacji jest udostępnienie danych surowych na portalu APA lub w innym publicznym repozytorium (Cooper, VandenBos, 2013). W tym kontekście zapewne warto by się zastanowić nad jeszcze dalej idącym rozwiązaniem, czyli publikowaniem w formie materiałów uzupełniających wszelkich użytych materiałów, które można opublikować (np. nie podlegają prawom autorskim ze strony osób trzecich). Przykładowo, skrypty komputerowe użyte do badań z wykorzystaniem komputerów, z pewnością same w sobie nie podlegają prawom autorskim (oczywiście, przy założeniu, że badacz dysponuje programem, który używa skryptu). Takie działania mogłyby w dużym stopniu ułatwiać przeprowadzanie replikacji, którym poświęcam następny podrozdział.

1.8. Replikacja danych w psychologii

(W tym podrozdziale będę omawiał tylko problemy psychologii związane z replikacjami dokładnymi, a więc takimi, w których warunki są tak bardzo zbliżone do oryginalnego badania, jak to tylko możliwe. Replikacje nigdy nie są identyczne, gdyż odbywają się na innych badanych, a jeśli nawet na tych samych badanych, to z umysłem zmienionym przez udział w oryginalnym badaniu. Replikacjom konceptualnym (konceptyjnym), a więc takim, gdzie celowo modyfikujemy elementy procedury, próbę, bodźce, sposoby operacjonalizacji itp., żeby lepiej poznać sam efekt, a nie zasadniczo dowieść jego prawdziwości poświęcam w dużej części rozdział drugi.)

Być może wszystkie problemy psychologii fałszywej pozytywnej nie miałyby żadnego praktycznego znaczenia, gdyby w psychologii sprawnie działał mechanizm replikacji. Gdyby standardem były regularne replikacje badań, Stapel mógłby zostać szybciej zdemaskowany, lub robiłbym mniej błyskotliwą, ale przynajmniej uczciwą karierę. O ważności replikacji pisze się od zawsze, jednak trudno nie oprzeć się wrażeniu, że z praktyką jest znacznie gorzej. Czasopisma, szczególnie dobre czasopisma, nie są

chętne, aby publikować replikacje dokładne. Wymowny jest przykład zamieszczenia po publikacji w JPSP kontrowersyjnego artykułu Bema (2011). Dowodził on jakoby występowania prekognicji, czyli zjawiska polegającego na tym, że fakty z przyszłości oddziałują na umysł w *przeszłości*. Grupa badaczy pod kierunkiem Richarda Wisemana dosyć szybko przystąpiła do replikacji (oczywiście nieudanej), a następnie skierowała do JPSP artykuł. Redaktor Elliot Smith nie tylko odrzucił replikację, ale wręcz nie wysłał ich do recenzji, odpowiadając, że „to czasopismo nie publikuje replikacji [dokładnych], niezależnie od tego czy są udane, czy nie. Nie jesteśmy Journal of Bem Replication” (za: Aldhous, 2011). Wiseman swoją replikację opublikował w końcu w PLoS One (Ritchie, Wiseman, French, 2012) natomiast redakcja JPSP też nie była do końca konsekwentna, bo opublikował jednak nieudaną replikację eksperymentu Bema, ale przeprowadzoną przez Galaka, LeBoeufa, Nelsona i Simmonsa (2012). Być może redaktorzy podjęli ten krok kierując się poczuciem przyzwoitości i nieformalnym naciskiem opinii publicznej. Wielkość próby w badaniu Galaka również była niespotykana jak na badania eksperymentalne (ponad 3 tys. osób). Tak czy owak, faktem jest, że poza wyjątkowymi sytuacjami JPSP obecnie nie publikuje replikacji (słowo „replication” lub „replicate” w tytule artykułu w tym czasopiśmie jest użyte zaledwie 28 razy, z czego 22 przypadki pochodzą z lat 1965-1990). (Co jednak się zmieniło w połowie 2014, już po pierwszej redakcji tego podrozdziału; patrz rdz. 1.10)

Badacze, którzy decydują się replikować badania innych, mogą, niestety, działać na niekorzyść własnej kariery. Jak napisała Barbara Spellmann:

Udane replikacje są niepublikowalne; czasopisma odrzucają je mówiąc „ale przecież już to wiemy”. Nieudane replikacje też są niepublikowane, nauczyliśmy się tego pierwszego dnia na studiach. Słyszałam, że uzasadnieniem takiej praktyki jest to, że „jest bardzo wiele powodów, dlaczego dobre opublikowane badanie może nie zostać zreplikowane” (Spellmann, 2012, s. 58).

W historii psychologii istniały przynajmniej 3 czasopisma, które deklarowały chęć publikowania replikacji (i badań z nieistotnymi wynikami): *Replications in Social Psychology*, *Representative Research in Social Psychology*, oraz *Journal of Articles in Support of the Null Hypothesis*. Dwa pierwsze już nie istnieją i nigdy nie zrobiły szczególnej kariery. *Replications* funkcjonował przez 3 lata (1979-1982), abstrakty jego artykułów nie są dostępne w żadnej znanej mi bazie elektronicznej. *Representative Research* działał znacznie dłużej (1970-2005), ale publikował średnio

nieco ponad 7 artykułów rocznie (nie tylko replikacje). Ostatnie wymienione czasopismo ciągle działa, ale publikuje od jednego (!) do 7 artykułów rocznie. Funkcjonuje bardziej na zasadzie internetowej ciekawostki np. nie jest wyliczany jego IF, co oznacza, że np. w polskich warunkach publikacje tam nie mają żadnego znaczenia dla rozwoju kariery.

Systematyczną próbę zbadania częstości replikacji w psychologii podjęli Makel, Plucker, 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*”. Stwierdzono, że takie słowa pojawiają się w około 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. Rzeczywiste replikacje stanowiło 68 % artykułów, ostateczny wskaźnik replikacji skorygowano więc do 1,07 % ($1,6 \% \times 68 \%$). Jakkolwiek należy dodać kilka zastrzeżeń. Około 1/3 replikacji były to replikacje w tym samym artykule, a dodatkowe 20 % były to replikacje wykonane przez autora/ów oryginalnego badania. Nie są to raczej replikacje uznawane przez ekspertów za najbardziej wartościowe (źródła). Trudno oczekiwać, żeby autorzy często wysyłali do czasopism sprzeczne dane w obrębie jednego manuskryptu, albo żeby podkopywali swoją pracę, publikując nieudane replikacje. Istotnie, jak się okazało, spośród replikacji wykonanych przez tego samego autora, odsetek udanych wynosił przeszło 90 %, a jeśli replikację wykonywał niezależny badacz, wówczas wskaźnik powtórzenia wyników spadał do 64 %. Notowano też pewne fluktuacje w zależności od dekady, przed 1950 praktycznie nie spotykano replikacji, a w latach 1990-2010 były nieco częstsze niż wcześniej (ok. 1,5 %).

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 dokonując faktycznej replikacji, nieświadomie powtarzają badania o zbliżonych hipotezach, które zostały już przeprowadzone, albo też umyślnie nie wspomina o wcześniej przeprowadzonych podobnych badaniach, żeby dodać swojemu odkryciu nimbu nowości (przyczyniając się tak czy owak do przyrostu wiedzy). W przypadku doboru słów jedyna alternatywa, jaka przychodzi na myśl to „repetition” (czasami też niektóre badania określano jako

„repeat study”, ale to w starszej literaturze). Sytuacje nieświadomej replikacji mogą występować wcale nierzadko, nikt nie zna całej literatury, a prowokacja Petersa i Ceci (1982) pokazała, że redaktorzy nie potrafili nawet rozpoznać artykułów sprzed kilku lat z ich własnych czasopism. Tak czy owak nawet, gdyby replikacji było 5 razy więcej niż wynika ze wskazań Makel i współpracowników, byłoby ich ciągle relatywnie niewiele na tle badań „nowatorskich”.

Autorzy nie poczynili, niestety, rozróżnienia na replikacje dokładne i konceptualne w swojej analizie. Choć omawiają we wprowadzeniu to rozróżnienie, to jednak w sekcji „metoda” nie ma informacji o tym, jakie były kryteria uznania artykułu za replikację. Przypuszczam, że liczba replikacji konceptualnych jest wielokrotnie wyższa niż owe 1 % (por. np. Wojciszke, 2006), nawet, jeśli badacze opisują swoich badań tymi słowami. Interesujące byłoby też zbadanie tekstów spoza setki najbardziej prestiżowych czasopism, choć moje osobiste wrażenie jest takie, że słabe czasopisma nie mają większej liczby replikacji (jedynie większą liczbę słabych badań).

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. Dodatkowo nawet, gdy sporadycznie replikacje się ukazują, są to częściej replikacje udane. Może jest tak, że w psychologii badania są tak przenikliwe zaprojektowane, że replikacje po prostu muszą się udać, znacznie bardziej prawdopodobne wydaje się jednak, że redaktorzy niechętnie publikują nieudane replikacje, a łaskawszym okiem patrzą na udane replikacje konceptualne.

Warto jednak odnotować, że choć po wybuchu afery Stapela czasopisma nie publikują jeszcze masowo replikacji, to jednak w ostatnich dwóch latach pojawiło się kilka artykułów z nieudanymi replikacjami paru często cytowanych eksperymentów psychologicznych. Szczególne zasługi dla publikacji takich artykułów ma czasopismo *PLoS One*. Omówię pokrótce trzy takie projekty, z czego dwa pochodzą właśnie z *PLoS*.

Jedno z takich badań dotyczyło szczególnie rozpowszechnionego paradygmatu badawczego tj. przymowania behawioralnego. Paradygmat ten ma liczne odmiany, tym niemniej na ogół rzecz sprowadza się do tego, że subtelnie eksponowane prymy

zmieniają w jakiś sposób zachowanie badanego (na ogół proste reakcje np. odpowiedzi na pytania). Prymowanie może polegać na ekspozycji słów (np. w zadaniu układania zdań z rozrzuconych słów) albo obrazków, ale często zakłada się też, że myślenie o określonych rzeczach jest samo w sobie prymą. Klasyczne badania w tym obszarze przeprowadzili Bargh, Chen, i Burrows (1996). Osoby, którym prezentowano słowa związane ze starością (typu „zmarszczki”, „emerytura”, a w kontekście amerykańskim były to dodatkowo „bingo” albo „Floryda”), następnie wolniej przechodziły do następnego pokoju w porównaniu do grupy kontrolnej (eksperyment 2). Badanie, choć cytowane do tej pory ponad 3300 razy (!), było do niedawna tylko dwa razy replikowane, na stosunkowo nielicznych próbach i replikacje te tylko częściowo potwierdzały efekt (za: Bakker i in., 2013). Dwie próby replikacji podjęli ostatnio niezależnie Doyen, Klein, Pichon, Cleeremans (2012), oraz Pashler, Harris, Coburn (2011). Żaden z zespołów nie zdołał odtworzyć oryginalnego efektu pomimo stosunkowo wysokiej mocy (odpowiednio 120 i 66 osób badanych; w oryginalnym badaniu 30 osób). Dodatkowo Doyen i in. (2012) pokazali, że efekt zależny w dużej mierze od oczekiwań eksperymentatora. Jeśli czas przejścia mierzony był stoperem, a eksperymentator oczekiwał dłuższego czasu, rzeczywiście taki notował. Obiektywny pomiar czasu przy pomocy automatycznego stopera, uruchamianego przez przekroczenie niewidocznego lasera, niwelował różnice.

W innym znanym badaniu w paradygmacie prymowania, badani, którym eksponowano koncept chuligana stadionowego wykazywali następnie nieco niższe wyniki w teście wiedzy ogólnej. Osoby, prymowane konceptem profesora, wykazywały wyższą wiedzę (Dijksterhuis i van Knippenberg, 1998). (Prymowanie polegało na wyobrażaniu sobie takich osób i opisywaniu przez kilka minut, czym się charakteryzują). Shanks i in. (2013), przeprowadzili 9 eksperymentów, w części zastosowano identyczną procedurę, a w części zbliżoną. Ogólnie meta-analiza wszystkich danych nie potwierdziła występowania efektu. Oryginalne badanie było cytowane ponad 600 razy.

Kolejne badanie (ściślej: program badań) również był kierowany przez Apa Dijksterhuisa. Zaproponował on teorię nieświadomego myślenia, która zakłada, wbrew elementarnej intuicji, że w pewnych sytuacjach myślenie jest najskuteczniejsze wtedy, kiedy nie myślimy świadomie, ale pozwalamy nieświadomym procesom przetwarzania informacji wykonać całą pracę (Dijksterhuis i

Nordgren, 2006). Taki sposób myślenia miał być szczególnie skuteczny w przypadku problemów złożonych. Przykładowo w szandarowym badaniu opublikowanym w *Science* (Dijksterhuis, Bos, Nordgren, Van Baaren, 2006; cytowane ponad 800 razy) badani wybierają jeden z czterech samochodów, a każda opcja opisana jest 12 cechami. Badany otrzymuje informacje pojedynczo, każdą przez 8 sekund. Po okresie przyswajania danych, przez 4 minuty świadomie myśli nad problemem lub rozwiązuje anagramy (a więc w zamyśle myśli nieświadomie). W pierwszym wypadku około 20 % badanych wybrało najlepszą opcję (co jest na poziomie przypadku), w drugim około 60 %. Byłby to więc bardzo silny efekt, nawet jeśli trudno mówić o trafności ekologicznej tej procedury (np. badani nie mieli w czasie myślenia dostępu do informacji, a trudno przypuszczać, że 48 jednostek informacji zostało efektywnie zapamiętanych przez kilka minut, abstrahując już zupełnie od faktu, że nikt w ten sposób nie podejmuje decyzji). Inne badania w tym stylu zostały opublikowane w szeregu prestiżowych czasopismach np. w JPSP (Dijksterhuis, 2004) albo PS (Dijksterhuis, Bos, Van der Leij, Van Baaren, 2009). Ogólne przesłanie tych badań było następujące: „masz trudny problem, nie myśl o nim, prześpij się z tym, rozwiązanie przyjdzie samo”. Było ono na tyle medialnie nośne, że eksperymenty Dijksterhuisa były często opisywane w prasie (Ulanowski, 2006).

Zanim przejdę dalej, pozwolę sobie na małą osobistą dygresję. Badania te są mi w pewien sposób bliskie, gdyż efekt próbowałem zreplikować w swojej pracy dyplomowej na studiach magisterskich (Budzicz, 2008). Nie udało mi się powtórzyć jego wyników (konkretnie: Dijksterhuis i in., 2006, eksperyment 1), pomimo dołożenia starań, żeby warunki były jak najbardziej zbliżone do oryginalnego eksperymentu. Nie było różnicy między grupą myślącą i poddawaną dystrakcji, a dodatkowo grupa, której w czasie myślenia eksponowano informacje, wypadła wyraźnie lepiej. Wyniki zostały wysłane do jednego z polskich czasopism jako krótki raport. Artykuł został odrzucony. Z pewnością nie był doskonały i recenzenci wytknęli kilka wad. Jednak szczególnie zaskoczyło mnie, że jedna z uwag recenzenta mówiła, że „wyniki zbyt odbiegają od wyników oryginalnych, co musi wskazywać na błędy w metodologii badania”.

Wracając jednak do „głównego” obiegu naukowego. Huizenga, Wetzels, van Ravenzwaaij, Wagenmakers (2012) nie zdołali powtórzyć wyników badania Dijksterhuisa. Przeprowadzili 4 eksperymenty, łącznie obejmujące 480 badanych.

Nieświadome myślenie nie powodowało polepszenia wyników, a długość nieświadomego myślenia nie wpływała na jakość decyzji. Nieświadome myślenie nie produkowało też najlepszej subiektywnie decyzji (badani post factum świadomie wazyli istotność poszczególnych informacji). Nieudane replikacje przeprowadzili też Acker (2008), González-Vallejo, Lassitera, Bellezza i Lindberga (2008), Calvillo i Penalosa (2009); konceptualne replikacje kwestionujące zasadnicze uzasadnienie teoretyczne przeprowadzili m.in. Lassiter, Lindberg, González-Vallejo, Bellezza, Phillips (2009) i Newell, Wong, Cheung, Rakow (2009) oraz Waroquier, Marchiori, Klein, Cleeremans (2010). Interesującym jest, że efekt ten spotkał się z licznymi próbami replikacji (często nieudanymi), podobnie jak kontrowersyjne efekty Bema (2011). Niestety, typowy, mało kontrowersyjny efekt, prawdopodobnie nie może liczyć na taką uwagę społeczności badaczy.

Sam Dijksterhuis opublikował meta-analizę badań nad nieświadomym myśleniem. Wg niej efekt ten występuje i jest umiarkowanie silny ($g = 0.224$) (Strick i in., 2011). Dijksterhuis w swojej meta-analizie nie zrobił rozróżnienia, jaka jest siła efektów uzyskiwana przez jego zespół i przez niezależnych badaczy. Jest to o tyle istotne, że wcześniejsza o kilka lat (siłą rzeczy dużo mniejsza) meta-analiza przeprowadzona przez Ackera (2008) pokazała, że efekt „wychodzi” przede wszystkim Dijksterhuisowi i jego zespołowi (por. ryc. 4: Acker, 2008). Dość symptomatyczne jest w tym kontekście to, że Dijksterhuis w podsumowaniu swojej meta-analizy zachęcał badaczy do prób replikacji odkrytego przez niego efektu, ale sugerował szukanie własnych procedur jego skutecznej replikacji. „[...] Uważamy, że lepiej jest dostosować nasz paradygmat do warunków badanych niż próbować dokładnie zreplikować nasze eksperymenty na osobach w innych krajach i na innych kontynentach”. (s. 759; Strick i in., 2011). Oczywiście, trudno się nie zgodzić z sugestią, że czasami określone efekty psychologiczne są uwarunkowane kulturowo (por. Henrich, Heine, Norenzayan, 2010; a także rdz. 2.7), a procedury nie należy bezmyślnie kopiować bez uwzględnienia takiego kontekstu. Jednak zalecenie Dijksterhuisa może być również bardzo wygodną wymówką do usprawiedliwiania braku sukcesu replikacji przez niezależnych badaczy. Nigdy nie ma pewności, czy brak udanej replikacji wynika z niewystępowania oryginalnego efektu, czy może z tego, że badacz nie dołożył dostatecznych starań do adaptacji procedury. Rozsądniejsze wydaje się, żeby najpierw spróbować szeroko udokumentować

występowanie efektu, a dopiero potem starać się zrozumieć kulturowe uwarunkowania, szczególnie w sytuacji, gdy proponowany efekt jest tak odległy od dotychczasowej teorii, zdroworozsądkowej intuicji czy ma potencjalnie duże, potencjalnie groźne, implikacje praktyczne. Dodam, że Wagenmakers i in. (2012) badali identycznie jak Dijksterhuis holenderskich studentów, a mimo tego nie powtórzyli jego wyników. Kontaktowałem się w swego czasu z Dijksterhuisem pytając m.in. czy istnieją jakieś badania, w których nie wykazano występowania efektu. Dijksterhuis odpisał: „Nikt nie publikuje wyników z nieistotnymi wynikami i my także. Mamy dane, gdzie myślący nieświadomie nie mieli przewagi nad myślącymi świadomie, ale nigdy nie udało nam się pokazać czegoś odwrotnego” (Dijksterhuis, osobista komunikacja, 2008). Ufam, że druga część jest prawdziwa. Pierwsza część pokazuje opisywaną w tym rozdziale, wypaczającą obraz rzeczywistości „kulturę sukcesu”.

Warto także zaznaczyć, że porażka replikacji nie oznacza automatycznie, że wyniki oryginalnego badania są błędne. Autorzy replikacji też mogą popełniać błędy, albo replikacja może się nie udać przez czysty przypadek. Wydaję się jednak, że jeśli replikacja ma większą moc (szczególnie poprzez większą grupę lub bardziej rzetelne narzędzia), to jej wyniki powinny być preferowane.

Sporadycznie zarówno oryginalne badania jak i nieudana replikacja są równocześnie prawdziwe, ponieważ badacze nie uwzględniają jakiegoś elementu procedury (próby, operacjonalizacji itd.), który jest ważny, ale nie został zareportowany w oryginalnym badaniu, a więc i replikujący go nie uwzględnili (por. Open Science Collaboration, 2012). Słabe efekty psychologiczne mogą szczególnie być podatne na różne subtelne czynniki. Jednak w przypadku silnych efektów nieznaczne odstępstwa od procedury nie powinny zasadniczo niwelować efektu. Oryginalny efekt spowolnienia ruchów pod wpływem prymowania „konceptem” starości (eksperyment 2: Bargh i in., 1996) był wyraźny tj. wynosił w liczbach bezwzględnych ponad sekundę (na 9 metrowy spacer), a w kategoriach siły efektu około $d = 0,9$. Podobnie efekt wyboru najlepszej opcji w sytuacji nieświadomego myślenia z eksperymentu opublikowanego w Science był silny ($d = 0,7$; eksperyment 1; Dijksterhuis i in., 2006). Wielkości oryginalnego efektu prymowania konceptem profesora (w porównaniu do sekretarki)

na wykazywaną wiedzę wynosiła też około $d = 0,7$ ³⁵ (eksperyment 1: Dijksterhuis, van Knippenberg, 1998). A przypomnę, że wszystkich tych wyników nie udało się zreplikować.

Oczywiście, wspomniane nieudane replikacje są tylko kroplą w morzu wszystkich oryginalnych badań i nie jest pewne czy zapoczątkują jakościowo nowy trend. Podkreślę jeszcze raz, że dotyczą nie jakiś obskurnych badań opublikowanych w nieznanych czasopismach, ale rzekomo mocnych efektów, często cytowanych, o statusie nieomal faktów. Baza psychfiledrawer zawiera³⁶ 66 raportów z replikacji dokadnych badań z dobrych czasopism, z czego 45 (68 %) jest nieudanych. Bakker i in. (2013) posunęli się na tyle daleko, że zaproponowali, żeby APA utworzyła specjalną grupę roboczą, która stworzyłaby listę badań obecnych w podręcznikach (o statusie „faktów”), jednak w rzeczywistości opartych na danych pochodzących z jednego artykułu lub serii badań o słabej mocy, a więc wymagających potwierdzenia przez niezależnych badaczy.

Wydaje się, że istnieje olbrzymi rozdźwięk między powszechnym przekonaniem o pozytywnych skutkach replikacji, a jej niską częstością stosowania. W rdz. 1.10 pokażę, gdzie zostały zainicjowane zmiany w tym kierunku.

1.9. Podsumowanie najważniejszych wniosków o rzetelności danych w psychologii

Krótko podsumowując to, co zostało do tej pory powiedziane:

- Afera Stapela pokazała, że mechanizm wykrywania nierzetelności w badaniach psychologicznych przez *peer-review* oraz – przede wszystkim – przez systematyczne replikacje jest poważnie ograniczony.
- Prawdopodobnie bardzo niewielka liczba oszustów jest wykrywana w psychologii.
- Zdecydowana większość psychologów przyznaje się do stosowania jakichś metod wybiórczego lub arbitralnego analizowania i prezentowania danych. Zachowanie takie jest zasadniczo uznawane za dopuszczalne.

³⁵ Siłę efektów wyliczyłem szacunkowo z dostępnych w artykule danych, gdyż wszystkie te artykuły nie zawierają miar wielkości efektów.

³⁶ Stan na czerwiec 2015

- Nieudane badania są praktycznie nieobecne w najlepszych czasopismach, choć psychologowie przyznają, że niewychodzące eksperymenty są częstsze niż udane. Olbrzymi odsetek udanych badań w literaturze sugeruje, że ma miejsce mechanizm wybiórczego publikowania i/lub podkręcania wyników przy pomocy wątpliwych praktyk. Dowodami szczególnie o tym świadczącymi są większa liczba udanych badań w psychologii niż w naukach przyrodniczych, negatywna korelacja między próbą, a wielkością efektu w badaniach, oraz bardzo niska moc badań w psychologii oraz rozkłady wartości p w literaturze odbiegające od oczekiwanych.
- Dokładne replikacje są czymś niezwykle rzadkim w psychologii. W ostatnich latach kilka silnych efektów, często cytowanych, mających nieomalże status faktów, nie udało się zreplikować.
- Trudność wzajemnej kontroli społeczności badaczy dodatkowo zwiększa to, że większość badaczy niechętnie wysyła dane surowe do niezależnych analiz, pomimo tego, że jest to ich obowiązkiem. Analiza samych danych przetworzonych w artykułach często wskazuje na błędy np. w wyliczaniu wartości p .

Cytowane na samym początku rozdziału teksty o kryzysie nie są więc raczej przesadą. Mimo wszystko pojawiają się też głosy, że z psychologią nie jest jednak tak źle (np. *European Association of Social Psychology*, 2012). Pashler i Harris (2012) rozprawili się z trzema często wykorzystywanymi mitami.

Pierwszy mit zakłada, że choć w psychologii dokładne replikacje są czymś nieczęstym, to jednak replikacje konceptualne są wcale nierzadkie, a przy tym pozwalają nam zwiększyć prawo do generalizacji efektów. Jednak Pashler i Harris zwracają uwagę, że opublikowanie nieudanej replikacji konceptualnej może być jeszcze trudniejsze niż opublikowanie nieudanej replikacji dokładnej, gdyż brak sukcesu nie wskazuje tutaj bezpośrednio na nieprawdziwość oryginalnego efektu, ale może być wynikiem zbyt dalekiego odejścia od oryginalnej procedury (np. nietrafnej manipulacji). Udaane replikacje konceptualne mogą podlegać szczególnie silnej presji na wybiórcze publikowanie (ang. *publication bias*), gdyż nierzadko są recenzowane przez badaczy z oryginalnego artykułu. Przy wielkiej liczbie prób replikacji

konceptualnych w badaniach o niskiej mocy, wiele z nich wyjdzie przez czysty przypadek, co może potem sprawiać wrażenie „zalewu” udanych replikacji.

Drugi popularny mit mówi, że choć nauka błądzi, to na długą metę ma wielką zdolność do samokorekty i w ostatecznym rozrachunku fałszywe prawidłowości zostaną wykryte. Krytykując ten mit autorzy przedstawili wyniki własnej analizy polegającej na wyszukaniu psychologicznych artykułów naukowych z frazą „failure to replicate”. Mediana czasu pomiędzy oryginalnym badaniem, a nieudaną replikacją wynosiła 4 lata, a tylko w 10 % artykułów ten czas wynosił więcej niż 10 lat. Spośród (nielicznych) opublikowanych replikacji, większość dotyczy badań sprzed niewielu lat. Im więcej lat upływa, tym coraz mniejsze prawdopodobieństwo, że ktoś zreplikuje dane badanie i ewentualnie je skoryguje. Badacze rzadziej sięgają do starszych roczników, a i sama liczba nowych badań rośnie lawinowo.

Inna wersja tego mitu głosi, że choć większość badań nie zostanie wprost skorygowana, to nastąpi innego rodzaju korekta, polegająca na tym, że dyscyplina zacznie badać te same fenomeny w inny, bardziej trafny sposób, a stare badania zostaną po prostu zapomniane. Ze względu na panujące w psychologii „mody” (por. też Rozin 2007), autorzy uważają to za mało prawdopodobne. Podają przykłady dziedzin, które tętniły kiedyś życiem i wygenerowały sporo, też zreplikowanych, badań (na przykład wpływ wyobrażeń na pamięć długoterminową), po czym nagle, bez wyraźnego powodu zainteresowanie naukowców tymi problemami ulotniło się. W dalszym ciągu są one jednak szeroko omawiane w podręcznikach akademickich. Korekta występujących tam błędów jest mało prawdopodobna, ze względu na brak owego zainteresowania. Aby wzmocnić ten argument przywołują też niedawny artykuł z *Nature* (Begley i Ellis, 2012). Badacze raka opisywali swój wieloletni program rozwoju leków w oparciu o kilkadziesiąt klasycznych badań w dziedzinie. Ostatecznie tylko niewielką ich ilość byli w stanie zreplikować. Trudno przypuszczać, żeby w psychologii, mierzącej się z jeszcze większą niespójnością teoretyczną i arbitralnością metod te problemy były mniejsze. Nieudane replikacje eksperymentów o statusie faktów (rdz. 1.7) zdają się to potwierdzać.

Pashler i Harris obalili też trzeci, który mówi, że ze względu na poziom α 5 %, liczba błędów fałszywych pozytywnych będzie w nauce relatywnie niewielka. Jak wynika z wcześniejszych analiz przy szerokich „elastycznych” możliwościach analizy danych (Simmons i in., 2011), oraz deklarowanym nierzadkim stosowaniem tychże

możliwości (John i in., 2012) liczba błędów fałszywych pozytywnych może być większa, być może znacznie większa niż owe 5 %.

1.10. Skutki kryzysu „poststapelowskiego”

Po fali artykułów z krytyką typowych praktyk badawczych i publikacyjnych, przyszedł w roku 2013 i szczególnie 2014 liczne deklaracje czasopism o gotowości do zmiany. Omówię kilka sygnałów gotowości do zmiany w uprawianiu nauki, oraz kilka projektów, które mają „nadrobić” stracony czas.

W ostatecznym rozrachunku mechanizmem kontroli nierzetelności w nauce (i nieświadomego błędzenia) jest mechanizm replikacji dokonywanych przez niezależnych badaczy. Nie zaskakuje więc, że duża część głosów dotyczy zwiększenia roli replikacji w psychologii (Asendorpf i in., 2013; Nosek, Spies i Motyl, 2012; Francis, 2012), oczywiście replikacji dokładnych, gdyż replikacji konceptualnych, wydaje się, że jest niemało.

Jest to o tyle ważne, że nie sądzę, żeby istniały absolutnie skuteczne sposoby wykrywania fałszerstwa i przekłamań na etapie recenzji artykułów. U Stapela znaleziono post factum szereg nieprawidłowości, ale można sobie wyobrazić badacza, który ma większą matematyczną świadomość i do tworzenia danych używa np. programów generujących liczby pseudolosowe o zadanych wartościach.

Jak powszechnie wiadomo „system” nie zachęca badaczy do systematycznych replikacji (mniejsza szansa na publikację, wymóg „nowatorstwa” ze strony instytucji przyznających granty itp.), z drugiej wydaje się, że same czasopisma nie patrzą na nie przychylnym okiem, zakładając, że artykuły replikacyjne będą miały średnio niższą cytowalność (zapewne słusznie).

Pewne wyłomy zostały jednak już zrobione. Czasopismo *Perspectives on Psychological Science* (5. 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 utworzony ma zostać 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 przy

pomocy metod meta-analitycznych, a sam artykuł ma zostać opublikowany niezależnie od wyników. Autorzy oryginalnego badania również będą zaproszeni do współuczestnictwa. Jak podkreśla redakcja nacisk zostanie położony na siłę efektu, a nie na dychotomiczne decyzje „replikacja udana-nieudana”, gdzie kryterium sukcesu jest tylko istotność statystyczna wyników. Pierwszy z opublikowanych artykułów ukazał się w drugiej połowie 2014 roku i dotyczył efektu verbal overshadowing, polegającego na tym, że werbalny opis określonego obrazu (tutaj: twarzy) utrudnia później jego czysto wizualną identyfikację (ang. *verbal overshadowing*). Efekt został potwierdzony, choć wyniki ostatecznie okazały się być znacznie słabsze niż w oryginalnym tekście (Alogna i in., 2014). W grudniu 2014 kolejna trwają 2 kolejne szerokie replikacje, a w kolejnej trwa nabór badaczy.

Równie wielką rewolucją wydaje się ogłoszona w połowie 2014 roku decyzja JPSP o publikacji replikacji. Specjalne artykuły replikacyjne mają być publikowane tylko w wersji elektronicznej, co jednak zapewne nie będzie 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 (przy czym 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 *JEP: General*. Inne czasopismo psychologii społecznej *Journal of Experimental Social Psychology* w nowej polityce publikacyjnej również zapowiedziało gotowość publikacji replikacji, choć nie określiło tak szczegółowo wymogów, jakie stawia takim artykułom.

Psycholog społeczny Brian Nosek postanowił zebrać zespół, który systematycznie powtórzyłby badania w trzech czołowych czasopismach psychologicznych, bez względu na ich treść: *PS*, *JPSP*, oraz *JEP: Learning, Memory, and Cognition*. Wg pierwszych doniesień projekt miał obejmować *wszystkie* (!) artykuły empiryczne opublikowane w 2008 roku w tych trzech czasopismach (Bartlett, 2012). Informacje te budziły we mnie mieszanię szacunku i powątpiewania, bowiem rocznik 2008 tych trzech czasopism obejmował odpowiednio 206, 168 i 119 artykułów (a więc zakładam, że przynajmniej ponad 1000 badań). W rzeczy samej późniejsze doniesienia mówiły już tylko o 30 badaniach z każdego czasopisma (Open Science Collaboration, 2012). Być może od początku planowano zreplikować relatywnie

niewielką liczbę badań, a doniesienia o próbie replikacji wszystkich były prasowym zniekształceniem. Tym niemniej, rzeczywiste czy nie, ale to obniżenie poprzeczki jest dosyć wymowne, bowiem pokazuje, że replikacja badań tylko z trzech czasopism (spośród setek) z tylko jednego rocznika nastrocza zbyt dużych trudności logistycznych, żeby ją urzeczywistnić. Wzmacnia to tezę, że olbrzymia większość badań nigdy nie zostanie zreplikowana, a więc jakaś liczba fałszerstw, przekłamań lub artefaktów w nich zawartych nie zostanie skorygowana (por. też Pashler i Harris, 2012). Tak czy owak, gdyby nawet zreplikowano kilkadziesiąt losowych badań, byłby to pierwszy (wg mojej wiedzy) projekt nakierowany na ogólną replikowalność badań psychologicznych. Będzie to prawdopodobnie najlepszy możliwy sprawdzian rzetelności badań psychologicznych, dający bardziej trafne wyobrażenie o stanie nauki niż badania nad QRP. Jakiegolwiek nie będą wyniki, przyniesie to pożytek naszej dyscyplinie. Gdyby liczba udanych replikacji była znacząca, pokaże to, że psychologia, pomimo złożoności przedmiotu, „elastyczności” metod i tendencji do wybiórczego publikowania, potrafi generować systematycznie rzetelną wiedzę. Gdyby liczba udanych replikacji była niska (np. poniżej 50 %) być może spowoduje to tyle duży szok, że zapoczątkuje znaczące zmiany w praktykach badawczych i publikacyjnych (sam Nosek w dwóch artykułach, bardziej „publicystycznych”, zaproponował już na czym takie zmiany miałyby polegać: Nosek i Bar-Anan, 2012; Nosek, Spies i Motyl, 2012).

Pojawiają się sugestie, że czasopisma powinny otworzyć swoje łamy nie tylko na replikacje, ale też 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 ilość 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. Wydaje się więc, że rozsądny kompromis zaproponowały dwa czasopisma: *Attention, Perception & Psychophysics* (Registered Reports and Replications in Attention, Perception & Psychophysics, 2013) i *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 dodatkowo oceniają, czy istnieją mocne przesłanki, żeby wątpić w rzetelność oryginalnego badania). Decyzja będzie też uwarunkowana spodziewaną mocą badania, która powinna wynosić minimum 90 %. Jeśli decyzja będzie pozytywna, autorzy mają rok na przeprowadzenie badań i przesłanie kompletnego manuskryptu (z wynikami i ich omówieniem), a wspomniane czasopisma mają zamiar je opublikować, niezależnie od tego, jakie będą ostatecznie rezultaty.

Jak wspomniałem, jest to rozsądny kompromis między publikowaniem wszystkiego, a publikowaniem tylko istotnych wyników (występującym obecnie), na którym powinny skorzystać i czasopisma i badacze. Czasopisma mogą sobie na wstępie wybrać, co je interesuje i nie muszą akceptować wszystkiego. Badacz nie będzie miał gwarancji akceptacji projektu, ale jeśli zostanie zaakceptowany będzie miał gwarancję, że dane zostaną opublikowane. Czasopisma dzięki temu zyskają w miarę pewne dane, bo badacz nie będzie miał żadnej motywacji, żeby dane fałszować. Sam badacz nie będzie musiał obawiać się marnowania zasobów na potencjalnie niepublikowalne badania, oraz będzie mógł się skupić na solidnym zrobieniu badania o odpowiedniej mocy. Nie będzie miał żadnego interesu, żeby „kombinować”, czy to poprzez wątpliwe praktyki badawcze, czy to poprzez robienie serii badań i wybierania tylko tych, które wyszły. Niestety, te dwa czasopisma obejmują tylko dwie dosyć wąskie subdyscypliny psychologiczne, ale mam nadzieję, że inne czasopisma pójdą tym śladem.

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. Wg mojej wiedzy jednak żadne czasopismo nie zdecydowało się póki co na takie rozwiązanie. Oprócz zredukowania roli istotności wyników przy podejmowaniu decyzji o publikacji, procedura taka może moim zdaniem sprzyjać innemu pozytywnemu zjawisku. Dużo mówi się o tym, że brak istotności przekreśla szanse na publikację. Ale nie wspomina się o tym, że może występować też odwrotny efekt: istotność wyników może sprzyjać badaniom, które z różnych innych względów są niskiej jakości. Być może jest to jedna z

przyczyn, dlaczego publikuje się wiele badań ze słabymi efektami lub arbitralnymi decyzjami. Spekulowałbym, że recenzenci nierzadko myślą „jest to naciągane, ale nie ma tu oczywistych błędów; coś mimo wszystko wyszło i niech już społeczność pozna te wyniki”.

Również sam 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 i Task Force on Statistical Inference, 1999) wyraźnie postuluje się, żeby większy nacisk kłaść na referowanie przedziałów ufności i wielkości efektów (uzasadnienie matematyczno-metodologiczne tego problemu wykracza poza łamy tej pracy, odsyłam szczególnie do artykułów Cohena, 1994/2006 oraz Cumminga, 2014).

Wysuwane są też postulaty, które mają potencjalnie zminimalizować arbitralność analiz prowadzącą do „psychologii fałszywie pozytywnej”. Szczególnie sugeruje się pójście śladem medycyny i szerokie prejestrowanie 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 prejestrowanie badań. Zwraca się uwagę na to, że znaczna część badań jest w pewien sposób prejestrowana, 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.

Aby zredukować tworzenie psychologii „fałszywie pozytywnej” Simmons, Nelson i Simonsohn (2012) postulują, żeby czasopisma wprowadziły wymóg określania przez autorów czy zraportowali wszystkie zmienne, warunki, sposoby ustalania wielkości próby oraz usuwania przypadków odstających. Ich zdaniem w każdej sekcji poświęconej metodzie, autorzy mogliby dodawać standardową informację „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, to musieliby skłamać, żeby temu zaprzeczyć. Zdaniem DeScioli, Christnera i Kurzbana (2011) jest wielka psychologiczna różnica między nie mówieniem prawdy, a kłamaniem. Od 2014 roku autorzy chcący publikować artykuły empiryczne w PS muszą dostarczyć powyższe informacje do redakcji. Żeby umożliwić autorom przedstawienie maksymalnie pełnego obrazu badania, sekcje metoda i rezultaty nie są zaliczane do limitu znaków w tekście (Association for Psychological Science, 2014).

Zachęca się też badaczy, żeby post factum ujawniali wszystkie istotne informacje o przeszłych opublikowanych badaniach. Idąc w sukurs tym zaleceniom (i nie czekając na reakcję redakcji), grupa psychologów utworzyła stronę PsychDisclosure.org (ang. disclosure = ujawnianie), która daje autorom tekstów empirycznych możliwość ujawnienia takich dodatkowych informacji jak to, czy opisano w artykule wszystkie zmienne zależne, grupy eksperymentalne oraz sposoby wyłączenia obserwacji i ustalania wielkości próby (LeBel i in., 2013). Twórcy serwisu skontaktowali się z połową (wybraną losowo) autorów artykułów w 2012 roku z 4 czołowych czasopism³⁷, których poprosili o ujawnienie stosownych informacji na stronie. Motywacją miało być zwiększenie wartości informacyjnej artykułu. Wysłano zapytanie tylko do połowy autorów ze względu na wymogi komisji etycznej (osoby postronne nie mają możliwości ustalenia, czy nieobecni na stronie autorzy artykułów nie ujawnili informacji, czy może nie zostali o to poproszeni).

Informacje ujawniło nieco ponad 46 % ankietowanych. W kilku przypadkach ujawnione informacje wskazują na wątpliwe praktyki badawcze. Przykładowo, połowa przyczyn (8/19) nie ujawnienia informacji o wszystkich grupach eksperymentalnych to brak istotnych efektów, choć w jeszcze większej liczbie przypadków powody nie zostały jasno określone (9/19). Znacznie częściej nie informowano o wszystkich wykorzystanych zmiennych (łącznie 73 przypadki), z czego kilkanaście procent wynikało z braku istotnych efektów (11 przypadków). Większość takich sytuacji to brak związku zmiennej z pytaniem badawczym (47/73), choć oczywiście pojawia się pytanie, po co w ogóle zmienną mierzono od samego początku. Autorzy artykułów

³⁷ *Psychological Science; Journal of Personality and Social Psychology; Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP:LMC); Journal of Experimental Psychology: General (JEP:G)*

sporadycznie też wskazywali życzenie redakcji jako powód usuwania zmiennych i warunków z raportu.

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 obowiązek³⁸ udostępniać takie dane, okazuje się, że gdy przychodzi, co do czego, większość tego nie robi (rdz. 1.7). W pewien sposób trudno się dziwić. Analiza danych przez niezależnego naukowca może prowadzić do zakwestionowania rzetelności lub kompetencji badacza. Można też twórczo odczytać z nich coś nowego, w takim wypadku jednak splendory spadną na kogo innego. Więc w najlepszym wypadku niewiele się (osobiście) zyskuje, a w najgorszym wypadku traci. Badacze niestety rozumieją tą 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” (str. 1876). Postuluje się, żeby dostarczenie danych surowych było równoległe z wysłaniem manuskryptu, a dane te powinny być później ogólnodostępne (poza szczególnymi sytuacjami). Praktycznie wszystkie najważniejsze czasopisma mają edycje elektroniczne i wydaje się, że technicznie istnieją wszelkie możliwości, żeby czasopisma udostępniały dane w Internecie. Istnieje wiele repozytoriów, w których można już teraz zdeponować i/lub udostępniać swoje dane (przy czym zdeponowanie nie musi być automatycznie równoznaczne z publicznym udostępnieniem np. datadryad.org; easy.dans.knaw.nl; zenodo.org)

Omówię na koniec krótko zmiany zadeklarowane przez niektóre czasopisma. Czasopisma *The Journal of Social Psychology* (Grahe, 2014) i PS (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 (ang. *badges*) informujące o tym, czy badacz wykonał te zalecenia. *Archives of General Psychology*, nowe czasopismo APA w formule Open-Access (Cooper, VandenBos, 2013) wymaga, aby badacz udostępniający artykuł zdeponował dane w otwartym repozytorium (podobne wymogi ma *PLoS One*,

³⁸ Ten obowiązek określają wymogi redakcyjne czasopism oraz punkt 8.14 kodeksu etycznego APA.

czołowe czasopismo *open access*). Szereg 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. Szereg zaleceń dotyczących publikacji wydało *Journal of Experimental Social Psychology* (JESP Editorial Guidelines, 2014). Autorzy są m.in. zachęceni do szczerego opisywania procesu generowania hipotez, unikania „łowienia p” i 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 wszystkim służyć ich objaśnieniu. Możliwe jest w nim publikowanie danych z artykułów wcześniej opublikowanych w innych czasopismach.

Informacje zawarte w tym rozdziale wskazują na to, że istnieje pewna gotowość zmiany. Otwartym pozostaje pytanie, czy ta gotowość zamieni się w systematyczną zmianę praktyk badawczych.

1.11. Nierzetelność niezależna od badacza

Czasami jednak, pomimo najlepszej woli, badacze nie będą podawali prawdziwych danych o świecie. Dlatego na koniec tego rozdziału chciałbym omówić często cytowany³⁹ artykuł Ioannidis (2005), zatytułowany bardzo prowokacyjnie „Dlaczego większość opublikowanych wyników badań jest fałszywa”. Sam Ioannidis reprezentuje medycynę, jak i w swoim artykule podaje przykłady z medycyny, nie widzę jednak żadnych przeszkód, żeby jego analizy nie odnieść do psychologii. Tytułowa teza opiera się na założeniu, że prawdopodobieństwo prawdziwości rezultatu badawczego zależy od trzech głównych czynników: mocy statystycznej badania ($1 - \beta$), założonego poziomu istotności statystycznej (α), oraz – o czym zwykle nie wspomina się w tekstach metodologicznych – proporcji prawdziwych testowanych hipotez do fałszywych testowanych hipotez w danej dyscyplinie nauki (Ioannidis oznacza jako R – zapewne od „ratio”). Omówię krótko ten ostatni czynnik,

³⁹ Cytowany ponad 2300 razy

gdyż pojęcie istotności statystycznej jest dobrze znane, a o mocy pisałem wcześniej (rdz. 1.4).

Ustanawiając poziom alpha (czyli poziom istotności statystycznej) o wartości 0,05 badacze zwykle intuicyjnie przyjmują, że pewna ilość badań (owe 5 %) da wynik fałszywy pozytywny (tj. występowanie efektu, którego w rzeczywistości nie ma) przez zupełny przypadek (oczywiście, zakładając brak innych artefaktów oraz świadomych zniekształceń danych). To założenie nie uwzględnia jednak proporcji prawdziwych do fałszywych testowanych hipotez. Choć nie są znane mi dane, ile testowanych w psychologii hipotez jest zgodna z rzeczywistością (czyli „prawdziwa”), zakładam intuicyjnie, że jest znacznie więcej fałszywych niż prawdziwych. Jeśli na 1000 hipotez odsetek prawdziwych wynosi 10 %, wówczas wśród 900 testowanych fałszywych hipotez w około 45 przypadkach popełnimy błąd fałszywy pozytywny ($0,05 * 900$). Zakładając, że nawet wykryjemy wszystkie 100 rzeczywiście prawdziwych efektów, w takim wypadku liczba błędów fałszywych pozytywnych jest sporo większa niż 5 % (tj. $45 / (100+45) = 31 \%$). Jeśli wśród wszystkich hipotez rzeczywiście prawdziwych jest 5 % albo 1 % wtedy liczba błędów fałszywych pozytywnych wyniesie odpowiednio 48 % i 83 %⁴⁰. Oczywiście badania psychologiczne nie mają idealnej mocy statystycznej (tj. nie zawsze wykrywają zależność, która rzeczywiście istnieje), dlatego ten odsetek będzie zwykle większy (przy założeniu mocy statystycznej na poziomie 0,7, dla wartości rzeczywiście prawdziwych hipotez 10 %, 5 % i 1 %, odsetek błędów fałszywych pozytywnych to odpowiednio 39 %, 57 % i 87 %). Wartości te można obliczyć z wyprowadzonego przez Ioniddesa wzoru na rzeczywistą trafność badania (*positive predictive value*, PPV), który ma postać:

$$PPV = (1 - \beta) * R / (R - \beta * R + \alpha)$$

Wartość $1 - PPV$ jest to odsetek błędów fałszywych pozytywnych wśród wszystkich wyników pozytywnych w danej nauce przy założeniu określonych wartości α , β i R .

⁴⁰ Na podobnej zasadzie, jeśli test na określoną chorobę rzadko daje wynik fałszywy pozytywny, ale jeśli choroba jest bardzo rzadka, to większość wyników pozytywnych będzie fałszywie pozytywna. Ludzie jednak ignorują rozpowszechnienie choroby i zniekształcenie to nazywane jest przez psychologów poznawczych „ignorowaniem ogólnych prawidłowości” (por. Tyszka, 2000, rdz. 9.3)

Dalej Ioannidis wprowadza pojęcie „zniekształcenia” (oznaczane „ u ”; ang. *bias*), a więc kombinacji różnych świadomych i nieświadomych błędów w procedurze, zbieraniu i analizie danych, które powodują, że wynik badania jest pozytywny⁴¹, gdy w rzeczywistości nie powinien być, przy czym, oczywiście nie dotyczy to opisanej wyżej sytuacji, gdy wynik jest pozytywny przez czysty przypadek (w psychologii owe „zniekształcenia” nazywamy zwykle artefaktami, choć wspomniane wcześniej wątpliwe praktyki badawcze opisywane przez Simmonsa i in. (2011) z pewnością również zaliczają się do tej kategorii). Przyjmując że u to odsetek wyników prezentowanych jako prawdziwe, choć w rzeczywistości błędnych przez zniekształcenia, wtedy wzór na PPV jest następujący:

$$PPV = ([1 - \beta] \cdot R + u \cdot \beta \cdot R) / (R + \alpha - \beta \cdot R + u - u \cdot \alpha + u \cdot \beta \cdot R)$$

Aby lepiej uzmysłowić znaczenie tego wzoru przygotowałem w tabeli 1.4 wyliczenia odsetka wyników fałszywych pozytywnych (spośród wszystkich badań pozytywnych) w zależności od różnych poziomów u i R , przy założeniu „tradycyjnego” poziomu $\alpha = 0,05$ oraz (dosyć optymistycznego) poziomu $\beta = 0,30$ (moc=0,7). Dla bardziej realistycznego poziomu $\beta = 0,5$ (moc=0,5) odsetek błędów fałszywych pozytywnych wzrasta o dalsze 2 % (przy $R=0,2$; $u=0,5$) do 7 % (przy $R=0,2$; $u=0,5$)

⁴¹ Za literaturą przedmiotu sformułowania „wynik pozytywny” używam w tym kontekście jako synonimu „wynik istotny statystycznie”, a „wynik negatywny” jako „wynik nieistotny statystycznie”. Nie ma to więc nic wspólnego z określeniem korelacja pozytywna/negatywna (tj. wprost- lub odwrotnie-proporcjonalna)

TABELA 1.4. Odsetek badań z wynikami fałszywymi pozytywnymi w zależności od odsetka rzeczywiście prawdziwych hipotez i odsetka badań zniekształconych przez błędy

R (odsetek rzeczywiście prawdziwych hipotez w stosunku do fałszywych hipotez)	u (odsetek badań pozytywnych zniekształconych przez świadome i nieświadome błędy)	Odsetek badań z wynikami fałszywymi pozytywnymi wśród wszystkich badań z wynikami pozytywnymi (1 – PPV)
0,2	0,1	50 %
0,2	0,3	68 %
0,2	0,5	76 %
0,4	0,1	33 %
0,4	0,3	51 %
0,4	0,5	61 %

Źródło: wyliczenia własne na podstawie Ioannidis (2005)

Jak więc widać z tabeli teza w tytule artykułu o fałszywości większości opublikowanych badań jest zasadna. Oczywiście wadą takiego rozumowania jest to, że wielkości R i u możemy tylko szacować. Oszacowanie wartości R nawet na poziomie 0,2 (na każde 2 rzeczywiście prawdziwe hipotezy przypada 10 rzeczywiście fałszywych hipotez) wydaje mi się optymistyczne: zdolność ludzkiego umysłu do generowania hipotez o prawidłowościach psychologicznych prawdopodobnie wielokrotnie przekracza rzeczywistą liczbę prawidłowości jakie stworzyła natura, a być może większość hipotez jest tak czy owak zafałszowana przez nieznaną większość zmiennych wpływających na ludzkie zachowanie, ogólniejszych zasad rządzących ludzkim umysłem czy słaby rozwój teorii w psychologii. Jeśli chodzi o wielkość wartości „u” pewną wskazówką co do jej poziomu w psychologii mogą być cytowane analizy Simmonsa i in. (2011) – wynika z nich, że badacz, które stosuje wszystkie sztuczki ma 61 % szans otrzymać „istotny” wynik przy braku jakichkolwiek efektów, a – jak już wspominałem – z badań ankietowych wynika, że niektóre z nich są stosowane przez większość badaczy. Do tego dochodzi bliżej nieokreślona liczba badań, gdzie występują – nieświadomie dla badacza – artefakty wynikające z niedoskonałości procedury, odgadywania hipotezy przez badanego, nietrafności narzędzia, komunikacji między badanymi i wielu innych (Rosenthal i Rosnow, 2009).

Sam Ioannidis dostarcza sugestii, co do tego, jakiego rodzaju procedury będą zwiększały prawdopodobieństwo wyników fałszywych pozytywnych (FP):

- a) im więcej zespołów testuje daną hipotezę, tym większe prawdopodobieństwo, że chociaż jeden z nich znajdzie wyniki FP,
- b) im mniejsze grupy badane tym mniejsza moc statystyczna badań, a więc większa szansa wyników FP (wówczas mniej wyników rzeczywiście pozytywnych jest odkrywanych, a więc wśród wszystkich pozytywnych wzrasta proporcja fałszywie pozytywnych),
- c) im słabsze siły efektów w danej nauce, tym większe prawdopodobieństwo błędów FP (analogicznie jak wyżej),
- d) im większa liczba testowanych zmiennych i większa dowolność w wyborze „ciekawych” wyników tym większa liczba błędów FP (wniosek identyczny jak u Simmonsa i in., 2011). Jak trafnie zauważa Ioannidis komercyjne programy do „data mining-u” reklamują się swoją zdolnością do odnajdywania istotnych zależności poprzez „drenowanie” danych.
- e) im większe motywacje finansowe i inne rodzaje presji w danej dziedzinie nauki tym większe prawdopodobieństwo wyników FP (tą pierwszą prawidłowość potwierdza doskonale zestawienie przygotowane przez Stroebe i in. (2012), gdzie zdecydowana większość udokumentowanych fałszerstw miała miejsce w naukach biomedycznych; można spekulować na ile występujący na zachodnich uniwersytetach nacisk „publish or perish” skłania niektórych badaczy do fałszerstw, żeby ułatwić sobie robienie kariery⁴²).

Ioannidis zaleca na te problemy kilka remediów. Po pierwsze testy statystyczne powinny mieć większą moc statystyczną, szczególnie poprzez liczniejsze grupy, lub przeprowadzanie meta-analiz. Dla równowagi dodaje, że badania o bardzo licznych próbach wiążą się z ryzykiem odkrywania efektów, które pomimo statystycznej istotności, mają praktyczną istotność zbliżoną do zera. Kolejnym zaleceniem jest rejestracja badań przed ich wykonaniem (postulat powtórzony m.in. przez Simmonsa i in., 2011), tak, aby ograniczyć arbitralność decyzji badacza i „wyławianie” ciekawych zależności spośród wszystkich testowanych. Wreszcie Ioannidis zaleca

⁴² Puszki Pandory na temat „jaka powinna być polityka oceny kadry na uczelniach” w tej pracy otwierać nie zamierzam. Brak jakiegokolwiek nacisku, jak naucza doświadczenie, skutkuje brakiem sensownej aktywności, więc pewnie jakaś równowaga między jednym i drugim jest pożądana.

poświęcać większą wagę prawdopodobieństwu prawdziwości hipotezy (R w jego równaniach) i na podstawie jej hipotetycznej wartości opierać projekty badawcze. W psychologii, w moim przekonaniu, porada ta ma niewielkie zastosowanie praktyczne relatywnie do medycyny, o której głównie pisze Ioannidis. Trudno jest mi sobie wyobrazić realistyczne oszacowanie wartości R dla psychologii. Testowane zmienne często nie podlegają bezpośredniej obserwacji, a ich pomiar jest zapośredniczony przez subiektywne procesy psychologiczne, możliwości artefaktów są większe, a i same rozumienie bardziej elementarnych procesów leżących u podstaw zachowania jest dużo bardziej ograniczone niż w medycynie (testując nowy lek można potencjalnie opierać się na znanych prawidłowościach z zakresu chemii organicznej, czy biologii komórki). Tak czy owak świadomość tego, że wartość R wywiera wpływ na prawdopodobieństwo prawdziwości wyników pozytywnych psychologombadaczom na pewno nie zaszkodzi.

ROZDZIAŁ 2

PROBLEMY TRAFNOŚCI ZEWNĘTRZNEJ BADAŃ W PSYCHOLOGII SPOŁECZNEJ

2.1. Wprowadzenie

Pozwolę sobie na początek tego rozdziału na kilka obserwacji z osobistej historii poznawania psychologii jako nauki. Jak chyba każdy student psychologii, na pierwszych latach studiów wiedzę o badaniach psychologicznych czerpałem przede wszystkim z książek akademickich i popularnonaukowych (kontakt z „normalnymi” artykułami naukowymi przyszedł znacznie później). Siłą rzeczy była to więc wiedza przede wszystkim o tzw. „klasycznych” badaniach. Książką, która wywarła wtedy na mnie duży wpływ była „40 prac badawczych, które zmieniły oblicze psychologii” (Hock, 2003). Praktycznie wszystkie opisane w niej badania oceniałem wówczas (i dziś podobnie) jako bardzo interesujące poznawczo. Procedury, zmienne zależne i sposób ich pomiaru były w nich łatwo intuicyjnie zrozumiałe, nawet dla laika. Wyniki były wyraźne i mówiły coś istotnego o naturze ludzkiej. Były to też (co mogę stwierdzić dopiero z dzisiejszej perspektywy) najczęściej badania w których zmienne zależne miały naturę behawioralną (lub jakoś „obiektywną” np. zmiana rozmiaru mózgu u szczurów pod wpływem stymulacji środowiskowej). Jedyne badania z listy czterdziestu, gdzie wykorzystano przede wszystkim narzędzia kwestionariuszowe, dotyczyło odziedziczalności osobowości.

Później, gdy zacząłem czytać naukę bez „filtra dydaktycznego”, tj. artykuły w naukowych czasopismach, początkowo doznałem dużego zawodu. Badania publikowane w nawet dobrych czasopismach często testowały mało inspirujące, poznawczo trudne do ogarnięcia hipotezy. Nierzadko miałem poczucie dużej arbitralności decyzji dotyczących zastosowanych procedur, obliczeń czy sposobów przedstawiania danych (kilka przykładów podałem w rozdziale 1; dalsze w tym rozdziale). Wyniki, choć zawsze istotne statystycznie, bardzo często wskazywały na słabe efekty. I co najważniejsze, rzadko badano zachowanie jako takie. Wszelakie kwestionariusze, deklarowane postawy, hipotetyczne odpowiedzi na dane sytuacje itp. były najczęściej spotykane. Zaobserwowałem też, że standardową grupą badaną są studenci i od tej reguły wyjątki są nieczęste.

Różne formy danych samoopisowych (czy mówiąc inaczej „introspekcyjnych”) budziły we mnie pewien sceptycyzm. Zasadniczo wynikało to (i wynika) z przekonania, że niewiele fundamentalnych odkryć zostało dzięki nim dokonanych. Patrząc na problem z innej strony, różnymi formami danych „introspekcyjnych” ludzkość dysponuje od tysiącleci, ale do XX wieku postęp w systematycznej wiedzy o zachowaniu był bardzo niewielki.

Przez pewien czas większą estymą darzyłem tzw. miary utajone (ang. *implicit measures*) oparte na czasach reakcji (stąd np. Budzicz, 2012b), wierząc, że mają doskonałą trafność wewnętrzną i przyzwoitą zewnętrzną, a przede wszystkim są trudne do zniekształcania przez badanego (świadomego i nieświadomianego). Kilka nieudanych badań z ich wykorzystaniem i dalsze lektury zmieniły mój stosunek również do takich miar.

Żeby nie być gołosłownym przedstawię przykładowe badanie, które dobrze obrazuje mój zawód nauką „z górnej półki”.

W artykule opublikowanym w JPSP, zespół badaczy badał wpływ lękowych sytuacji na siłę przekonań religijnych (McGregor, Nash, Prentice, 2010). Wg teorii lękowe sytuacje (np. porażka w zadaniu poznawczym), będą skłaniały jednostkę do przekierowania uwagi na te dziedziny życia, w których mogą poczuć się „kompetentni” i niejako zrekompensować sobie niepewność. Zdaniem autorów taką dziedziną jest między innymi sfera przekonań i doświadczeń religijnych (teorię przedstawiam w dużym uproszczeniu).

Przeprowadzono 3 eksperymenty, w których część badanych poddawano sytuacji lękotwórczej (w dwóch badaniach polegało to na tym, że badani czytali niezrozumiały tekst, w trzecim opisywali myśli i emocje związane z trudnościami w relacji z jakąś bliską osobą). Następnie mierzono przekonania religijne. W każdym eksperymencie używano jednak innych zmiennych zależnych, a dodatkowo czasami miały one kilka komponentów. Pytano więc m.in. o wiarę w Boga i stopień zgadzania się z podstawowymi dogmatami religijnymi (badanie 1), siłę „żarliwości” (ang. *zeal*) religijnej wyrażającej się np. w braku wątpliwości czy gotowości walki o religię (badanie 2), oraz o stopień internalizacji niektórych przekonań religijnych, religijny ekstremizm oraz religijny szowinizm (badanie 3). Dodatkowo mierzono szereg zmiennych dyspozycyjnych, z których niektóre zestawiano w nadrzędne skale.

Ogólnie tylko w jednym z trzech badań wykazano prosty efekt polegający na wzroście przekonań religijnych pod wpływem manipulacji eksperymentalnej. W dwóch pozostałych badaniach nie podano informacji o tym, czy taki prosty efekt zadziałał (można przypuszczać, że nie). Natomiast we wszystkich 3 badaniach uzyskano istotny efekt wpływu manipulacji eksperymentalnej na siłę przekonań religijnych, ale tylko w interakcji z dodatkowymi zmiennymi dyspozycyjnymi, w każdym badaniu innymi (te konstrukty to: awersja do lękowej niepewności; motywacja dążeńiowa, poczucie wpływu na realizację codziennych celów⁴³). We wstępie nie ma informacji o potencjalnym wpływie zmiennych dyspozycyjnych na efekt, dopiero pod sam koniec części teoretycznej, przed opisem badań wspomina się, że postanowiono również sprawdzić wpływ zmiennych dyspozycyjnych na rzeczony efekt. Wywołuje to poczucie, że zmienne dyspozycyjne jako współzmiennne są dodane cokolwiek na siłę i teoria jako taka ich nie przewiduje.

Dodatkowo uzyskane efekty interakcji są częściowo zgodne z intuicją np. osoby o wyższej awersji do lękowej niepewności mają wyższą siłę przekonań religijnych pod wpływem manipulacji. Ale bywają też problematyczne w interpretacji np. osoby o niższej awersji do lękowej niepewności pod wpływem sytuacji lękowej mają niższą siłę przekonań religijnych. W tym ostatnim przypadku bardziej przekonujący byłby wniosek, że osoby o niskiej awersji nie będą zmieniały siły przekonań religijnych. Niepokojące jest też, że siła efektów interakcji zmniejszała się wraz ze wzrostem liczebności grupy badanej ($R^2 = 0,13$; $R^2 = 0,09$; $R^2 = 0,03$; dla N odpowiednio 59; 96 i 120), choć interpretacje tych danych utrudnia fakt, że każdorazowo testowano inne zmienne zależne w interakcji z innymi zmiennymi dyspozycyjnymi.

Badanie to w moim przekonaniu skupia jak w soczewce wszystkie słabości wielu badań psychologii społecznej (niektóre już omawiane w rozdziale 1):

- ciekawa teoria zakładająca względnie proste, intuicyjnie uchwytnie efekty, które jednak w kontakcie z empirią nie są już tak przekonujące (np. działają tylko w interakcji z dodatkowymi zmiennymi),
- prezentowanie głównie efektów, które działają, a inne potencjalnie ciekawe dane (np. wpływ prosty manipulacji na zmienną zależną) z niezrozumiałych przyczyn są pominięte,

⁴³ Nie wiem, czy te terminy zostały przyswojone w polskiej literaturze, więc podaje własne tłumaczenie. W oryginale ang.: anxious uncertainty aversion; approach motivation, empowered temporal goal engagement.

- arbitralne decyzje (np. o sumowaniu niektórych zmiennych dyspozycyjnych),
- duża liczba raportowanych zmiennych (i ich różnorodność), z których zasadniczo większość układa się w „działającą” całość, co musi wzbudzać podejrzenia o wybiórczą prezentację danych,
- do wyników dorobiona następnie w podsumowaniu odpowiednia „obudowa” teoretyczna, w której niewiele jest miejsca na opisywane tu wątpliwości,
- wielka złożoność procedury badawczej poprzez rozliczne zmienne i analizy. Ogarnięcie badania wymaga dużego wysiłku poznawczego, a sama złożoność może – spekulowałbym – zniechęcać Czytelników (i recenzentów) do bardzo wnikliwej analizy i krytyki.
- używanie tylko zmiennych samoopisowych.

To tylko jeden z przykładów, gdy kontakt z „powszechną” nauką jest poznawczo dużo mniej atrakcyjny, niż kontakt z „klasycznymi” badaniami. Podkreślę raz jeszcze, że jest to nauka z czasopisma uznawanego za niezwykle prestiżowe (JPSP) (nie będę już zajmował miejsca przykładami, ale kontakt z czasopismami „z niskiej półki” jest na ogół jeszcze mniej inspirujący).

Dlatego miłą niespodzianką było dla mnie, gdy trzech badaczy o uznanym dorobku w psychologii społecznej wyraziło podobny zawód nad nauką w tekście na łamach PoPS (Baumeister, Vohs, Funder, 2007). Zatytułowany został on prowokująco „Psychologia jako nauka o samoopisach i ruchach palców” i zdobył niemałe uznanie w społeczności badaczy (ponad 480 cytowań). („Ruchy palców” w tym kontekście to różne metody oparte na czasach reakcji, czyli miarach utajonych). Tytułowy wniosek oparto na przeglądzie literatury, dodam, że stosunkowo niedużym. Autorzy przeanalizowali kilka numerów JPSP i pokazali, że rzeczywistego zachowania jest we współczesnej psychologii bardzo niewiele. Olbrzymią większość stanowią jego pośrednie miary takie jak tytułowe samoopisy i techniki oparte na czasach reakcji („ruchy palców”). W analizowanych numerach JPSP (z 2006 roku) mniej niż 20 % empirycznych badań polegało na obserwacji i pomiarze rzeczywistego badania. Pozostałe badania wykorzystywały rozmaite skale szacunkowe, szczególnie w postaci samoocen, czasami wymagano od badanych, aby wyobrazili sobie hipotetyczne sytuacje i określili, jak by zareagowali, a kiedy indziej proszono o

sprawozdanie z przeszłych zachowań. Autorzy w ten sposób podsumowali swoją analizę:

Tak oto wygląda „zachowanie” w czołowym czasopiśmie psychologii społecznej i osobowości: oceny, oceny i jeszcze raz oceny, okazjonalnie jakiś wybór, czytanie i wypełnianie testu oraz przekreślanie litery „e” (Baumeister i in., 2007, s. 398).

Dodatkowo przeanalizowano również zawartość numerów JPSP z wcześniejszych dekad. Stwierdzono, że spadek znaczenia rzeczywistego zachowania rozpoczął się w latach 80', bo jeszcze w latach 70' zachowanie występowało w ponad 80 % badań. Autorzy przedstawili kilka hipotez na temat przyczyn takiego stanu rzeczy. Po pierwsze, może to wynikać z pewnego oportunizmu psychologów. Poproszenie badanych o wypełnienie kwestionariusza lub zareagowanie na serie bodźców na ekranie komputera jest logistycznie mniej czaso- i kosztochłonne niż zaaranżowanie eksperymentów behawioralnych, co ma znaczenie szczególnie w sytuacji, gdy prestiżowe czasopisma wymagają kilku badań (np. Carver, 2004), oczywiście z samymi udanymi wynikami. Dla dobra kariery jest to więc strategia dużo mniej ryzykowna. Dodatkowo, ich zdaniem, komisje etyczne na uniwersytetach wykazują niejednokrotnie nadmierne przewrażliwienie na zagrożenia etyczne badania: „z punktu widzenia komisji etycznej jest mniej inwazyjne zapytać badanego ile zjadł, niż obserwować ile rzeczywiście zjadł [...]. Analogicznie jest mniej inwazyjne zmierzenie „uprzedzeń” poprzez szybkość naciskania klawiszy w odpowiedzi na bodźce na komputerze niż zaobserwowanie jak badany wchodzi w interakcje z osobą, wobec której mógłby być uprzedzony” (Baumeister i in., 2007 s. 400).

Podobne wnioski wyciągnął Furr (2009), jednak były one poświęcone obecności zachowania w psychologii osobowości. Przeanalizował trzy roczniki (1997; 2002; 2007) JPSP (sekcji poświęconej osobowości) oraz *Journal of Personality*. Różne formy danych behawioralnych występowały w ok. 20-25 % badań, przy czym bezpośrednio obserwowane zachowanie w 2,5 %, a raportowane bieżące zachowanie (np. w postaci prowadzonych dzienników) w około 1 %. Większość „danych behawioralnych” (ok. 13 %) stanowiły retrospektywne sprawozdania z przeszłych zachowań robione przez badanego lub osoby dobrze go znające, trudno więc tak naprawdę mówić tu o danych behawioralnych sensu stricto. Zdecydowana większość badań nie zawierała żadnych danych behawioralnych, ale „specjalność” psychologów, czyli korelacje testów z testami.

Nadmienię też, że Furr wprowadza definicję zachowania, która będzie mi przydatna w dalszej części, a która, jak się wydaje, czyni sprawne rozróżnienie pomiędzy rzeczywistym zachowaniem, a jego pośrednimi wskaźnikami. Zachowanie to więc „wypowiedzi (wyłączając raporty werbalne w kontekście testowania psychologicznego) lub ruchy, które są potencjalnie dostępne dla uważnego obserwatora korzystającego z normalnych procesów zmysłowych” (Furr, 2009, s. 372). Ta definicja wyłącza z „zachowania” wewnętrzne procesy fizjologiczne lub neuronalne, ale także milisekundowe różnice w czasach reakcji, które są niedostępne dla obserwatora korzystającego z normalnych procesów zmysłowych. Poprzez uwzględnienie tylko obserwowalnych elementów zachowania, kładzie szczególny nacisk na rolę zachowania w interakcjach społecznych.

Pozwolę sobie w tym momencie na małą dygresję. Założeniem, które przyjmuję apriori w tej pracy jest takie, że przedmiotem zainteresowania psychologii jest tak definiowane zachowanie. Szczególnie zachowanie interesujące jest oczywiście dla psychologów społecznych, oraz przedstawicieli wszystkich obszarów psychologii stosowanej. Istnieją obszary badań, których „zachowanie” zasadniczo nie interesuje, jak np. badanie elementarnych procesów przetwarzania informacji w mózgu, szczególnie ich neuronalnej podstawy. Jednak większość psychologii w moim silnym przekonaniu sprowadza się zawsze w ostatecznym rozrachunku do zachowania np. badamy kwestionariuszowo zróżnicowanie osobowości, ale robimy to głównie, żeby móc przewidywać, jak określone cechy osobowości będą się objawiały w codziennych kontekstach; pytamy o postawy, gdyż zwykle implikują one określone zachowania, lub gotowość do podejmowania pewnych zachowań w szczególnych sytuacjach; badamy inteligencję przy pomocy testów, ale tak naprawdę interesuje nas zdolność do rozwiązywania problemów poznawczych w pracy, w szkole lub w codziennym życiu; mierzymy poziom depresji u pacjentów po to, żeby zapobiegać psychicznemu cierpieniu, ale też przewidywać niekorzystne zjawiska np. wycofanie społeczne lub próby samobójcze itd.

Badania nad zachowaniem są nie tylko dużo atrakcyjniejsze poznawczo. Zaryzykuję tezę, że są przede wszystkim tym, co jest esencją tej nauki i jej ostatecznym celem.

2.2. Przykłady alternatywnego podejścia

Baumeister, Vohs i Funder (2007) przeanalizowali tylko po dwa numery JPSP z każdej dekady, a Furr (2009) trzy roczniki dwóch czasopism, więc jest to stosunkowo skromna analiza (jednak moje odczucia z lektur szerszej grupy czasopism i roczników są podobne). Zakładając, że te analizy są trafne, nie wyklucza to oczywiście tego, że dane behawioralne pojawiają się czasami we współczesnych czasopismach. Tak się na pewno dzieje, i zanim przejdę do dalszych rozważań przytoczę kilka ciekawych przykładów z ostatnich lat, które udało mi się „wyłowić” z ogromu danych niebehawioralnych. Posłużą one do zilustrowania kilku ważnych koncepcji w dalszej części.

- Keizer, Lindenberg i Steg (2008) sprawdzali, czy przestrzeganie przez innych norm społecznych wpływa na to, czy ludzie sami chętniej je łamią (a więc czy polityka „zero tolerancji dla najmniejszych wykroczeń” jest skuteczna). W serii eksperymentów aranżowano otoczenie tak, żeby wskazywało na to, że normy są lub nie są przestrzegane. Przykładowo, obok napisu „zakaz malowania ścian” znajdowały się ściany zamalowane graffiti lub zupełnie czyste. Przypadkowi przechodnie byli osobami badanymi. Mieli okazję zachować się w sposób zgodny z normami lub niezgodny (np. w jednym z badań mogli ukraść banknot 5 euro z listu wystającego ze skrzynki, lub wrzucić list do skrzynki bez naruszania koperty; w innym mogli sobie skrócić drogę pomimo zakazu przechodzenia itp.). Ostatecznie, w 6 niezależnych eksperymentach, w których zastosowano różne formy manipulacji i pomiaru zmiennej zależnej, stwierdzono, że otoczenie wskazujące na przestrzeganie norm sprzyja ich przestrzeganiu.
- Francuski psycholog Nicolas Gueguen testował szereg hipotez (na ogół inspirowanych ewolucyjnie) dotyczących tego, co zwiększa szanse na powodzenie u osobników płci przeciwnej, przynajmniej jeśli chodzi o etap umówienia się na randkę. Mężczyźni-pomocnicy badacza próbowali zdobyć na ulicach numer telefonu przechodzących dziewczyn. Jak się okazało, odnosili większy sukces jeśli: wychodzili z luksusowego samochodu (Gueguen, Lamy, 2012); byli ubrani w strój strażaka (Guéguen, 2009) albo mieli ze sobą gitarę (Guéguen, N., Meineri, S., & Fischer-Lokou, 2013).
- Ten sam badacz sprawdzał również, co przyciąga uwagę mężczyzn. Co ciekawe, zmienna zależna w takich badaniach miała inną postać: mierzono nie liczbę zdobytych numerów telefonów, ale liczbę podchodzących mężczyzn w klubie

nocnym lub czas, jaki upłynął, zanim kobietę zagadał pierwszy mężczyzny od wejścia do baru. Jak się okazało, kobiety przyciągały większą uwagę mężczyzn, jeśli: były wyzywająco ubrane (Guéguen, 2011); miały intensywny makijaż (Gueguen, 2008a); miały powiększone piersi przy pomocy dodatkowych wkładek (Gueguen, 2007) lub uśmiechały się do mężczyzn (Gueguen, 2008b). W zbliżonym tonie inny zespół obserwował w klubie nocnym jakie cechy i zachowania kobiet powodują, że podchodzi do nich więcej mężczyzn. Okazało się, że wpływ ma większa ekspozycja ciała, szczególnie dekoltu oraz zmysłowe ruchy taneczne (Hendrie, Mannion, Godfrey, 2009).

- Neville (2012) badał związki między nieuczciwością akademicką polegającą na kupowaniu prac dyplomowych, a poziomem zaufania społecznego w poszczególnych amerykańskich stanach. Wykorzystano dane firmy Google dotyczące tego, jak często użytkownicy w wyszukiwarce internetowej w zależności od stanu wpisywali hasła w rodzaju „pisanie prac dyplomowych”. W stanach, w których częściej szukano takich usług był wyraźnie niższy poziom zaufania społecznego. Takiego efektu nie było jednak dla etycznie neutralnych, choć też akademickich, zapytań w rodzaju „standardy APA jak stosować” (kontrolowano oczywiście dodatkowe zmienne, np. liczbę studentów w poszczególnych stanach).
- Zespół badaczy w szeregu badań wyposażali osoby w niewielkie dyktafony (Electronically Activated Recorder – EAR) uruchamiające się w przypadkowych momentach i nagrywające to, co się aktualnie dzieje w otoczeniu. Badania takie trwały zwykle kilka dni. Badani nie wiedzieli, kiedy dyktafony działają, ale po skończonym badaniu mogli odsłuchać wszystkie nagrania i ewentualnie usunąć wybrane fragmenty. Metoda ta pozwoliła odkryć szereg ciekawych prawidłowości, zarówno przewidywalnych, jak i zaskakujących. W jednym z badań (opublikowanych w *Science*) okazało się wbrew odwiecznym stereotypom, że kobiety w ciągu dnia nie mówią dużo więcej słów niż mężczyźni (Mehl, Vazire, Ramírez-Esparza, Slatcher, Pennebaker, 2007). W innym badano związki osobowości (mierzonej kwestionariuszowo) i rozmaitych zachowań. Przykładowo, okazało się, że ekstrawertycy częściej przebywają w towarzystwie, ale nie rozmawiają częściej przez telefon. Poziom neurotyzmu nie wpływa na częstość płakania, ale neurotyczne kobiety rzadziej się śmieją. Osoby otwarte na

doświadczenie częściej przebywają poza domem, choć dotyczy to tylko mężczyzn itd. (Mehl, Gosling, Pennebaker, 2006).

- Wang i Repetti (2014) nagrywali na wideo zachowania badanych w normalnych domowych kontekstach, sprawdzając między innymi zależność między poziomem wsparcia między małżonkami (na podstawie nagrań) a ilością symptomów depresyjnych (na podstawie miar samoopisowych). W innym badaniu sprawdzali, jak stres w pracy wpływa na interakcje między członkami rodzinami w ciągu kilku godzin po powrocie do domu z pracy (Wang, Repetti, Campos, 2011). Stres w pracy sam w sobie nie miał wpływu na pogorszenie (nagrywanych) interakcji, ale zanotowano taki wpływ, gdy poziom neurotyzmu był podwyższony u jednego z małżonków. Warto dodać, że ochotnicy za tak znaczące naruszenie ich prywatności otrzymywali niebagatelną gratyfikację w wysokości 1000 dolarów, co niestety pokazuje, że dane behawioralne mogą się wiązać ze znacznie wyższymi nakładami.
- Nolan, Schultz, Cialdin, Goldstein, Griskevicius (2008) sprawdzali skuteczność różnych rodzajów komunikatów skłaniających do oszczędzania prądu. Wszystkich badanych informowano o tym, w jaki sposób mogą zaoszczędzić energię. W zależności od grupy wskazywano na korzyści z tego płynące takie jak ochrona środowiska, osobisty interes („taka a taka oszczędność na rachunku”) czy deskryptywne normy społeczne (przykładowo „77 % osób w Twojej okolicy oszczędza energię”). Badacze pytali o to, jak skuteczny wydaje się określony komunikat, ale nie ograniczali się tylko do tego. Odczytywali liczniki prądu i sprawdzali, którzy badani rzeczywiście zaczęli oszczędzać. Sami badani sądzili, że najskuteczniejsze są komunikaty, w których odwołujemy się do korzyści dla środowiska lub korzyści finansowych, ale w rzeczywistości okazało się, że zdecydowanie najskuteczniejsze jest odwoływanie się do norm społecznych.
- Cobey, Stulp, Laan, Buunk, Pollet (2013) oraz Pawlowski, Atwal, Dunbar (2008) w niezależnych badaniach obserwowali zachowanie rowerzystów i pieszych. Szczególnie interesowały ich różnice międzypłciowe w częstości podejmowania ryzykownych zachowań (takich jak przejeżdżanie na strzeżonym przejeździe kolejowym, (nie)używanie świateł po zmroku, przechodzenie na czerwonym świetle itp.). Wyraźnie częściej mężczyźni zachowywali się w sposób bardziej ryzykowny.

Podany zbiór badań jest dosyć szeroki, obejmuje zarówno eksperymenty (przede wszystkim terenowe), quasi-eksperymenty, jak i badania korelacyjne. Łączy je to, że przynajmniej jedną ze zmiennych było zachowanie (w tej czy innej formie). Badania takie jak tutaj opisane odróżniają się w kilku istotnych kwestiach od „tradycyjnych” badań psychologicznych opartych na kwestionariuszach i innych pośrednich wskaźnikach zachowania. Są one, na ogół, koncepcyjnie znacznie prostsze. W badaniach behawioralnych ograniczenia praktyczne wymuszają mocno uproszczony model analizy statystycznej. W eksperymencie terenowym często badacze są w stanie stworzyć dwa-trzy warunki eksperymentalne i mierzyć tylko jedną zmienną zależną. W badaniach laboratoryjnych, niejednokrotnie badani otrzymują plik kwestionariuszy, istnieje wiele warunków eksperymentalnych, współzmiennych i zmiennych zależnych. A wielka liczba testowanych zależności, jak już wspominałem, jest najprostszym sposobem na „złwienie” jakichś statystycznie istotnych efektów (rdz. 1.2-1.4).

Takie badania, jak tutaj opisane, używają też zmiennych wyrażonych w intuicyjnie łatwiej uchwytanych jednostkach, szczególnie dla niespecjalistów. Przykładowo, informacja o tym, jaki procent kobiet zgodził się umówić na randkę (np. Gueguen, 2009) jest łatwiejsza w interpretacji niż informacja, że, dajmy na to, kobiety oceniały określonych hipotetycznych mężczyzn jako bardziej atrakcyjnych o 0,39 punktu w likertowskiej skali 11 punktowej (np. tak jak w badaniu Bressan, Stranieri, 2008).

Za największą zaletę takich badań uważam jednak ich relatywnie niskie nasycenie teorią. Wymaga to z mojej strony nieco szerszego wyjaśnienia.

2.3. Problem uteoretyzowania obserwacji

W filozofii nauki występuje tzw. problem uteoretyzowania obserwacji (ang. theory-ladenness of observation; np. Grobler, 2006, rdz. 3; Okasha, 2002, rdz. 5; Kukla, 2008). Głównym jego założeniem jest to, że nie istnieje coś takiego jak „czysta obserwacja”. Każdy akt obserwacji jest zależny od wielu dodatkowych założeń. Dajmy na to, dokonując obserwacji „empirycznej” o treści „słońce weszło dziś o 5:43”, dokonujemy szeregu niekoniecznie wyartykułowanych założeń, np. że narzędzie pomiarowe (zegar) rzetelnie wskazuje godzinę, nie ulegamy aktualnie halucynacjom, a w atmosferze nie występują zjawiska zmieniające zasadniczo obraz

itp. Oczywiście, w zaawansowanej nauce mamy na ogół do czynienia z dużo bardziej wyrafinowanymi obserwacjami. Za twierdzeniem „prąd płynie w obwodzie” kryje się sporo teorii na temat natury świata czy działania aparatury pomiarowej (por. Okasha, 2002, rdz. 5). Nie wspominam tu nawet o niewyobrażalnie złożonych eksperymentach (np. akcelerator LHC w CERN: Lincoln, 2010), w których ilość wiedzy potrzebnej do skonstruowania i obsługi aparatury oraz interpretacji danych wymaga współpracy setek specjalistów. Na ekranie badaczy nie pojawia się po prostu informacja „znaleziono bozon Higgsa” – wyciągnięcie takiego wniosku jest niemożliwe w oderwaniu od olbrzymiej liczby dodatkowych założeń.

Zdaniem niektórych filozofów nauki (szczególnie Kuhna, por. Kukla, 2008) konsekwencją problemu uteoretyzowania jest to, że nie ma obserwacji prawdziwych „obiektywnie”, ale że obserwacje są prawdziwe tylko w kontekście jakichś teorii, czy szerzej: w kontekście określonego paradygmatu badawczego (swego czasu polemizowałem z najbardziej skrajnymi przedstawicielami tego poglądu: Budzicz, 2011). Nie uważam, żeby ten argument miał praktyczne znaczenie w codziennej działalności badaczy.

O ile zapewne faktycznie „czyste obserwacje” nie występują, o tyle teorie potrzebne do uzasadnienia obserwacji mogą być zupełnie innymi teoriami, niż te, które te obserwacje mają potwierdzać (lub falsyfikować). Przykładowo, badając różnice międzypłciowe w reakcjach na zdradę, Kuhle (2011) wyodrębnił rodzaje reakcji na zdradę (na podstawie wcześniejszych teorii) i zoperacjonalizował ich wskaźniki. Jego obserwacje były więc zależne od pewnych sformalizowanych założeń, ale założenia te były czymś zupełnie niezależnym od testowanej głównej hipotezy mówiącej, że mężczyźni wykazują wyższy poziom zazdrości seksualnej, a kobiety emocjonalnej. Jest całkowicie możliwe, że w obrębie zaproponowanej przez Kuhle operacjonalizacji, badania wskazałyby na odwrotny efekt lub na brak różnic.

Pomimo tych zastrzeżeń problem uteoretyzowania obserwacji ma pewne znaczenie dla naszych rozważań. Zakładam, że istnieją obserwacje o różnym stopniu „uteoretyzowania”. Szczególnie ma to miejsce wtedy, gdy potwierdzenie podstawowej hipotezy badawczej jest możliwe tylko wtedy, gdy przyjmiemy bardzo wiele dodatkowych założeń dotyczących operacjonalizacji zmiennych, manipulacji eksperymentalnej czy trafności procedury. Porównajmy przykłady, jak mierzyć „agresywność”:

- a. liczba bójek, które inicjuje osoba w okresie jednego roku,
- b. ilość aplikowanych drugiemu badanemu sesji nieprzyjemnego białego szumu (jak u Andersona i Dilla, 2000),
- c. wydłużenie czasu reakcji na słowa związane z agresją w emocjonalnym zadaniu Stroopa (Gollwitzer, Denzler, 2009; przykładowo użyto słów takich jak „atak”, „nienawiść”, „przemoc”).

W przykładzie a. operacjonalizacja zmiennej „agresywności” odpowiada zapewne z grubsza temu, jak laicy wyobrażają sobie koncept „agresywności”. W kolejnych przypadkach trzeba przyjąć więcej założeń. W przykładzie c. takie założenia mogłyby być następujące:

- emocjonalny efekt Stroopa jest ogólnie trafny i odzwierciedla zmiany w naszym systemie poznawczym, które potencjalnie przekładają się na rzeczywisty wzrost agresywności,
- nawet, jeśli emocjonalny efekt Stroopa jest trafny, to zakładamy, że trafne są też konkretne słowa użyte w danym badaniu i rzeczywiście aktywizują one schemat agresji,
- pomiar efektu Stroopa nie wchodzi w interakcję z procedurą np. badani domyślają się celu i reagują inaczej na słowa związane z agresją itp.,
- nawet, jeśli procedura emocjonalnego efektu Stroopa jest zasadniczo trafna tj. koreluje dodatnio i niezerowo z aktywizacją agresywności w umyśle, to nie musi to automatycznie oznaczać, że zaobserwowane różnice mają jakiegokolwiek realne przełożenie na rzeczywistą agresywność, szczególnie w sytuacji, gdyby się okazało, że korelacja emocjonalnego efektu Stroopa z agresywnością jest bardzo słaba itd.

Inny przykład, tym razem związany nie z trafnością pomiaru jakiejś zmiennej, ale z trafnością manipulacji. Załóżmy, bez wikłania się w dywagacje teoretyczne, że istnieje coś takiego jak motywacja romantyczna, wyrażająca się m.in. zwiększoną gotowością do nawiązywania intymnych relacji z osobnikami płci przeciwnej. W jaki sposób możemy ją wzbudzić? W literaturze można znaleźć takie sposoby:

- I. poprzez możliwość flirtu z atrakcyjnym osobnikiem płci przeciwnej obecnym w pokoju w którym odbywa się eksperyment (rzekomo drugim badanym, w

rzeczywistości pomocnikiem eksperymentatora) (tak jak u Lydon, Menzies-Toman, Burton, Bell, 2008);

- II. polecenie wyobrażania sobie flirtu z atrakcyjną osobą płci przeciwnej (tak jak u Griskeviciusa, Cialdiniego i Kenricka, 2006);
- III. krótkotrwałe, sublimalne ekspozycje słów związanych z miłością (tak jak u Mikulincer, Shaver, Gillath, Nitzberg, 2005).

Znowu, wraz z kolejnym „poziomem” narasta liczba założeń, jakie trzeba przyjąć, żeby być przekonanym, że manipulacja jest trafna. Dajmy na to w punkcie II dodatkowe założenia jakie trzeba przyjąć to:

- badani rzeczywiście wyobrażają sobie rzeczy zgodnie z poleceniem eksperymentatora (a nie „bujają w obłokach”);
- wyobrażanie sobie w porównaniu do bezpośredniej obecności atrakcyjnego osobnika płci przeciwnej nie aktywizuje jakichś rejonów mózgu, które zasadniczo zmieniają naturę efektu;
- wyobrażanie ma dostateczną siłę (w porównaniu do rzeczywistej interakcji), żeby zaktywizować efekt itp.

W przypadku naszych bardziej „realistycznych” sposobów operacjonalizacji lub manipulacji (a i I) badacze oczywiście też przyjmują rozmaite założenia, np. że badani nie domyślają się, że atrakcyjna osoba z którą rozmawiają jest pomocnikiem eksperymentatora, ten pomocnik jest „w typie” osób badanych itp., ale jak się wydaje, liczba potencjalnych założeń jest mniejsza, a dodatkowo te zmienne i rodzaje manipulacji bardziej realistycznie oddają codzienne konteksty. Rażenie białym szumem jest raczej niecodzienną sytuacją. Jednak biorąc pod uwagę wymagania etyczne, mimo wszystko może to być bardziej „realistyczny” pomysł niż pomiar agresji poprzez dostępność poznawczą określonych słów mierzoną emocjonalnym zadaniem Stroopa lub testem decyzji leksykalnych. Gdyby się pojawiły jakieś dane, które stawiają pod znakiem zapytania całą procedurę emocjonalnego testu Stroopa, automatycznie dyskwalifikowałoby to trafność takich badań. Ewentualna falsyfikacja procedury rażenia białym szumem wydaje się być mniej prawdopodobna (oczywiście badanie może okazać się nietrafne z szeregu innych powodów).

Tak więc podsumowując ten problem: uteoretyzowanie obserwacji jest czymś nieuniknionym w nauce, powinniśmy jednak dążyć do możliwie jego najniższego poziomu. Psychologia najprawdopodobniej nie jest na takim poziomie rozwoju, żebyśmy musieli posługiwać się tak wysoce uteoretyzowanymi obserwacjami jak np. w fizyce cząstek wysokich energii.

Największej zalety niskiego poziomu uteoretyzowania obserwacji upatruję w tym, że takie obserwację mogą być łatwo inkorporowane przez różne teorie. Ewentualna falsyfikacja teorii, które zainspirowały badanie, nie musi oznaczać automatycznej falsyfikacji trafności takich danych, jeśli są one oparte na metodach stosunkowo mało uteoretyzowanych. Przykładowo, w klasycznym badaniu psychologii społecznej, mężczyźni, którzy przechodzili przez wiszący most, byli zagadywani przez atrakcyjną pomocniczkę badaczy (Dutton, Aron, 1974). Dawała im swój numer telefonu. Mężczyźni tacy następnie próbowali się częściej z nią umówić, niż mężczyźni zaczepiani w bardziej „spokojnych” okolicznościach. Wyniki tych badań interpretowano w ten sposób, że mężczyźni, którzy przeszli przez most, nie rozumieli źródła swojego pobudzenia i przypisywali je spotkaniu z kobietą. Nie musi to być jednak jedyne wyjaśnienie. Alternatywna hipoteza może opierać się na tzw. teorii ewolucji historii życiowych, która wskazuje, że zagrażające środowisko aktywizuje nasz popęd seksualny. Dotyczy to szczególnie mężczyzn, którym do odniesienia sukcesu reprodukcyjnego wystarcza samo zapłodnienie. Zagrożenie nie musi aktywizować równie silnie popędu u kobiet, gdyż ewentualny sukces reprodukcyjny wymaga u nich znacznie więcej czasu i energii niż samo zapłodnienie. Istotnie, w jednym badaniu przymowanie konceptem śmierci, zwiększało dostępność treści seksualnych, ale tylko u mężczyzn (Gillath, Landau, Selcuk, Goldenberg, 2011⁴⁴) (analogicznie u Duttona i Arona efekt nie występował u kobiet).

W każdym razie stosunkowo niski poziom uteoretyzowania obserwacji z badania Duttona i Arona jest jego wielką zaletą, gdyż obserwacje te mogą być potencjalnie inkorporowane przez rozmaite teorie wyjaśniające naturę ludzką i jej szczegółowe mechanizmy. W tym badaniu, o czym rzadziej się wspomina, mierzono również inną zmienną zależną: poziom seksualizacji treści wypowiedzi na podstawie swobodnej „interpretacji” obrazka w teście projekcyjnym. Tutaj, poziom uteoretyzowania jest

⁴⁴ Podaję to badanie jako przykład, choć nie przekonuje mnie w tym tekście jedna ze zmiennych zależnych (badanie 2), jak również sposób manipulacji eksperymentalnej (w badaniu 1 i 2).

dużo wyższy, trzeba wręcz zacząć od elementarnego problemu trafności takich testów (szeroka dyskusja np. Lilienfeld, Wood, Garb, 2000). Wydaje się natomiast, że próba umówienia się na randkę przez telefon raczej nie zostanie nigdy zakwestionowana jako wyznacznik motywacji romantycznej (i/lub seksualnej) w przeciwieństwie do wyników testu projekcyjnego (podobnie jak przechodzenie przez wiszący most jako manipulacja służąca indukcji lęku⁴⁵; np. w porównaniu do wyświetlania wyrazów oznaczających śmierć, tak jak u Gillath i in., 2011 badanie 2). Efekt ten można opisać bez odwoływania się do teorii wyższego rzędu tak: „przechodzenie przez most wiszący skłania mężczyzn do częstszych prób umówienia się z atrakcyjną kobietą”. Takie proste „prawo obserwacyjne”, które mogą następnie próbować wyjaśnić teorię wyższego rzędu (np. „nierozumienie źródła pobudzenia powoduje przypisywanie go bodźcom, które są najłatwiej dostępne”). Niepowodzenie jednej teorii wyższego rzędu nie przekreśla jednak podstawowego prostego efektu.

Jestem przekonany, że rozwój psychologii nie może się obejść bez dużego, w miarę rzetelnego kanonu faktów, które są relatywnie nisko uteoretyzowane, a więc mogą być inkorporowane przez wiele teorii. Dobrze udokumentowane fakty są czymś zasadniczo najpewniejszym w nauce, i mogą trwać nawet wtedy, gdy teorie je wyjaśniające ulegną zmianie. Przykładowo MacLeod (1991) zebrał najważniejsze wnioski płynące z 50 lat badań nad efektem Stroopa, przedstawił najważniejsze spory teoretyczne o jego naturę oraz stworzył listę kilkunastu pytań, na które przyszłe pokolenia badaczy powinny odpowiedzieć. Choć efekt Stroopa sam w sobie jest dosyć prosty do eksperymentalnego wykazania i łatwo się go replikuje, jego wyjaśnienie jest przedmiotem wielu teoretycznych sporów. Jest wysoce pewne, że nasze obecne rozumienie natury efektu Stroopa nie jest ostateczne i ulegnie w przyszłości zmianie. Nie sądzę natomiast, że samo występowanie efektu Stroopa zostanie kiedykolwiek zakwestionowane. O jego systematycznej powtarzalności przekonałem się wielokrotnie prowadząc zajęcia dydaktyczne. Przyszłe pokolenia badaczy raczej nie zakwestionują samego efektu, choć mogą zakwestionować jego wyjaśnienie. Oczywiście zademonstrowanie efektu Stroopa też nie jest „czystą obserwacją”, tym niemniej jego występowanie można sprowadzić do takiego opisu, w którym nie będziemy musieli odwoływać się do żadnych teorii ani pojęć „wyższego”

⁴⁵ Geniusz tego badania polega na tym, że wykorzystali sytuację, na którą badani sami się decydowali, uniknięto więc wielu dylematów etycznych.

rzędu typu kontrola poznawcza, hamowanie itd. (oto jak taki opis mógłby brzmieć: „jeśli poprosimy badanych o nazywanie kolorów, jakimi są napisane słowa, wówczas w wypadku słów oznaczających kolory, badani będą wykonywali zadanie średnio dłużej, jeśli bodźce są niespójne [słowo oznacza inny kolor niż czcionka, jakim jest napisane], niż jeśli bodźce będą spójne, lub neutralne [np. kolorowy pasek]”).

Naszej nauce potrzeba jak najwięcej „efektów Stroopa”, a więc efektów, które są rzetelne, łatwo powtarzalne i w miarę możliwości silne. Ich brak zakotwiczenia w złożonych paradygmatach teoretycznych może być paradoksalnie zaletą, gdyż efekty takie mogą stanowić bazę dla rozwoju przyszłych teorii i metateorii, natomiast one same nie będą zależne od sukcesu określonych teorii wyższego rzędu. Jestem przekonany, że badania oparte o zmienne behawioralne (tak jak zdefiniował je np. Furr, 2009) z większym prawdopodobieństwem mogą spełniać taką rolę niż badania z wykorzystaniem pośrednich wskaźników zachowania lub nadmiernie skomplikowanych procedur. Oczywiście badania behawioralne nie są absolutną „prawdą objawioną” i mogą zawierać jakieś niezidentyfikowane artefakty, być wynikiem przypadku czy oszustwa. Niemniej opisywane w rdz. 2.2 efekty „behawioralne”, gdyby zostały systematycznie potwierdzone, wydaje się, że mogą być swobodnie zaadoptowane przez szereg teorii wyższego rzędu. Co równie ważne, obecne i przyszłe teorie wyższego rzędu będą musiały jakoś uwzględniać te obserwacje. Przykładowo, niezależnie od tego, czy przyjmiemy tezę o biologicznej, społeczno-kulturowej lub mieszanej determinacji różnic międzypłciowych w podejmowaniu ryzyka, udana teoria będzie musiała uwzględniać fakt wyższej ryzykowności zachowań rowerzystów u mężczyzn i jakoś go objaśniać.

Jeszcze raz podkreślę, że nie chciałbym, żeby moja obrona metod behawioralnych była odbierana jako cudowne panaceum na wszelkie możliwe ograniczenia metody naukowej. Nawet, gdy mamy do czynienia z realistycznym eksperymentem terenowym, mogą się zdarzyć artefakty, przypadek wypaczający wynik, lub wyniki będą wysoce niereprezentatywne i będą mówiły nam niewiele o tym, jak postępuje typowy człowiek (np. określone miejsca w mieście przyciągają osoby o określonych cechach dyspozycyjnych; okres historyczny, w którym przeprowadza się eksperyment ma jakieś szczególne cechy, które wpływają na wyniki; niekontrolowane przez badacza cechy miejsca badania wywierają wpływ na badany efekt itd.). W

naucze takiej jak psychologia zbiór potencjalnych moderatorów jest trudny, a wręcz niemożliwy, do ostatecznego ustalenia.

Podkreślę raz jeszcze, że dane behawioralne, albo aplikacja określonych efektów do codziennych kontekstów może nie być w ogóle interesująca dla niektórych obszarów badań. Natura niektórych odkryć jest taka, że opisują jakieś elementarne prawidłowości dotyczące umysłu (np. efekt Stroopa) i jako takie są ciekawe same z siebie i nie wymagają natychmiastowego przełożenia do świata codziennego zachowania. W tym zresztą przykładzie efekt Stroopa jest sam w sobie czynnością skrajnie nienaturalną (czytanie koloru czcionki, zamiast słowa) i trudno byłoby wprost przełożyć go na sferę codziennego doświadczenia. W żaden sposób nie przekreśla to jego ważkiego znaczenia dla rozumienia elementarnych procesów poznawczych.

Oczywiście, metody behawioralne nie są też ratunkiem przed fałszerstwem, lub psychologią fałszywie pozytywną, co pokazał zarówno Stapel w swoich bardzo „behawioralnych” badaniach w *Science* (Stapel, Lindenberg, 2011), jak i Sanna (Sanna, Chang, Miceli, Lundberg, 2011). Gdyby nie fałszywość tych badań, mogłyby spokojnie zostać umieszczone w rdz. 2.2. jako ciekawe przykłady badań behawioralnych.

Tak więc moim celem nie jest tu apel „porzućmy całkowicie nierealistyczne warunki i nie-behawioralne miary”. W psychologii jest miejsce dla tego rodzaju badań i pełnią swoje role poznawcze. Wydaje się natomiast, że dużo większy udział miar bardziej behawioralnych i bardziej realistycznych warunków jest czymś pożądanym, szczególnie w psychologii społecznej, gdzie – w olbrzymiej większości badań – opisywane efekty mają dosyć oczywiste implikacje dla codziennych kontekstów (przykłady zaprezentuję w rdz. 2.8.).

Jeśli przyjmiemy założenie, które już zarysowałem wcześniej, że w psychologii tym co nas ostatecznie interesuje jest na ogół zachowanie, to powinniśmy ostatecznie dążyć do danych behawioralnych. Gdyby jednak było tak, że istnieje bezpośrednie, wysoce trafne przełożenie między zachowaniem, a pośrednimi wskaźnikami zachowania, w takim wypadku moglibyśmy zrezygnować z danych behawioralnych na rzecz takich typów danych, które są najłatwiejsze do uzyskania (a więc jeśli ludzie rzeczywiście prawie zawsze robiliby to, co deklarowali, to uzyskanie deklaracji jest łatwiejsze, a przy tym nie zmniejsza naszej pewności). Sęk w tym, że istnieje w

psychologii bogata tradycja kwestionowania trafności metod samoopisowych. Od pewnego czasu pojawia się coraz więcej głosów kwestionujących trafność miar utajonych (np. testu utajonych skojarzeń) jako wyznaczników ludzkiego zachowania. Dwie następne rdz. poświęcam na omówienie najważniejszych prac w tym obszarze. Mam nadzieję, że będą one dostatecznie przekonywały, że niestety nie można w psychologii w drodze do „zachowania” iść na skróty.

2.4. Ograniczenia trafności metod samoopisowych

Znaczna część psychologii jest oparta na danych samoopisowych w tej czy innej postaci. Są stosunkowo łatwe do uzyskania, zwykle również dostarczają mniejszej liczby dylematów etycznych niż inne rodzaje danych (co nie znaczy, że żadnych). Tradycja ich używania jest prawie tak stara jak sama psychologia, niemała ich część (szczególnie kwestionariusze) jest oparta na złożonych teoriach psychometrycznych, również nakierowanych na przeciwdziałanie oszukiwaniu w badaniach (np. Ziegler, MacCann, Roberts, 2011). Moim celem nie jest kwestionowanie roli danych samoopisowych w psychologii. Chciałbym jednak przedstawić najważniejsze badania wskazujące na to, że do takich danych należy podchodzić zawsze z pewną zdrową dozą sceptycyzmu i opieranie rozwoju teorii psychologicznych tylko na tego rodzaju danych jest zdecydowanie niedostateczne. Zaznaczę także, że w tej rdz. nie czynię rozróżnienia między różnymi typami danych samoopisowych (takich jak postawy, opisy przeszłych zachowań, deklarowanie hipotetycznych zachowań, wyniki ustrukturyzowanych kwestionariuszy itd.). Nie wykluczam, że niektóre z nich mogą być nieco bardziej trafne niż inne, jednak zagadnienie to wykracza poza zakres tej pracy.

Chciałbym zacząć z pewnego „meta-poziomu”. Od pewnego czasu są rozwijane teorie, wskazujące na to, że „prawdziwa” samowiedza nie zawsze jest celem wartościowym dla naszego umysłu (Henriques, 2003; Bering, Shackelford, 2004; Von Hippel, W., Trivers, 2011; Kurzban, 2010). Wymagania poznawcze życia w grupie są istotnym motorem ewolucji poznawczej. Dunbar wykazał, że stosunek kory nowej do reszty mózgu jest bardzo wysoko skorelowany ze wielkością grupy społecznej ($r = 0,76$; Dunbar, 1993). Obecność wielu innych osobników własnego gatunku, z którymi wchodzi się w długotrwałe złożone interakcje, wiąże się z licznymi wymaganiami dla

umysłu. Trzeba między innymi wiedzieć, kto jest członkiem grupy, na kim można polegać, z kim wcześniej się dobrze współpracowało, jaka jest relacja między innymi członkami grupy, w jaki sposób się zachować, żeby wywrzeć określone wrażenie itd. Dokonując w swoim życiu mnóstwo tego rodzaju „obliczeń” społecznych w sposób zupełnie „instynktowny”, nie zdajemy sobie na ogół sprawy z tego, jakie bardzo są wymagające dla systemu poznawczego. Pojawienie się języka naturalnego jeszcze bardziej skomplikowało sytuację. Znaczenia nabrały nie tylko informacje rejestrowane bezpośrednią percepcją, ale też informacje przekazywane werbalnie. Można było uzyskiwać więcej informacji o innych, pojawiły się też możliwości manipulowania tymi informacjami, żeby zdyskredytować rywali, lub polepszyć swój wizerunek. Ma to swoje niebanalne implikacje. Nawet dziś olbrzymia większość informacji jakie ze sobą wymieniamy w codziennych rozmowach to plotki (por. Dunbar, Marriott, Duncan, 1997). Zdaniem niektórych badaczy wymusiło to poznawczy „wyścig zbrojeń”. Rozwijały się te umiejętności, które pozwalają wpływać na stany mentalne innych poprzez mowę. Zdaniem Henriquesa (2003) uzasadnianie własnego zachowania było kamieniem milowym ewolucji poznawczej samoświadomości (przedstawiłem jego teorię w: Budzicz, 2010).

W rzeczy samej wiele eksperymentów psychologii społecznej zaświadcza o tym, że mówiąc o sobie rozmijamy się z prawdą, oględnie rzecz ujmując. Klasyczna praca w tym obszarze nosi wymowny tytuł „Totalitarne Ego. Fabrykowanie i zmienianie historii osobistej” (Greenwald 1980) i omawia szereg dobrze udokumentowanych efektów. Przykładowo te efekty to wiara w większą moc sprawczą niż rzeczywiście posiadana albo przypisywanie sukcesów dyspozycjom, a porażek niefortunnym zbiegom okoliczności. Większość z nas uważa, że jest ponadprzeciętnie dobrymi kierowcami, liderami, że ma wyższą inteligencję niż średnia w populacji, jest bardziej empatyczna, kieruje się w swoich sądach obiektywnymi danymi (gdy inni kierują się emocjami lub ideologią) itd. W przykładowym badaniu 68 % pracowników naukowych z jednego uniwersytetu uznało, że należy do najlepszej ćwiartki dydaktyków na tej uczelni (Cross, 1977). Popularne omówienie różnorodnych badań zawiera książka wymownym tytule „Błądzą wszyscy, ale nie ja” (Tavris, Aronson, 2008).

Co jednak istotne, ludzie nie tylko mówią o sobie nieprawdziwe rzeczy, ale też w nie na ogół autentycznie wierzą. Dlatego biolog ewolucyjny Robert Trivers poszedł krok dalej i rozwinął teorię ewolucji samooszukiwania (Von Hippel, Trivers, 2011). W

dużym uproszczeniu zakłada ona, że rzeczywiste okłamywanie innych jest poznawczo wymagające. Trzeba rejestrować nie tylko „prawdę”, ale i utrzymywać w pamięci informacje, kogo się okłamało i w jaki sposób. Prostsze byłoby w takim kontekście posiadanie od razu zniekształconego obrazu rzeczywistości w głowie, bez potrzeby świadomego „okłamywania”.

Idee te, w szerszej skali społeczeństwa, rozwija Robert Kurzban w książce „Dlaczego każdy (inny) jest hipokrytą” (wydanej też po polsku). Z punktu widzenia interesu jednostek opłacalne jest propagowanie pewnych postaw, a w rzeczywistości zachowywanie się zupełnie inaczej. Klasyczny przykład to konserwatyzm obyczajowy: w kategoriach adaptacyjnych dla mężczyzn jest często opłacalnie pochwalanie konserwatyzmu oraz przedstawianie siebie jako osoby obyczajowo konserwatywnej. Można narzucać żonom, żeby same były konserwatywne (tym samym zmniejszać ryzyko niepewności ojcostwa), oraz przekonywać je, że samemu jest się pewnym długotrwałym partnerem. Równocześnie przy sprzyjających okazjach można zachowywać się w sposób wysoce „niekonserwatywny”. Liczne przypadki konserwatywnych polityków przyłapanych na zdradach małżeńskich zdają się wspierać taką intuicję. Przy czym wkładem Triversa i Kurzban jest wskazanie, że tacy mężczyźni często są rzeczywiście wysoce przekonani o swoim konserwatyzmie obyczajowym. Ewentualne skoki w bok mogą racjonalizować lub wypierać z pamięci.

Wkład powyższych badaczy polega na wskazaniu i uzasadnieniu tego, że trafna i prawdziwa wiedza o sobie samym niekoniecznie ma jakąś wartość adaptacyjną, a szczególnie nie musi się pojawiać w kontekście przedstawiania siebie w toku interakcji społecznych. Oczywiście mogą istnieć również wielorakie ograniczenia związane z ewolucją wydajnych systemów pamięciowych. Ograniczenia ludzkiej pamięci w ogóle są szeroko opisane w literaturze i nie będę tutaj ich omawiał (np. Schachter, 2003; Loftus, 2005).

Przejdę natomiast do przykładów konkretnych badań, które potwierdzają to, że nasza samowiedza niejednokrotnie rozmija się z rzeczywistością. Niewątpliwie najśłynniejszym artykułem samoopisowych sceptyków jest tekst Nisbetta i Wilsona (1977) z *Psychological Review*⁴⁶ (ponad 9000 cytowań, a przy tym, wydaje się, że

⁴⁶ Artykuł na pewno się nie zdezaktualizował, pomimo tego, że liczy już ponad 4 dekady. Nowsza książka Wilsona rozwija te idee i jest zatytułowana znamienne: „Strangers to ourselves”.

popularność nie maleje, ale rośnie, np. tylko w latach 2009-2015 zanotowano około 3100 cytowań).

Autorzy zaprezentowali szereg badań, w których pokazali, że niejednokrotnie osoby badane mają bardzo znikome pojęcie o swoich procesach mentalnych. W szczególności są nieświadomi bodźców, które w istotny sposób wpływają na ich zachowanie, nie są świadomi, że ich reakcje powstają w odpowiedzi na dane bodźce (nawet jeśli są świadome występowania danych bodźców) lub też przypisują znaczenie bodźcom, które nie mają żadnego znaczenia.

Przykład pierwszej sytuacji, to badanie, w których osoby miały dokonać „decyzji konsumenckiej”, wybierając jeden z czterech przedmiotów rozłożonych przed ich oczami. Były one praktycznie identycznie. Stwierdzono bardzo silne preferowanie artykułu najbardziej po prawej (wybierany cztery razy częściej niż ten najbardziej po lewej). Badani pytani, czemu wybrali określony artykuł, dawali różne odpowiedzi, ale nikt nie wskazał na położenie (a były to, dla przypomnienia, praktycznie identyczne produkty).

Sytuacja braku świadomości wpływu bodźca (pomimo świadomości samego bodźca), była pokazana w innym badaniu, w którym uczono osoby par słów. Część osób, uczyła się par słów, które generowały bardzo konkretne skojarzenia, tak jak w sytuacji „ocean-księżyc” (skojarzenie: przyływ). Następnie proszono osoby o wygenerowanie różnych nazw, np. nazwy dla detergentu. Osoby, które uczyły się pary „ocean-księżyc” dużo częściej wymyślały „Przyływ”. Choć badani pamiętali wcześniej wyuczone słowa, to jednak nie wskazywali spontanicznie na nie jako na przyczynę generowanych skojarzeń, ale wymyślali inne przyczyny (np. „Przyływ to dobra nazwa dla detergentu”).

Badani też czasami przypisują znaczenie bodźcom, które nie mają znaczenia. W jednym z badań Nisbetta i Wilsona badani oglądali film, który w jednym warunku miał słabą ostrość, w innym warunku na korytarzu pracowały głośno maszyny elektryczne, a w jeszcze innym film był oglądany bez żadnych dystraktorów. Badani odpowiadali m.in. jak interesujący był film, albo jak sympatyzowali z głównym bohaterem. Pytano następnie o to, w jaki sposób hałas lub ostrość mogły wpływać na ocenę. Zdaniem badanych wpływały, choć okazało się, że tak naprawdę nie miały żadnego wpływu. W zbliżonym badaniu wycięcie fragmentów opowiadania też nie

wpływało na oceny, choć badani, którym zaprezentowano potem te wycięte części twierdzili post factum, że gdyby te fragmenty były, ich oceny by się zmieniły.

(Oczywiście, nie możemy wiedzieć na pewno w przypadku pojedynczych osób, że ich oceny by się nie zmieniły. Tym niemniej średnie wyniki grup wskazywały na to, że określone bodźce nie miały wpływu na średnie wyniki grupy).

Najogólniejszy wniosek z tych badań jest taki, że instynktownie poszukujemy wyjaśnień naszego zachowania, szczególnie na podstawie poznawczo najłatwiej dostępnych informacji. Te wyjaśnienia mogą być trafne, ale mogą też dramatycznie rozmijać się z rzeczywistością. Wydaje się, że w codziennym życiu faktycznie rzadko spotykamy sytuację, w której osoby mówią „nie wiem, dlaczego tak się zachowałem”. Bardziej naturalną sytuacją jest uzasadnianie swojego zachowania, z równoczesnym przekonaniem o prawdziwości tych uzasadnień. Wyniki tych eksperymentów dobrze uzupełniają wnioski z klasycznych już badań Gazzanigi (omówienie np. Gazzaniga, 2005) nad osobami z przeciętym ciałem modelowym. Gdy eksponować takim osobom do prawej „niejęzykowej” półkuli mózgu proste polecenia w rodzaju „wstań”, osoby je wykonują. Gdy jednak zapytać, dlaczego robią, to co robią, generują spontanicznie rozmaite wytłumaczenia o prawdziwości których byli głęboko przekonani (np. „wstałem, bo chciałem rozprostować nogi”), ale które nie miały nic wspólnego z rzeczywistym powodem (czyli bodźcem eksponowany do prawej półkuli mózgu). Mówiąc prościej: „językowa”⁴⁷ półkula część naszego mózgu jest przyzwyczajona do tego, żeby generować wyjaśnienie zachowania, nawet wtedy, gdy do rzeczywistego powodu nie ma dostępu.

Przykładowo, na ograniczenia ludzkiej samowiedzy wskazuje szeroki program badań w paradygmacie „afektywnego prognozowania” (ang. *affective forecasting*) pod kierunkiem Daniela Gilberta i Timotheo Wilsona (przegląd: Wilson, Gilbert, 2003). Ogólny wniosek z tych badań jest taki, że ludzie mają tendencje do przeceniania siły emocji pod wpływem pozytywnych i, szczególnie, negatywnych zdarzeń. W sztandarowym badaniu tego zespołu, rekrutowano ludzi w bardzo różnych sytuacjach życiowych, którzy przewidywali siłę swoich emocji pod wpływem takich wydarzeń jak zerwanie związku, niedostanie awansu, przegrana kandydata w wyborach i oceny te zestawiano z szacunkami własnych emocji tych samych (lub innych osób), gdy

⁴⁷ Dodam dla uzupełnienia, że rozróżnienie na procesy prawo- i lewopółkulowe jest mocno medialnie przereklamowane por. Geake, 2008

rzeczone wydarzenia miały później miejsce. Osoby w 6 niezależnych badaniach wyraźnie przeceniały siłę negatywnych emocji i, w mniejszym stopniu, siłę pozytywnych emocji relatywnie do emocji rzeczywiście później przeżywanych.

Określone, łatwo dostępne poznawczo, uzasadnienia mogą zasadniczo zmieniać nasze spostrzeganie rzeczywistości i nas samych. W jednym z najbardziej znanych takich badań, Strack, Martin i Schwarz (1988) pytali osoby badane o częstotliwość chodzenia na randki i poczucie subiektywnego dobrostanu. Okazało się, że częste chodzenie na randki bardzo sprzyja dobrostanowi. Korelacja jest, jak na badanie psychologiczne, bardzo wysoka (około $r = 0,6$). Odkrycie to było jak najbardziej zgodne z intuicją i na tym można by sprawę zamknąć. Jednak okazało się, że gdy zmieniono kolejność pytań, tj. najpierw pytano o poczucie subiektywnego dobrostanu, a potem o częstotliwość randek, korelacja spadła praktycznie do zera (dokładnie $r = -0,1$; nieistotna). Podobny efekt uzyskano, gdy badano związek zadowolenia z małżeństwa i zadowolenia z życia. W zależności od kolejności pytań mogłaby być to znikoma zależność ($r = 0,18$) lub niezwykle silna ($r = 0,67$) (Schwarz, Strack, & Mai, 1991).

Wtrącę małą dygresję w tym miejscu. Niejednokrotnie miewam wrażenie, że w artykułach empirycznych analiza potencjalnego „zanieczyszczania” zmiennych samoopisowych przez inne pytania i/lub sytuacje eksperymentalne jest w niewielkim stopniu podnoszona przez samych badaczy. A wspomniana przed chwilą łatwość, z jaką badani potrafią zmieniać swoją opinię musi zmuszać do takich analiz.

Przykładowo Strelau i Zawadzki (2005) badali związki między cechami temperamentu (wyodrębnionymi rzecz jasna w Regulacyjnej Teorii Temperamentu Strelaua), a nasileniem objawów zaburzenia stresu pourazowego (PTSD) u ofiar powodzi i katastrof górniczych. W 5 niezależnych próbach cechy temperamentu wyjaśniały od kilkunastu do kilkudziesięciu procent zmienności PTSD (często więcej niż siła traumatycznego wydarzenia). Wyniki te uzyskano na dosyć dużych próbach (łącznie przebadano 1093 osoby), ale zmienne miały charakter wyłącznie samoopisowy, co zmusza do zastanowienia, czy jakaś część tego związku nie jest zwiększona podobną treścią pytań oraz wpływaniem treści jednych pytań na odpowiadanie na inne. Kwestionariusz FCZ-KT zawiera pytania o m.in. uporczywe wracanie do przeszłych wydarzeń (w skali perseweratywności). Jedną z cech PTSD jest dokładnie to samo. Nawet, jeśli natura tych zachowań jest każdorazowo inna

(indukowana dyspozycją lub sytuacyjnie), to jednak nie można wykluczyć interpretacji, że odpowiadania na pytania kwestionariusza FCZ-KT wpłynęło na odpowiadanie na pytania kwestionariusza mierzącego PTSD (lub na odwrót, gdyż kolejność w jakiej badani wypełniali narzędzia nie została określona).

Wracając do problemu trafności badań samoopisowych. Szereg innych czynników zmieniających sposób odpowiadania badanych, czynników merytorycznie niezwiązanych z główną hipotezą, zawiera przegląd Schwarza (1999). Wymienia on m.in.

- wpływ skrajnych wartości danej skali na odpowiadanie: jeśli pytanie o to, jak udane badani mają życie było w skali $<0,10>$ wówczas 13 % wskazywała maksymalną wartość. Jeśli jednak skrajne wartości były opisane jako $<-5,5>$ wówczas maksymalna wartość była wybierana przez 34 % osób. W innym wypadku, gdy pytano osoby o rozmaite dolegliwości psychosomatyczne inne były odpowiedzi, jeśli skala rozciągała się od punktów „dwa razy w miesiącu lub rzadziej-kilka razy dziennie lub częściej”, a inne jeśli rozciągała się od punktów „nigdy- dwa razy w miesiącu lub częściej” (w pierwszym przypadku 62 % wybierało odpowiedź dwa razy w miesiącu lub częściej, w drugim wypadku analogiczny odsetek wynosił 39 %).
- różnica między wyborem odpowiedzi a spontanicznym generowaniem odpowiedzi: jeśli badanych proszono o wybór z listy „do jakiej najważniejszej rzeczy musisz przygotować swoje dziecko” 61 % wybierało odpowiedź „żeby było autonomiczne w swoim myśleniu”, ale tylko 4 % osób tak odpowiadało spontanicznie, gdy zadawano pytanie otwarte;
- odpowiadanie na pytania o których nie ma się kompletnie pojęcia: 30 % osób udzieliło odpowiedzi na pytanie o swój stosunek do fikcyjnej „Ustawy o handlu artykułami rolniczymi uchwalonej w 1978 roku”;
- spełnianie oczekiwań badacza: badanych pytano o wyjaśnienie przyczyn zjawiska masowych morderstw. Badacz przedstawiał się jako przedstawiciel „Instytutu Badań nad Osobowością” lub „Instytutu Badań Społecznych”. W zależności od „afiliacji” badani bardziej akcentowali uwarunkowania dyspozycyjne lub społeczne.

Z nowszych badań samoopisowych sceptyków bardzo duże wrażenie robi pomysłowy eksperyment Halla, Johansson i Strandberga (2012). Badanym rozdano kwestionariusz, w którym ustosunkowywali się do różnych twierdzeń dotyczących kwestii światopoglądowych (np. „Przeglądanie Internetu i maili na szeroką skalę przez służby wywiadowcze powinno być dozwolone jako metoda walki z międzynarodową przestępczością i terroryzmem”, inne pytania dotyczyły m. in. konfliktu Izrael-Palestyna, legalności prostytucji czy problemu uniwersalności/relatywności zasad moralnych). Badani ustosunkowywali się do tych twierdzeń na skali 1-9. Badany przewracał kartkę, żeby przejść do następnej strony. Dzięki specjalnemu papierowi oryginalne pytania odklejały się i zostawały twierdzenia przeciwne. Więc w powyższym przykładzie byłoby to „Przeglądanie Internetu i maili na szeroką skalę przez służby wywiadowcze nie powinno być dozwolone jako metoda walki z międzynarodową przestępczością i terroryzmem”. Badani następnie wracali do wcześniejszych pytań i proszono ich o uzasadnienie ich odpowiedzi. Tak więc jeśli ktoś „bardzo zgadzał” się z określonym twierdzeniem, nagle był konfrontowany z faktem, że popierał twierdzenie zupełnie przeciwne. Część badanych domyśliła się, że coś jest nie w porządku (41 %) lub stwierdziła, że musieli się pomylić przy zaznaczaniu (4 %). Jednak, pozostali, a więc ponad 50 % osób, jak gdyby nigdy nic, zaczęły uzasadniać swoje odpowiedzi. I to pomimo tego, że chwilę wcześniej odpowiadali zupełnie inaczej!

Warto też wspomnieć o wynikach meta-analiz dotyczących związku postaw i zachowań. Wg mojej wiedzy, żadna z takich meta-analiz nie stwierdziła, że związek ten jest zerowy. Z drugiej jednak strony siła zależność była zwykle taka, że postawy wyjaśniały dużo mniej niż połowę wariancji. I tak badacze szacują siłę takiego związku na: $r = 0,38$ (Krauss, 1995); $r = 0,27$ (Armitage, Conner, 2001) lub około $r = 0,41$ (Wallace, Paulson, Lord, Bond, 2005). Często cytowany przegląd (Ajzen, Fishbein, 1977) nie zawiera jednej wyliczonej wartości związku, ale stwierdza, że w 32 % analizowanych badań nie znaleziono istotnych zależności między postawami a zachowaniem, a w dalszych 40 % zależność ta była słaba ($r < 0,4$). W meta-analizie dotyczącej związku między postawą względem używania prezerwatyw, a rzeczywistym zachowaniem również stwierdzono umiarkowany związek ($r = 0,45$; Albarracin, Johnson, Fishbein, Muellerleile, 2001). W tym przeglądzie wszystkie zmienne samoopisowe łącznie wyjaśniały 30 % wariancji zachowania. Identyczną

ilość wariancji przewidywały zmienne samoopisowe w domenie przewidywania przyszłej aktywności fizycznej (Hagger, Chatzisarantis, Biddle, 2002).

Badania w różnych domenach cechują się skrajnie różnymi wartościami związku zmienne samoopisowe-rzeczywiste zachowanie (Ajzen, 2000). Szczególnie istotnym moderatorem związku wydaje się być konkretność postawy (Ajzen, Fishbein, 1997) np. popieranie ekologiczności jest słabym predyktorem datków na Greenpeace. Poza tym stwierdza się, że związki między deklarowanymi postawami a zachowaniem są silniejsze wtedy, gdy postawy są poznawczo bardziej dostępne i wtedy, gdy badani są pewni swoich postaw, a dzieje się tak, gdy badani mają częsty kontakt z obiektem postawy i mają regularnie okazje, żeby przedstawić swoją postawę (Glasman, Albarracín, 2006). W badania eksperymentalnych, które nierzadko dotyczą kwestii mało związanych z życiem badanych, ten warunek niekoniecznie musi być spełniony. Podobnie trafność postaw zmniejszała odczuwana presja społeczna na to, żeby deklarować „pożądany” rodzaj postawy (Wallace, Paulson, Lord, Bond, 2005).

Mając świadomość ograniczeń metod samoopisowych warto zadać pytanie, co można robić, gdy ze względów praktycznych i etycznych, nie można badać zachowania w sposób inny niż samoopisowy (pierwsze przykłady z brzegu: agresja fizyczna, zachowania seksualne, nadużywanie substancji psychoaktywnych itd.). Można zdać się na metody samoopisowe, ale można też wykazać się ponadprzeciętną kreatywnością w pozyskiwaniu danych. Kuhle (2011) wykazał się taką kreatywnością w obszarze, który jest wysoce wrażliwy etycznie mianowicie dotyczącym zdrady w związku. Kuhle testował hipotezę o ewolucyjnie uwarunkowanych różnicach międzypłciowych w rodzajach zazdrości o romantycznego partnera. tj. u mężczyzn wyższa ma być zazdrość seksualna, u kobiet emocjonalna. Analizował treści osobliwego amerykańskiego show telewizyjnego, w którym osoba podejrzewająca partnera o zdradę zgłasza się do takiego programu, a producenci wysyłają „w teren” ekipę detektywów, która ma to zbadać. Jeżeli stwierdzi się rzeczywisty fakt zdrady, to w programie zdradzający i zdradzany są konfrontowani ze sobą. Kuhle przejrzał archiwalne nagrania programu i sprawdzał reakcję osób. Zdradzane kobiety wyraźnie częściej pytały o emocjonalne aspekty zdrady („czy ją kochasz?”), a zdradzani mężczyźni częściej o aspekty seksualne („czy przespałeś się z nim?”). Trudno ze względów etycznych wyobrazić sobie badanie, w którym osoby badane są konfrontowane z rzeczywistością lub

domniemaną zdradą partnera. Pomysłowa procedura Kuhle pozwoliła jednak w jakimś stopniu obejść te ograniczenia (dodam, że sama idea takiego programu jest skrajnie wątpliwa etycznie; jeśli jednak program został już zarejestrowany, to bilansu etycznego nie zmieni to, że badacz przeanalizuje jego treść i przedstawi wyniki w postaci zbiorczych danych).

Dane samoopisowe wymagają mniejszej kreatywności, przynajmniej jeśli chodzi o zaprojektowanie procedury. Udań procedury eksperymentalne czasami wymagają długotrwałego doskonalenia (przy czym czasami poszukiwanie „działającej” procedury jest rodzajem psychologii fałszywie pozytywnej), tymczasem w badaniu samoopisowym wystarczy badanym rozdać plik kwestionariuszy, lub wysłać link do ankiety internetowej. Oczywiście, bywa, że niebanalnej kreatywności wymaga stworzenie solidnego narzędzia badawczego.

Dla mnie osobiście dwa argumenty przeciwko opieraniu rozwoju psychologii tylko na metodach samoopisowych są najbardziej przekonujące. Jednym jest refleksja historyczna wskazująca na to, że najważniejsze badania w historii naszej nauki nie miały takiej natury (Hock, 2003; Wertheimer, 2012; Hergenhahn, Henley, 2013). Drugim argumentem jest przekonanie, że człowiek ma introspekcyjny dostęp do informacji od zarania cywilizacji, a nawet wcześniej. Mimo to wydaje się, że w ostatnim stuleciu nastąpił niemały skok w rozumieniu natury ludzkiej w porównaniu do wcześniejszych stuleci. Zapewne daleko nam jeszcze do pełnego zrozumienia natury ludzkiej, ale trudno nie oprzeć się wrażeniu, że próby zrozumienia zachowania człowieka, poprzez pytanie o jego zachowanie nie jest najlepszą strategią.

Opracowań i badań sceptyków samoopisów jest znacznie więcej (Ross, Conway, 1986; Menon, 1993; Piedmont, McCrae, Riemann, Angleitner, 2000; Robinson, Clore, 2002; Fan i in., 2006; Nave, Sherman, Funder, 2008) i wyczerpujące ich omówienie wymagałoby wielokrotnie więcej miejsca. Teza, jaką chciałem tu dowieść, jest taka, że samoopisy nie wyjaśniają 100 % rzeczywistego zachowania i należy zachować wobec nich ostrożny sceptycyzm, a tam, gdzie to możliwe należy badać zachowanie sensu stricto. Nie twierdzę równocześnie, że samoopisy mają zerową trafność, lub nie należy ich nigdy stosować (szczególnie, gdy ze względów praktycznych lub etycznych nie ma lepszych rozwiązań). Wydaje się jednak, że teza ta nie jest specjalnie kontrowersyjna, dlatego nie będę jej dalej rozwijał.

2.5. Ograniczenia trafności metod mierzących zmienne utajone

Osobną kategorię stanowią techniki badające utajone postawy lub dostępność określonych schematów poznawczych, czyli baumeisterowskie „ruchy palców”. Będę je dalej nazywał miarami utajonymi (ang. *implicit measures*). Nie są one tak znane jak metody samoopisowe więc pozwolę sobie omówić je nieco dokładniej i podać różne przykłady ich zastosowań, żeby pokazać, że występują wcale nierzadko w psychologii (choć oczywiście dużo rzadziej niż metody samoopisowe czyli miary „jawne”).

Techniki te w zdecydowanej większości opierają się na reagowaniu w odpowiedni sposób na bodźce (najczęściej słowne). Określony wzór czasów reakcji wyznacza poziom hipotetycznej postawy i/lub dostępności określonych schematów w umyśle (w tym najpopularniejsze, a więc test decyzji leksykalnych, emocjonalny test Stroopa, test utajonych skojarzeń, test afektywnego primingu). Wyjątkiem jest technika uzupełniania słów, gdzie badani uzupełniają rebusy słowne. Przykładowo osoba musi znaleźć litery do rebusu „s_x”, tak żeby powstał sensowny wyraz. Może uzupełnić je w języku angielskim na dwa sposoby: „sex” lub „six”. W tym przykładzie, jeśli badany rozwiązuje szereg rebusów wskazując słowa związane z seksualnością zamiast neutralnych, wskazuje to jakoby na wyższą poznawczą dostępność schematów seksualnych i/lub wyższą motywację seksualną. W jednym z badań (opublikowanych w JPSP) mężczyźni wachając rzeczy noszone przez owulujące kobiety częściej rozwiązywali zadanie w ten sposób niż gdy wachali takie same rzeczy noszone przez nieowulujące kobiety, albo zapachy całkowicie neutralne (Miller, Maner, 2011; badanie 1). Co ciekawe, te – dosyć oryginalne – bodźce w badaniu Millera i Manera w niewielkim stopniu przekładały się na jawnie deklarowany poziom pobudzenia seksualnego.

Jednak, jak wspominałem, prawie wszystkie miary utajone oparte są o czas reakcji. Bodaj najpopularniejszą z takich technik jest test utajonych skojarzeń (Implicit Association Test – IAT). Do jej stosowania wystarcza zwykły komputer, z pomocą łącza internetowego badania można też przeprowadzać w domu badanych. Jak wskazują twórcy tej metody dzięki dedykowanym stronom internetowym przebadano ponad milion osób (za: Blanton, Jaccard, 2006).

W technice tej badany uczy się reagować na przeciwstawne bodźce (np. na słowa przyjemne i nieprzyjemne; na atrakcyjne i nieatrakcyjne twarze itd.) jedną z dwóch rąk. W następnych blokach badani muszą jedną ręką reagować na dwie kategorie bodźców (np. na nieprzyjemne słowa i atrakcyjne twarze lewą ręką, a prawą na przyjemne słowa i atrakcyjne twarze, które są tym razem przemieszane w obrębie bloku). Zestawia się tak bodźce, żeby wzór czasów reakcji wskazywał na poziom utajonej postawy np. jeśli osoba ma pozytywne nastawienie do osób atrakcyjnych, będzie szybciej reagowała, gdy musi zareagować tą samą ręką na słowa przyjemne. Będzie reagowała wolniej, jeśli musi użyć tej samej ręki zarówno, żeby wskazać twarz atrakcyjną i bodźce nieprzyjemne (wyczerpujący opis stosowania tej techniki zawiera np. Lane, Banaji, Nosek, Greenwald, 2007).

Mniej znane, ale też szeroko obecne w literaturze narzędzia to test decyzji leksykalnych i emocjonalne zadanie Stroopa. Konceptualnie są dużo prostsze do wyjaśnienia. W teście decyzji leksykalnych badani muszą zdecydować, czy określony ciąg liter jest słowem ich macierzystego języka. Krótszy czas reakcji (decyzji leksykalnej) wskazuje na wyższą poznawczą dostępność określonego słowa, a więc w domyśle, większą dostępność schematu poznawczego. W emocjonalnym zadaniu Stroopa, podobnie jak w klasycznym zadaniu Stroopa, badani muszą nazywać kolor słów, jednak eksponowane słowa nie oznaczają kolorów (tak jak w „klasycznym” Stroopie), ale określone koncepty, schematy poznawcze itp. Te dwie metody omówiłem szerzej w innej publikacji (Budzicz, 2012b).

Czwarta najpopularniejsza (wg mojego rozeznania) metoda to tzw. priming afektywny. Zasadniczo badany musi określać walencję bodźca (np. zdecydować czy słowo „radość” oznacza coś pozytywnego czy negatywnego). Przed określaniem walencji bodźca pokazywany jest bodziec docelowy. Jeśli w badaniu tym chodzi o utajone postawy względem Afroamerykanów, to badani zobaczą twarze białe i czarne. W zamyśle, jeśli osoba jest uprzedzona do Afroamerykanów, zobaczywszy twarz czarną będzie dłużej kategoryzowała walencję bodźca pozytywnego, a walencję bodźca negatywnego (analogicznie, gdyby osoba była uprzedzona do osób białych).

W mojej wcześniejszej publikacji (Budzicz, 2012b) wymieniłem kilkanaście badań z najlepszych czasopism psychologii społecznej, gdzie stosowano test decyzji

leksykalnych i emocjonalne zadanie Stroopa. Prawdopodobnie nie jest to bardzo znana część psychologii, więc pokażę tutaj kilka przykładów ich stosowania.

Hafer (2000) wyświetlała badanym film opowiadający o brutalnej napaści na młodego człowieka. Manipulowała zakończeniem, część badanych dowiadywała się, że sprawca ucieka wymiarowi sprawiedliwości za granicę, a część dostawała informacje, że sprawca zostaje schwytany i osądzony. Następnie badanych poddawano emocjonalnemu zadaniu Stroopa. Badani w grupie „bez sprawiedliwości” około 20 ms wolniej nazywali kolor neutralnych słów (np. „sortować”, „poszerzyć”). Jednakże największe różnice (ponad 80 ms dłużej) miały miejsce w przypadku słów związanych ze sprawiedliwością (np. „uczciwy”, „nierówny”). W drugim badaniu o podobnej procedurze osoby badane dodatkowo wypełniały kwestionariusz w którym m.in. określały jak bardzo przypisywały winę ofierze przemocy. Wyniki umiarkowanie ($r = 0,27$) korelowały z wynikami interferencji słów związanych ze sprawiedliwością. Badanie to nie udzielało oczywiście odpowiedzi, która miara (samoocena vs. EST) lepiej przewidywałaby behawioralne reakcje (np. gotowość do niesienia pomocy) na niewinne ofiary przemocy i wypadków. Correia z zespołem (2007) stosując analogiczną procedurę zreplikowali te wyniki, a dodatkowo pokazali, że efekt wydłużenia czasu reakcji na słowa związane ze sprawiedliwością dotyczy tylko osób z tej samej grupy etnicznej (tj. w badaniu portugalscy studenci wykazywali tylko takie reakcje, jeśli ofiary były opisane jako Portugalczycy, ale nie wykazywali takich reakcji dla Cyganów).

W artykule Gollwitzer i Denzler⁴⁸ (2009) opisano eksperyment, w którym badani grali z drugim rzekomym badanym w grę ekonomiczną o rzeczywiste pieniądze. Na koniec gry fikcyjny partner dokonywał podziału wygranej, robił to jednak w sposób bardzo niesprawiedliwy (znaczną część wspólnie zarobionych pieniędzy zatrzymywał dla siebie). Następnie badani uczestniczyli w teście decyzji leksykalnych, w którym mierzono ich poznawczą aktywizację agresji poprzez leksykalne klasyfikowanie słów takich jak: „atak”, „nienawiść”, „przemoc”. Po tym zadaniu dawano badanym możliwość zemsty poprzez zaaplikowanie drugiemu badanemu żmudnego nudnego zadania. Ponad 30 % osób skorzystało z tej opcji. Na koniec ponownie przeprowadzono test decyzji leksykalnych z takimi samymi bodźcami. W przypadku

⁴⁸ Tytuł artykułu jest wysoce „medialny”: „Co czyni zemstę słodką? Widok cierpiącego sprawcy czy przekazanie wiadomości?”

„braku zemsty” poznawcza dostępność słów związanych z agresją nie zmieniła się. W przypadku „zemsty”, gdy druga osoba deklarowała, że rozumie swój błąd poznawcza dostępność tych słów zmniejszyła się, a więc hipotetyczna „agresywność” spadła, natomiast wzrastała, gdy osoba nie dostrzegała, że zadanie jest karą za jej niewłaściwe zachowanie. Podsumowując, wniosek z tych bardzo pomysłowych badań jest następujący: zemsta zmniejsza agresję, ale tylko, gdy osoba na której się mścimy rozumie dlaczego ją spotkała.

Opisane powyżej badanie ma niemały stopień realizmu i logistycznie jest dosyć skomplikowane. Jednak nie zawsze badacze tworzą tak realistyczne sytuacje. Przykładowo Mikulincer, Gillath, Shaver (2002) badali jak prymowanie słowami związanymi z zagrożeniem („porażka”, „separacja”) lub neutralnymi („czapka”) będzie wpływała na rozpoznawanie imion ważnych osób w życiu badanego (ang. *attachment figures*), imion krewnych, imion znanych osób i imion nieznanych osób (każda osoba badana zawczasu wymieniała imiona odpowiednich osób). Prymowanie polegało na bardzo krótkiej (podprogowej) ekspozycji słowa. Korzystano z testu decyzji leksykalnych, jak i emocjonalnego zadania Stroopa. Prymowanie słowami oznaczającymi „porażkę” i „separację” zwiększało poznawczą dostępność imion bliskich osób, ale nie miało wpływu na poznawczą dostępność imion krewnych, znanych osób i nieznanych osób. Wzór wyników był zgodny z oczekiwaniami dla obydwu technik. Choć wyniki są teoretycznie ciekawe, warto zauważyć, że procedura ta miała bardzo niski poziom realizmu. Dodatkowo posiadając dużą pracownię komputerową i studentów obowiązkowo uczestniczących w badaniu, można przebadać setki osób w ciągu jednego dnia.

Belles, Kunde, Neumann (2010) testowali hipotezę mówiącą, że kobiety wcześniej dojrzewające płciowo mają bardziej negatywne nastawienie do mężczyzn (częściowo uzasadnianą ewolucyjnie, a częściowo zjawiskiem społecznego uczenia). W kolejnych trzech badaniach użyli trzy testy mierzące utajone postawy: priming afektywny, test utajonych skojarzeń, oraz test ich autorstwa, polegający na szacowaniu odległości od źródła dźwięku (którym był męski lub damski głos). Kobiety wcześniej dojrzewające płciowo miały bardziej negatywną postawę względem mężczyzn mierzoną primingiem afektywnym i testem utajonych skojarzeń, oraz oceniały głosy męskie jako oddalone dalej. Były to, jak twierdzą autorzy, dosyć wyraźne efekty (korelacje na poziomie $r = 0,39-0,52$). Co interesujące, w żadnym z

trzech badań nie zapytano wprost kobiet o to, jaką mają postawę względem mężczyzn. Hipotezę można było testować przy pomocy bardziej pracochłonnych metod, choć konceptualnie bardzo prostych (np. umożliwiając kobietom interakcję z obcym mężczyzną i oceny ich reakcji przez sędziów kompetentnych).

Badań z wykorzystaniem testu decyzji leksykalnych są setki (zapytanie w PsycINFO o „IAT” albo „Implicit Association Test” daje 1600 wyników⁴⁹). Znaczna ich część dotyczy nasilenia hipotetycznych utajonych postaw w stosunku do określonych obiektów (na ogół grup społecznych) pod wpływem określonej manipulacji eksperymentalnej lub w zależności od określonych cech dyspozycyjnych. Czasami jednak tezy są teoretycznie znacznie bardziej interesujące. Rudman i Phelan (2010) wykorzystali wersję testu utajonych skojarzeń, który w zamyśle ma mierzyć nieświadome aspiracje poprzez zestawienie wzorów czasów reakcji na słowa w rodzaju „ja”, „mój”, „siebie” w kontekście słów oznaczających przywództwo lub uległość. Dodatkowo zastosowano „genderowy” IAT, gdzie badani rozpoznawali imiona żeńskie i męskie w kontekście słów oznaczających ciepło i władzę. Badano tylko kobiety, którym prezentowano opisy mężczyzn i kobiet zgodne ze stereotypem (np. mężczyzna prezes i kobieta nauczycielka w szkole) lub niezgodne (np. kobieta wzięty makler i mężczyzna pielęgniarz). Dodatkowa grupa kontrolna uczyła się informacji o zwierzętach. Prymowanie typowymi rolami zwiększało siłę nieświadomych postaw zgodnych ze stereotypem w „genderowym” IAT (tj. szybsze reagowanie na słowa związane z kobietami w kontekście ciepła, a wolniejsze władzy), nie miało natomiast wpływu na „aspiracyjny” IAT. Prymowanie nietypowymi rolami paradoksalnie zmniejszało „aspiracje” mierzone testem utajonych skojarzeń, ale tylko relatywnie do grupy kontrolnej, a nie do grupy typowej. Sukces paradoksalnie obniżał aspiracje.

Hipotezy i wyniki są teoretycznie bardzo ciekawe i ważne, a samo badanie technicznie poprawne ale zdaje się powielać wszystkie typowe grzechy badań w psychologii społecznej ostatnich lat: metody o nieznanej trafności, ogrom mierzonych zmiennych, skomplikowane analizy statystyczne, raczej niezrozumiałe dla laików, a koniec końców efekty o umiarkowanej sile. Niezrozumiałe jest też pominięcie w ogóle badania mężczyzn. Choć badanie jest opisane jako krążące wokół tematu aspiracji zawodowych kobiet, byłoby ciekawe sprawdzić, jakie efekty uzyskiwaliby mężczyźni

⁴⁹ Stan na grudzień 2014.

(a może byli badani, ale ich wyniki psuły obraz rzeczy, więc o nich nie wspomniano?). Ogólnie badaniu trudno jednak odmówić tego, że może być wskaźnikiem, że „coś jest na rzeczy” i nie można mieć pretensji do redaktorów, że je opublikowali. Zakładam jednak, że przy odrobienie pomysłowości można by stworzyć dużo prostszy eksperyment behawioralny z wykorzystaniem bardziej czytelnych zmiennych. Warto tu wspomnieć, że prowadzony przeze mnie zespół studentów próbował zreplikować niektóre wyniki (dotyczące wpływu prymowania stereotypem na utajone „aspiracje”) i nie uzyskał istotnych wyników.

Zaprezentowane tu przykłady badań mają obrazować, że część psychologii społecznej obficie korzysta z metod badających utajone zmienne. W artykułach czasami miary utajone były używane jako jeden z dowodów występowania określonych prawidłowości i prezentowano szereg dodatkowych danych z wykorzystaniem innych metod (np. Mikulincer, 1998; Schimel, Hayes, Williams, Jahrig, 2007; Kay, Jost, 2003). Często, niestety, do wsparcia jakiejś hipotezy miary utajone były jedynymi wykorzystanymi metodami (np. Kramer, Goldman, 2003; Mikulincer, Gillath, Shaver, 2002; Hafer, 2000; Correia, Vala, Aguiar, 2007; Gollwitzer, Denzler, 2009). Na podstawie lektury tych artykułów mogę stwierdzić, że teoretyczne uzasadnienie użycia takich technik jest na ogół minimalne. Zwykle polega na stwierdzeniu, że „metoda X to często wykorzystywana technika w badaniach” i tu następuje kilka odnośników do literatury.

Zakładałbym, że popularność, jaką cieszą się metody badające utajone postawy i dostępność konstruktów, może wynikać z trafności wewnętrznej, jaką zdają się zapewniać (np. minimalna interakcja na linii badacz-badany), oraz z ich relatywnej łatwości i szybkości stosowania. Przeprowadzenie badania z ich użyciem nie wymaga interakcji z drugim człowiekiem, prawdopodobnie więc nie ma możliwości subtelного przekazywania oczekiwań przez eksperymentatora i ogranicza to zjawisko samospełniającego się proroctwa (oczywiście interakcja i potencjalne zakłócenia mogą być na etapie wcześniejszym). Wykorzystanie komputerów oznacza też, że można w relatywnie krótkim czasie przeprowadzić badanie znacznej liczby osób, szczególnie jeśli uczelnia posiada duże pracownie komputerowe (co jest raczej standardem w cywilizowanym świecie). Wyniki można w szybki zautomatyzowany sposób przetwarzać, więc mają tę przewagę nad ankietami papier-ołówek, że nikt nie musi wprowadzać danych ręcznie. Z kolei typowe badania

stricte behawioralne wymagają zaangażowania eksperymentatorów lub obserwatorów, więc liczba osób do przebadania na raz jest ograniczona. Dodatkowo, spekulowałbym, techniki te bardzo mogą sprzyjać wszelakim sposobom „łowienia” p: używa się bardzo wielu słów jako bodźców, które można zestawiać w interesujące „wskaźniki”, lub wybiórczo porównywać z innymi kategoriami słów.

Kluczowym pytaniem dotyczącym tych metod jest oczywiście pytanie o ich trafność i rzetelność (psychometryczną). Zaczniemy od rzetelności. Choć standardem w artykułach empirycznych jest podawanie rzetelności narzędzi kwestionariuszowych, to w badaniach z wykorzystaniem metod utajonych nie podaje się na ogół tych wskaźników. Matematycznie i konceptualnie nic nie stoi na przeszkodzie, żeby to robić. Przykładowo w emocjonalnym zadaniu Stroopa bada się czas reakcji na wiele słów związanych z określonym konstruktem (np. związane z agresją: złość, wściekłość, pobicie, nienawiść itp.) i w takim wypadku czasy reakcji na poszczególne słowa, przez analogię do kwestionariuszy, można potraktować jako „pozycje testowe”.

LeBel i Paunonen (2011) zebrali artykuły (por. tabela 2 w ich artykule), gdzie wyliczano rzetelność poszczególnych narzędzi. Na ogół nie jest ona bardzo wysoka. Przykładowo omawiany tutaj test primingu afektywnego ma trafność połówkową na poziomie $r < 0,2$, test uzupełniania słów $r < 0,3$, z kolei rzetelność testu decyzji leksykalnych wynosi $\alpha < 0,5$. Wg cytowanego tam artykułu (tj. Gawronski, Deutsch, Banse, 2011) rzetelność wewnętrzna testu IAT jest przyzwoita $\alpha = 0,6-0,9$. Inne teksty kwestionują jednak taki stan rzeczy. Przykładowo Rezaei (2011) przebadął korelację test-retest dwóch wersji tego zadania i uzyskał umiarkowany wynik $r = 0,53$ (jeśli badani nie mieli treningu w używaniu narzędzia wynosiła ona tylko $r = 0,32$). Co ciekawe korelacja była zbliżona niezależnie od tego, czy czas jaki upłynął między pierwszym a drugim pomiarem wynosił jeden tydzień czy 15 minut.

W większość przypadków rzetelność tych narzędzi jest niska lub umiarkowana. Poziom rzetelności ma istotne konsekwencje, szczególnie dla replikowalności wyników. Silny efekt, ale mierzony przy pomocy nierzetelnego narzędzia, ma małe szanse na replikację. LeBel i Paunonen (2011) przeprowadzili stosowne symulacje i stwierdzili, że w przypadku silnego efektu o sile 0,8 (d Cohena), dla grup 30 osobowych przy narzędziu o przyzwoitej rzetelności ($\alpha = 0,8$) prawdopodobieństwo

udanej replikacji wynosi około 80 %. Jeśli jednak rzetelność narzędzia wynosi 0,3, to przy tych samych pozostałych parametrach, szansa replikacji spada do 40 %.

W ostatecznym rozrachunku najważniejsze są pytania o trafność. Literatura na ten temat jest raczej ograniczona relatywnie do częstości ich stosowania. Znalazłem tylko 2 meta-analizy ilościowe dotyczące trafności IAT oraz jeden artykuł poświęcony trafności emocjonalnego zadania Stroopa w grupach klinicznych (mający charakter bardziej jakościowy). W przypadku emocjonalnego zadania Stroopa (dla grup nieklinicznych), testu decyzji leksykalnych oraz zadania primingu afektywnego nie znalazłem żadnej meta-analizy sprawdzającej ich trafność. Zaznaczę, że pracując nad artykułem (Budzicz, 2012b) przeczytałem bardzo dużo tekstów wykorzystujących te narzędzia i jest dla mnie skrajnie nieprawdopodobne, żebym mógł przeoczyć ewentualne meta-analizy.

W przeglądzie badań dotycząca trafności emocjonalnego zadania Stroopa w badaniu grup klinicznych (Williams, MathewsMacLeod, 1996) ogólny wniosek jest taki, że metody te mają pewną trafność, to znaczy osoby o określonych zaburzeniach w inny sposób przetwarzają słowa związane ze specyfiką ich zaburzenia (np. osoby cierpiące na arachnofobię – słowa związane z pajakami; osoby depresyjne – słowa związane z odrzuceniem itd.). Przegląd ten nie zawiera jednak kwantyfikacji ich trafności np. w postaci jakiejś miary wielkości efektu.

Meta-analizę trafności IAT przeprowadzili twórcy narzędzia (Greenwald, Poehlman, Uhlman, Banaij, 2009), a następnie niezależny zespół (Oswald, Mitchell, Blanton, Jaccard, Tetlock 2013). Obydwie opublikowano na łamach JPSP.

Nie będzie zaskoczeniem, że pierwsza meta-analiza wskazała na pewną trafność IAT w przewidywaniu zmiennej kryterialnej w rozmaitych obszarach. Średnia korelacja pomiędzy wynikiem IAT a ową zmienną wyniosła 0,27, i była zwykle nieco niższa niż korelacja między zmienną kryterialną a postawą mierzoną jawnie (średnia korelacja 0,36). Wskazano jednak, że IAT ma szczególnie wysoką trafność relatywnie do miar jawnych w obszarze uprzedzeń rasowych. Tutaj IAT lepiej przewidywał zmienną kryterialną niż miary jawne (korelacje odpowiednio 0,24 i 0,12). Lepiej radził też sobie w obszarze innych niż rasowe uprzedzeń międzygrupowych (np. etnicznych) (odpowiednio: 0,2 i 0,12). Gorzej przewidywał zmienne kryterialne w

obszarach zachowań konsumenckich, postaw politycznych, cech osobowości czy nadużywania alkoholu i narkotyków.

Średnia korelacja pomiędzy IAT a zmienną kryterialną na poziomie 0,27 to oczywiście nie jest silny efekt. Z drugiej jednak strony biorąc poprawkę na złożoność ludzkiego zachowania i wielkość szumu w badaniach psychologicznych, jest to wielkość, której nie można zignorować.

Oczywistym pytaniem jest, czym są owe „zmienne kryterialne”, z którymi korelowano jawnie deklarowane postawy oraz postawy utajone. Sprawdziłem wyrywkowo 5 artykułów z cytowanej meta-analizy. Przy doborze kierowałem się różnorodnością tematyczną, oraz przyzwoitą renomą czasopisma. Wydaje się, że ta niewielka próbka daje pewien obraz, czym jest owa zmienna kryterialna:

1. Obszar uprzedzeń rasowych: szybkość i dokładność kategoryzowania twarzy z niejednoznaczną rasą jako mających pochodzenie europejskie lub afroamerykańskie w zależności od ekspresji mimicznej (które wyrażały szczęście lub wściekłość) (Hugenberg, Bodenhausen, 2004);
2. Obszar uprzedzeń wobec osób otyłych: deklarowana chęć nawiązania kontaktów z przedstawioną otyłą lub nieotylą osobą (np. gotowość do wspólnej pracy nad projektem, wspólnej nauki, zostania przyjacielem takiej osoby) (Brochu, Morrison, 2007);
3. Obszar uprzedzeń płciowych: ocena kompetencji aplikujących do pracy kobiet i mężczyzn (badani wcielali się w rolę „rekrutera”) (Rudman, Glick, 2001);
4. Obszar uprzedzeń etniczno-rasowych: przejawianie (na podstawie deklaracji) różnych negatywnych zachowań względem członków innych grup etnicznych i rasowych (np. nieprzyjemne odzywki, wrogie gesty, niezapraszanie takich osób na spotkania towarzyskie itd.) (Rudman, Ashmore, 2007);
5. Obszar postaw konsumenckich: deklarowana gotowość do zakupienia określonych typów produktów (Vantomme, Geuens, De Houwer, De Pelsmacker, 2005).

Na podstawie tej listy wyraźnie widać, że „zmienna kryterialna” to na ogół samoopisy, lub inne formy „utajonych” zmiennych (jak w punkcie 1). Jedynie zmienna kryterialna w pkt. 3 zbliża się do zachowania jako takiego, choć oczywiście poziom realizmu jest

niewielki. Dodam, że korelacje pomiędzy metodami jawnymi i utajonymi też są bardzo niskie (zwykle $r < 0,2$), co też budzi wątpliwości, bo trudno przypuszczać, żeby miary jawne tak skrajnie odbiegały od niejawnych.

I to już powinno wystarczyć do zakwestionowania trafności i sensowności stosowania IAT. Jednak pojawiła się nowsza meta-analiza, napisana przez autorów niezwiązanych z nurtem badań IAT. Jest wyraźnie bardziej sceptyczna. W przeciwieństwie do wcześniejszej meta-analizy skoncentrowano się tylko na trafności IAT w obszarze dyskryminacji rasowej i etnicznej. Uwzględniono w niej nowsze badania, dodano badania wcześniej pominięte, wyliczono ponownie wielkości efektów. Główna zmiana polegała na tym, że nie uśredniono wielu efektów z jednej próby, tak jak robili Greenwald i in. (2009). Ostatecznie nowa meta-analiza przyniosła wyraźnie niższe wskaźniki korelacji dla uprzedzeń rasowych i etnicznych (odpowiednio 0,15 i 0,12), nie stwierdzono też, żeby IAT lepiej przewidywał zmienne kryterialne niż miary jawne.

Badacze podzieli też zmienne kryterialne na kategorie (poglądy dotyczące polityki społecznej, zachowania interpersonalne, spostrzeganie osób, czas reakcji, mikrozachowania⁵⁰ i aktywność mózgu). Nie stwierdzono, żeby IAT lepiej przewidywał wyniki w zależności od tego, jaka była natura zmiennej kryterialnej⁵¹. Badań, w których zmienną kryterialną były zachowania interpersonalne było najmniej obok czasów reakcji, co też pokazuje, że mój krótki przegląd zmiennych kryterialnych może być trafny. Tak czy owak w przypadku zachowań interpersonalnych i mikrozachowań, a więc kategorii najbardziej zbliżonych do „zachowania” trafność IAT była bardzo niska (odpowiednio 0,13 i 0,07). Jest to o tyle zastanawiające, że teoretycznie trafność powinna tu być bardzo wysoka. Badacze, jak twierdzą, zwracali też uwagę na częstą arbitralność doboru wskaźników do tworzenia zmiennych kryterialnych, co oczywiście idzie w sukurs wnioskowi z rozdziału 1.

Ogólnie jednak poza IAT metody utajone wydają się być słabo przetestowane pod kątem trafności. Ich krytyka zresztą jest dużo bardziej fundamentalną. Zasadniczą zaletą tych metod wg badaczy ich stosujących ma być to, że ich stosowanie

⁵⁰ Czyli niewerbalne i subtelne werbalne oznaki uprzedzeń np. nieznaczna zmiana tonu głosu, postawa ciała itp.; mierzone zarówno przez sędziów kompetentnych, jak i przez osoby wchodzące w interakcje.

⁵¹ IAT okazał się trafniejszy w przypadku jednej zmiennej kryterialnej: danych przy użyciu technik neuroobrazowania. Jest to jednak najprawdopodobniej pochodną wybiórczego publikowania, gdyż artykułów z takimi metodami było relatywnie niewiele i wielkości efektów były w nich średnio wyraźnie wyższe.

obchodzi ograniczenia wynikające ze stosowania metod samoopisowych, a więc podatność na świadome i nieświadome zniekształcenia, szczególnie w przypadku tematów o wrażliwej naturze (np. uprzedzenia rasowe). Miary takie mają odzwierciedlać nieświadome lub introspekcyjne niedostępne reprezentacje, mają być zasadniczo odporne na wpływ lęku przed oceną (a więc jednostka nie może ich rozmyślnie modyfikować, żeby przedstawić się w lepszym świetle), oraz mają odzwierciedlać głębokie, mało podatne na zmiany schematy w umyśle, a więc mają być dużo mniej zmienne (też w sensie psychometrycznym) niż dane samoopisowe. Gawronsky, LeBel i Peters (2007) rozprawiają się z tymi trzema twierdzeniami, wskazując, że nie ma dowodów na to, lub twierdzenia te są wręcz sfalsyfikowane.

Głównym argumentem na rzecz miar utajonych są niskie korelacje, jakie wykazują z tradycyjnymi miarami samoopisowymi. Jednak Gawronsky i in. (2007) wskazują, że fakt taki może wynikać z wielu różnych czynników. Jednym z nich może być wysoki poziom błędu pomiaru, jakimi miary utajone się cechują. Wymieniają pięć prac empirycznych wskazujących na bardzo niską wewnętrzną spójność psychometryczną takich miar. Dalej wskazują, że same w sobie mogą w niewielkim stopniu korespondować z miarami jawnymi. Przykładowo, test IAT zawsze mierzy preferencję określonych obiektów względem innych (np. osób białych względem czarnych), co nie musi się przekładać na absolutną wielkość preferencji mierzoną przy pomocy bardziej bezpośrednich miar.

Twierdzenie, że zmienne utajone są odporne na wpływ społecznych oczekiwań, również jest wątpliwe. Implikowałoby to, że korelacja pomiędzy zmiennymi utajonymi a jawnymi powinna maleć wraz ze wzrostem społecznych oczekiwań (np. w kwestiach uprzedzeń rasowych). Tymczasem tak się nie dzieje. W jednym z badań sędziowie oceniali poziom społecznych oczekiwań w różnych obszarach. Nie znaleziono związku z meta-analitycznymi oszacowaniami wielkości korelacji miary jawne-utajone. Gdy kontrolowano spontaniczność reakcji (samoopisowych) korelacja taka była wręcz pozytywna, a więc im mniej badani mogli się zastanowić nad własną reakcją tym bardziej rosła siła korelacji między dwoma rodzajami miar w obszarach wrażliwych (Hofman, Gawronski, 2005). Inną implikacją byłoby to, że świadome wpływanie na reakcje utajone powinno być zasadniczo niemożliwe. Choć faktycznie, istnieje sporo badań, gdzie próby świadomego zmieniania takich miar są mniejsze

niż miar jawnych, to jednak nie są niezerowe, a dodatkowo Gawronsky (2005) podają przykłady dwóch badań empirycznych, gdzie takie wpływanie było wyraźne.

Istnieje wreszcie twierdzenie, że jawne postawy są bardziej podatne na zmianę, jednak taka zmiana nie wymazuje utajonych postaw, które tak czy owak mogą się ujawniać w zachowaniu lub fizjologii (np. student na liberalnym uniwersytecie deklaruje brak uprzedzeń rasowych, ale wykazuje fizjologiczne objawy stresu np. większe przewodnictwo skóry, w obecności osób innych ras). Gawronsky i in. stwierdzają, że przegląd literatury nie pozwala jednoznacznie rozstrzygnąć tego sporu. Z jednej strony wskazują 2 badania empiryczne, gdzie istotnie stwierdzono, że próby eksperymentalnego oddziaływania na postawy zmieniały tylko postawy jawne, a nie utajone, ale równocześnie w 3 innych badaniach nie znaleziono takich różnic. Jednak, co kluczowe, w 4 badaniach okazało się, że określone programy zmiany postaw wpływały na zmiany rzekomych postaw utajonych, ale nie wpływały na zmiany postaw jawnych. Wyniki te trudno wyjaśnić w kontekście przyjętych twierdzeń o naturze takich miar.

W starszych rocznikach JPSP (np. z lat 70') znajduję sporadyczne używanie testów projekcyjnych (np. Dutton, Aron, 1974), w nowszych nie występują nigdy. Pojawiają się za to regularnie miary utajone. Zaryzykowałbym tezę, że miary utajone są odpowiednikami testów projekcyjnych, skrojonymi dla wymagań XXI wieku. Pod pewnymi względami nawet lepszymi. Z jednej strony nie mają całej ezoterycznej otoczki, z drugiej strony użycie komputerów, „obiektywnych” czasów reakcji dodaje im nimbu poważnej naukowości. Nie bez znaczenia jest pewnie też to, że ich stosowanie jest mniej czasochłonne, samo zliczenie wyników wymaga tylko minimalnej biegłości informatycznej i przy odpowiedniej infrastrukturze oraz zapleczu badanych studentów można w krótkim czasie badać wiele osób (lub przeprowadzać wiele względnie niewielkich badań, w których na pewno „złowi się” jakieś p – rdz. 1.4). Dodatkowo, niektóre są luźno oparte na prostych procesach poznawczych, których naukowość jest trudna do zakwestionowania (np. efekt Stroopa, Macleod 1989 ; priming afektywny: Fazio, Sanbonmatsu, Powell, Kardes, 1986)

Wydaje się, że tak jak testy projekcyjne, miary utajone wywodzą się zasadniczo z tego samego pragnienia ominięcia ograniczeń metod samoopisowych i dotarcia do źródła „podświadomych przekonań” (w jakimś stopniu są więc dowodem na sceptycyzm części środowiska naukowego na jawne postawy i samoopisy – patrz

rdz. 2.4). Dodatkowo, mogą być atrakcyjne dla badaczy, bo z jednej strony samo ich przeprowadzanie i zliczanie wyników jest względnie proste, z drugiej jednak strony zaprojektowanie badania, dobór bodźców i interpretacja wyników wymagają pracy „eksperta”, podobnie więc jak testy projekcyjne, do ich stosowania potrzeba wiedzy niedostępnej zwykłym śmiertelnikom.

Zapewne techniki badające zmienne utajone mają swoje miejsce w nauce i chyba nie warto ich całkowicie odrzucać, tym bardziej, że niektóre wydają się mieć jakąś niezerową trafność. Wydaje się jednak, że ciągle jest to krok w niewłaściwym kierunku – badanie nie zachowania, ale jego pośrednich wskaźników. I są to metody, która wydają się doskonale wpisywać w strategię badawczą polegającą na robieniu wielu niedużych, organizacyjnie nieskomplikowanych badań, a więc obarczonych niskim ryzykiem braku publikacji i braku postępu w karierze naukowej. Dla mnie ostatecznym argumentem przeciwko miarom utajonym jest to, że przy ich pomocy nie dokonano żadnego przełomowego odkrycia, choć niektóre są z nami już ponad dwie dekady.

2.6. Studenci jako osoby badane – potencjalne ograniczenia trafności

Kolejnym ograniczeniem trafności badań psychologicznych jest dominacja studentów w roli osób badanych. W literaturze problem ten jest poruszany od dekad. Najstarsza praca do jakiej dotarłem, zawierające krytykę zbyt częstego wykorzystania studentów jako osób badanych, pochodzi z lat 50' XX wieku (Farber, 1952). W kolejnych dekadach temat regularnie powracał (Smart, 1966; Schultz, 1969; Higbee, Lott, Graves, 1976; Greenberg, J. 1987 itd.). Wydaje się, niestety, że bardzo niewiele z tych głosów krytycznych wyniknęło i fundamentalnie nic się nie zmieniło.

Powody badania studentów wydają się być oczywiste: przy relatywnie niedużym wysiłku i kosztach, można przebadać duże grupy badanych w krótkim czasie (szczególnie w badaniach typu papier-ołówek). Jeśli zachętą są dodatkowe punkty zaliczeniowe, nie trzeba się angażować w żmudne nakłanianie potencjalnych badanych do wzięcia udziału. Jako osoba, która miała na studiach epizod w działach sprzedaży, wiem, że namawianie ludzi do czegoś, do czego nie mają ochoty, jest wysoce wyczerpujące psychicznie. W przypadku studentów „zmotywowywanych” dodatkowymi punktami do zaliczenia, problem ten zupełnie znika.

W badaniu studentów nie ma nic złego, jeśli dane takie służą do wstępnych weryfikacji hipotez lub opracowania procedur badawczych. Gorzej, jeśli na studentach badania się kończą. Najlepsze czasopisma, które jak pokażę za chwilę, publikując w olbrzymiej większości dane zebrane na studentach, dają badaczom nieformalną wskazówkę, że można zrobić karierę bez konieczności kosztownego szukania badanych poza murami uczelni. Uporczywe trzymanie się studentów jako osób badanych jest niestety kolejnym argumentem za tezą, że kultura badawcza psychologii bardziej premiuje publikowalność niż wartość poznawczą.

Organizacyjna i logistyczna wygoda korzystania z grup studentów może na długą metę ograniczać możliwości rozwoju nauki. Oczywiście, wykorzystanie studentów jako osób badanych nie musi koniecznie oznaczać nietrafności badań. Tym niemniej jest to grupa w pewnych aspektach nieco inna od ogółu populacji, co zostało szeroko udokumentowane w literaturze. Te różnice mogą znacząco zniekształcać obraz natury ludzkiej, jaki wypływa z badań.

Dla osoby czytającej czasopisma psychologiczne absolutna dominacja studentów jako osób badanych jest czymś oczywistym, mimo to przedstawię wyniki kilku systematycznych analiz. Sears (1986) zbadał, jaki procent artykułów empirycznych z trzech flagowych czasopism psychologii społecznej (JPSP, JESP, PSPB) z rocznika 1980 zawiera dane zebrane na studentach. Było to odpowiednio 70 %, 81 % i 81 % studentów (z czego około 2/3 to studenci psychologii lub zajęć z psychologii). Dodatkowo w dalszych 8 % (JESP) i 12 % (JPSP) artykułów badanymi byli to też studenci, ale nieamerykańscy. Tak więc dorośli, inni niż studenci, stanowili od 11 % do 19 % badanych. W obrębie trzech sekcji JPSP najwięcej studentów notowano w „Attitudes and Social Cognition” (92 %), najmniej w „Personality and Individual Differences” (70%), natomiast „Interpersonal Relations and Group Processes” miał wynik pośredni (81 %). Różnice te wynikają prawdopodobnie z tego, że badania w obrębie osobowości i różnic indywidualnych częściej korzysta się z kwestionariuszy, stąd można pozyskiwać dane w domach badanych, i nie jest wymagana żadna specjalna infrastruktura laboratoryjna, jak w przypadku eksperymentów.

Nowsza analiza Henry’ego (2008) pokazuje wręcz pogłębienie trendu.

Przeanalizowano, kto wchodził w skład prób w badaniach z tych samych czasopism z lat 1990-2005 (uwzględniono tylko badania dotyczące stygmatyzacji i uprzedzeń). Analiza ta pokazała wykorzystanie studentów na poziomie 87 % (JPSP) i 93 %

(PSPB). W przypadku JESP wręcz nie odwoływano się do innych prób (98 %). W 6 innych czasopismach z psychologii społecznej⁵² studenci stanowili od 70 do 93 % badanych. Arnett (2008) przeprowadził wyliczenie dla rocznika 2007 JPSP i wykazał, że studenci psychologii stanowili 67 % badanych w próbach amerykańskich i 80 % w nieamerykańskich.

Gallander-Wintre, North, Sugar (2001) zestawili takie analizy przeprowadzone w 18 artykułach tj. łącznie 31 wyliczeń dla różnych czasopism i różnych roczników z lat 1949-1995. W pięciu studenci stanowili mniej niż 60 % badanych. Najczęściej, w dwudziestu przypadkach, wyliczony odsetek studentów jako badanych wynosił ponad 70 %.

Omówię teraz dane, że studenci potrafią się różnić w zasadniczy sposób od ogółu populacji. Klasycznym tekstem jest tutaj analiza Searsa (1986). Zebrał on najważniejsze wnioski dotyczące tego, jak studenci różnią się od ogólnej populacji. Nie ogranicza się ona tylko do listy wniosków z badań różnic między studentami i ogólną populacją, ale też wskazuje na wynikające z tego „paradygmaty” w psychologii społecznej. Sears pokazuje, że w porównaniu ze starszymi dorosłymi studenci mają słabiej skryształizowane postawy, gorzej wyartykułowane sądy na temat samego siebie, większe skłonność do ulegania autorytetom, bardziej niestabilne relacje rówieśnicze i wyższe wyniki w miarach inteligencji płynnej. Idąc dalej przekonuje, że fakt, że badania na studentach stanowią bazę, na której buduje się wiedzę psychologii społecznej, przyczynia się do trwania poglądów takich jak to, że ludzie łatwo ulegają wpływom społecznym, łatwo zmieniają postawy (a przy tym zachowują się w sposób niespójny z nimi), oraz to, że obraz siebie budują na podstawie poglądów innych, a nie introspekcji.

Badacz idzie nawet jeszcze dalej i twierdzi, że przesunięcie psychologii społecznej w kierunku procesów kognitywnych jako determinantów zachowania (kosztem emocji, dyspozycji osobowościowych, norm grupowych czy własnych interesów), które dokonało się w latach 60' i 70', mogło wynikać ze stosowania tej niereprezentatywnej bazy badanych. Studia wyższe są aktywnością, w której premiowane jest logiczne myślenie, analityczne zdolności czy ogólna racjonalność. Również studenci to osoby wyselekcjonowane na podstawie przyzwoitych wyników testów zdolności takich jak

⁵² *Journal of Applied Social Psychology, Basic and Applied Social Psychology, Group Processes and Intergroup Relations, Social Cognition, British Journal of Social Psychology, European Journal of Social Psychology*

SAT. Ta atmosfera racjonalności i „poznawczości” miała zdaniem Searsa nieopatrzenie przenikać do laboratoriów psychologicznych, mając bardzo daleko idące reperkusje właśnie w postaci zmiany paradygmatu badań.

Inny badacz Peterson (2001) przeprowadził meta-analizę drugiego stopnia, polegającą na zestawieniu wyników meta-analiz badań (z psychologii społecznej i dziedzin pokrewnych), w których porównywano wielkości różnych efektów psychologicznych dla grup studentów i nie-studentów. Peterson chciał pokazać, gdzie wiadomo, że wyniki studentów i nie-studentów systematycznie się różnią.

Choć wiele wyników było zbliżonych, w wielu wypadkach siła efektów dla studentów i reszty populacji była znacząco różna, lub nawet miała inny kierunek. Przykładowo, związek między empatią, a zachowaniami prospołecznymi był wyższy u studentów ($r = 0,3$), niż u niestudentów ($r = 0,09$); studenci wykazywali znacząco większy efekt ($d = 0,47$) próżniactwa społecznego (zmniejszanie wysiłku, gdy działa się z innymi) niż pracownicy organizacji, u których zaobserwowano wręcz (słaby) efekt o przeciwnym kierunku ($d = -0,19$); studenci dużo więcej zyskiwali na grupowych treningach otwartości i empatii ($d = 1,8$) niż ogólna populacja ($d = 0,9$); przemoc w telewizji przyczyniała się do zwiększania agresywnych zachowań u studentów w większym stopniu ($r = 0,37$) niż u innych dorosłych ($r = 0,19$) itp. Kilkanaście meta-analiz dotyczy bardzo szerokiego spektrum zagadnień, zapewne dlatego, w przeciwieństwie do Searsa, Peterson ograniczył się do sprawozdania ilościowych różnic w danych, bez głębszej analizy teoretycznej, stwierdzając jedynie, że „nie było jednoznacznego wzoru tych różnic” (s. 458). Wydaje się jednak, że te meta-analizy idą w sukurs wielu twierdzeniom Searsa, pokazując na przykład, że studenci łatwiej ulegają niektórym wpływom (np. treningów otwartości, telewizji). Wyniki niektórych przytoczonych tu meta-analiz wydają się mocnym argumentem na rzecz ostrożności w interpretacji wyników badań opartych tylko na studentach. Dane te, choć często są zbliżone, czasami jednak znacząco odbiegają od wyników ogólnej populacji, zwykle w ten sposób, że w próbach studenckich efekty są wyraźnie większe.

Ograniczenia trafności wynikającej z badania studentów w konkretnym obszarze uprzedzeń grupowych szeroko analizuje Henry (2008). Wskazuje, że uniwersytety to miejsca, w których promowane są raczej wartości liberalne i egalitaryzm, tak więc studenci powinni wykazywać mniejszy poziom uprzedzeń, lub chociaż bardziej subtelnie go maskować. Przebadał 277 studentów oraz 430 osób z ogólnej populacji

pod kątem szeregu samoopisowych zmiennych mierzących deklarowane uprzedzenia względem różnych grup (kobiet, Latynosów, Afroamerykanów itd.), rozmaite postawy i ideologie. Studenci istotnie wykazywali więcej pozytywnych uczuć względem tych grup oraz ogólnie wyższy poziom liberalnych postaw. Jednak zanotowano nie tylko takie proste różnice. Co szczególnie różniło populację studencką i ogólną to interakcję pomiędzy zmiennymi. I tak, o ile ciepłe postawy względem grup w populacji ogólnej korelowały z poziomem ogólnego wsparcia dla różnorodności, o tyle u studentów takich efektów nie było. Zdaniem autora, może to świadczyć o tym, że studenci dostosowują swój poziom deklarowanych uprzedzeń grupowych do oczekiwań, i nie ujawniają do końca swoich „autentycznych” przekonań. Idąc dalej, pokazał, że u studentów ideologia był znacznie lepszym mediatorem związku pomiędzy wsparciem nierówności, a sprzeciwem wobec akcji afirmatywnej. Zdaniem autora, to nie jest rzeczywista zależność, ale wskaźnik tego, że studenci mają lepsze wyczucie, jak odpowiadać, żeby ich kwestionariuszowe odpowiedzi były ideologicznie bardziej spójne. Te i inne prawidłowości, zdaniem Henry’ego powodują, że literatura w psychologii uprzedzeń siłą rzeczy musi być ograniczona, jeśli nie uwzględniamy nie-studentów. W podsumowaniu napisał:

Czy można wiedzę o psychologii uprzedzeń zebraną na studentach generalizować do ogólnej dorosłej populacji? Na pewno, jakąś część można. Problem jednak leży w tym, że w tym momencie nie możemy być pewnym, którą część i w jakim stopniu. Te pytania są ze swej natury empiryczne, i jako naukowcy nie powinniśmy o tym nigdy zapominać (Henry, 2008, s. 67).

Osobnym problemem może też być to, że studenci są na ogół osobami o relatywnie niewielkim doświadczeniu życiowym, niewielkiej wiedzy o własnych preferencjach i obrazie świata mało skonfrontowanym z rzeczywistością. Może mieć to szczególnie znaczenie w tych obszarach badawczych, w których zachowania i preferencje mogą się zmieniać wraz z wiekiem i doświadczeniem. Obszarem takim wydają mi się być preferencje cech u partnerów płciowych. Olbrzymia większość literatury jest oparta na badaniu studentów. Przykładowo w szeregu często cytowanych badań (Gangestad, Garver-Apgar, Simpson, Cousins, 2007; Haselton i Miller, 2006; Li i in. 2013; Li, Valentine, Patel, 2011; Edlund, Sagarin, 2010) dosyć daleko idące wnioski na ten temat opierano tylko na badaniu tej grupy. W wielkim badaniu Bussa (1989) dotyczącym preferencji przy doborze partnera płciowego przebadano ponad 10 tys. osób z 34 krajów z 5 kontynentów. Niewiele jest przykładów w psychologii społecznej

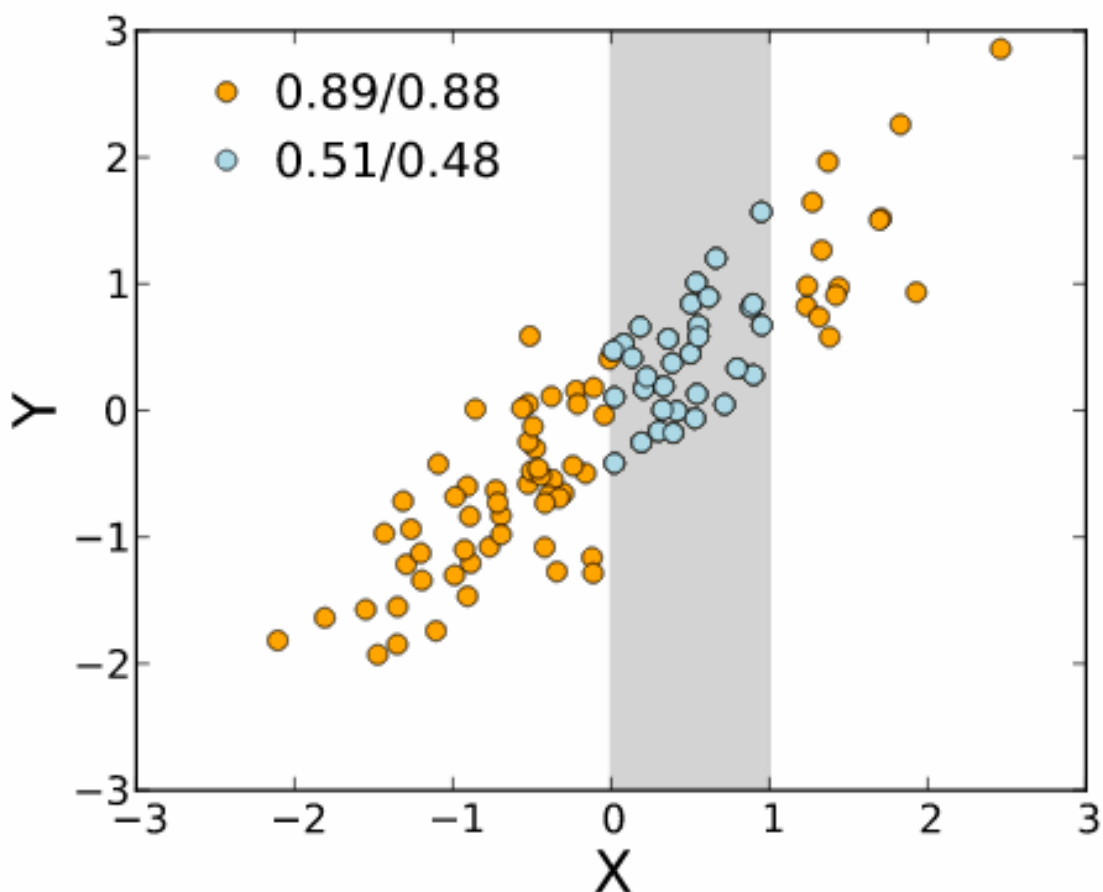
badani obejmujących tak szerokie spektrum kulturowe. Niestety, duża część tego wysiłku została jakoś zmarnowana, gdyż byli to w olbrzymiej większości bardzo młodzi badani (mężczyźni: średnio 23,5 lat; kobiety: 22,5 lat; przy odchyleniach standardowych około 2,5-3,0 roku). Podobnie Meston i Buss (2007; popularne omówienie: Meston, Buss, 2010) prowadzili dosyć duże badania na temat przyczyn uprawiania seksu. Zebrali bardzo grupę badanych, która liczyła ponad 1500 osób. Grupą tą byli studenci o średnim wieku 19 lat... Na tych badaniach oparto potem niezliczone popularne artykuły. Krytykę dominacji studentów jako badanych z perspektywy rozwojowej przedstawili Gallander-Wintre i in. (2001).

Studenci są wreszcie, co oczywiste, grupą mniej zróżnicowaną pod kątem pewnych cech niż ogólna populacja. Statystycznie patrząc, taka sytuacja będzie powodowała, że ewentualne zależności w grupie studenckiej będą słabsze, niż byłyby w całej populacji (por. rycina 2.1). Wydaje się, że z taką sytuacją możemy mieć do czynienia w badaniach nad zależnością między inteligencją, a wynikami egzaminów na różnych poziomach edukacji. O ile w ogólnej populacji szkolnej notuje się bardzo silne korelacje między inteligencją ogólną a wynikami testów szkolnych (Deary, Strand, Smith, Fernandes, 2007), o tyle w populacji studenckiej takie zależności są wyraźnie słabsze (Rhode, Thompson, 2007).

Dodatkowym ograniczeniem trafności wynikającym z badania studentów jest pozalaboratoryjna komunikacja między nimi. Ten problem nie występuje, jeśli badani nie mają możliwości kontaktowania się ze sobą, bo np. są rekrutowani na ulicy. Natomiast w znacznej części artykułów z psychologii społecznej podawana jest informacja, że studenci otrzymywali za badania „course credit” (punkty zaliczeniowe) – można więc zasadnie zakładać, że są z tej samej grupy lub roku. Studenci mogą przekazywać sobie informacje o szczegółach badaniach, zapewne szczególnie w przypadku, gdy procedura okazuje się być wysoce zaskakująca, wzbudzająca silne emocje lub określone zachowanie może wiązać się z dodatkową gratyfikacją. Bardzo rzadko spotykam się z tym, żeby problem ten był podnoszony na łamach artykułów. Dodatkowo, rzadko spotkałem się z informacją, jak długo trwał cały cykl badania (a nie sesja pojedynczej osoby). Jest to o tyle istotne, że można sobie wyobrazić, że jeśli wszyscy badani są przebadani w trakcie jednej sesji, lub w ciągu jednego dnia, problem pozalaboratoryjnej komunikacji może być minimalny. Badacze też nie mają

w zwyczaju informować czy kolejność uczestniczenia w badaniu była źródłem dodatkowej wariancji (co mogłoby wskazywać na komunikację pozalaboratoryjną).

RYCINA 2.1. Siła określonej zależności dla całego spektrum i dla wycinka danych



Źródło: przedruk z https://en.wikipedia.org/wiki/Correlation_and_dependence

Na długo zanim jeszcze przystąpiłem do pracy nad tą rozprawą, przeczytałem artykuł z eksperymentem, w którym badano problem komunikacji studentów poza laboratorium. Badani uczestniczyli w “eksperymentencie psychologicznym”, w którym mieli m.in. odgadnąć cyfrę. Dla wszystkich badanych cyfra do odgadnięcia była taka sama. Niektórzy badani w kolejnych dniach badania nad wyraz szybko odgadywali cyfrę, co wskazywałoby na to, że otrzymali informację od wcześniejszych badanych. Z pewnym wstydem przyznaję, że pomimo usilnych starań nie mogłem już zlokalizować tego badania, choć sama jego treść wyraźnie utkwiała mi w pamięci⁵³.

⁵³ „Lekcja” ta nauczyła mnie tego, żeby prowadzić spis wszystkich przeczytanych artykułów.

Problem wszechobecności studentów jako badanych jest podobny jak problem braku replikacji dokładnych w psychologii. Prawie każdy mówi, że to złe, ale niewiele z tego wynika (a przynajmniej do niedawna niewiele wynikało; patrz rdz. 1.10). Żeby zwiększyć trafność badań psychologicznych w tym względzie wystarczy zastosować konceptualnie bardzo proste rozwiązanie: replikować badania na populacjach innych niż studenckie. Niestety, wymaga to wyższych kosztów, a przy tym, jak się wydaje, nie daje to dużej przewagi przy ubieganiu się o publikacje w niezłych czasopismach.

W ostatnich latach można zauważyć trend polegający na coraz częstym zbieraniu danych poprzez Internet. Przykładowo Buhrmester, Kwang, Gosling (2011) pokazali, że przy pomocy serwisu „Amazon Mechanical Turk” można zdobyć w przeciągu dnia kilkaset wypełnionych ankiet po kosztach jeden dolar amerykański za osobę lub nawet mniej.

Nie wydaje mi się, żeby ten trend – jeśli się szeroko upowszechni – był rozwiązaniem dużo lepszym niż badanie studentów. Choć użytkownicy Internetu z pewnością są grupą szerszą niż studenci, pewnym jest, że nie jest to cała populacja (nie każdy korzysta z Internetu, i zapewne nie każdy, kto korzysta odwiedza strony, gdzie znajdują się badania internetowe). W kilku badaniach stwierdzono, że w badaniach internetowych uzyskuje się podobne wyniki, jak w tradycyjnych badaniach (Riva, Teruzzi, Anolli, 2003; Crump, McDonnell, Gureckis, 2013; Dandurand, Shultz, Onishi, 2008). Problem w tym, że na ogół „tradycyjne” badania są badaniami na studentach.

Nawet, jeśli faktycznie okaże się, że badani w internecie nie odróżniają się tak bardzo od ogólnej populacji, możliwości zdobywania danych tą drogą będą siłą rzeczy ograniczały się do różnych rodzajów danych samoopisowych, w szczególności trudno sobie wyobrazić pozyskiwanie tą drogą danych behawioralnych. Mam też poważne wątpliwości, czy takie dane mogą być wysoce trafne, biorąc pod uwagę rozproszenie uwagi, jakie często towarzyszą korzystaniu z Internetu w domu (np. Zwarun, Hall, 2014). W badaniu Crumpa, McDonnella i Gureckisa (2013) względnie proste zadania poznawcze były porównywalnie wykonywane przez Internet i w laboratorium, jednak w przypadku bardziej złożonych poziom wykonania u osób wykonujących je na swoich komputerach w domu wyraźnie spadał.

Znamienne jest przy tym, że dane internetowe, to kolejny rodzaj danych, które można zaliczyć do umownej kategorii „po linii najmniejszego oporu”.

2.7. Badanie osób z cywilizacji zachodniej – potencjalne ograniczenia trafności

Problemem jest nie tylko dominacja studentów w badaniach, ale też fakt, że są to przede wszystkim studenci amerykańscy. Sporadycznie pojawiają się studenci z innych krajów zachodnich (w tym Izraela). Jeffrey Arnett (2008) nazwał ten problem „porzuconymi 95 %”, co ma odzwierciedlać fakt, że olbrzymia większość opublikowanych badań odbywa się na osobach, które stanowią tylko 5 % światowej populacji. Przeanalizował on narodowość badanych w 6 czołowych czasopismach APA⁵⁴ w latach 2003-2007. 82 % badanych pochodziło z USA lub krajów anglosaskich (Wielkiej Brytanii, Kanady, Australii i Nowej Zelandii). Dalsze 13 % pochodziło z nieanglojęzycznych krajów europejskich, 3 % z Azji (głównie Japonii i coraz częściej Chin), po 1 % z Ameryki Łacińskiej i Izraela, mniej niż 1 % z Afryki i Bliskiego Wschodu. Obrazowo wyliczono, że amerykański student ma 4000 razy większe szanse, żeby wziąć udział w badaniu psychologicznym niż osoba spoza zachodniego kręgu cywilizacyjnego (Henrich, Heine, Norenzayan, 2010).

Nie stanowiłoby to dużego problemu, gdyby wiadomym było, że kontekst kulturowo-społeczny nie wywiera zasadniczego wpływu na zachowanie. Problem w tym, że istnieje, aż nadto dowodów, że taki kontekst wywiera wpływ i to niejednokrotnie znaczący.

Weźmy przykład względnie prostego efektu psychologicznego, jakim jest gra w ultimatum. Polega ona na tym, że jeden z dwóch graczy ma do rozdzielenia sumę pieniędzy. Proponuje podział pieniędzy np. 60 % - 40 %. Drugi gracz go akceptuje, lub odrzuca. Odrzucenie oznacza, że żaden z graczy nie otrzymuje nic. Okazuje się, że w tego typu grze, osoby w przypadku bardzo niesprawiedliwych podziałów (np. 90 % - 10 %), badani odrzucają je, choć patrząc z punktu widzenia czysto ekonomicznej „racjonalności” powinni zadowolić się jakąkolwiek nagrodą większą od zera. Tak jednak nie następują i badani wolą zrezygnować z niezerowej, ale niewielkiej części, żeby ukarać nieuczciwego dzielącego. U zachodnich studentów gracze oferują drugiemu badanemu zwykle część pieniędzy w granicach 43%-48% (Gintis, Bowles,

⁵⁴ *Developmental Psychology, JPSP, Journal of Abnormal Psychology, Journal of Family Psychology, Health Psychology, Journal of Educational Psychology*

Boyd i Fehr, 2007). Okazuje się jednak, że w innych kręgach kulturowych występują znacząco odmienne zachowania. Gintis i in. (2007) przebadali kilkanaście niezachodnich społeczności z całego świata. I tak:

- W niektórych społecznościach w Papui Nowej Gwinei osoby oferowały dla drugiej strony ponad 50 %, ale te oferty były i tak odrzucane. Ma to najpewniej związek z kulturą podwyższania statusu przez rozdawanie kosztownych prezentów. Otrzymanie kosztownego prezentu oznacza wywyższenie cudzej pozycji społecznej, więc osoby nie akceptują takiej „szczodrości”.
- W społeczności Ache z Południowej Ameryki zdecydowana większość ofert była wyższa niż 40 %. Nie były one nigdy odrzucane. Dla kontrastu w sąsiedniej grupie Hadza oferty były niskie i częste były odrzucenia. W pierwszej z tych grup dominuje zwyczaj równego dzielenia mięsa z polowania niezależnie od tego, kto był myśliwym, w drugiej, choć podział grupowy występuje, osoby mają tendencje, żeby zdobycze ukrywać i zachowywać tylko dla siebie i najbliższej rodziny.
- W grupie Tsiname ze wschodniej Afryki oferty były niskie, ale też praktycznie nie występowały odrzucenia. W grupie tej występuje bardzo niski poziom współpracy na poziomie powyżej najbliższej rodziny, nie ma rozbudowanych sankcji społecznych, a więc też zwyczaju karania osób łamiących normy społeczne.

Na tym przykładzie widać, że nawet zachowanie w relatywnie prostej grze ekonomicznej może być bardzo zróżnicowane oraz zdeterminowane przez złożone wzory interakcji społecznych. Nie jest tutaj przedmiotem zainteresowania, jakie jest źródło tych różnic (por. np. Schaller, Murray, 2008), tylko sam fakt ich występowania.

W 2010 w *Behavioral and Brain Sciences* ukazał się tekst, który można uznać za manifest przeciwko ograniczaniu badań tylko do członków kultury zachodniej⁵⁵ (Henrich, Heine, Norenzayan, 2010). Autorzy ukuli na badanych w psychologii nazwę WEIRD (ang. dziwny), a więc Western, Educated, Industrialized, Rich i Democratic (z Zachodu, wykształceni, z krajów uprzemysłowionych, bogatych i demokratycznych). Główną tezą ich manifestu jest twierdzenie, że choć w artykułach na ogół nie mówi się wprost, że odkryte prawidłowości dotyczą całej ziemskiej populacji, to na ogół też nie wskazuje się potencjalnych ograniczeń wynikających z próby. Panuje raczej

⁵⁵ Ponad 1700 cytowań.

kultura niejawnego zakładania, że badane zjawiska generalizują się na członków innych kultur.

Heinrich zbierają też przykłady kilkunastu efektów, gdzie udało się udokumentować występowanie poważnego międzykulturowych różnic w działaniu mechanizmów psychologicznych. Wymieniają m.in., że:

- słynna iluzja Mullera-Lynna nie jest iluzją dla osób żyjących w technologicznie prostych społecznościach, gdzie osoby w toku ontogenezy nie spotykają się z równoległymi i zbiegającymi się liniami (takimi jak np. pokoju);
- zniekształcenia poznawcze, które są dobrze udokumentowane w kulturze zachodniej (np. olbrzymia większość osób jest przekonana o własnej ponadprzeciętności), nie występują w wielu krajach azjatyckich, gdzie raczej normą jest zaniżanie własnych osiągnięć i znaczenia;
- amerykańskie dzieci dłużej i chętniej pracują nad zadaniami, które mogą same wybrać, z kolei wiele azjatyckich dzieci bardziej preferuje zadania wybrane dla nich przez dorosłych, którym ufa;
- podstawowy błąd atrybucji, czyli przekonania o większej roli czynników dyspozycyjnych w determinowaniu zachowania w porównaniu do czynników sytuacyjnych, dotyczy głównie osób z cywilizacji zachodniej. W Japonii istnieje tendencja do przeceniania znaczenia sytuacji w zachowaniu.

Osoby badane najczęściej w Ameryce, a więc liberalni studenci, członkowie klasy średniej, różnią się także od innych, dużo rzadziej, badanych osób z amerykańskiego społeczeństwa (np. niestudujących konserwatystów). Takie osoby mają m.in. większą tendencję do racjonalizowania swoich wyborów w sytuacji dysonansu, mają większe poczucie własnej wyjątkowości, albo możliwości wpływania na swój los. Osoby z tzw. rodzin robotniczych, w przeciwieństwie do amerykańskich studentów bardziej zakotwiczeni w swojej sieci społecznej, a w swojej moralności częściej uwzględniali interesy społeczności (a nie tylko jednostki). To tylko wybrane przykłady, które przedstawili Heinrich i in. Niestety, takich różnic może być znacznie więcej, gdyż możliwe, że wiele efektów nie jest zwyczajnie powtarzanych na członkach populacji ogólnej (o innych kulturach nie wspominając).

Można oczywiście zadać pytanie, czemu badanie innych społeczności jest takie ważne. Czy dane o samych Amerykanach, choć nie generalizują się na inne populacje, nie dają mimo wszystko wiedzy o rzeczonych Amerykanach i nie pozwalają na budowanie – na tej podstawie – skutecznych interwencji, przynajmniej w stosunku do Amerykanów? W moim przekonaniu psychologia powinna być bardziej reprezentatywna z kilku względów. Wiedza o zróżnicowaniu behawioralnym na całym globie jest wartością samą w sobie. Dobre opisanie tego zróżnicowania może być bardzo istotnym elementem budowania metateorii zachowania ludzkiego, pozwoli nam bowiem stwierdzić, które elementy naszej natury są względnie trwałe i niezmiennie, a więc prawdopodobnie ukształtowane filogenetycznie, a które elementy naszej natury są w większej mierze plastyczne, wynikające z socjalizacji i innych mechanizmów uczenia się. Przykładowo, badania zdają się wskazywać, że teoria umysłu jest częścią naszego biologicznego wyposażenia, bowiem wszędzie na świecie jej rozwój rozpoczyna się w dosyć zbliżonym wieku (Wellman, Cross, Watson, 2001).

Badanie zróżnicowania behawioralnego pozwala nam też podejmować próby jego wyjaśnienia. We wspomnianym badaniu Buss (1989) i jego ankieterzy przebadali ponad 10000 osób z 38 kultur. Pokazano, że niektóre różnice międzypłciowe dotyczące preferencji partnerów płciowych są uniwersalne (np. wszędzie mężczyźni preferowali młodsze kobiety, a kobiety starszych mężczyzn; wszędzie mężczyźni przywiązywali większą wagę do atrakcyjności, a kobiety do pozycji społecznej partnera itp.). Później okazało się, że pomimo uniwersalności tych różnic płciowych, sama waga przykładana do poszczególnych cech jest różna. I tak, Eagly i Wood (1999) wskazują na to, że różnice międzypłciowe w wadze przykładanej do zamożności partnera są warunkowane siłą ekonomiczną kobiet: im ta siła jest większa, tym mniejszy nacisk kobiet na zasobny portfel. Dzięki szerokiej, międzykulturowej bazie danych empirycznych można lepiej rozumieć uwarunkowania zmienności zachowania.

Badając samych Amerykanów nie wiedzielibyśmy ani o behawioralnym zróżnicowaniu wynikającym z kultury, ani o uniwersalizmach natury ludzkiej, ani nie posiadalibyśmy danych do badań korelacyjnych między zachowaniem a cechami środowiska. Nasza wiedza o naturze ludzkiej byłaby uboższa, a nasze możliwości stworzenia sensownej meta-teorii psychologicznej dużo bardziej ograniczone.

Dominacja zachodniego świata dotyczy zresztą nie tylko tego, skąd wywodzą się badani. Arnett (2008) przeprowadził wyliczenia, w których pokazał, że autorzy tekstów, redaktorzy i komitety redakcyjne są najczęściej złożone z Amerykanów, sporadycznie Europejczyków a badacze spoza tzw. „zachodu” stanowili 1-5 % pierwszych autorów lub współautorów (w zależności od czasopisma). W tabeli 2.1 zaprezentowałem stosowne dane dla czasopisma JPSP. Heinrich i in. (2010) dodatkowo wskazują, że 70 % cytowanych prac w psychologii pochodzi z USA. I jest to najwyższy wskaźnik z wszystkich przebadanych dziedzin (np. w chemii analogiczny odsetek wynosi 37 %). Jest to tym bardziej problematyczne, że o ile w chemii czy astrofizyce trudno oczekiwać istotnego kulturowego wpływu na przedmiot badania, o tyle w psychologii (i innych naukach społecznych) istnieją przesłanki, żeby spodziewać się takiego wpływu. Pytanie, na ile „zachodni” sposób myślenia kształtuje (i ogranicza) rozwój psychologii leży poza zainteresowaniem tej pracy, tym niemniej warto tu odnotować, że taki wpływ może istnieć.

TABELA 2.1. Odsetek autorów i członków redakcji oraz osób badanych w JPSP

	Osoby z krajów anglojęzycznych (w nawiasie samo USA)	Osoby z Europy (bez krajów anglojęzycznych)	Pozostali
Pierwsi autorzy tekstów	78 % (65 %)	18 %	3 %
Pozostali autorzy	81 % (69 %)	15 %	4 %
Pochodzenie badanych	74 % (62 %)	19 %	8 %
Redaktorzy naczelni i zastępcy	79 % (63%)	16 %	5 % (wszyscy: Izrael)
Członkowie rad naukowych czasopism	89 % (75 %)	10 %	0,5 %

Źródło: opracowanie na podstawie danych w Arnett (2008)

Warto też zauważyć, że paradoksalnie pochodzenie badacza może wchodzić w interakcje z jego tendencjami do „wygładzania” danych. Fanelli (2010b) badał, jak często w pracach amerykańskich badaczy potwierdzona jest wyjściowa hipoteza (dane pozytywne). Zaobserwował nieoczekiwaną rzecz: występowało wyraźne zróżnicowanie częstotliwości wyników pozytywnych w zależności od stanu. Trudno racjonalnie wytłumaczyć, dlaczego badacze w jednych stanach mieliby większe szczęście w generowaniu trafnych hipotez badawczych. Eksplorując temat głębiej okazało się, że badacze ze stanów, w których występuje duża konkurencja na rynku akademickim (tj. duża liczba badaczy i publikacji w przeliczeniu na wielkość populacji) częściej zgłaszali prace z wynikami pozytywnymi. Zestawiając to z powszechnie znanymi faktami, że w USA panuje bardzo silna konkurencja na rynku akademickim, oraz że amerykańscy badacze dominują w publikacjach w psychologii (por. tabela 2.1, oraz Arnett, 2008) można zadać zasadne pytanie, na ile kanon wiedzy psychologicznej jest z tej racji dodatkowo zniekształcony.

Postulat zwiększenia zróżnicowania kulturowego badanych ten jeszcze trudniejszy do realizacji niż postulat badania nie tylko studentów. O ile można jeszcze wyobrazić, że badacz, zamiast tradycyjnego sięgania po studentów, przebadają część dorosłej populacji swojego miasta w rozsądnych kosztach, o tyle dodatkowe przebadanie osób z kilkunastu dodatkowych krajów może leżeć poza zasięgiem olbrzymiej większości badaczy. Rozwiązaniem wydaje się być promowanie współpracy międzynarodowej. Wspominany już w rdz. 1.10 projekt rejestrowanych replikacji w PoPS wpisuje się w ten postulat. W pierwszym opublikowanym badaniu (Alogna i in., 2014) zgłosiły się zespoły badawcze nie tylko z USA, ale też np. z Czech, Polski czy Nowej Zelandii. Jest to pewien (choć ciągle niedostateczny) postęp w stosunku do badania populacji tylko amerykańskiej. Niestety populacja Afryki, Azji i Ameryki Południowej jest ciągle silnie niedoreprezentowana.

2.8. Hipotetyczne przykłady zwiększania trafności zewnętrznej badań

Ostatnią część tego rozdziału chciałbym poświęcić przykładom konkretnych działań, jakie można by podjąć, żeby zwiększyć trafność zewnętrzną badań psychologicznych. O ile to, jak zapobiegać wąskiej „bazie” (studentów i mieszkańców

cywilizacji zachodniej) jest oczywiste, o tyle to, jak wychodzić poza badania samoopisowe i oparte na czasach reakcji wymaga większej kreatywności.

Aby ukonkretnić postulat „więcej zachowania w psychologii” przeanalizuję teraz kilka przykładowych badań z jednego numeru JPSP. Przedstawię je w wersji uproszczonej i nie będę skupiał się na wszystkich subtelnościach teoretycznych, gdyż przykłady te mają służyć tylko zobrazowaniu idei, a nie omawianiu konkretnych badań. A więc w roczniku 2010 mogliśmy byli zapoznać się z takimi oto badaniami:

1. McGregor, Nash, Prentice (2010) (omawiany już szerzej w rdz. 2.1)

- a. Zakładany efekt: niemożność zrealizowania konkretnych celów będzie zwiększała zaangażowanie w zachowania religijne
- b. Sposób dowodzenia występowania efektu: osoby, które otrzymywały nierozwiązywalne zadanie wykazywały wyższy wzrost przekonań religijnych (mierzonych kwestionariuszowo) w porównaniu do osób z grupy kontrolnej (uwzględniano również pewne zmienne dyspozycyjne, ale dla tych rozważań to nieistotne)

2. Bergsieker, Shelton, Richeson (2010)

- a. Zakładany efekt: osoby białe w interakcji z osobą czarną będą przede wszystkim podkreślały swoją sympatyczność; natomiast osoby czarne w obecności osób białych swoją kompetencję (inteligencję, wiedzę itp.), gdyż obydwie grupy chcą przeciwdziałać stereotypom i oczekiwaniom drugiej strony
- b. Sposób dowodzenia efektu: 1) proszono o wyobrażenie sobie interakcji z osobą tej samej rasy lub innej rasy i odpowiadanie na pytanie o to jakie elementy zachowania byłyby ważne 2) badanie ankietowe na temat zachowania względem przyjaciela innej rasy lub tej samej 3) zaaranżowanie interakcji z osobą z innej lub takiej samej rasy, a następnie ocena przez ślepych sędziów kompetentnych zachowań na podstawie nagrań na taśmach wideo

3. Piff, Kraus, Côté, Cheng, Keltner (2010)

- a. Zakładany efekt: osoby z niższych klas społecznych są bardziej prospołeczne (co wyrażą się m.in. większą szczodrością, zaufaniem i gotowością do pomagania)
- b. Sposób dowodzenia efektu: trzy eksperymenty, w tym dwie gry ekonomiczne w których mierzono gotowość do dzielenia się i zaufanie do partnera gry (np. poprzez dzielenie się potencjalnymi punktami, bez gwarancji ich odzyskania); eksperyment w którym mierzono gotowość do pomocy rzekomemu drugiemu badanemu; badanie kwestionariuszowe w którym indukowano poczucie przynależności do różnych stopni drabiny społecznej (poprzez poznawcze porównywanie się z osobami stojącymi nisko lub wysoko), a następnie wypełniano ankietę mierzącą „ile należyłożyć na dobroczynność”. Z wyjątkiem ostatniego badania przynależność do klasy społecznej badano poprzez kwestionariuszowe samooceny (na marginesie: jedno z badań autorstwa tego zespołu zakwestionował potem Francis, 2012c)

Wszystkie powyższe artykuły empiryczne z pewnością zaliczają się do „górnej półki”.

Aby dowieść zakładanych efektów przeprowadzano kilka niezależnych badań (minimum 4), stosowano alternatywne, czasami pomysłowe metody, nieco modyfikowano warunki i próbowano wyeliminować alternatywne wyjaśnienia.

Trafność wewnętrzna wydaje się być bez zarzutu. Podbudowa teoretyczna jest szeroka (m.in. cytowanych jest sporo wcześniejszych badań). W moim przekonaniu artykuły jako takie z pewnością zasługują na druk, a poznawczo wskazują, że istnieją dobre wstępne przesłanki, żeby przypuszczać, że hipotetyczne efekty są prawdziwe.

Zarazem mając świadomość ograniczeń trafności powyższe artykuły powinny stanowić początek pracy nad dowodzeniem rzeczonych efektów. Rzetelna praca empiryczna wymagałaby teraz, aby niezależny zespół badaczy zreplikował owe badania (aby wykluczyć występowanie jakiś niezidentyfikowanych artefaktów, względnie wykluczyć oszustwo), następnie warto by sprawdzić „tradycyjne” zagrożenia trafności zewnętrznej: czy inne osoby, inne warunki i sposoby manipulacji zmiennymi lub inne okoliczności kulturowe nie zmieniają zasadniczo obrazu wyników.

Zakładając nawet, że hipotetyczne efekty są trafne zewnętrznie (np. inne grupy społeczne zasadniczo uzyskiwałyby podobne wyniki) w dalszym ciągu owe badania relatywnie niewiele mówią nam o tym jak owe efekty mają się do „życia codziennego”. Przykładowo w badaniu nad zachowaniami religijnymi manipulacja eksperymentalna podnosiła wynik o ok. pół punktu na pięciopunktowej skali postaw religijnych. Choć można wyliczyć wielkość efektu relatywnie do grupy kontrolnej, myślę jednak, że powie nam to bardzo niewiele jak bardzo porażka w życiu codziennym będzie zwiększyła uczestnictwo w praktykach religijnych. Analogicznie nie wiemy czy 90 ms (tak jak u Hafer, 2000) różnicy w zmodyfikowanym zadaniu Stroopa przekłada się na radykalne zmniejszenie wrażliwości na ofiary niesprawiedliwości czy niewielkie? Może też być tak, że obecny w laboratorium efekt, w codziennym życiu wskutek interakcji z innymi zmiennymi rozmywa się zupełnie, a może wprost przeciwnie, potęguje się. Oczywiście, rozmaite modyfikacje służą nie tylko potwierdzeniu/falsyfikacji wyjściowego efektu, ale również zrozumieniu jego uwarunkowań (w tym ograniczeń występowania).

Oto kilka pomysłów jak mogłyby wyglądać kontynuacje badań nad hipotetycznymi efektami:

1. Niepowodzenia zwiększają zachowania religijne (McGregor, Nash, Prentice, 2010): pomimo tego że hipoteza implikuje konsekwencje behawioralne badano tylko postawy. Być może ów efekt jest krótkotrwały i nie przekłada się na zasadnicze zmiany zachowania? Być może nie jest on uniwersalny, ale wynika tylko ze specyficznych funkcji religii w określonym kręgu kulturowym? Na początku spróbowałbym powtórzyć eksperyment, tym razem jednak zaangażować jakąś behawioralną zmienną zamiast pomiaru postawy. Przykładowo po eksperymencie do badanych podchodziłby pomocnik badacza i zapraszał do wzięcia udziału w jakiejś formie praktyk religijnych. Dalej można sobie wyobrazić podłużne nawet kilkuletnie badania, w których monitoruje się rozmaite aspekty życia badanych oraz zaangażowanie w praktyki religijne (badani za stosowną gratyfikacją np. regularnie wypełniają ankiety w Internecie). Warto byłoby też przebadać różne grupy społeczne mniej lub bardziej „odnoszące sukcesy”, względnie osoby o różnych stopniach sukcesu w obrębie grup społecznych pod kątem ich aktywności religijnej.

2. Interakcje międzyrasowe (Bergsieker, Shelton, Richeson, 2010): w jednym z badań pojawia się nagrywanie rzeczywistej interakcji i analizowanie jej przez sędziów kompetentnych (co jest z pewnością krokiem w dobrym kierunku). Dalsze badania mogłyby wyglądać w ten sposób: obserwowanie i przysłuchiwanie się interakcjom międzyrasowym w kawiarniach, lub innych publicznych miejscach (analogicznie jak u Dunbara, Marriott i Duncana, 1997), korzystanie z rzeczywistych zarejestrowanych danych (np. rozmaitych talk show). Interesującym rozwinięciem teoretycznym i sprawdzianem teorii, byłyby sytuacje, gdy osoba nie powinna się przejmować swoim stereotypem, bo powinien on ulec odwróceniu (np. osoba czarna stoi wysoko w hierarchii organizacji, a biała nisko). Szczególnie warto podkreślić, że badanie wykonano na amerykańskich studentach i jak już wspomniano prawdopodobnie jest to grupa o nastawieniu mało rasistowskim na tle reszty społeczeństwa (Henry, 2008). Zastanawiające jest, więc czy takie same efekty zaobserwowano by u osób z grup o innych wartościach.
3. Klasa niższa bardziej skłonna do pomagania (Piff, Kraus, Côté, Cheng, Keltner, 2010): to badanie zasługuje na pochwałę w stosunku do poprzednich, gdyż 3 z 4 badań używało jako zmiennej zależnej zachowania, co prawda w kontekście laboratoryjnym, ale jednak. Jedno z badań uwzględniało procesy wewnętrzne. I tu można sobie wyobrazić szereg prób uzupełnienia danych, szczególnie o eksperymenty terenowe (osoby na ulicy ulegają np. jakiemuś wypadkowi; sprawdza się osoby, które były skłonne do pomocy; badanie można dodatkowo urozmaicać przeprowadzając je w miejscach, gdzie zwykle przebywa klasa wyższa np. ulice z luksusowymi sklepami, i tam gdzie zwykle przebywa klasa średnia lub niższe). Wydaje się, że również można spróbować pozyskać liczne dane archiwalne o rzeczywistych zachowaniach (przykładowo: skłonność osób z klas niższych do działania w organizacjach charytatywnych; ilość udzielanych pożyczek znajomym itd.). Dane pochodzące z ogólnonarodowych szerokich badań (takich jak „Diagnoza społeczna”) wydają się również potencjalnie doskonałym punktem wyjścia do dalszego sprawdzania tej hipotezy.

Pogłębianie badań nad określonymi efektami powinno pomóc wygenerować lepiej udokumentowaną i tym samym pewniejszą wiedzę, ale również przy tej okazji badacze powinni z większą szansą odkryć istotne moderatory hipotetycznych

efektów. Poza tym badanie tego samego efektu przy pomocy rozmaitych metod może pomóc lepiej naturę niektórych pośrednich miar zachowania (takich jak czasy reakcji), a tym samym przyszłe badania, które z jakichś względów (np. etycznych) będą wykorzystywały tylko owe miary, będą mogły być z większym stopniem pewności generalizowane na zachowania pozalaboratoryjne.

Powyższymi przykładami starałem się pokazać (mam nadzieję, że w sposób udany), że rozszerzanie spektrum danych o dane behawioralne jest jak najbardziej możliwe. Niestety, jest też bardziej pracochłonne i potencjalnie bardziej ryzykowne, gdyż może oznaczać zainwestowanie wielu zasobów w potencjalnie niepublikowalne badania. Jest to jednak problem mentalnościowy, a nie problem niemożliwych do obejścia ograniczeń w zwiększaniu trafności zewnętrznej badań psychologicznych.

ROZDZIAŁ 3

KUMULACJA WIEDZY W PSYCHOLOGII

3.1. „Metakryzys” psychologii – wprowadzenie

Niewątpliwie psychologia jest jedną z podstawowych nauk empirycznych, a większość jej nurtów współdzieli z innymi podstawowymi naukami (szczególnie fizyką, chemią i biologią) najważniejsze założenia epistemologiczne, w szczególności to, które zakłada występowanie poznawalnych na drodze empirycznej praw rządzących fenomenami, które są w obrębie zainteresowania danej nauki (co – moim zdaniem – pozytywnie odróżnia ją od niektórych postmodernistycznych nurtów humanistyki zakładających brak wiedzy obiektywnej i antyempiryzm, por. Budzicz, 2011). Dyscyplina posiada ugruntowaną pozycję instytucjonalną na uczelniach i cieszy się pewnym poważaniem w oczach opinii publicznej (np. psychologowie regularnie są pytani o opinię w mediach). Analiza powiązań pomiędzy czasopismami naukowymi z wykorzystaniem zaawansowanych metod naukometrycznych wykazała, że psychologia jest jedną z 7 „centralnych nauk” (ang. *hub science*) tj. „metanauką”, wokół której koncentrują się inne dyscypliny wiedzy (Boyack, Klavans, Börner, 2005). Psychologowie generują corocznie olbrzymią ilość wiedzy pochodzącej z badań empirycznych (przykładowo baza PsycINFO zawiera abstrakty z prawie 2,5 tys. czasopism, ogólnie prawie 3,5 mln abstraktów), a wiele czasopism pod względem rygorów edytorskich w niczym nie ustępuje najbardziej prestiżowym czasopismom fizycznym czy biologicznym.

Pomimo tego regularnie pojawiają się głosy poważnie kwestionujących kierunek, w jakim zmierza psychologia jako całość, oraz zdolność psychologii do rozwiązywania istotnych problemów teoretycznych i praktycznych. Rozmaici autorzy zwracają szczególnie uwagę na to, że psychologia ciągle nie dorobiła się unifikującej metateorii, która wyznaczałaby ramy do rozumienia fenomenów psychologicznych. Takimi metateoriami są przykładowo mechanika kwantowa w fizyce, a w bliższej psychologii biologii - neodarwinowska synteza (która nie wyjaśnia wszystkich fenomenów biologicznych, ale daje podstawy do tego, żeby eksperci od mchów, płetwali i fizjologii wątroby mieli wspólną płaszczyznę porozumienia, nawet jeśli ich praca powierzchownie wydaje się być zupełnie różna). Inne zarzuty dotyczą braku

ujednoliconych standardów badania zjawisk psychologicznych, dużego stopnia separacji pomiędzy poszczególnymi subdyscyplinami (a nawet różnymi obszarami w obrębie subdyscyplin), pojawiających się i przemijających mód na metody i obszary badawcze oraz znacznego rozdziewku pomiędzy badaniami podstawowymi a praktyką. Kilka reprezentatywnych cytatów:

Charakter psychologii w coraz większym stopniu manifestuje się w szybkim przyroście wąsko zorientowanych i kompulsywnie wyizolowanych obozów, przyroście, który zdaje się nie mieć żadnych granic. Dążymy do odpowiadania na małe pytania, zamiast na duże, nie widzimy całego lasu tylko poszczególne drzewa (Bevan, 1991, s. 475)

Kiedy spojrzymy na psychologię przez soczewkę unifikacji, zobaczymy naukę wypełnioną nadmiarowymi pojęciami, zasadami, odkryciami, teoriami, testami, terapiami i temu podobnymi. Akceptacja nadmiarowości prowadzi w naszej płodnej nauce do niemożliwej do ogarnięcia, chaotycznej złożoności i braku efektywności. Taki sposób działania byłby nie do zaakceptowania w fizyce, chemii czy biologii, jest on wbrew podstawowemu wymaganiu parsymonii w teorii naukowej. Nie mogłoby być tak, że jedna grupa fizyków używałaby podstawowych pojęć masy, grawitacji, protonu, neutronu i kwarku, podczas gdy druga grupa używała innych pojęć dla tych samych zjawisk, a ta różnica w terminologii byłaby uzasadnionym powodem istnienia dwóch odrębnych teorii i braku komunikacji. Ale tak się dzieje w psychologii, co skutkuje zdumiewającym bałaganem i nieefektywnością. (Staats, 1999, s. 7)

[...] występujące w psychologii rozliczne i zróżnicowane programy badawcze, nie są zwykle ani koherentnie powiązane ze sobą, ani w istotny sposób związane z dociekaniem badawczymi w powiązanych dziedzinach naukowych. Zamiast tego jesteśmy świadkami niepohamowanej akumulacji różnorodnych faktów, odkryć i teorii, które zwyczajowo nie integrują się z konkurencyjnymi enklawami badawczymi i frakcjami teoretycznymi. W skrócie, psychologia funkcjonuje jako mało dojrzała nauka. (Rand, Ilardi, 2005, s. 7)

Praktycy w psychologii są w dalszym ciągu w stanie utrzymać względną niezależność od psychologii jako nauki. Przykładowo, pomimo braku przekonującego empirycznego wsparcia, wielu profesjonalistów w dalszym ciągu korzysta z technik podobnych do tych rozwiniętych przez Freuda prawie wiek temu. W rzeczy samej wielu psychologów klinicystów i doradców aktywnie kwestionuje twierdzenie, że terapie psychologiczne wymagają empirycznej walidacji (Rand, Ilardi, 2005, s. 11).

Nie ma poważniejszego ryzyka [zawodowego] dla badaczy zajmujących się „modnymi” podejściami. Mody ostatecznie przemijają. Tworzą entuzjazm, częściowo poprzez wyolbrzymianie swoich możliwości. Psychoanaliza, behawioryzm i neuronauka poznawcza twierdzą, że wyjaśniają większość ludzkiego zachowania i, że są najlepszym podejściem. To przyciąga wyznawców, ale naraża je na zasadniczą krytykę i ostatecznie porzucenie.

Wartościowe „mody”, które mogą realistycznie wyjaśnić 20 % pytań, sprzedaje się jako obietnice wyjaśnienia praktycznie wszystkich”. (Rozin, 2007, s. 755)

Jako, że ta dysertacja jest w szczególności sposób poświęcona psychologii społecznej, warto w tym kontekście przywołać też szereg cytatów z artykułu Elmsa (1975) poświęconych tylko tej subdyscyplinie. Wiele wypowiedzi brzmi nad wyraz współcześnie:

Psychologia społeczna jest teraz w „okresie kryzysu”, w sensie takim sensie, w jakim Kuhn używał tego terminu w książce „Struktura rewolucji naukowych”. Wydaje się, że straciliśmy rozeznanie, co do istotnych problemów, które mamy badać i modeli, których mamy używać w naszych badaniach i teoriach (Berkowitz, za: Elms, 1975, s. 967)

Nasi najlepsi naukowcy poszukują sensownego paradygmatu. Trudno z góry przewidzieć, które kierunki okażą się ślepyimi zaułkami, a które nie (Smith, za: Elms, 1975, s. 967)

Od czasu do czasu badacze ożywiają określone tematy czy obszary, które wydają się nowe i ważne; ale prędkiej czy później okazują się być jałowe, więc są porzucane (Moscovici, za: Elms, 1975, s. 967)

Psycholodzy społeczni regularnie dokonują w laboratorium odkryć, w których efekty główne i proste interakcje wydają się niezawodnie generalizować, a potem okazują się być „błędnyimi ognikami”, bo nie można ich zreplikować konceptualnie, albo okazują się być produktem interakcji wyższego rzędu z jakimiś trywialnymi zmiennymi, które są charakterystyczne tylko dla określonych warunków eksperymentalnych (Janis, za: Elms, 1975; s. 968)

[...] w eksperymentalnej psychologii społecznej jest mnóstwo odkryć efektów polegających na złożonych interakcjach wielu zmiennych, jednak proste efekty pojawiają się rzadko. Nawet tam gdzie je uzyskujemy, często wynika to po prostu z braku uwzględnienia zmiennych E, F, G itd., których obecność pokazałaby interakcję” (Campbell, za: Elms, 1975 s. 968)

Tak więc głosów krytycznych na temat psychologii nie brakowało na długo przed początkiem afery Stapela. Chciałbym teraz przyjrzeć się przykładom konkretnych sytuacji w psychologii, gdzie pojawiły się duże problemy w odpowiadaniu na ważne pytania badawcze.

3.2. Konkretnie przykłady „metakryzysu”

Niedawno w Sądzie Najwyższym Stanów Zjednoczonych miała miejsce rozprawa dotycząca prawa lokalnych legislatur do ograniczania dystrybucji brutalnych gier komputerowych (Sąd Najwyższy USA, 2011; Ferguson, 2013). Przeciwnicy takich zakazów wskazywali na to, że gry powinny korzystać z pierwszej poprawki do

konstytucji, gwarantującej wolność wypowiedzi. Zwolennicy argumentowali, że wolność wypowiedzi można ograniczać w imię wyższego dobra, czyli tutaj w imię zmniejszania przestępczości nieletnich. Zasadnicza część rozprawy była poświęcona statusowi naukowemu hipotezy, że brutalne gry wzmagają przemoc i inne problemy psychologiczne.

W procesie uczestniczyło wielu badaczy i instytucji naukowych. Amerykańskie Towarzystwo Psychologiczne (APA) jednoznacznie wskazywało na szkodliwe skutki gier komputerowych, a więc konsekwentnie podtrzymywało opinię, którą wygłosiło w kilka lat wcześniejszym raporcie (APA, 2005). W toku rozprawy okazało się jednak, że sprawa nie jest taka jednoznaczna. Omawiane w pierwszych dwóch rozdziałach problemy rzetelności i trafności danych zostały niestety przez sąd uwypuklone.

Przykładowo, sędziowie zauważyli m.in. ograniczenia wynikające ze stosowania niektórych miar agresywności (rozdział 2). Jeden z sędziów przytoczył przykład badania, w którym dzieci grały w brutalną grę, a następnie miały wypełnić łamigłówki słowne z wieloma rozwiązaniami. Okazało się, że grający badani częściej wstawiali potem literę „d” w słowie „explo_e” (a więc „explode”) niż literę „r” („explore”) (jest to rzecz jasna przykład miary utajonej rzekomo mierzącej poznawczą dostępność pojęć związanych z agresją: rdz. 2.5). Ten i inne przykłady zdaniem Sądu wskazują na to, że w wielu badaniach występuje wyraźny rozdzźwięk między tym jak mierzona jest agresja, a tym jak jest postrzegana w powszechnej opinii. „Zapobieganie tego typu fenomenom [uzupełnianie niewłaściwych liter w słowach – przyp. ŁB] [...] nie jest dla sądu przekonującą wartością” (za: Sąd Najwyższy USA, 2011, s. 13).

Podnoszony był argument, że jeden z najchętniej wykorzystywanych w tego typu badaniach sposobów operacjonalizacji „agresywności”, polegający na tym, że badani rażą nieprzyjemnym białym szumem innych badanych znajdujących się w drugim pomieszczeniu (podobnie jak w eksperymencie Milgrama w rzeczywistości nikogo nie rażą, jednak o tym nie wiedzą). Przykładowo badani, którzy „wygrywają” sesje gry mają możliwość zaaplikować szum o różnej sile i długości, takich sesji odbywa się kilkanaście (kto wygrywa daną sesję jest zwykle ustawione przez badaczy).

Ferguson (2013) zwraca uwagę, że z danych surowych wyciąga się wskaźniki w bardzo arbitralny sposób w różnych artykułach. Jako zmienne zależne były wykorzystywane np. liczba sesji, w których badani zaaplikowali najgłośniejsze szумы, pierwiastek kwadratowy z długości szumów pomnożony przez intensywność,

tylko sumy intensywności szumów, albo średnie z intensywności i długości osobno. Czasami analizowano wszystkie sesje, a czasami dzielono je na 3 kolejne bloki i badano wyniki w każdym z nich (i jak można się domyślić działały tylko niektóre bloki). Różnice te nie wynikały bynajmniej z braku komunikacji pomiędzy badaczami stosującymi nowatorskie narzędzie, gdyż podane tutaj przykłady operacjonalizacji pochodzą z artykułów, w którym zawsze jednym ze współautorów był jeden badacz, Andersona⁵⁶, nomen omen, znany zwolennik tezy o wysokiej szkodliwości gier (np. Anderson, Dill, 2000) i główny twórca wspomnianego raportu APA (2005). Co oczywiste, arbitralność w wyborze danych, z których utworzy się zmienną zależną, może być strategią łowienia wyników istotnych statystycznie (rozdział 1).

Znacznie poważniejszy zarzut sądu dotyczył jednak tego, że w obrębie społeczności badaczy istnieją duże rozbieżności w ocenie szkodliwości. Istnieją z jednej strony meta-analizy wskazujące na jednoznacznie szkodliwy wpływ gier komputerowych (Anderson i in., 2010), z drugiej strony zespoły innych badaczy wykazują, że efekt ten jest bardzo słaby i prawdopodobnie wynika z wybiórczego publikowania (Ferguson, 2007; Ferguson i Kilburn, 2009; Sherry, 2001). Wybór badań i wybór akcentów może determinować wyniki takich analiz. W badaniach korelacyjnych uzyskuje się wyraźniejsze efekty, choć oczywiście nie ma możliwości kontroli wszystkich zmiennych. W badaniach eksperymentalnych uzyskuje się znacznie słabsze efekty, a same zmienne są mocno pośrednimi miarami agresji. Nawet, gdy badacze potrafią zgodzić się, że średni efekt z konkretnego zbioru badań kształtuje się na poziomie dajmy na to $r = 0,15$, nie ma zgody czy jest to wartość zawyżona (np. niewielki czas ekspozycji i ograniczone możliwości reakcji w laboratorium) czy zaniżona (np. rodzaj korelacji pozornej, bo w badaniu korelacyjnym nie uwzględnia się wielu zmiennych), oraz jak interpretować praktyczne znaczenie takiej wielkości (Ferguson, 2013, s. 64).

Pomimo tych wszystkich zastrzeżeń i niejednoznaczności, jak już wspominałem, APA przedstawiła kilka lat temu raport jednoznacznie wskazujący na szkodliwość gier komputerowych. Ferguson zwrócił jednak uwagę, że autorami raportu byli przede wszystkim badacze, którzy wcześniej prowadzili badania wskazujące na

⁵⁶ Dodam na marginesie, że przeglądając abstrakty artykułów Andersona poświęcone problematyce wpływu gier komputerowych na agresywność, nie znalazłem żadnego wskazującego na brak takiego wpływu. Wydaje się absolutnie nieprawdopodobne, że przy takim umiarkowanie silnie efekcie nie wykonał wielokrotnie badań potwierdzających hipotezę zerową.

negatywne działanie gier, a nie dopuszczono do głosu badaczy bardziej umiarkowanych. Najlepiej natomiast, gdyby sprawą zajęli się badacze w ogóle nie zaangażowani wcześniej w temat, co oczywiście nie miało miejsca.

Zachodzi tu też ewidentny konflikt interesów. APA z jednej strony jest organizacją naukową, ale z drugiej strony reprezentuje interesy grupowe psychologów, więc może zyskiwać na wskazywaniu występowania problemów, których rozwiązaniem powinni zajmować się psycholodzy.

Ostatecznie Sąd Najwyższy miał przekonać argument wskazujący na systematyczne zmniejszanie brutalnej przemocy wśród nieletnich w ciągu ostatnich kilkunastu lat, pomimo równoległe znaczącego wzrostu sprzedaży gier. Przychylił się do zdania, że obostrzenia w sprzedaży takich gier są niekonstytucyjne i przysługuje im ochrona na mocy pierwszej poprawki.

Proces i wyrok są wg mojego nieformalnego rozeznania praktycznie nieznane w środowisku polskich psychologów, pomimo tego, że był to niewątpliwie ważny sprawdzian możliwości psychologii w odpowiadaniu na istotne problemy społeczne. „Egzaminatorem” na tym sprawdzianie była, co nie zdarza się często, niezależna instytucja, ciesząca się dużym prestiżem. Ze względu na wytknięte duże niespójności w literaturze, a jednocześnie dogmatyzm grupy badaczy i APA, test ten nie został chyba zdany pozytywnie.

Przykłady szkodliwości (lub braku) gier komputerowych nie jest zresztą jedynym przypadkiem, gdy w społeczności badawczej nie ma zgody, co do zasadności określonego efektu i to nie tylko na poziomie pojedynczych badań empirycznych, ale na poziomie meta-analiz. Nie jest trudno znaleźć kilka przykładów.

I tak, grupa psychologów ewolucyjnych w meta-analizie opublikowanej na łamach PB wskazuje, że faza cyklu miesięcznego u kobiet ma wpływ na preferencje co do partnera płciowego (Gildersleeve, Haselton, Fales, 2014). Inna grupa badaczy, wywodzących się z nurtu zakładającego dużą rolę socjalizacji i kultury w kształtowaniu zachowania, przekonuje z kolei w meta-analizie opublikowanej na łamach Emotion Review, że wpływu takiego nie ma (Wood, Kressel, Joshi, Louie, 2014; przy okazji wytykają badaniom w tym nurcie m.in. wysoką arbitralność w ustalaniu, które dni cyklu zaliczają się do „płodnych”, por. też Gelman, Loken, 2014).

Nguyen i Ryan (2008) na łamach *Journal of Applied Psychology* przeprowadzili meta-analizę efektu zagrożenia stereotypem, polegającego na pogorszeniu wyników zadań matematycznych u kobiet, którym „przypomniało” się o tym, że kobiety osiągają niższe wyniki w takich testach. Efekt, choć niezbyt silny, jest ich zdaniem rzetelny i systematycznie powtarzalny. Wkrótce potem, dwa niezależne zespoły (Stoet, Geary, 2012; Flore, Wicherts, 2014) opublikowały swoje meta-analizy. Podobne były wnioski, jeśli chodzi o siłę efektu, jednak w obydwu odkryto daleko idące anomalie w danych wskazujących na wybiórcze publikowanie.

W ostatnich latach dużą popularnością cieszył się efekt „wyczerpania siły ego” polegający na tym, że jakikolwiek wysiłek woli (np. podejmowanie decyzji, powstrzymywanie się od jakiejś przyjemnej czynności) powoduje pogorszenie wytrwałości w następnej czynności, nawet zupełnie niezwiązanej z pierwszą (np. rozwiązywanie zadań matematycznych). Opublikowana na łamach PB meta-analiza wskazywała, że efekt taki jednoznacznie występuje (Hagger, Wood, Stiff, Chatzisarantis, 2010). Jednak meta-analiza niezależnego zespołu (Carter, McCullough, 2014) wskazuje na słabszy efekt oraz pokazuje dowody na wybiórcze publikowanie. Co ciekawe, analiza Hagger i in. (2010) nie podejmuje w ogóle problemu wybiórczego publikowania.

Wspominałem już o dwóch meta-analizach z odmiennymi wnioskami dotyczącymi wyższej trafności testu utajonych skojarzeń w porównaniu do jawnych miar w obszarze uprzedzeń rasowych (rdz. 2.5).

Tak jak na poziomie pojedynczych badań może zachodzić arbitralność w wyborze danych do analizy, podobnie na poziomie meta-analiz badacze mogą się różnić przyjmowanymi założeniami dotyczącymi tego, które badania uwzględnią w końcowych analizach, jaką wagę przypiszą poszczególnym danym, czy jak zinterpretują końcowy rezultat. Badacze mogą również różnić się w zakresie powagi podejścia do problemu wybiórczego publikowania (Ferguson, Brannick, 2012).

Intuicyjnie wydaje się, że jeśli w nauce na jakieś twierdzenie jest „glejt” w postaci meta-analizy, jest to praktycznie równoznaczne z tym, że dane twierdzenie jest prawdziwe (por. np. artykuł w *Gazecie Wyborczej* gloryfikujący meta-analizy: Orlński, 2013b). Okazuje się, że w psychologii niekoniecznie musi to mieć miejsce.

Większość autorów meta-analiz jest równocześnie autorami badań empirycznych

jakimś dziwnym trafem pokazujących wnioski w tym samym kierunku jak ich późniejsze meta-analizy. Jest czymś pozytywnym, że w psychologii nie ma dogmatów i nawet meta-analiza z PB może zostać zakwestionowana. Jednak sam fakt, że takie sprzeczne meta-analizy się pokazują, moim zdaniem, pośrednio wskazuje na to, że mamy do czynienia z pewnym kryzysem psychologii w odpowiadaniu na kluczowe problemy poznawcze i praktyczne.

Przy okazji meta-analizy o wpływie cyklu miesięcznego na preferencję płciową Ferguson, który specjalizuje się w problemie wybiórczego publikowania, wyraził przekonanie że meta-analizy nie są w stanie odpowiadać na ideologicznie nacechowane pytania:

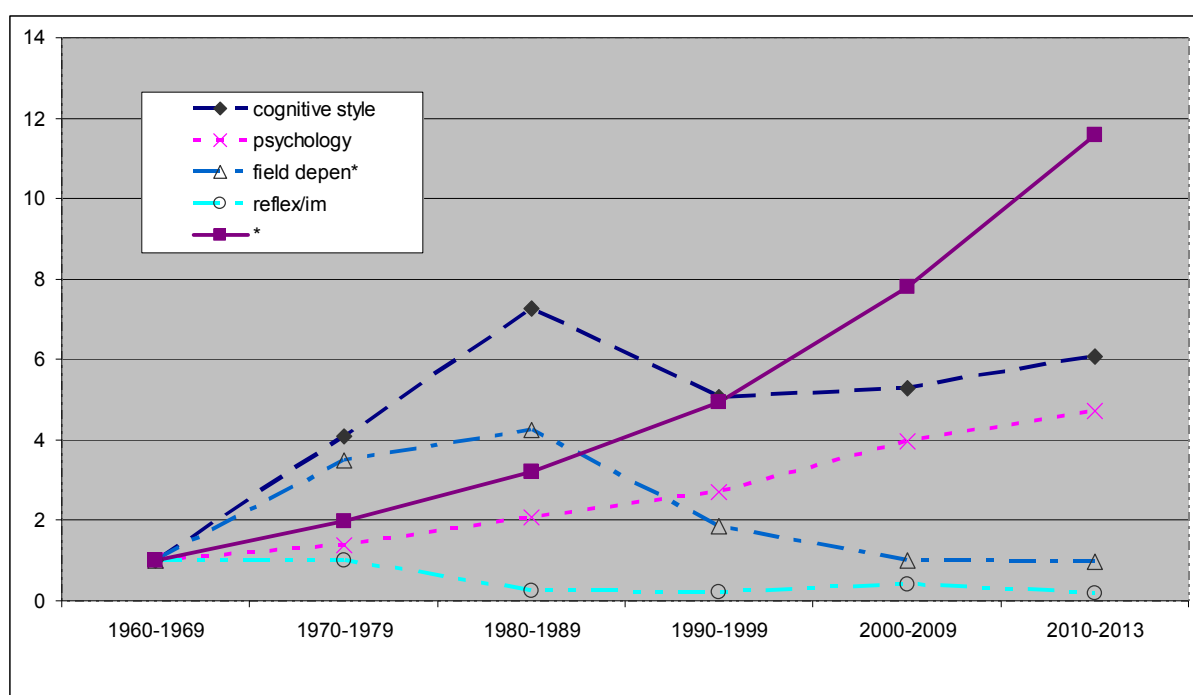
W wielu punktach decyzyjnych, szczególnie dotyczących tego, jakie dane wybrać, jak je zakodować, w jaki sposób wyciągać wielkości efektów i w jaki sposób interpretować badania, istnieje dużo miejsce na to, żeby autorzy meta-analizy wtrącili swoje własne przekonania. Meta-analizy są zwykle chwalone przez tych, którzy zgadzają się z ich wnioskami i odrzucane jako błędne przez tych, którzy się nie zgadzają. Jako takim, meta-analizom nie udało się zastąpić artykułów przeglądowych jako bardziej obiektywnym źródłem wiedzy. (Ferguson, 2014, s. 251)

Kolejny przykład „meta-kryzysu” psychologii dotyczy zdolności psychologii do kumulowania wiedzy. Przyjmuje się, w dużym uproszczeniu, że w nauce są przeprowadzane badania, które są albo systematycznie potwierdzane i wchodzi do kanonu podstawowych faktów, albo są falsyfikowane. Spotyka się czasami twierdzenia, że problem z psychologią nie polega na tym, że ani jedno, ani drugie nie ma zasadniczo miejsca. Raczej wybuchają mody na obszary i metody, po czym, po pewnym czasie mody te zanikają, a pojawiają się mody na nowe obszary dociekań (Rozin, 2007, Pashler, Harris, 2012). W literaturze pozostaje duża liczba danych, które nie są weryfikowane, ani tym bardziej falsyfikowane.

W moim przekonaniu przykładem takiego obszaru jest tematyka stylów poznawczych. Trudno nie odnieść wrażenia, że był taki okres, gdy badacze prześcigali się w wymyślaniu nowego stylu. Przegląd Kozhevnikovej (2007) omawia szerzej 13 stylów, ale podaje też odwołania do prac, gdzie zebrano omówienie 40 różnych stylów. Większość stylów omawianych przez Kozhevnikovą jest prawdopodobnie nieznanych dla psychologów akademickich. Tak czy owak zarówno tych bardziej znanych (np. zależność/niezależność od pola), jak i mniej znanych (np. zasięg skanowania) nie spotyka się już dużo w nowszej literaturze. Przeprowadziłem

przegląd bazy psycINFO pod kątem wybranych słów kluczowych związanych ze stylami. Jak widać na wykresie 3.1 tematyka stylów poznawczych powoli zanika (szczególnie jest to widoczne przy badaniach nad konkretnymi stylami), szczególnie relatywnie do wzrostu liczby publikacji w ogóle. Matczak, 2000 w swoim rozdziale przeglądowym poświęconym stylom poznawczym wymienia przede wszystkim teksty z lat 60 i 70.

WYKRES 3.1. Relatywna częstość występowania określonych słów w tytule lub abstrakcie w bazie PsycINFO (okres 1960-69 = 1).



* - oznacza wszelkie możliwe rekordy, a więc ogólny wzrost liczby publikacji

Przestudiowałem kiedyś część literatury, chcąc dowiedzieć się, czy tak bogaty „urobek” (kilkadziesiąt stylów) został jakoś zintegrowany w jednolitą teorię, a może raczej sfalsyfikowany, skoro zainteresowanie nimi spada. Trudno o definitywną odpowiedź na takie pytanie, ale wg mojej najlepszej wiedzy w tematyce stylów poznawczych nie doszło do zasadniczego przełomu lub powszechnie akceptowanego konsensusu (por. np. Kozhevnikova, 2007; Sternberg, Grigorenko, 2001, Ridding, 1997). Odnosi się raczej wrażenie, że tematyka po prostu się znudziła i ją porzucono.

Tematyka stylów poznawczych przestała być „modna” z jednym małym wyjątkiem. Od dłuższego czasu Robert Sternberg rozwija własną teorię stylów poznawczych (np. Sternberg, 1996). W przeciwieństwie do innych teorii, ta jest faktycznie obecna w nowszych numerach dobrych pism pokroju „Personality and Individual Differences” i innych (autorem bardzo wielu prac jest uczeń Sternberga Li-fang Zhang). Sama teoria i badania nad nią są na tyle znamiennym przypadkiem, niejako „psychologią w pigułce”, że poświęcę im kilka słów.

Zasadniczo w tej teorii Sternberg uznał, że umysł jest zorganizowany podobnie jak rząd i występują w nim różne funkcje (wykonawca, legislator, sędzia), sposoby orientacji na problemy (monarchiczny, oligarchiczny, anarchiczny, hierarchiczny), podejścia do zastanych metod (liberalny i konserwatywny), orientacja względem otoczenia (zewnątrzny, wewnętrzny) oraz różny nacisk na poziom analizy (ogólny i lokalny). Każdy styl mierzy się przy pomocy samoopisowych pytań kwestionariuszowych, mających bardzo ogólną naturę (np. „Kiedy podejmuję decyzję, lubię porównywać różne punkty widzenia”; „Lubię zadania, w których moja rola jest jasno zdefiniowana”). Nie spotkałem się nigdy z testami mierzącymi style poprzez sposób/skuteczność wykonywania zadań (tak jak w teście figur ukrytych, który ma mierzyć zależność od pola). Tak więc do bogatej listy wcześniejszych stylów Sternberg dodał kolejnych 13 stylów.

Wszystko to brzmi na pozór bardzo atrakcyjnie, chciałem jednak dociec czy poza atrakcyjną metaforą były jeszcze jakieś dane (szczególnie empiryczne), które skłoniły Sternberga do utworzenia takiej, a nie innej metafory organizacji stylów poznawczych. W jego publikacjach nie znalazłem takiej odpowiedzi (m.in. Sternberg, 1996; Sternberg, Grigorenko, 2001; Zhang, Sternberg, 2006; Sternberg, Zhang, 2005). Odnosi się raczej wrażenie, że koncepcja powstała całkowicie z powietrza. Oczywiście, samo w sobie nie wyklucza to jej sensowności, wszak Sternberg mógł mieć jakąś błyskotliwą intuicję. Tym niemniej, trochę ironizując, zastanawia mnie, czy inny badacz nie mógłby stworzyć innej atrakcyjnej metafory np. umysł jako samochód, a więc istnieje styl przyspieszania, hamowania, zmienia (umysłowych) biegów itd. Wierzę, że do tej metafory (i do wielu innych) też można dorobić atrakcyjną historijkę oraz skonstruować poważnie wyglądające narzędzia psychometryczną.

W nauce odróżnia się kontekst odkrycia od kontekstu uzasadnienia i w ostatecznym rozrachunku przyjmuje, że najważniejszy jest ten ostatni. Przejrzałem więc trochę badań z wykorzystaniem teorii Sternberga. Powstało samoopisowe narzędzie mierzące 13 stylów, choć w tekstach wskazuje się na to, że ma bardzo umiarkowaną spójność (α na poziomie średnio około 0,65, np. Zhang, 2004a; Zhang, 2005). Trudno nie odnieść wrażenia, że wyodrębnione czynniki mogą się ujawniać w analizie czynnikowej bardziej wskutek podobieństwa pytań, niż leżących u ich podstaw źródeł organizacji umysłu.

Analizując dalej kilkanaście badań z wykorzystaniem tej teorii stylów poznawczych Sternberga stwierdziłem, że program badawczy w dużej mierze polega na korelowaniu ich z wszelkimi możliwymi innymi zmiennymi samoopisowymi np. postrzeganym środowiskiem pracy (Zhang, 2005), poczuciem własnej wartości (np. Zhang, 2001), deklarowanym natężeniem problemów psychologicznych (Chen, Zhang, 2010), cechami osobowości (Zhang, 2006), rozwojem tożsamości (Zhang, 2008), preferowanymi stylami uczenia (Zhang, 2004b, Zhang) czy poglądami na to, czym jest kreatywność (Zhu, Zhang, 2011). Jedyne rodzaje danych „z życia”, jaki jest wykorzystywany w tego typu badaniach, to oceny akademickie. Ale i tutaj efekty są niewielkie⁵⁷ i słabe (Zhang, 2004a; Bernardo, Zhang, Callueng, 2002). Nie spotkałem natomiast żadnego badania, w którym obserwowano by sposoby wykonywania zadań poznawczych (w laboratorium czy bardziej naturalnych warunkach) i próbowano te dane zestawiać z wynikami samoopisowego kwestionariusza mierzącego styl poznawczy wg teorii Sternberga. Osobna kwestia, że same pytania o style są czasami tak mało konkretne, że być nie zdziwiłbym się, gdyby drobne różnice w nastawieniu umysłowym bardzo zmieniały wyniki.

Teoria stylów poznawczych Sternberga nie ma jakiegoś strategicznego znaczenia we współczesnej psychologii. Ale omawiam ją dlatego, że jest przykładem tego, że duże środki zostały zaangażowane do przeprowadzenia kilkudziesięciu badań, które poza korelacjami testów z testami niewiele mówią. Badania takie są z sukcesem publikowane w przyzwoitych czasopiśmie. Jak w soczewce zdaje się ten program badawczy skupiać problemy z rozdziału 2 i być może też rozdziału 1. Przypuszczam,

⁵⁷ Przykładowo w jednym z takich badań (Zhang, 2004a) zestawiano oceny z 6 przedmiotów z 13 stylami poznawczymi (łącznie 78 porównań). Znalezione 3 istotne korelacje (na poziomie $p < 0,05$) i jest to zapewne wynik na poziomie przypadku. Mimo to autor w artykule omawiał te wyniki tak jakby opisywana zależność rzeczywiście miała miejsce.

że ostatecznie teoria ta dołączy do reszty stylów w „lamusie historii”, gdy ostatni badacze stracą nią zainteresowanie.

W sukurs twierdzeniom o meta-kryzysie psychologii zdaje się też przychodzić niedawny artykuł Greenwalda (2012), w którym przedstawił on 12 ważnych sporów teoretycznych w psychologii. Większość z nich wywodzi się jeszcze z lat 60 i 70. Tabela X zawiera ich listę. Zdaniem Greenwalda do dziś żaden z nich nie został rozstrzygnięty (w przypadku jednego, o naturę reprezentacji wyobrażeń w umyśle, większość badaczy opowiada się za tym, że spór jest rozstrzygnięty, choć ciągle są głosy przeciwne). Kryterium nierozstrzygnięcia sporu jest brak zasadniczego konsensusu w najnowszej literaturze.

Greenwald pokazuje więc, że czasami postęp w psychologii odbywa się w bardzo powolnym tempie. Jego tekst jest o tyle wart wspomnienia, że spory te dotyczą, wydawać by się mogło, konceptualnie bardzo prostych problemów, wokół których istnieje niezły kanon danych empirycznych, co do których nie ma większych kontrowersji (co chyba na ogół nie jest normą w naszej nauce). Z drugiej jednak strony warto zastrzec, że nierozstrzygalność tych sporów może też wynikać z ich dużej poznawczej złożoności, gdyż często dotyczą one pytań o elementarny sposób funkcjonowania umysłu, a takich pytań nie da się wprost łatwo przetestować. Wierzę jednak, że pytania, jakie na ogół spotyka się w literaturze (np. „czy zagrożenie stereotypem pogarsza poziom wykonania w zadaniach matematycznych”) nie mają tak wyrafinowanej złożoności teoretycznej i możliwe jest na nie bardziej pewna odpowiedź.

Przykłady podane przez Greenwalda świadczą o umiarkowanej dojrzałości psychologii w sensie zdolności do znajdowania odpowiedzi na ważne pytania teoretyczne. Mogą jednak świadczyć o pewnej *dojrzałości* rozumianej jako zdolność do powstrzymania się od sądu, gdy dane na to nie pozwalają. To, że 12 sporów przedstawionych przez Greenwalda jest nierozstrzygniętych, może wskazywać na to, że psychologowie nie spieszą się z odpowiedzią na jakieś pytanie, gdy stopień pewności tej odpowiedzi byłby umiarkowany. I jest to objaw jak najbardziej zdrowej nauki.

TABELA 3.1. *Nierozstrzygnięte spory teoretyczne w psychologii wg Greenwalda (2012)*

Czy język i kultura wpływają na proces kategoryzacji obiektów?
Jaką strukturę w umyśle mają emocje: dwubiegunową czy też istnieją dwie niezależne dymensje: pozytywna i negatywna?
Czy zjawisko zmiany postaw pod wpływem wykonywania czynności, które są niezgodne z tymi postawami (jak w eksperymentach Festingera) daje się najlepiej wyjaśnić poprzez zjawisko dysonansu poznawczego, zmiany samopostzegania czy też wywierania wrażenia na innych?
Czy przeszukiwanie w pamięci krótkotrwałej odbywa się równolegle czy szeregowo?
Czy uczenie utajone odbywa się poprzez asocjacje czy przyswajanie zasad?
Czy mentalne reprezentacje wzrokowe mają naturę analogową czy propozycjonalną?
Czy proces prymowania semantycznego odbywa się poprzez rozprzestrzeniającą się aktywację czy poprzez bezpośredni związek prym i słów docelowych (ang. <i>compound cueing</i>)?
Który model (cech, egzemplarzy, prototypów, klasyczny) najlepiej opisuje proces kategoryzacji?
Czy ludzie podejmują działania altruistyczne kierując się motywacją zewnętrzną czy wewnętrzną?
Czy proces zmiany śladów pamięciowych pod wpływem subtelnych sugestii (np. jak w eksperymentach pamięciowych Loftusa) odbywa się w ten sposób, że oryginalne ślady pamięciowe ulegają zmianie, czy w ten sposób, że powstają niezależne ślady pamięciowe, które „wypierają” oryginalne informacje?
Czy podczas wydawania sądów przy niedostatecznej ilości informacji przeważają heurystyki i zniekształcenia poznawcze czy racjonalne analizy?
Co ma pierwszeństwo w umyśle: afekt czy poznanie?
Która teoria wyjaśnia fakt rozdzielenia zjawisk pamięciowych w umyśle: modułów, procesów czy „progów”?

Fundamentalnie jednak powyższe trudności mogą brać się z tego, że psychologia nie zajmuje się takimi relatywnie prostymi fenomenami jak fizyka i chemia. Wybitny biolog ewolucyjny Ernst Mayr (2004) wskazuje, że świat fenomenów biologicznych jest nieporównywalnie bardziej złożony niż świat fenomenów fizycznych. Najprostsza komórka jest systemem o stopniu złożoności niespotykanym nigdzie w świecie zjawisk fizycznych (stąd lepiej rozumiemy przeszłą i przyszłą ewolucję wszechświata niż np. budowę komórek). Idąc dalej, można zasadnie zakładać, że 100 mld komórek mózgu wchodzących w interakcje ze sobą, innymi mózgami, oraz środowiskiem są systemem o jeszcze większej złożoności. Dlatego, psycholodzy raczej nie będą odnosić tak spektakularnych sukcesów jak fizycy. Tematowi dojrzałości nauki poświęcam następny podrozdział.

3.3. Porównanie dojrzałości psychologii na tle innych nauk (szczególnie przyrodniczych)

W przytoczonych w rdz. 3.1 cytatach mała dojrzałość psychologii jest wprost (lub domyślnie) zestawiana z ideałem nauki bardziej „rozwinętej”, i przykładem może być tutaj fizyka. Cechy fizyki takie jak precyzyjne przewidywania (a więc i łatwość falsyfikacji), spójność pojęć, wysoce precyzyjne narzędzia pomiarowe, łatwość wyodrębniania poszczególnych zmiennych i ich wpływu na efekt itp. są z pewnością czymś pożądanym, mogą jednak być nieosiągalne dla psychologii w jakiegokolwiek przewidywalnej przyszłości, nie ze względu na ograniczenia poznawcze psychologów, ale ze względu na dużo bardziej złożony i efemeryczny przedmiot badania. Dlatego w tym podrozdziale zajmę się krótko hipotezą hierarchii nauk oraz przeglądem danych empirycznych jej poświęconej. Warto wiedzieć, jak psychologia ma się na tle nie tylko „twardych” nauk, ale też innych nauk „miękkich”, szczególnie społecznych i humanistycznych.

Ideę hierarchii nauk zaproponował już Auguste Comte w XIX wieku. Intuicyjnie można łatwo dostrzec, że fizyka radykalnie odróżnia się od psychologii, gdyż mniejszą rolę odgrywają w niej uwarunkowania kulturowe, jest bardziej zmatematyzowana czy ogólnie „trafniejsza” (np. porównując precyzję, z jaką kieruje się sondy kosmiczne na odpowiednie trajektorie do precyzji przewidywania zachowania człowieka). Fizycy mają też większy konsensus, co jest ważne w ich nauce i zwykle wskażą np. na ważność mechaniki kwantowej i teorii względności. Psycholodzy nie mają tak dużej zgody, o czym świadczą np. skrajnie różne oceny psychoanalizy czy myślenia ewolucyjnego w psychologii. Zaproponowałbym jeszcze jeden, znany mi z doświadczenia wskaźnik dojrzałości. Będąc studentem psychologii miałem poczucie, że poszczególne przedmioty na studiach łączą się ze sobą w niewielkim stopniu i wiedza z jednego przedmiotu nie jest raczej wspominana/używana na przedmiotach w następnych semestrach. Wg informacji, jakie posiadam, sprawa ma się na nich zgoła inaczej np. na technologii żywienia, gdzie zaawansowana wiedza z chemii organicznej z pierwszych lat jest warunkiem absolutnie niezbędnym do rozumienia treści na zajęciach w późniejszym etapach studiów⁵⁸.

Gdy mówimy więc o „dojrzałości” nauki zakładamy, że nauki dzielą się na bardziej i mniej dojrzałe. Przy czym „dojrzałość” dotyczy w tym kontekście wartości naukowej i

⁵⁸ Wiedzę tę pozyskałem od informatorki, której dziękuję. Sam nie jestem w stanie ocenić jej trafności.

stopnia pewności generowanych idei, a nie „dojrzałości” instytucjonalnej (umocowania na uczelniach, liczby naukowców itd.). Co ciekawe, intuicje te można uchwycić też przy pomocy bardziej zobiektywizowanych metod rzeczywiście wskazujących na występowanie hierarchii nauk. Bodaj najciekawsza z takich prób została podjęta przez Simontona (2004). Zebrał on badania dotyczące głównych 5 nauk, wykorzystując rozmaite mierzalne wskaźniki ich „dojrzałości”. Co świadczy o poziomie dojrzałości jest bardziej kwestią intuicji, a nie „obiektywnej” prawdy, tym niemniej gdy już ustalimy jakie są wskaźniki dojrzałości i je sensownie zoperacjonalizujemy, można je już w miarę obiektywnie zbadać. Simonsohn podzielił wskaźniki dojrzałości na główne i pomocnicze. Główne to:

- stopień konsensusu wśród ekspertów dotyczący tego, kto wniósł istotny wkład w rozwój nauki (im wyższy, tym bardziej dojrzała nauka),
- stosunek praw do teorii w podręcznikach akademickich (im więcej praw, tym większa dojrzałość),
- ilość młodych naukowców z ponadprzeciętną liczbą cytowań (im wyższy tym bardziej dojrzała nauka, gdyż w dojrzałszej nauce młodzi naukowcy przekazując światu wartościowe odkrycia nie muszą aż tak bardzo posiłkować się prestiżem, tytułami lub nagrodami branżowymi),
- produktywność naukowców po dłuższej przerwie w karierze (im bardziej dojrzała nauka tym mniejsza produktywność, gdyż badacz potrzebuje więcej czasu, żeby zapoznać się z najaktualniejszym stanem wiedzy),
- stopień konsultacji zoperacjonalizowany jako liczba badaczy, którzy wymienieni są w podziękowaniach (zgodnie z teorią porównań społecznych, im ludzie są mniej pewni swojej wiedzy, tym częściej powołują się na „społeczny dowód słuszności”; mniejsza liczba przywołań innych badaczy byłaby więc wskaźnikiem większej dojrzałości),
- rozpowszechnienie wykresów (im większe tym większy stopień dojrzałości),
- koncentracja cytowań (jeśli badacze odwołują się często do kilku artykułów można przyjąć, że panuje wśród nich konsensus, co do tego, że rzeczony

artykuły stanowią istotny wkład w rozwój nauki; w mniej dojrzałych naukach większe rozproszenie cytowań byłoby wskaźnikiem braku konsensusu⁵⁹).

Simonton wyodrębnił również 5 pomocniczych wskaźników, takich jak subiektywnie spostrzegana przez ekspertów „twardość” nauki, szybkość cytowań (w bardziej dojrzałych naukach potrzeba mniej czasu na ocenę, co stanowi wartościowy wkład), językowa płynność badaczy mówiących o swojej dyscyplinie (wskaźnik oparty o ciekawe odkrycie Schachtera i in., którzy wykazali, że w bardziej dojrzałych dziedzinach wykładowcy mówią bardziej płynnie o swoich obszarach, pomimo tego, że nie ma różnicy w poziomie płynności przy zwykłej towarzyskiej rozmowie), przewidywalność badań (im bardziej dojrzała nauka, tym częściej określone odkrycia zostały wcześniej przewidziane), oraz wiek laureatów nagrody Nobla.

Zebrawszy rozrzucone po literaturze naukometrycznej opracowania na temat wielkości w/w wskaźników, Simonton stworzył jeden metawskaźnik, który wykazywał niezłą wewnętrzną spójność. Nauki układają się pod względem tak zoperacjonalizowanej dojrzałości w przewidywalny sposób: fizyka, chemia, biologia, psychologia, socjologia, przy czym fizyka i chemia były do siebie względnie zbliżone, w połowie skali znajdowała się biologia z blisko umiejscowioną psychologią, socjologia wyraźnie odstawała.

Nietrudno zauważyć, że kolejność ta pokrywa się ze stopniem matematyzacji poszczególnych nauk, stopniem niezależności nauk od uwarunkowań ideologicznych i kulturowych, stopniem złożoności badanych przez nie obiektów oraz ich następstwem czasowym i przyczynowo-skutkowym w historii wszechświata (tj. materia elementarna, materia złożona, życie, umysł, społeczeństwo/kultura).

Fanelli i Glanzel (2013) przeprowadzili analizę wg innej metodologii, która przyniosła jednak bardzo podobne wyniki (i którą nazwali „bibliometrycznym dowodem na hierarchię nauk”). Przeanalizowano bazę 29 tysięcy artykułów, podzielonych na 12 dyscyplin naukowych⁶⁰. Autorzy wyróżnili szereg prostych właściwości tekstów, pod kątem których zostały automatycznie przeanalizowane przez komputry. Założono między innymi, że w bardziej dojrzałych naukach artykuły powinny średnio:

⁵⁹ Ten wskaźnik budzi moje największe wątpliwości, gdyż można by argumentować, że w mało dojrzałej nauce badacze często przywołują jakieś teksty, kierując się względami niemerytorycznymi typu „prestż” określonego badacza w środowisku

⁶⁰ W porównaniu z podziałem Simontona (2004) dodano kilka nauk (np. ekonomię, matematykę), a niektóre podzielono na subdyscypliny (np. biologię podzielono na biologię molekularną, biochemię, zoologię i botanikę).

- mieć większą liczbę autorów, gdyż nowe odkrycia w tych naukach wymagają współpracy dużych zespołów i osób o dosyć wąskich specjalizacjach,
- być krótsze, gdyż mniej miejsca potrzeba na wprowadzanie do tematu, wyjaśnianie i uzasadnianie planu badawczego,
- ale mieć dłuższy tytuł, gdyż więcej słów potrzeba na wyartykułowanie konkretnego i jasnego problemu badawczego,
- mieć mniejszą liczbę źródeł do których się odwołują w bibliografii, z tych samych względów co punkt b (Simonton (2004) też stwierdził, że w mniej dojrzałych naukach badacze częściej używają „społecznego dowodu słuszności”, wymieniając większą liczbę osób w podziękowaniach. Możliwe, że cytowanie większej liczby tekstów jest rodzajem podobnego mechanizmu, dodatkowo gwarantującego przychylność większej liczby potencjalnych recenzentów),
- zawierać mniej odwołań do monografii, gdyż koncentrują się na konkretnych zagadnieniach, więc nie trzeba odwoływać się do ogólnych podręczników,
- posiadać nowsze odwołania w bibliografii, gdyż łatwiej budować nowe badania na wcześniejszych, spory teoretyczne są szybciej rozwiązywane i następuje w miarę ciągła kumulacja wiedzy (wskaźnik identyczny jak u Simontona),
- dzielić ze sobą większą liczbę cytowanych źródeł, jeśli artykuły są o podobnym profilu. Podobnie Simonton (2004) stwierdził, że cytowalność artykułów w bardziej dojrzałych naukach ma większą wariancję, gdyż istnieje kilka tekstów kluczowych, na którą duża część prac się powołuje; tymczasem w naukach o mniejszym stopniu dojrzałości panuje mniejszy konsensus, co jest istotnym wkładem w rozwój dziedziny, a więc i większe rozbieżności w wyborze tekstów, na które warto się powoływać,
- rzadziej używać pierwszoosobowych zaimków osobowych, gdyż twierdzenia są bardziej obiektywne

Wszystkie te przewidywania się potwierdziły⁶¹ i nauki utworzyły następującą „hierarchię”: matematyka, astrofizyka, fizyka, chemia, biologia molekularna, biochemia, zoologia i botanika, psychologia i psychiatria, ekonomia, nauki społeczne, a na końcu nauki humanistyczne i sztuka. Opracowanie to jest oczywiście czysto

⁶¹ Oczywiście w obrębie poszczególnych kryteriów znajdowały się tzw. wyjątki od reguły, ale były one nieznaczne.

ilościowe, trudno zresztą wyobrazić sobie jakościową analizę takiej liczby tekstów (także ze względu na ograniczone kompetencje pojedynczej osoby, aby oceniać teksty z tak różnych dziedzin, być może rozwój sztucznej inteligencji ułatwi w przyszłości takie analizy). Można się zastanawiać nad każdym kryterium, oraz nad tym na ile takie względnie proste analizy mogą oddawać złożoność nauki. Tym niemniej pomiędzy tymi kryteriami występowała bardzo duża zgodność, tj. wyższy wynik określonej na jednym, na ogół wpływał na wyższy wynik na innym.

Konsensus jest istotnym czynnikiem w obydwu rankingach. W nauce w ostatecznym rozrachunku ocena wartości badań i teorii opiera się na zdaniu ekspertów (nie mamy póki co komputerów generujących „obiektywną” wartość jakiegoś wycinka nauki). Dlatego warto tu jeszcze wspomnieć o danych wskazujących na umiarkowany konsensus w psychologii.

Z jego braku badacze często posilkują się kryteriami zgoła niemerytorycznymi, takimi jak (rzekomy) prestiż (lub jego brak) autorów tekstów oraz instytucji z których się wywodzą. Peters i Ceci (1982) przeprowadzili prowokację, polegającą na wysłaniu do dwunastu prestiżowych czasopism psychologicznych artykułów, które ukazały się w tych samych czasopismach kilkadziesiąt miesięcy wcześniej. Zmieniono nieznacznie tytuły i abstrakty, oraz podpisano je nazwiskami nieznanych autorów z nieznanych instytucji o podejrzanych nazwach (polskim odpowiednikiem mógłby być „Instytut Holistyczny Człowieka z Pcimia Dolnego”). Artykuły te zostały bez wyjątku odrzucone, jednak najczęściej (75%) nie ze względu na wykrycie próby oszustwa, ale z powodu, jak twierdzili redaktorzy, „poważnych błędów metodologicznych”. Nie stanowiło to jednak przeszkody, aby te same artykuły opublikować kilka miesięcy wcześniej, gdy autorzy byli znani i pochodzili z prestiżowych uniwersytetów. Oczywiście, prowokacja ta dawałaby więcej informacji o psychologii na tle innych nauk, gdyby równolegle autorzy wysyłali artykuły do czasopism biologicznych, chemicznych czy fizycznych.

Innym pośrednim dowodem na niższy konsensus (nieuwzględnionym jednak przez Simontona) jest bardzo wysoki stopień odrzutów występujący w dobrych czasopismach psychologicznych, szczególnie na tle dużo bardziej dojrzałej fizyki (Suls, Martin, 2009). W psychologicznych czasopismach „wysokiej półki” nawet 80-90 % artykułów jest odrzucanych. Może to świadczyć o tym, że w fizyce między autorami tekstów i recenzentami panuje większa zgodność, co do tego, co stanowi

wartościowy wkład w naukę, a więc autorzy są pierwszymi cenzorami swoich potencjalnych tekstów, i mają trafniejsze rozeznanie, które badania mają szansę być opublikowane.

Podsumowując ten podrozdział, w obydwu rankingach pozycja psychologii jest powodem do relatywnego optymizmu. Choć znalazły się za naukami przyrodniczymi (co było do przewidzenia), to jednak okazały się nieformalnym liderem nauk „miękkich”. Być może przyczyną jest wskazywany przez niektórych fakt, że psychologia na tle innych nauk społecznych jest raczej wolna od postmodernistycznego relatywizmu epistemologicznego. Psychologia znalazła się natomiast poniżej innych nauk przyrodniczych i nie jest to specjalnie zaskakujące. Rankingi te zdają się potwierdzać intuicje, wskazujące na to, że przedmiot psychologii nie pozwala uzyskiwać tak szybkiego postępu wiedzy.

Mając pełną pokory świadomość ograniczeń, w dalszej części zastanowię się, czy mimo wszystko badacze nie mogliby lepiej wykorzystywać możliwości, jakie mają. Przedstawię propozycję modelu rozwoju psychologii. Zacznę od przyjrzenia się intrygującej propozycji Rozina, który twierdzi, że problemy psychologii mogą brać się z tego, że zbyt szybko chce osiągnąć status dojrzałej nauki.

3.4. Dojrzałość nauki a metoda – argumentacja Rozina

Zanim przedstawię myśl Rozina, pozwolę sobie (nie pierwszy raz) na pewne osobiste refleksje. Kilka lat po studiach psychologicznych dostrzegam, że wykładowcy olbrzymią większość czasu poświęcali omawianiu tzw. „teorii”. I to teorii dosyć wysokiego poziomu, a w bardzo niewielkim stopniu koncentrowali się na danych empirycznych, wspominając o nich marginalnie, albo wręcz w ogóle nawet ich nie wymieniając⁶² (modelowym przykładem takiego postępowania jest kultowy w pewnych kręgach podręcznik rozwoju patologii osobowości „Style charakteru” Stevena Johnsona). W pewien sposób trudno się dziwić – zebrać sporo solidnych danych potwierdzających daną teorię jest trudno, i na pewno jest to trudniejsze niż samo sformułowanie teorii i jej atrakcyjne przedstawienie (podobnie jak kariera polegająca na „dogłębnym rozumieniu Freuda” może z większą szansą zakończyć

⁶² Pamięć ludzka jest podatna na wielorakie zniekształcenia i być może w rzeczywistości nie było tak źle. Jednak przegląd sylabusów wiele lat po studiach wskazuje mi, że artykuły empiryczne są praktycznie niespotykane w nich, podobnie jak meta-analizy.

się sukcesem, niż kariera polegająca na poszukiwaniu dowodów na tegoż twierdzenia). Na tej samej zasadzie „Inteligencja emocjonalna” Golemana była dużym sukcesem komercyjnym, gdyby jednak Goleman podjął normalną pracę badacza uniwersyteckiego, żmudnie zbierającego dane, domyślam się, że efekty nie byłyby tak spektakularne, a jego sytuacja finansowa nie byłaby tak dobra.

Ta pewna „teoriocentryczność” może dotyczyć nie tylko wykładowców jednego z uniwersytetów, ale może się generalizować do psychologii w ogóle. Nie w takim sensie, że psycholodzy zapominają o danych. Oczywiście zbierają dane empiryczne i to monstrualną ilość. Wydaje się jednak, że olbrzymia większość danych powstała przez „testowanie” wcześniej wyartykułowanych modeli teoretycznych, a przynajmniej tak jest przedstawiana. Często do niespodziewanych odkryć jest dorabiana ideologia post factum (Kerr, 1998). Jednak wydaje się, że nawet te niespodziewane odkrycia biorą się z testowania konkretnych hipotez, a nie obserwowania rzeczywistości w sposób bardziej eksploracyjny.

Spotkałem się w życiu uniwersyteckim z niepisaną zasadą, że prawdziwy psycholog-badacz musi mieć przed badaniem złożony model teoretyczny, który wyjaśnia określoną badaną działkę. Model ten oczywiście może okazać się błędny, i to jest akceptowane, ale nie jest raczej mile widziane podejście na zasadzie „nie wiem jeszcze co wpływa na X, ale gdy poobserwuje to zachowania, mam nadzieję to wywnioskować”. Wydaje się, że ten sam przymus stoi też za zjawiskiem przedstawiania wyników badań jako przewidzianych teoretycznie od początku, pomimo tego, że tak naprawdę teoria została dorobiona do nich, gdy badacz już znał wyniki, tym samym nadając „przewidującej” je teorii nieuprawnionej wartości (Kerr, 1998).

Dlaczego problemem może być posiadanie daleko idących teorii na wstępie i program badawczy polegający na testowaniu kolejnych hipotez? Dlatego, że może to ograniczać pole widzenia rozumianego zjawiska (Greenwald, Pratkanis, Leippe, Baumgardner, 1986). Przystępując do badania korelacyjnego lub eksperymentu, siłą rzeczy wybieramy bardzo niewielki podzbiór wszystkich możliwych zmiennych. Jeśli wyniki są „istotne”, może to nas zwalniać z chęci szukania dalszych moderatorów, o próbach falsyfikacji nie wspominając. Oczywiście, jeśli dysponujemy dobrze przetestowanymi teoriami i chcemy wyjaśnić jakieś nieścisłości eksperyment nie jest niczym złym. Ale w nauce o niższym stopniu rozwoju nie musi to być dobra strategia.

Darwin nie stworzył teorii doboru naturalnego poprzez projektowanie badań, które ją przetestują. Zbierał ogromną ilość danych (przede wszystkim z zapisu kopalnego, z efektów doboru sztucznego i biogeograficznego rozmieszczenia organizmów), po to, żeby znaleźć w tych danych wzór. Efektem była książka „O powstaniu gatunków”, jedno z najwybitniejszych ludzkich osiągnięć intelektualnych w dziejach (Darwin 1859/2009).

Myśl, że psychologia może wyprzedzać swój czas poprzez nadmierne polegania na eksperymencie i testowaniu formalnych teorii, stanowi istotę inspirującego artykułu, badacza o biologicznym „backgroundzie” Paula Rozina (2001). Rozin upatruje słabości psychologii społecznej w nierobieniu tego, co jest podstawą mało rozwiniętej nauki, a więc szerokiej, dogłębnej obserwacji i opisu fenomenów (w rozdziale 3.10 przeanalizuje pewne nieścisłości związane z pojęciem „obserwacji” i „opisu” w postulacie Rozina, ale w tej chwili nie jest to istotne). Oczywiście, Rozin nie kwestionuje roli eksperymentów i modeli jako takich, ale twierdzi, że psychologowie za szybko do nich zmierzają pomijając rzeczy bardziej podstawowe (przywołuje metaforę Ascha, który stwierdził, że nie można przyspieszyć czasu, przesuając wskazówki). Oddając głos Rozinowi:

Przedwczesne próby sztucznego zwiększenia dojrzałości nauki ograniczają kreatywność, zamykają oczy na ważne nowe fenomeny, powodują, że dziedzina będzie się angażowała w długie programy badawcze, które ostatecznie będą miały niewiele wspólnego z podstawowym celem badań (tj. światem społecznym) oraz ogólnie odciągają badaczy od świata realnego świata, gdzie wszystko się zaczęło. (Rozin, 2001, s. 5)

Patrząc historycznie – zdaniem Rozina – badania w naukach przyrodniczych zwykle zaczynały się od próby systematycznego opisu odkrytego zjawiska, jego uniwersalności (i anomalii), a dopiero potem następowały próby znalezienia regularności, czy wreszcie praw. Jako przykład podaje sformułowanie prawa gazów Boyle’a i praw genetyki Mendla. Badania na których te prawa oparto nie były pierwotnie motywowane chęcią weryfikacji określonego modelu, ale raczej próbą jak najdokładniejszego opisania interesujących fenomenów.

Rozin nie ogranicza się tylko do uwag o naturze historyczno-publicystycznej, ale też przeprowadza analizę zawartości kilku numerów JPSP oraz trzech czasopism biologicznych z górnej półki (*Cell*, *American Journal of Physiology*, *Development*). Wynika z niej, że w JPSP w około 70 % badań uczestniczą tylko studenci

amerykańscy; bardzo rzadko wymienia się religię, rasę, religię badanych, albo to w jakiej porze roku odbywało się badanie; większość badań inspirowanych jest wyartykułowanym modelem lub hipotezą (70 % w obszarze poznania społecznego); posiada statystyki inferencyjne (100%!) i nie zawiera żadnych obserwacji. Dla porównania badania w czasopismach biologicznych bardzo rzadko są inspirowane modelem teoretycznym (ok. 20 %). Najczęstszą inspiracją jest chęć opisu procesu „jak X wpływa na Y” (np. 62 % artykułów w czasopiśmie *Cell*) lub „jak X zmienia się w czasie od t_1 do t_2 ” (43 % w *Development*). Statystyki inferencyjne spotyka się relatywnie rzadko (0 % badań w czasopiśmie *Cell*, 14 % w *Development*, ale 85 % w *American Journal of Physiology*). Dodatkowo czasopisma biologiczne regularnie replikują określone prawidłowości na podstawie danych dotyczących wielu gatunków (biologiczny odpowiednik używania różnych grup społecznych i kulturowych w psychologii).

Autor wytyka psychologii, szczególnie społecznej, inne dobrze znane błędy m.in. korzystanie w badaniach z bardzo jednorodnych i „niszowych” grup (studentów psychologii i członków cywilizacji zachodniej), które są następnie generalizowane na całą populację ludzką, nie uwzględnianie kontekstów i nie wymienianie ich w badaniach z psychologii społecznej (np. pora dnia, pora roku, miejsce itd. itp.), tak jakby badania z tej dziedziny były oderwane od reszty rzeczywistości i te informacje nie miały żadnego znaczenia. Ponownie oddając głos Rozinowi:

W przeciwieństwie do odnoszących sukcesy nauk przyrodniczych czy psychologii percepcji, badania w psychologii społecznej nie zostały poprzedzonej gruntowną obserwacją i zbieraniem informacji o istotnych fenomenach oraz opisem uniwersalnych stałych. W bardziej dojrzałych naukach, które psychologia chciałaby naśladować, większy nacisk kładzie się na opis fenomenów, jak również w mniejszym stopniu polega się na eksperymentach. Takie nauki, szczególnie nauki biologiczne, zwracają mniejszą uwagę na modele i hipotezy, a większą na dane, w przeciwieństwie do „dowodów” lub „badań definitywnie potwierdzających hipotezę”. W badaniach nad zachowaniem człowieka w sytuacjach społecznych, w których efekty kontekstualne są liczne, a istotne procesy złożone, zebranie danych, które jednoznacznie potwierdzą hipotezę jest niezwykle trudne (Rozin, 2001, s. 3).

Uwagi Rozina skłaniają do refleksji. W nauce o zachowaniu zwierząt, nie ma możliwości, żeby rozdać „badanym” plik kwestionariuszy lub posadzić przed komputerem i kazać naciskać guzik. Badacze muszą więc obserwować i czasami spotykamy badania oparte na setkach lub tysiącach godzin obserwacji (przykładowo,

badanie transmisji kulturowej pewnych zwyczajów u makaków japońskich oparto na 2 tys. godzin obserwacji: Huffman, Nahallage, Leca, 2008). Analogicznych badań jest w psychologii jak na lekarstwo. Czasami się zdarzają np. badanie Fraley i Shaver (1998), w którym obserwowano pożegnania ludzi na lotnisku (analizowano je potem w kontekście teorii przywiązania), albo badanie Dabbs, Hargrove, Heusel, (1996), gdzie przy okazji innych badań opisano zaobserwowane różnice w zachowaniu członków bractw studenckich, uważanych za „kujonów” i „impreszowiczów” (zasadniczo badano poziom testosteronu, ale raporty behawioralne też stanowią integralną część artykułu). Ale nawet w powyższych przypadkach nie można mówić o tysiącach godzin obserwacji. Ostatnio ukazało się kilka badań, w których na podstawie danych z ich telefonów komórkowych analizowali wzory ludzkiego przemieszczania się (Song, Qu, Blumm, Barabási, 2010). Jest to niewątpliwie potężna ilość danych behawioralnych, ale badacze zasadniczo w olbrzymiej większości pracy wyręczył komputer (podobne rzeczy można powiedzieć o rozlicznych badaniach, w których analizuje się zachowanie ludzi na portalach społecznościowych: Wilson, Gosling, Graham, 2012). Brakuje w psychologii zwykłej, tradycyjnej pracy antropologicznej, a tego typu praca, która jest wykonywana w pokrewnych dyscyplinach (np. socjologii) nie przebija się do psychologów.

Rzecz jasna, Rozin nie mówi „róbmy wszystko tak samo jak badacze komórek”. Zmusza raczej do zastanowienia, czy to co robimy teraz zawsze dobrze nam służy. Nie od dziś wiadomo, że nadmierne przywiązanie do modeli teoretycznych może ograniczać rozwój (Greenwald, Pratkanis, Leippe, Baumgardner, 1986; modelowym, wręcz „klinicznym”, przypadkiem takiego upośledzenia jest bodajże psychoanaliza). Podobnie trudno nie zgodzić się z Rozinem w stwierdzeniu, że stosowanie eksperymentu, gdy wiemy mało o badanym zjawisku, zapewne niewiele poszerzy naszą wiedzę. Jeśli eksperymentalnie stwierdzimy, że X wywiera wpływ na Y (na ogół w badaniach psychologicznych ów „X” wywiera wpływ na kilka, kilkanaście procent wariancji „Y”), w dalszym ciągu niewiele wiemy co oprócz X może jeszcze wpływać na Y, w jaki sposób X wchodzi w interakcję z innymi zmiennymi wywołującymi Y, nie mówiąc już o najbardziej elementarnym pytaniu: jaki podstawowy proces powoduje, że X wywołuje Y. Jest zapewne możliwe, że jakiś genialny badacz wygeneruje i przetestuje wszystkie rozsądne hipotezy co do zmiennych wywołujących Y, choć przypuszczam, że jest to bardzo mało

prawdopodobne. Sądzę raczej, że etap gruntownego opisu określonego zachowania z większym prawdopodobieństwem zaowocuje trafnymi hipotezami, co do potencjalnych moderatorów i zmiennych istotnie modyfikujących dany efekt niż sytuacja, gdy badacz angażuje się tylko w badanie eksperymentalne z celem potwierdzenia lub falsyfikacji gotowego modelu, a wiedzę o efekcie czerpie „z głowy” i wcześniejszych teorii (które, jak wiadomo, w naszej nauce są różnej jakości).

W innym tekście Rozin (2006) zauważa, że w podręcznikach akademickich niezwykle rzadko znajdujemy odpowiedzi na pytania typu: „o czym zwykle ludzie rozmawiają”, „w jaki sposób spędzają czas wolny”, „na co wydają pieniądze”, „jak są sposoby okazywania sympatii innym”, „jakie są typy argumentów używanych w dyskusjach” itd. Analizując skrupulatnie kilkanaście podręczników akademickich z psychologii społecznej, rozwojowej i wprowadzeń do psychologii, Rozin pokazał, że dużo więcej miejsca zajmuje w nich temat agresji i uprzedzeń niż temat jedzenia, religii czy czasu wolnego razem wzięte (tematy agresji i uprzedzeń są społecznie bardzo ważne, ale na pewno nie tak powszechne jak dalej wymienione). Mówiąc ściślej, te ostatnie tematy zwykle nie pojawiają się w ogóle. Problem oderwania psychologii od codziennych ludzkich zachowań jest logicznie patrząc inną kwestią niż kwestia metod, ale zapewne mającą wspólne źródła. Bez podstawowej pracy obserwacyjnej trudno udzielić odpowiedzi na powyższe „banalne” pytania. Możliwe są co prawda ankiety, ale mają dobrze znane ograniczenia (rdz. 2.4). Sądzę, że solidna wiedza opisowa o podstawowych obszarach ludzkiego zachowania mogłyby stanowić użyteczny korpus dla budowania teorii na temat ludzkiego zachowania (coś na kształt korpusów językowych w językoznawstwie), zapewne będzie też łatwiej doprowadzić do u wspólnienia języka badając rzeczy tak podstawowe.

Wspominałem wcześniej o badaniu, w którym zachowanie badanych nagrywano w ich domach (Wang, Repetti, 2014) oraz o badaniu, w którym analizowano reakcje na zdradę w programach rozrywkowych (Kuhle, 2011). Te dwa badania zainspirowały mnie do poszukania badań, w których ludzkie zachowanie byłoby systematycznie monitorowane przez dłuższy okres czasu (dni i tygodnie).

Teoretycznie, badacze nie musieliby nawet sami przeprowadzać takich badań, istnieje wszak sporo tzw. „reality show” i innych materiałów, w których monitorowano długotrwale zachowanie. Ludzie w takich programach mogą zachowywać się często w sposób mało naturalny, a i sama próba może mocno odbiegać od typowej w

populacji. Tym niemniej, jestem w stanie sobie wyobrazić dziesiątki hipotez dotyczących międzysytuacyjnej zmienności/stałości zachowania, które można by sprawdzić dzięki takim widowiskom. Zapewne nawet programy Ewy Drzyzgi mogłyby stanowić przebogate źródło danych behawioralnych na temat tego, jak ludzie wyjaśniają swoje zachowanie itp.

Innym źródłem danych behawioralnych mogą być badania wykonywane w ramach badania kosmonautów. Jednym z takich projektów był Mars 500, gdzie grupa 6 mężczyzn spędzała 500 dni w warunkach, które miały symulować wyprawę na Marsa. Badani byli cały czas obserwowani i nagrywani. O tych badaniach dowiedziałem się z popularnonaukowego filmu o planowanej wyprawie na Marsa, i od razu pomyślałem, że potencjalnie stanowi ono niewyczerpane źródło danych behawioralnych.

Tak więc przejrzałem bazę PsycINFO używając różnych kombinacji pytań⁶³. Choć pojawiały się badania wokół niektórych programów (np. Oksanen, 2014 analizował jak jest przedstawiane uzależnienie w „show”, w którym uczestnicy wychodzili z nałogów), to nie znalazłem żadnych badań, w których wykorzystywano by systematyczne dane obserwacyjne z reality show. W bazie PsycINFO znalazłem łącznie 8 artykułów z badaniami z projektu Mars 500. 5 z nich dotyczyło zmiennych fizjologicznych (np. wzory snu), 2 zmiennych poznawczych (np. percepcja przestrzeni i rozwiązywanie zadań poznawczych), tylko jedno dotyczyło zmiennych stricte behawioralnych, a dokładnie (Tafforin, 2013). Wątpliwe przy tym, żeby dotarło do wielu psychologów, gdyż ukazało się w czasopiśmie *Aviation, Space, And Environmental Medicine*. Żadne z tych badań nie ukazało się w czasopiśmie stricte psychologicznym.

Te przykłady mają pokazać, że istnieją potencjalnie możliwości, żeby uzyskać dużo podstawowej, opisowej wiedzy na temat ludzkiego zachowania i pośrednio obejść wielorakie ograniczenia praktyczne i etyczne⁶⁴ wynikające z badania studentów w dosyć krótkich sesjach laboratoryjnych. W ostateczności jestem w stanie sobie wyobrazić, że sami psycholodzy mogliby zorganizować eksperyment polegający na ciągłej obserwacji przez tydzień, dwa grupy dobrze opłacanych ochotników, który

⁶³ M.in. „reality show”; „Big Brother” w połączeniu z kombinacjami w rodzaju „data analysis”; „daily behavior”; „behavior analysis” itp.

⁶⁴ „Reality show” są same w sobie wątpliwe etycznie, tym niemniej skoro już zostały przeprowadzone, moralny bilans się nie zmieni jeśli ktoś przeanalizuje dane.

kosztowałby promil tego, co wielkie projekty w naukach przyrodniczych jak CERN czy sekwencjonowanie ludzkiego genomu.

Brakuje mi szerokiej wiedzy z historii rozwoju nauk przyrodniczych i filozofii nauki, żeby z dużym stopniem pewności ocenić postulaty Rozina (2001) i ich ewentualny skutek dla rozwoju nauki. Ma on jednak prawdopodobnie rację, mówiąc, że mamy system badawczy, który jest oparty na testowaniu wyartykułowanych hipotez i nie wykonujemy pracy obserwacyjnej. Pomimo tego, że prawdopodobnie mamy takie możliwości.

3.5. Kumulacja wiedzy w psychologii – wprowadzenie do modelu

Ze względów, które omówiłem w rdz. 3.3 nie sądzę, żebyśmy w psychologii mogli liczyć na tak szybką kumulację wiedzy, jak w naukach przyrodniczych. Nie licząc na szybkie i spektakularne efekty, oczywiście można zasadnie zadawać pytanie, czy obecne praktyki badawcze są optymalne z punktu widzenia możliwości, które mamy. Z całą pewnością kumulacja wiedzy nie będzie możliwa, jeśli większość badań będzie opartych na „łowieniu p” lub innych formach nierzetelności (rozdział 1). Tam, gdzie prawidłowości psychologiczne mają istotne przełożenie na codzienne zachowania, tam nie zaszkodzi również, żebyśmy mieli jak najwięcej danych behawioralnych, a nie danych opartych na zawodnych wskaźnikach (rdz. 2.4 i 2.5). Wszelkim subdyscyplinom psychologii przydałoby się też większa „pula” badanych, szerokie wykroczenie poza badanie studentów (rdz. 2.6) z krajów zachodnich (2.7).

Za punkt wyjścia przyjmuję więc założenie, że rzetelność i trafność badań psychologicznych są warunkami koniecznymi do kumulacji wiedzy (tzw. postępu). Nie są jednak wystarczające. Można sobie wyobrazić sytuację, że pomimo sensownej rzetelności i trafności badań, poszczególne odkrycia są ze sobą bardzo luźno powiązane, nie inspirują się wzajemnie i w niewielkim stopniu przekładają się na wzrost wiedzy w postaci dobrych teorii. Z jednej strony przypadkowe, niezorganizowane zbieranie danych może utrudniać budowanie teorii, które są czymś więcej niż ogólnymi historyjkami, nefalsyfikowalnymi, dopuszczającymi z góry masę wyjątków. Bez dobrych teorii konkretne odkrycia empiryczne są w ostatecznym rozrachunku tylko ciekawostkami, nie mającymi żadnego praktycznego znaczenia. Z drugiej strony nauka to nie tylko indukcyjne budowanie teorii z obserwowanych

zjawisk, ale też testowanie teorii poprzez przewidywanie niebanalnych odkryć. Przewidywanie zjawisk na podstawie teorii jest ich najlepszym sprawdzianem, pozwala je falsyfikować lub przynajmniej odróżniać teorie lepsze od gorszych (ze względu na przedmiot naszej nauki, nie sądzę, żeby większość teorii dało się skonstruować tak, żeby generowały bardzo precyzyjne przewidywania, a więc było łatwo falsyfikowalne i bardziej realistycznym celem wydaje się preferowanie teorii na podstawie ich siły relatywnie do innych teorii). Nie wierzę, żeby mogło istnieć dobre wyjaśnianie bez sensownego przewidywania. Oglądając programy informacyjne można odnieść wrażenie, że dyżurni politolodzy zawsze potrafią objaśnić dowolne zdarzenia na scenie politycznej, nie sądzę, żeby często udawało się im coś niebanalnego przewidzieć.

W dalszej części zarysuję model rozwoju nauki, który będzie podstawą przy empirycznej analizie praktyk badawczych w psychologii (rozdziały 4-6). Model ten jest nie tyle rozprawką z filozofii nauki, ile wstępem do operacjonalizacji potrzebnej na dalszych etapach. Zaznaczam, że nie jestem zawodowym filozofem nauki, więc zapewne mój model będzie bardziej intuicyjny, nie będzie tak zniuansowany jak prace filozofów nauki, mam jednak nadzieję, że będzie użyteczny na potrzeby naszej analizy.

Podstawą tego modelu są dwie koncepcje: prawidłowości różnych poziomów i przedmiotu psychologii. Najważniejsze tezy będę wyróżniał w tekście.

3.6. Model rozwoju psychologii #1 – prawidłowości niskiego i wysokiego poziomu

Przeanalizujemy najpierw jedno badanie z psychologii społecznej (inspirowane ewolucyjnie), które pozwoli zrozumieć ideę prawidłowości różnego poziomu.

Haselton i Buss (2000) w artykule w JPSP zaproponowali teorię zarządzania błędami (tłumaczenie dosłowne moje; ang. *error management theory*), która zakłada, że koszty błędów fałszywie pozytywnych i fałszywie negatywnych w codziennych ludzkich doświadczeniach nie zawsze są jednakowe. Przykładowo, idąc w lesie i słysząc szelest w krzakach, mniejszym błędem będzie założenie, że powoduje go drapieżnik (gdy w rzeczywistości jest to wiatr), niż założenie, że szelest jest wynikiem wiatru, podczas gdy w rzeczywistości w zaroślach ukrywa się drapieżnik. Idealnie

byłoby mieć zawsze stuprocentowo trafne przewidywania, ale w normalnym świecie jest to nieralistyczne, więc w sytuacji niedostatecznej ilości informacji ludzie będą skłaniali się w kierunku takich przewidywań, które mają mniejszy koszt.

Haselton i Buss swoją teorię postanowili empirycznie przetestować na przykładzie dostrzegania zainteresowania seksualnego u osób płci przeciwnej (później uzupełniono ją o wiele innych obszarów np. Haselton, Nettle, 2006). Przyjęto, że mężczyźni, jako płeć mniej inwestująca w ewolucyjnym sensie, będą w większym stopniu interpretowali niejednoznaczne sygnały ze strony kobiet (np. uśmiech) jako przejaw zainteresowania romantycznego, gdyż błąd polegający na niedostrzeganiu zainteresowania romantycznego, tam gdzie ono występuje jest bardziej kosztowny niż dostrzeżenie zainteresowania, tam gdzie go w rzeczywistości nie ma (co jest też pochodną teorii inwestycji rodzicielskich Triversa, 1972). Badani, kobiety i mężczyźni, czytali opis rozmaitych zachowań (fikcyjnych kobiet i mężczyzn) i oceniali na ile świadczyłyby one o zainteresowaniu romantycznym (przykładowo: „kobieta/mężczyzna podchodzi do nowego pracownika (kobiety/mężczyzny) w pracy, uśmiecha się szeroko i zaczyna serdeczną rozmowę”; „kobieta/mężczyzna utrzymuje długi kontakt wzrokowy z dopiero co poznaną osobą na przyjęciu” itp.). Badani mężczyźni w niejednoznacznych gestach dostrzegali wyższe zainteresowanie romantyczne niż kobiety.

Załóżmy, że wiemy, że efekt nie jest fałszerstwem, rodzajem artefaktu, wynikiem przypadku i jest replikowalny przez niezależne zespoły. Mamy jako taką pewność, że opisywane zjawisko występuje i chcemy zapytać, co właściwie wykazano w tym badaniu. Odpowiedź można udzielić na kilku poziomach. Oto przykłady:

- a. poziom najniższy opisuje literalnie, to co się dzieje, w tym wypadku odkryta prawidłowość mogłoby mieć taką postać „mężczyźni, jeśli przedstawić im listę (takich a takich) zachowań ze strony hipotetycznych obcych kobiet, oceniają średnio te zachowania jako świadczące o większym stopniu zainteresowania seksualnego niż kobiety i efekt ma (taką a taką) siłę”
- b. poziom nieco wyższy mógłby bardziej generalizować ponad obecną manipulację eksperymentalną, użyte bodźce, czy wręcz w ogóle metodę i mieć taką postać np. „mężczyźni w niejednoznacznych gestach przeszacowują zainteresowanie seksualne kobiet ich osobą”

- c. poziom jeszcze bardziej abstrakcyjny mógłby generalizować powyższe prawidłowości na jeszcze szersze spektrum zachowań i w takim wypadku „prawo” mogłoby brzmieć „W sytuacji, gdy umysł ludzki tworzy reprezentacje rzeczywistości, szczególnie w oparciu o niedoskonałe dane, spośród dwóch rodzajów błędów (fałszywych pozytywnych i fałszywych negatywnych) często jeden jest mniej kosztowny niż drugi. W takiej sytuacji ludzie będą mieli tendencję do interpretowania rzeczywistości tak, żeby uniknąć bardziej kosztownego błędu”.

(Zaproponowany tutaj ten konkretny podział na kolejne poziomy niekoniecznie musi być jedynym możliwym, albo najlepszym. Zapewne możliwe są inne klasyfikacje, podany przykład służy tylko ilustracji zasadniczej myśli.)

To nie jest koniec „poziomów” i możliwe są oczywiście dalsze poziomy teorii, w tym przykładzie poziomem czwartym mogłaby być jakaś „całościowa teoria zniekształceń poznawczych”, poziomem piątym jakaś metateoria opisująca poznanie społeczne, kolejnym „ogólna teoria poznania”, a jeszcze wyższym może metateoria opisująca zachowanie/naturę człowieka.

W języku potocznym prawidłowości niskiego poziomu są określane jako fakty, obserwacje, wyniki eksperymentów itd., natomiast prawidłowości wyższego poziomu jako teorie lub prawa. W filozofii nauki funkcjonują inne określenia na rozróżnienie różnych poziomów prawidłowości. Ernst Nagel przyjął podział na „prawa obserwacyjne” (ang. observational laws) i „właściwe teorie” (ang. proper theories) (za: Kuipers, 2007). Inne określenia na prawidłowości niskiego poziomu to m.in. prawa empiryczne, eksperymentalne lub fenomenologiczne, ogólne fakty, powtarzalne efekty, indukcyjne uogólnienia (za, Kuipers, 2007, s. 2). Wolę pozostać przy stosowaniu słowa „prawidłowości” (zamiennie z „prawa”), gdyż zakłada ono, że zjawiska (nieważne na jakim poziomie) są zasadniczo powtarzalne⁶⁵, i – do naszych celów – tyle filozofii nauki w zupełności wystarczy. Zakładam też pewną poznawczą otwartość, co do tego ile „poziomów” jest rzeczywiście w naturze oraz ile możemy stworzyć do naszych celów poznawczych. Nie chciałbym więc dekretować określonej liczby poziomów, ale będę używał rozróżnienia na prawidłowości wysokiego i niskiego poziomu.

⁶⁵ Przynajmniej uniwersalne i powtarzalne w takim zakresie zdarzeń jaki postulują. Prawidłowość „mężczyźni pod wpływem Y robią X” może zostać ograniczona przez następne odkrycia i okaże się, że tylko „mężczyźni, którzy są amerykańskimi studentami robią X”. Nie będzie więc taka prawidłowość już uniwersalna, w tym sensie, że dotyczy wszystkich, ale dalej będzie uniwersalna dla tych warunków jakie określa.

Prawidłowości na kolejnych poziomach różnią się od siebie pod kilkoma względami. Po pierwsze prawidłowości niższego poziomu są bardziej pewne. Zakładając, że badanie nie zostało sfałszowane, nie jest dziełem przypadku lub niezidentyfikowanym artefaktów, inni badacze powinni móc je powtórzyć, jeśli w miarę dokładnie odtworzą warunki. Natomiast prawidłowości wyższego poziomu są umysłową interpolacją prawidłowości niższego poziomu i ze swojej natury są bardziej spekulatywne. Przykładowo, z tego, że ludzie popełniają błędy w obszarze spostrzegania zainteresowania romantycznego, nie wynika wprost, że umysł dysponuje mechanizmem kalibrowania rodzajów błędów i wyboru odpowiedniego z nich, nie wynika też, że analogiczne prawidłowości będą miały miejsce dla obszaru np. spostrzegania zagrożenia fizycznego. Jeśli nawet prawidłowości dotyczące spostrzegania romantycznego są prawdziwe, nie implikuje to ich ewolucyjnego pochodzenia, mogły powstać jako wynik socjalizacji do ról płciowych itd. (uzasadnienie, w które nie wierzę, ale chodzi o przykład). Przykład ten ma pokazać, że określone prawidłowości niskiego poziomu mogą być czymś zasadniczo pewnym, ale ich wyjaśnienie (prawidłowości wysokiego poziomu), może być przedmiotem szerokiej dyskusji.

Mówiąc inaczej, jeśli prawidłowość niskiego poziomu jest prawdziwa, to różne alternatywne teorie (=prawidłowości wyższego poziomu) mogą ją objaśnić. Problem ten w filozofii nauki nazywa się „niedookreśleniem teorii przez empirię” (np. Douven, 2007). Mówiłem o tym więcej w rdz. 2.3.

Prawidłowości niższego poziomu są też bardziej konkretne i precyzyjnie przewidują, co powinno się wydarzyć, ale za to tylko w niezwykle ograniczonym zakresie sytuacji. Prawa wyższego poziomu są nam potrzebne, bo są ogólniejsze, a przez to pozwalają przewidywać większą liczbę sytuacji, z drugiej jednak strony w mniej precyzyjny sposób. Przykładowo teoria zarządzania błędami, nawet jeśli prawdziwa w domenie spostrzegania romantycznego, nie musi być prawdziwa w domenie, dajmy na to, spostrzegania zagrożenia przestępczością, gdyż ta może rządzić się jakąś inną teorią średniego zasięgu, czy w ostateczności być jednym z „wyjątków potwierdzających regułę”. Jeśli nawet spostrzegania zagrożenia przestępczością rządzi się podobnymi prawami, to określony konkretny eksperyment może tego nie wykazać, bo mogą się pojawić jakieś niezidentyfikowane problemy z manipulacją eksperymentalną, operacjonalizacją zmiennych, zgadywaniem hipotez przez

badanych itd. W związku z tym prawidłowości wyższego poziomu trudniej też empirycznie sfalsyfikować⁶⁶, natomiast próba falsyfikacji prawidłowości najniższego poziomu jest stosunkowo łatwa i polegałaby po prostu na dokładnym powtórzeniu eksperymentów (w tym przykładzie z badania Haselton i Bussa, 2000).

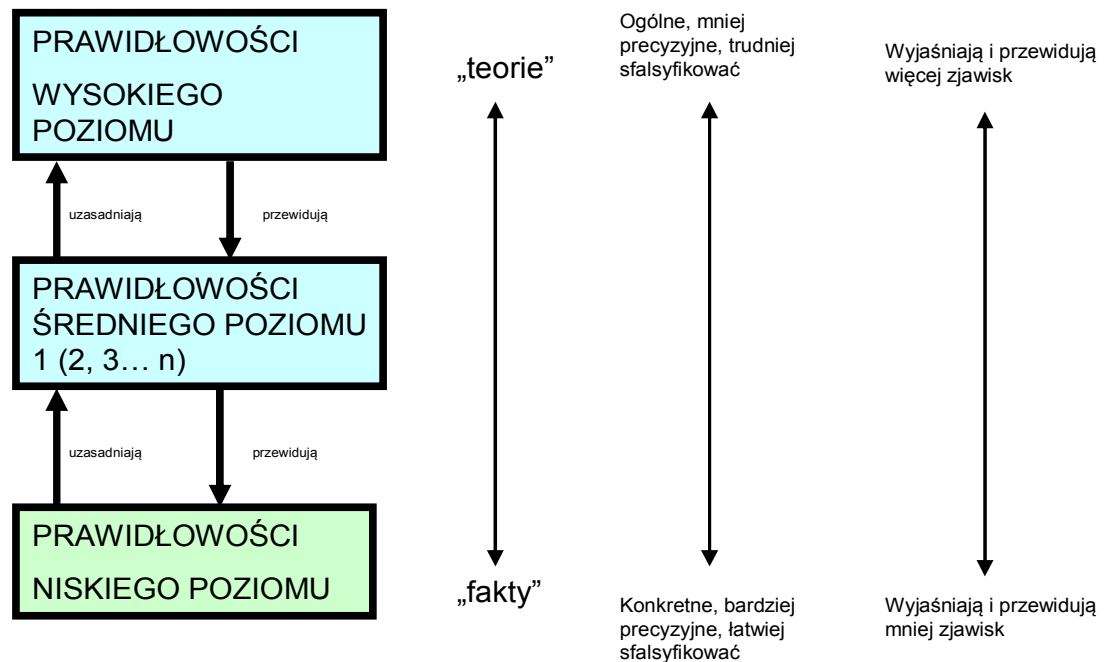
Prawidłowości niskiego poziomu w normalnej nauce uzasadniają prawidłowości wysokiego poziomu. Piszę o „normalnej” nauce, gdyż znam szereg teorii psychologicznych, które z powodzeniem funkcjonują na uniwersytetach, pomimo tego, że mają nikłe lub znikome uzasadnienie empiryczne i nie generują żadnych badań w sensownych czasopismach.

Prawidłowości wysokiego poziomu są nie tylko uzasadniane przez prawidłowości niższego poziomu, ale mogą też posłużyć do przewidzenia jakichś nowych niebanalnych odkryć i zaprojektowania konkretnego badania.

Choć prawidłowości niższego i wyższego rzędu łączy mocny związek, nie jest on ściśle symetryczny. Prawidłowości niższego poziomu mogą dalej istnieć, gdy zostaną z jakichś względów zakwestionowane prawidłowości wyższego poziomu. Jednak systematyczne i szerokie zakwestionowanie danych empirycznych z niższego poziomu powinno za sobą pociągać korektę teorii wyższego poziomu, które się na nich opierały. Gdyby okazało się, że z jakichś względów teoria zarządzania błędami nie może być prawdziwa, nie pociąga to za sobą automatycznie implikacji, że eksperymenty Haselton i Bussa (2000) są błędne. Trzeba dla nich poszukać po prostu nowego wyjaśnienia. Najważniejsze różnice dotyczące prawidłowości różnego poziomu podsumowuje Rycina 3.1.

⁶⁶ Studenci pytają mnie czasami o „badania, które sfalsyfikowały psychoanalizę”. Biorąc pod uwagę to, co napisałem w tekście, niestety, nie wierzę, żeby mogło być badanie, lub nawet grupa badań, które by to dokonały. Natomiast można wskazać, że psychoanaliza lub inna metateoria nie generuje nowych odkryć, bo jest praktycznie nieobecna w czasopismach głównego nurtu (por. np. Budzicz, 2012a)

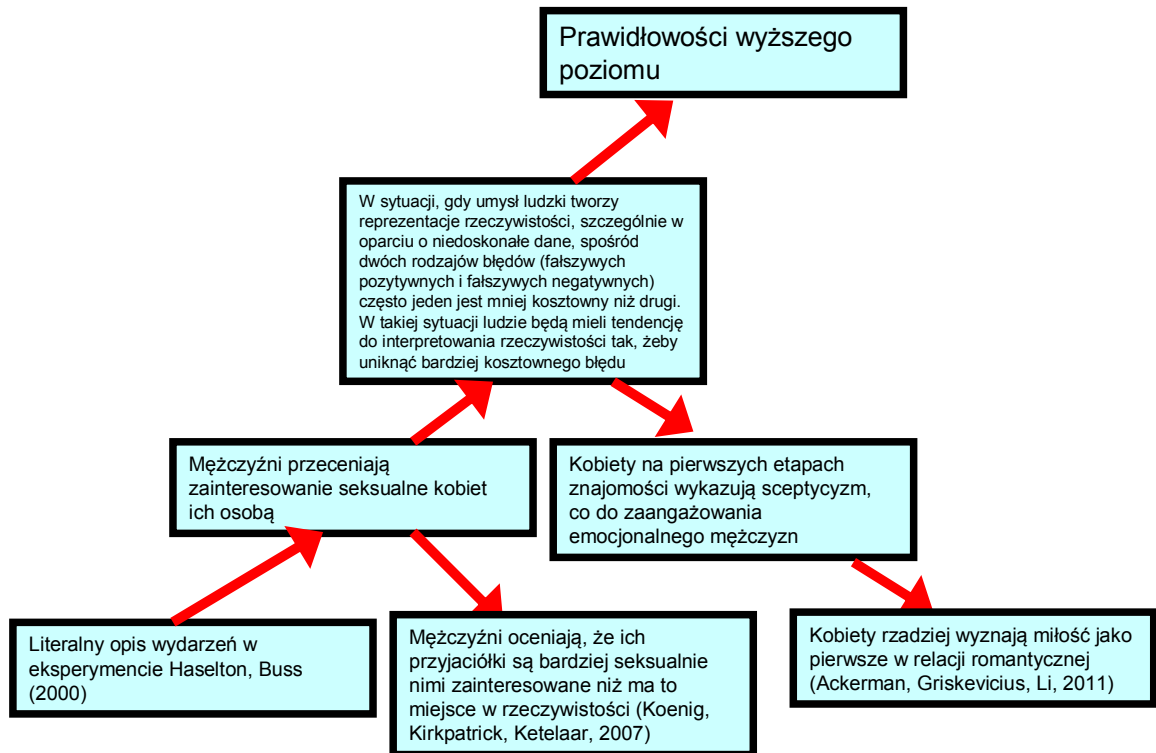
RYCINA 3.1. Najważniejsze różnice między prawidłowościami wysokiego i niskiego poziomu



Źródło: opracowanie własne

Prawidłowości wyższego poziomu, pomimo swojej bardziej spekulatywnej natury, mają tą podstawową zaletę, że pozwalają przewidywać występowanie nowych, wcześniej nieodkrytych prawidłowości niższego poziomu (jest to też test ich prawdziwości). Opierając się na prawidłowości z poziomu b (patrz wyżej s. 169) Koenig, Kirkpatrick i Ketelaar (2007) pokazali, że badani przez nich mężczyźni sądzili, że ich (konkretne) przyjaciółki wykazują większe zainteresowanie seksualne ich osobą niż rzeczywiście miało to miejsce (a jakie było naprawdę, było wiadomo, gdyż zapytano owe przyjaciółki o to zainteresowanie). Na podstawie prawidłowości c, skonstruowano też nową prawidłowość pośrednią, która mówiła o tym, że u kobiet powinien występować przeciwny mechanizm, tj. powinny bardziej sceptycznie oceniać poziom zaangażowania romantycznego u mężczyzn, szczególnie w początkowych etapach znajomości (gdyż potencjalnie mogą ponosić wszystkie koszty rozrodu, za to bez inwestycji partnera, który w międzyczasie zdąży się ulotnić). Ciekawym potwierdzeniem tej teorii było odkrycie, że w związkach romantycznych wbrew obiegowym opiniom, to kobiety rzadziej jako pierwsze wyznają miłość (Ackerman, Griskevicius, Li, 2011), oraz kobiety rzadziej ulegają zauroczeniu „od pierwszego spojrzenia” (Galperin, Haselton, 2009). Omawiane tutaj „drogi” rozwoju teorii i faktów przedstawiono na rycinie 3.2.

RYCINA 3.2. Schemat przykładowej sytuacji, gdy fakty budują teorie, a teorie pozwalają przewidywać fakty (uproszczony)



Przedstawiona tutaj teoria zarządzania błędami jest w literaturze określana jako jest teorią średniego zasięgu (*mid-level theories*). Zaznaczam, że nie należy ich mylić z prawidłowościami średniego poziomu. Określona teoria średniego zasięgu może być w niektórych sytuacjach zbiorem prawidłowości wysokiego poziomu. Teorie średniego zasięgu to teorie, które generują większość systematycznych programów badawczych w psychologii. Uczone na studiach główne „koncepty psychologiczne człowieka”, takie jak psychoanaliza, behawioryzm, podejście poznawcze itd., same w sobie wg mojej najlepszej wiedzy nie generują badań empirycznych, a przynajmniej ja się z takimi nie spotkałem. Spotkałem natomiast mnóstwo badań inspirowanych teorią opanowywania trwogi, teorią samodeterminacji, teorią nieświadomego myślenia, teorią inwestycji rodzicielskich, teorią ewolucji historii życiowych⁶⁷ itd. Teorie te czasami określają z jakiej „szkoły” się wywodzą, jednak większość nie

⁶⁷ Odpowiednio: *terror management theory, self-determination theory, unconscious thought theory, parental investment theory, life history theory*.

odwołuje się szeroko do jakiejś metakoncepcji, ale po prostu formułuje przewidywania w swoim wycinku rzeczywistości.

Cenione przeze mnie podejście ewolucyjne (Budzicz, 2012a) również nie jest metateorią, chyba, że za metateorię uznać zbiór bardzo ogólnych zasad w rodzaju „ewolucja kształtuje fenotyp psychologiczny”, „umysł nie zawsze jest dostosowany do współczesnego środowiska”, które jednak same w sobie nie generują wielu konkretnych przewidywań. O jej bogatym programie badawczym świadczy przede wszystkim zbiór wielu teorii średniego poziomu, takich jak teoria inwestycji rodzicielskich, teoria ewolucji historii życiowych, teoria konfliktu rodzic-potomstwo, czy wspomniana teoria zarządzania błędami. Podejście ewolucyjne mogłoby stanowić sensowną metateorię psychologiczną, gdyby formułowało zbiór praw wg których geny organizują mózg we współdziale środowiska i historii ewolucyjnej genów odpowiedzialnych za cechy behawioralne.

Z wymienionych przed chwilą względów jestem przekonany, że

TEZA 1 *Na tym etapie rozwoju nauki warto pragmatycznie skoncentrować się na tworzeniu sensownych teorii średniego zasięgu, które wyjaśniają ograniczony zbiór zjawisk, ale robią to na tyle precyzyjnie, żeby wygenerować konkretne przewidywania.*

Oczywiście, ostatecznym celem naszej nauki jest meta-teoria opisująca naturę człowieka (coś jak neodarwinowska synteza w biologii lub mechanika kwantowa i teoria względności w fizyce), myślę, że jednak ciągle nam do niej daleko.

3.7. Model rozwoju psychologii #2 – źródła postępu

Zanim przejdę dalej pozwolę sobie na kilka rozważań dotyczących źródeł postępu.

Nie będzie chyba kontrowersyjną teza, że istota kumulacji wiedzy w dowolnej nauce empirycznej to złożone współgranie faktów i teorii (czy używając wcześniejszej terminologii: prawidłowości niskiego i wysokiego poziomu). Warto w tym kontekście przytoczyć słowa ekonomisty Kennetha Bouldinga: „Teorie bez faktów są nagie, ale fakty bez teorii nic nie znaczą” (zapewne trawestował Kanta, który powiedział „doświadczenie bez teorii jest ślepe, ale teoria bez doświadczenia jest tylko intelektualną zabawą”). Nierozstrzygniętym sporem historyków nauki jest to, czy to

przede wszystkim teorie napędzają odkrycia empiryczne, czy może raczej odkrycia empiryczne powodują rozwój teorii. Zakładam, że związek ten jest obopólny tj. potrzeba pewnej liczby solidnych danych empirycznych, żeby stworzyć dobrą teorię średniego zasięgu (i pewnej liczby solidnych teorii średniego poziomu, aby stworzyć metateorię), zarazem jednak dobre teorie średniego zasięgu generują nowe przewidywania i odkrycia, które byłyby po prostu niemożliwe bez tych teorii. Przewidywanie niebanalnych zjawisk na podstawie teorii jest jej najlepszym sprawdzianem, oraz, postawiłbym tezę, niezwykle ułatwia skuteczne poszukiwania rozwiązań praktycznych problemów.

Zapewne nie jest logicznie wykluczone, aby w przypływie genialnego „natchnienia” stworzyć niezłą teorię średniego zasięgu bez opierania się na danych empirycznych, ale wydaje, że w nauce takiej jak psychologia jest to raczej mało prawdopodobne. Patrząc na inne nauki giganci nauk przyrodniczych tacy jak Einstein, Bohr czy Darwin tworząc swoje przełomowe teorie dysponowali potężnym korpusem solidnych danych empirycznych (np. Wróblewski, 2011; Coyne, 2009), choć z drugiej trzy przełomowe teorie biologa ewolucyjnego Roberta Triversa tj. altruizmu odwzajemnionego (Trivers, 1971), inwestycji rodzicielskich (1972), konfliktu rodzic-potomstwo (1974) wydają się być w niewielkim stopniu inspirowane empirycznie, a w większym stopniu wydedukowane z ogólniejszych zasad biologii ewolucyjnej. Olbrzymia większość danych empirycznych je wspierających została zebrana później.

Choć związek fakty-teoria jest niewątpliwie dwukierunkowy, na fakty kładłbym nieco większy nacisk. Głównym powodem jest to, że teorie, choć ważne, są zawsze konstrukcjami tymczasowymi, dopóki nie znajdziemy nic lepszego. Jak już podkreślałem, wszechstronnie przetestowane dane empiryczne są najmocniejszym elementem łańcucha wiedza, jaki mamy do dyspozycji. Pewne efekty można łatwo zreplikować (podawałem przykłady w rdz. 1.4). Zarazem wyjaśnienie często bardzo prostych efektów jest przedmiotem dużych teoretycznych kontrowersji (por. Greenwald, 2012). Nie zmienia to zasadniczego faktu, że efekty te są wg wszelkiego prawdopodobieństwa prawdziwe i niezależnie od tego, w jaki sposób zmieniają się teorie, one same będą dalej reprodukowalne przy zastosowaniu tych samych metod i tych samych warunków (przy przyjęciu kilku dodatkowych założeń np. że nie zmieni

się zasadniczo struktura ludzkiego mózgu, ale sędzę, że to kwestia raczej tysięcy lat, więc w praktycznej działalności nie ma to dużego znaczenia).

Oczywiście, dane empiryczne mają bardzo różny status. Nawet badania systematycznie zreplikowane mogą okazać się koniec końców rodzajem artefaktu. Szczególnym problemem jest, gdy dane są oparte na metodach bardzo mocno zanurzonych w teoriach dotyczących ich pomiaru i trafności (tak jak zadanie emocjonalnego Stroopa czy test utajonych skojarzeń). Są to niestety, nierzadko dane o wątpliwym statusie, jak starałem się przekonać w rozdziale 2.

Mimo bardziej spekulatywnej natury teorię wyższego poziomu są w ostatecznym rozrachunku niezmiernie ważne. Po pierwsze po to, żeby zintegrować multum prostych obserwacji, które bez nich są tylko niepowiązanymi ciekawostkami. Po drugie, żeby przewidywać nowe fakty, szczególnie biorąc pod uwagę, że wszystkich możliwych kombinacji zmiennych, moderatorów, mediatorów itd. jest potencjalnie nieskończenie wiele, a dzięki niezłym teoriom można „nieskończenie” ograniczyć do kilku rozsądnych hipotez. Po trzecie wreszcie po to, żeby rozwiązywać problemy praktyczne. Możliwe są rozwiązania bez jakichś rozbudowanego aparatu teoretycznego („zbiór technik X, Y, Z leczy depresję”), ale w przypadku nowych problemów lub nowatorskiego podejścia do starych problemów, coś musi kierunkować poszukiwania, a tym czymś może być sensowna teoria średniego (lub nawet wyższego) poziomu.

Podsumowując dotychczasowe rozważania w uproszczeniu model „postępu” nauki wyglądałby tak:

- a) zebrać fakty, dbając o ich wysoką rzetelność, możliwie szeroką generalizację, w miarę możliwości opierać się na dużym kanonie danych opisowych i obserwacyjnych
- b) z faktów wyekstrapolować prawdopodobne teorie
- c) rozstrzygnąć, które teorie są lepsze, szczególnie na podstawie tego, które potrafią precyzyjniej przewidywać rzeczywistość, lepiej generować nowe odkrycia oraz rozwiązywać problemy praktyczne (problem wyboru najlepszych teorii w filozofii nauki; patrz np. Lipton, 2003)
- d) używać dobrze uzasadnionych teorii do poszukiwania nowych faktów, które potem pozwolą ulepszyć stare teorie

- e) na podobnej zasadzie sensowne teorie średniego zasięgu integrować w meta-teorie, które z kolei pozwalają przewidywać nowe teorie średniego zasięgu, a te z kolei pozwolą przewidywać nowe teorie średniego zasięgu (tak jak to zrobił wspominamy już Trivers)

Postulowany tutaj model, choć w moim przekonaniu sensowny, nie opisuje wszystkiego. W szczególności nie opisuje inspiracji, jakie uzyskujemy dzięki odkryciom, teoriom i metodom zapożyczonym z sąsiednich (a czasami mocno oddalonych) subdyscyplin. Rozwijając szerzej te myśli:

- Problem nowych metod i narzędzi badawczych. Można podawać nieskończenie wiele przykładów z obszaru nauk przyrodniczych, w jaki sposób rozwój instrumentów takich jak mikroskop czy teleskop wpłynął na poszerzenie wiedzy o zbiór nowych faktów i wtórnie rozwój teorii (Wróblewski, 2011). Zdaniem Greenwalda (2012) to właśnie rozwój nowych narzędzi badawczych jest zasadniczym motorem postępu wiedzy. Podaje on interesujący przykład Nagród Nobla z fizyki, chemii i medycyny. W ostatnich 20 latach znaczna ich część (wg wyliczenia Greenwalda 82 %) została przyznana nie za rozwój teorii, ale za rozwój technik lub metod badawczych. Greenwald przywołuje też przykład swojego narzędzia IAT jako przykładu dużego wkładu w rozwój wiedzy, przykład ten jednak do końca mnie nie przekonuje (por. rdz. 2.5). Jest chyba jednak niekontrowersyjne, że niektóre narzędzia takie jak na przykład skanery fMRI (por. Logothetis, 2008) czy przenośne mikrofony (np. Mehl, Pennebaker, Crow, Dabbs, Price, 2001) pozwoliły nam uzyskać wiedzę, której innymi sposobami nie moglibyśmy uzyskać. Rozwój niektórych narzędzi (np. testów psychologicznych) leży w obrębie kompetencji psychologów, w rozwoju innych (np. narzędzi do obrazowania mózgu) jesteśmy zależni od inżynierów i teoretyków z innych dziedzin.
- Inspiracje dzięki teoriom z innych obszarów wiedzy. Czasami odpowiedź na niektóre pytania jest nierozzerwalnie związana z rozwojem teorii w pokrewnych obszarach. Przykładowo sensowność odpowiedzi na pytanie „jakie cechy osobowości sprzyjają agresji fizycznej” będzie wprost proporcjonalna do sensowności posiadanych przez nas teorii osobowości. Nawet największa liczba badań na temat związków „osobowości” i agresji nie pozwoli nam przeskoczyć ewentualnych ograniczeń w wiedzy o naturze osobowości. Czasami inspiracje

przychodzą z zupełnie nieoczekiwanej strony. Przykładowo, psychologia polityczna, w moim przekonaniu, dużo zyskała dzięki inspiracjom ze strony biologii ewolucyjnej, konkretnie dzięki teorii konfliktu między depresją inbredową a depresją outbredową⁶⁸. Avi Tuschman (2013) przeanalizował aspekty myślenia osób na spektrum lewica-prawica, wskazując, że niektóre z nich takie konserwatyzm obyczajowy, podporządkowanie autorytetom czy niechęć do obcych są rozstrzygnięciem konfliktu na rzecz jednej z tych dwóch strategii reprodukcyjnych⁶⁹. Przykłady można mnożyć np. gdyby powstały jakieś nowatorskie teorie rozumowania, mogłaby zyskać na tym wspomniana wcześniej teoria zarządzania błędami itd. Cała psychologia mogłaby prawdopodobnie zyskać dzięki bardziej wyrafinowanym teoriom funkcjonowania mózgu jako organu fizycznego (np. w jaki sposób możliwe interakcje neuronów ograniczają spektrum zachowań) czy teoriom rozwoju umysłu poprzez interakcje genów i środowiska. Liczne przypadki, gdy inspiracje z biologii pozwoliły lepiej zrozumieć dynamikę ludzkiego zachowania omawiają Kosslyn i in. (2002)

Osobne miejsce powinno być też poświęcone roli kreatywności badacza. Jest to zmienna mocno efemeryczna. Wywodzi się ona z intuicji, że nie wszystkie badania są w równym stopniu nowatorskie i „błyskotliwe”. Jakaś część badań powstała w określonym czasie nie dlatego, że wcześniej technologia na to nie pozwalała (lub teorie tego nie przewidywały), ale dlatego, że nikt po prostu na to nie wpadł. Efekt Stroopa został opisany w latach '30 XX wieku, choć zapewne nic technicznie nie stało na przeszkodzie, by badania te przeprowadzić w XVIII wieku, a nawet wcześniej, gdyż jedynymi narzędziami do ich przeprowadzenia są tablice z kolorowymi słowami i jakieś narzędzie mierzące czas (nie musi być bardzo dokładne, bo efekt jest bardzo silny). Pomysłowa, niesztaampowa procedura (manipulacja, pomiar zmiennej zależnej itd.) może bardzo wzmocnić określone twierdzenia. Moim ulubionym przykładem z psychologii społecznej jest procedura zastosowana przez Macrae, Milne, Bodenhausen (1994), która miała wskazywać na to, że stereotypy są narzędziami służącymi oszczędzaniu energii poznawczej. Osoby, uczące się stereotypowych informacji społecznych, lepiej równolegle zapamiętywały inne

⁶⁸ Czyli rozmnażaniem z relatywnie wysoko spokrewnionym i niespokrewnionymi osobnikami

⁶⁹ Choć rozmnażanie wsobne pomiędzy rodzeństwem lub rodzicami a potomstwem niesie na ogół jednoznacznie negatywne skutki dla dostosowania potomstwa, rozmnażanie pomiędzy osobami o umiarkowanym stopniu pokrewieństwa (np. kuzynowie 3.-4. stopnia) może nie obniżać dostosowania, a równocześnie promować geny odpowiedzialne za takie strategie.

informacje w tle zupełnie niezwiązane z treścią stereotypów (tutaj: wiadomości o Indonezji). Osoby pamiętały lepiej nie tylko stereotypowe informacje, ale też informacje podawane w tle. Choć pomysł ten, gdy już się go zna, wydaje się banalny, to jednak nikt wcześniej na niego wpadł, choć zapewne 50 czy 100 lat wcześniej były możliwości techniczne, żeby takie badanie przeprowadzić.

Głównym celem tego rozdziału jest pokazanie, że

TEZA 2 *Nie można całkowicie „racjonalnie” zaplanować „postępu” nauki, w jakimś stopniu zależy on od czynników zupełnie niezależnych od psychologów.*

Do tego tematu wrócę przy okazji jakościowego sprawozdania z analizy czasopism empirycznych, gdzie może uda mi się wskazać przykłady, gdzie inspiracje ze strony dosyć odległych teorii i narzędzi badawczych pozwoliły lepiej zrozumieć określone efekty.

3.8. Model rozwoju psychologii #3 – przedmiot psychologii

Kolejnym istotnym pytaniem jest to, co jest przedmiotem psychologii. Jest to pytanie o tyle kluczowe, że bez niego, nie będziemy wiedzieli, jaką „treść” powinny mieć odkrywane prawidłowości (różnych poziomów). Nie jest przy tym moim celem zdefiniowanie przedmiotu psychologii w postaci banalnych stwierdzeń w rodzaju „zrozumieć naturę człowieka”, ale raczej zdefiniować tak, żeby pomogło to w operacjonalizacji przy analizie teorii i badań. Zakładam, że najsensowniej będzie przedmiot psychologii wyabstrahować indukcyjnie z opisów badań psychologicznych.

Z treści artykułów empirycznych wnioskuję, że w olbrzymiej większości badań podstawowym celem jest wyjaśnienie jakiejś formy „reakcji”. Używam tu słowa „reakcja” w bardzo szerokim znaczeniu. „Reakcją” może być introspekcyjne sprawozdanie, „wynik” narzędzie psychometrycznego, deklarowana postawa, czas reakcji lub jakieś zachowanie w normalnym sensie tego słowa (tak jak w definicji Furra, s. 372). Niektóre subdyscypliny mogą być zainteresowane fizjologicznymi wskaźnikami, takimi jak aktywność mózgu, zasadniczo jednak olbrzymia większość psychologii skupia się na „reakcjach” dostępnych świadomości jednostce przynajmniej jeśli chodzi o sam fakt ich występowania (np. nawet jeśli badany nie zdaje sobie sprawy, że reaguje o kilkaset milisekund szybciej na określone bodźce, to przynajmniej ma świadomość, że wykonuje jakieś reakcje; natomiast w skanerze

fMRI nigdy nie wie, że w jego mózgu „zapala” się jakaś część). Oczywiście, jak podkreślałem w rozdziale 2, podtrzymuję tezę, że w psychologii społecznej to „zachowanie” rozumiane tak jak u Furra ma przede wszystkim znaczenie. Ale przyjmijmy, że „reakcję” czyli to „co objaśniane” zamierzamy objąć jak najszerszej.

Badania są przedstawiane jako sprawdzające wpływ albo czynników dyspozycyjnych, albo sytuacyjnych, albo ich interakcji na określoną reakcję (przykładowo w podręczniku Kenricka, Neuberga i Cialdiniego, 2006 kolejne podrozdziały są uszeregowane jako opisujące wpływ sytuacji, różnic indywidualnych i ich interakcji).

Konceptualnie nie jest jednak możliwe oddzielenie wpływu czynników wewnętrznych od zewnętrznych. Każdą sytuację ludzie przetwarzają przez swój system poznawczy (kamienie nie wypełnią ankiety, choćby nie wiadomo jak je prosić), podobnie niemożliwy jest do zbadania wpływ „czystej” osobowości bez wpływu sytuacji (nawet, jeśli zostawimy badanych samych w laboratorium, żeby zobaczyć, co będą robić, tak czy owak zainicjowaliśmy jakąś sytuację, w tym przykładzie nudy lub niepewności). Można więc powiedzieć, że w badaniach zawsze sprawdza się wpływ sytuacji w interakcji z czynnikami dyspozycyjnymi, nawet jeśli nie manipulujemy określoną kategorią lub nie uwzględniamy jej w modelu statystycznym.

Przyjmijmy jednak dla uproszczenia, że czynniki, którymi objaśnia się zachowanie „pochodzą” albo z wnętrza jednostki (wewnętrzne, dyspozycyjne), albo ze środowiska (zewnętrzne, sytuacyjne). Czynniki dyspozycyjnymi nie można zasadniczo manipulować, można je tylko statystycznie kontrolować i mieć nadzieję, że ich hipotetyczny wpływ na reakcje nie jest rodzajem korelacji pozornej.

Czynniki sytuacyjnymi można zarówno manipulować (o ile nie ma jakichś przeszkód natury technicznej lub etycznej), jak również kontrolować je statystycznie.

Sprawy komplikują się dużo bardziej, gdyż pod pojęciami „czynniki dyspozycyjne” i „czynniki sytuacyjne” kryje się potencjalnie bardzo wiele zmiennych o bardzo różnej naturze, które mogą wywierać złożony, trudny do uchwycenia wpływ, szczególnie mogą wchodzić ze sobą w wielorakie interakcje. Inne może być zachowanie człowieka w odpowiedzi na prowokację, jeśli spotka go w laboratorium, lub na ulicy; ten sam bodziec „sytuacyjny” może wywierać zupełnie różny wpływ w innym kontekście „sytuacyjnym”; wiele innych kontekstów tworzy jakieś bardziej złożone

interakcje, tak, że ludzkie zachowanie nie jest tylko prostą „sumą” poszczególnych kontekstów itp. Na zachowanie człowieka w laboratorium może wywierać wpływ nie tylko manipulowana i kontrolowana zmienna, ale też np. szerszy kontekst jego życia (trwa sesja egzaminacyjna lub niedawno rozstał się z dziewczyną), chwilowy stan fizjologiczny, przeczytana rano gazeta czy niedawne przemyślenia itd. Trudno manipulować takimi zmiennymi, a gdy je kontrolujemy, sam fakt pytania o nie, może wywierać wpływ na późniejsze zachowanie. Z istnienia wielu czynników badani mogą sobie po prostu nie zdawać sprawy.

Czynniki sytuacyjne to potencjalnie szeroka kategoria obejmująca zarówno bodźce w środowisku (na zewnątrz jednostki), jak i bodźce już nieobecne, ale wywierające wpływ na jednostkę (np. jakieś nastawienie zaindukowane określoną wcześniejszą sytuacją). Wywołują one określone mechanizmy przetwarzania informacji, które prowadzą do reakcji (mechanizmy przetwarzania informacji w najszerszym tego słowa znaczeniu, a więc obejmujące zarówno „racjonalne” kalkulacje, jak i tzw. procesy emocjonalne, motywacyjne, czy instynktowne reakcje itp.). Mechanizmy te mogą być względnie stałe lub tymczasowe, pochodzą jednak z umysłu jednostki. Różne subdyscypliny psychologii będą kładły akcenty na inne aspekty, szczególnie na to a) w jaki sposób te mechanizmy się uformowały (tu z kolei większy nacisk może być położony na wrodzone cechy naszego umysłu, tak jak w psychologii ewolucyjnej, lub na szeroko pojęte doświadczenia); b) jaka jest „natura” operacji poznawczych, które się dokonują (przykłady to: heurystyki, zniekształcenia i błędy poznawcze (bias), stereotypy, nawyki itp.) c) czy i jakie występują generalne „struktury” organizacji tych mechanizmów (przykładowo: osobowość; temperament; styl poznawczy itp.). Moim celem nie jest tu rozstrzygnięcie sporu jaka jest natura procesów w naszym umyśle, które warunkują zachowanie (w interakcji z bodźcami w otoczeniu), szczególnie, że – mam wrażenie – często fenomeny te są tworzone ad hoc i arbitralnie, co w dużej mierze wynika z tego, że obserwowane są tylko pośrednio (np. niektórzy badacze mogą nazywać określone mechanizmy „osobowością”, podczas gdy inni zakwalifikowaliby je jako zbiór heurystyk umysłowych)

Z czynnikami dyspozycyjnymi rzecz jest o tyle trudniejsza, że, o ile można jeszcze w miarę sensownie manipulować niektórymi czynnikami sytuacyjnymi w laboratorium czy eksperymencie terenowym, o tyle zmienne dyspozycyjne są niemanipulowalne

(nawet quasi-eksperymenty są rodzajem badania korelacyjnego), a w dodatku o ich strukturze mamy wyobrażenie mocno zapośredniczone przez teorie wyższego rzędu. Wyjaśnianie określonej reakcji poprzez czynniki dyspozycyjne, oznacza na ogół dużo wyższy poziom uteoretyzowania obserwacji. Przykładowo, znaczną większość manipulacji eksperymentalnych w psychologii społecznej daje się opisać terminami, w których nie odwołujemy się do żadnych teorii naukowych np. „gdy krzyczy się „pożar” badani zaczynają uciekać”. Jednak twierdzenie „gdy krzyczy się „pożar”, neurotycy szczególnie szybko zaczynają uciekać” niesie za sobą jakąś teorię na temat struktury osobowości i, co gorsza, masę założeń na temat jej pomiaru (i trafności tego pomiaru). Mimo wszystko jestem przekonany, że znaczną część terminów „dyspozycyjnych” można opisać bez odwoływania się do teorii wyższego poziomu, zostawiając sobie pewną dozę swobody, gdyby w przyszłości jakieś złożone teorie różnic indywidualnych uległy zasadniczej rewizji. Nie zmienia to jednak faktu, że psychologia różnic indywidualnych dorobiła się kilkudziesięciu badawczo mocno eksploatowanych czynników dyspozycyjnych (np. Leary, Hoyle, 2009), dodatkowo dalsze setki występuje rozrzuconych po rozmaitych publikacjach. Choć za niektórymi stoi monstrualna liczba badań (np. koncepcja wielkiej piątki) i trudno odrzucić myśl, że jakoś odzwierciedlają różnice w strukturze naszej psychiki, trudno nie powstrzymać się od wrażenia, że duża ich część służy tylko budowaniu kariery na nowym „odkryciu”, albo jest nową nazwą na wcześniej już opisany czynnik (patrz problem stylów poznawczych: rdz. 3.2). Znaczna część badań nad trafnością większości zmiennych ma naturę samoopisową, a ostatecznym dowodem na trafność wielu koncepcji różnic indywidualnych jest ich wyodrębnienie przy pomocy analizy czynnikowej. Samo w sobie nie przesądza to jeszcze o ich „rzeczywistości”, ale może brać się z tego, że ludzie mają pewne wyobrażenia o tym jak funkcjonują i zbiory takich wyobrażeń tworzą czynniki wyodrębniane potem statystycznie. Temat ten zasługiwałby na osobną dysertację, nie będzie chyba jednak kontrowersyjnym stwierdzeniem, że rozumienie i pomiar zmiennych dyspozycyjnych jest daleki od oczekiwanego ideału.

Podsumowując te rozważania można powiedzieć, że:

TEZA 3 *Prawidłowości niskiego poziomu określają przede wszystkim zależność (przyczynowo-skutkową lub korelacyjną) między zmiennymi sytuacyjnymi, dyspozycyjnymi lub ich interakcją, a szeroko pojmowaną „reakcją”.*

Z kolei prawidłowości wysokiego poziomu postulują określony mechanizm, który pośredniczy pomiędzy czynnikami dyspozycyjnymi i sytuacyjnymi, a reakcją. Mechanizm ten jest zasadniczo nieobserwowalny. Teoria zarządzania błędami jest wyrażoną w intuicyjnych twierdzeniach teorią opisu mechanizmu umysłowego „ważenia” różnych rodzajów błędów na podstawie ich kosztów (zakłada, że ten mechanizm jest w dużej mierze ukształtowany filogenetycznie). Dzięki postulowanemu mechanizmowi możemy znacznie więcej wyjaśnić i przewidzieć, gdyż, jak już wspomniano z prawidłowości najniższego poziomu bardzo niewiele wynika (co najwyżej, że w zbliżonych warunkach powinna zajść zbliżona reakcja; możemy być przy tym pewnie, że w tzw. „życiu” bardzo zbliżone warunki prawie na pewno nie znajdują).

Przedstawiona tutaj teoria zarządzania błędami jest opisowa, można powiedzieć intuicyjna. Nie jest to wada, szczególnie jeśli nic sensowniejszego nie możemy przedstawić. Tym niemniej teoria ta nie odnosi się ani do procesów informacyjnych, jakie przebiegają w mózgu, ani nie odnosi się do fizycznych odpowiedników tych procesów w mózgu. Z tego m.in. względu trudno byłoby ją zintegrować z innymi udokumentowanymi prawidłowościami błędów poznawczych (choć jak wskazują Haselton i Nettle, 2006 wiele takich błędów można wyjaśnić właśnie ich teorią), o innych teoriach dotyczących innych procesów poznawczych nie wspominając.

Przykładem teorii, która opisuje różne zjawiska psychologiczne z odwołaniem do elementarnych procesów przetwarzania informacji jest teoria ACT-R (Adaptive Control of Thought—Rational) stworzona przez Johna Andersona (np. Anderson, 1996; Anderson, Matessa, Lebiere, 1997; Anderson, 2014). Nie posiadam kompetencji, żeby ocenić jej trafność i przydatność dla psychologów społecznych. Wydaje się to jednak dobry przykład dążenia do tego, żeby zjawiska psychologiczne opisać przy pomocy elementarnych procesów przetwarzania informacji. Dążenie jest to o tyle pożądane, ponieważ

TEZA 4 *Prawidłowości wysokiego poziomu powinny dążyć do opisanie mechanizmów przetwarzania informacji (które są przyczyną zależności opisanych w tezie 3) przy pomocy elementarnych procesów przetwarzania informacji, dzięki czemu będzie potencjalnie kiedyś możliwa integracja różnych teorii średniego zasięgu, nawet opisujących prawidłowości z bardzo różnych obszarów.*

Idąc dalej zakładałbym, że:

TEZA 5 *Najwyższy poziom teorii powinien opisywać elementarne procesy przetwarzania informacji w kategoriach ich fizycznych odpowiedników.*

Ma to znaczenie o tyle, że pozwoli zintegrować wiedzę psychologiczną z wiedzą z bardziej rozwiniętych nauk przyrodniczych (szczególnie biologii molekularnej, chemii, fizyki), pozwoli w/w nauki wykorzystać do (częściowego) wyjaśniania ludzkiego zachowania, oraz da informacje o podstawowych obserwacyjnych wskaźnikach tych procesów. Zakładam jednak, że jest to perspektywa wysoce odległa.

W opisanym tu procesie rozwoju nauki prawa coraz wyższego poziomu opisują określone zjawisko niejako coraz „głębiej” (tj. najpierw poprzez intuicyjny opis procesu, potem przez elementarne procesy przetwarzania informacji i finalnie przez ich odpowiedniki fizyczne).

Można jednak zakładać, że prawa coraz wyższego poziomu opisują zjawisko coraz szerzej. Taką interpretację zarysowałem już wcześniej, przy okazji szczegółowego omawiania teorii zarządzania błędami. Przykładowo, teoria ta najpierw opisywała zniekształcenia poznawcze w obszarze spostrzegania intencji seksualnych, później rozszerzała się lub inkorporowała inne zniekształcenia np. związane z nabywaniem wstrętu na jedzenie, ze złudzeniem optymisty itd. (Haselton, Nettle, 2006). Można sobie wyobrazić, że teoria zarządzania błędami mogłaby być zintegrowana z innymi teoriami w jakąś jedną teorię zniekształceń poznawczych, ta mogłaby być zintegrowana z innymi teoriami dotyczącymi poznania w jakąś ogólną teorię poznania ludzkiego itd. Dlatego

TEZA 6 *Teorie wyższego poziomu mogą być związane z coraz „głębszym” opisywaniem danego mechanizmu, ale mogą być też związane z opisywaniem coraz szerszej klasy zjawisk*

W analizie empirycznej będę chciał zwrócić szczególną uwagę, czy jaki jest sposób rozszerzania wiedzy o danej prawidłowości.

3.9. Model rozwoju psychologii #4 – „nowatorskie” odkrycia vs. lepsze rozumienie efektu

Poruszę jeszcze jeden problem, który wg mojej oceny występuje i wpływa na ograniczenie możliwości rozwoju. Załóżmy, że udało nam się zebrać sporo danych empirycznych wspierających jakąś teorię średniego poziomu, dane te zostały niezależnie zreplikowane, ich występowanie jest uniwersalne kulturowo i wiadomo, że nie zależy od grupy społecznej lub kulturowej. Teoria średniego zasięgu wygenerowała przy tym sporo nowych przewidywań, które się potwierdziły. Z taką sytuacją, jak się wydaje, mamy do czynienia w szeregu domen ludzkiego zachowania (np. heurystyki poznawcze: Kahneman, 2012; zachowania polityczne: Tuschman, 2013; religia: Norenzayan, 2013; kooperacja społeczna: Cronk, Leech, 2012). W tych przypadkach jednak ciągle jedyne, na co możemy sobie pozwolić, to dosyć ogólne twierdzenia opisowe w rodzaju „mężczyźni przeceniają zainteresowanie romantyczne kobiet ich osobą”. Efekt ten nie jest jednak zerojedynkowy, to znaczy istnieje sporo mężczyzn, którzy przeceniają zainteresowania ich osobą w mniejszym stopniu niż niektóre kobiety, a i w obrębie mężczyzn istnieje duże naturalne zróżnicowanie tego zjawiska. Teoria sama w sobie nie mówi, którzy mężczyźni, w jakich okolicznościach, względem jakich kobiet będą wykazywali takie zniekształcenia. Niewiele też mówi, jak te efekty będą się miały do niezliczonej liczby innych efektów opisujących ludzką naturę.

Czasopisma i instytucje przyznające granty wymagają, aby badania były zawsze „nowatorskie”, tj. mocno niepodobne do dotychczasowych badań. Może to powodować to, że badacze będą ciągle gonili za nowymi odkryciami, zamiast starać się dogłębnie zrozumieć wszystkie istotne uwarunkowania reakcji w jednym badaniu. Szczególnie widać to w praktykach dotyczących wielkości efektów. Obecnie w dobrych czasopismach powszechnie raportuje się wielkości efektów, ale trudno nie odnieść wrażenia, że jest to rodzaj konwencji statystycznej, gdyż wielkości tych efektów nie są raczej poddawane analizie konceptualnej. W szczególności czasami można znaleźć relatywnie słabe efekty, takie, które wyjaśniają mały ułamek całej wariancji, a badacze kompletnie pomijają ten aspekt i piszą o efekcie w taki sposób jakby był znaczący. Tymczasem w codziennym życiu wyczuwamy intuicyjnie, że siła efektów ma wielkie znaczenie. Nie interesuje nas tylko czy palenie podnosi ryzyko zachorowania, ale jak bardzo; nie tylko czy partia polityczna X straciła poparcie wskutek skandalu Y, ale ile procent; nie tylko czy określona dieta pozwoli stracić wagę, ale czy kilkanaście kilogramów czy kilka gram itd. Większość badań jest

opisanych w konwencji „jest/nie ma”, a jeśli efekt jest (istotny statystycznie), dalej następuje dodawanie teorii do często miniaturowych efektów (przypominam o metaanalizie wielkości efektów w psychologii społecznej, gdzie najczęściej spotykany efekt to $r < 0,1$!; Richard i in., 2003).

Dlatego jeśli efekt jest słaby, dużo wariancji jest niewyjaśnione, badacze, zamiast przerzucać się do nowych badań, powinni spróbować zmaksymalizować wyjaśnianą wariancję poprzez dodawanie dodatkowych zmiennych dyspozycyjnych i sytuacyjnych. Jednak nie sądzę, aby dobra była sytuacja, znana mi z osobistych obserwacji, polegająca na projektowaniu od samego początku „wielkiego badania” obejmującego dziesiątki zmiennych. Taka strategia wydaje się sprzyjać potencjalnie „łowieniu p”, poprzez gigantyczną ilość zmiennych i interakcji, nie wspominając o potencjalnych artefaktach, wynikających z tego, że nie wiadomo ile sensownych danych się traci, poprzez to, że badani rozwiązują męczący plik kwestionariuszy. Lepsza wydaje się strategia polegająca na „atakowaniu” małych fragmentu problemu, „zdobywaniu przyczółka” w postaci szeregu badań z nieźle udokumentowanymi efektami i dopiero późniejszym rozszerzaniu poszukiwań.

TEZA 7 *W ostatecznym rozrachunku rozszerzanie rozumienia zjawiska służy zwiększeniu precyzji przewidywania, szczególnie poprzez określanie, jakie będzie prawdopodobieństwo lub siła danego zjawiska u konkretnego „Jana Kowalskiego” (przy założeniu, że możemy dosyć precyzyjnie poznać istotne zmienne wpływające na ten efekt, a ich pomiar nie zaburzy zasadniczo samego zachowania).*

Próby dosyć precyzyjnego przewidywania zachowania pojedynczych ludzi, wg mojego rozeznania, są w chwili obecnej w zasadzie nieobecne w psychologii społecznej. Nie warto się spieszyć, jeśli nie ma po temu odpowiednich możliwości, tym niemniej wydaje się, że na długą metę to najlepszy sprawdzian teorii psychologicznych (czy mówiąc górnolotnie: cel psychologii).

3.10. Model rozwoju psychologii – operacjonalizacja modelu

Oczywiście powyższymi rozważaniami nie wyczerpałem całego spektrum wszystkich możliwych problemów związanych z badaniem zachowania człowieka i innych złożonych zwierząt (np. Bermúdez, 2007; Botterill, Carruthers, 1999). Sprawy nie ułatwia wiele innych powodów np.

- a) znaczna część powtarzalnych procesów psychologicznych nie jest uniwersalna w takim sensie jak prawa fizyczne. Być może na najbardziej podstawowym poziomie przetwarzania informacji przez mózg występują procesy, które są uniwersalne dla wszystkich inteligentnych istot, ale nie wydają się, żebyśmy na ten poziom dotarli. Dobrze znany efekt polegający na instynktownym strachu przed pajakami (Öhman, Mineka, 2001) nie występuje u wszystkich ludzi i zapewne nie pojawiłby się na planecie, na której nie istnieją pająki. Niewykluczone, że – jako że pająki nie stanowią dziś większego zagrożenia – efekt ten przestanie występować w ciągu kilkuset lat. Oczywiście, co jeszcze lepiej udokumentowane, kontekst kulturowy wywiera szczególny wpływ ograniczający uniwersalność twierdzeń (patrz rdz. 2.7).
- b) człowiek nie jest tylko biernym odbiorcą sytuacji, ale również aktywnie poszukuje określonych sytuacji, inicjuje je lub w znacznym stopniu zwrotnie na nie wpływa. Rozpędzona w akceleratorze cząstka nie zmienia swojego „zachowania” pod wpływem informacji, że badacze mają określone teorie jak powinna się zachowywać. Ludzie potencjalnie mogą zmienić.

Przechodząc jednak do podsumowania rozważań z rdz. 3.6-3.9 i wcześniejszych rozdziałów, wydaje się, że w uproszczeniu psychologia powinna więc podążać w dwóch, uzupełniających się kierunkach.

Podstawowym kierunkiem jest budowa solidnego kanonu prawidłowości niskiego poziomu, faktów, które są solidnie udokumentowane (np. zreplikowane) oraz pokazują efekt na przykładzie codziennych zachowań (tam, gdzie taki związek jest postulowany). W miarę możliwości badania te powinny być jak najmniej uteoretyzowane⁷⁰, aby mogły stanowić podstawę, na których buduje się teorie wyższego rzędu, oraz aby mogły potencjalnie w przyszłości być adaptowane przez nowe teorie. Afera Stapela, oraz rozliczne nieprawidłowości w danych psychologicznych przedstawione w rozdziale 1, wskazują, że poziom rzetelności podstawowych odkryć jest ciągle niedostateczny. Konkretnie zalecenia dla zwiększania poziomu rzetelności i zmniejszania poziomu „uteoretyzowania” podzieliłbym na następujące kategorie. Badania powinny być:

⁷⁰ W sensie takim jak wcześniej przedstawiony, a nie w sensie, że jest zaoferowana teoria wyjaśniająca.

- powtórzone, szczególnie przez niezależne zespoły badawcze (replikacja ogólna = **RepOg**⁷¹)

- sprawdzone pod kątem trzech podstawowych moderatorów, które, jak wiadomo, wpływają na zachowanie, a więc grupy społecznej, wieku (=etapu rozwojowego) i grupy kulturowej (moderatory ogólne = **ModOg**) (oczywiście, takim podstawowym moderatorem jest też płeć, ale zakładam, że różnice płciowe są w miarę często sprawdzane, o czym zaświadcza m.in. niezliczone meta-analizy ich dotyczące publikowane w PB i nie tylko⁷²)

- sprawdzonych pod kątem moderatorów specyficznych dla określonego badania (inne bodźce, procedura itd.) (moderatory specyficzne = **ModSp**)

- starać się rozszerzyć rozumienie efektu, poprzez stopniowe dodawanie dodatkowych nowych zmiennych (rozszerzenie rozumienia = **RozszRoz**)

- tam gdzie postulowany efekt przekłada się na codzienne zachowania, tam poszukiwać dowodów w postaci takich właśnie zachowań, szczególnie danych z eksperymentów terenowych, nagrań lub innych rejestrów zachowań, wytworów ludzkich itd., a więc wszelkich możliwych danych, które rejestrują zachowanie bez ryzyka wystąpienia artefaktów laboratoryjnych. Jeśli stosuje się badania laboratoryjne, wszędzie tam gdzie to możliwe, należy dążyć do używania miar behawioralnych, a nie tylko opartych na pośrednich wskaźnikach zachowania (dane behawioralne = **DaneBeh**)

W tym kontekście warto też się zastanowić, jak „zoperacjonalizować” postulat Rozina dotyczący większej liczby obserwacji i opisu. Nie ukrywam, że z tym inspirującym postulatem, miałem jednak problemy, gdy przyszło do jego ukonkretnienia. Wydaje się, że zależy on w dużej mierze od tego, jak rozumiemy słowa „obserwacja” i „opis”.

Jeśli rozumiemy je częściowo, tak jak Rozin, czyli jako brak eksperymentu i skomplikowanych procedur, byłoby to nieco ograniczające. Omówione w rozdziale 2.8 badania choć były eksperymentalne i do tego dosyć złożone, zawierają jednak wnioski, które można przełożyć na dosyć proste zdania obserwacyjne z niewielką ilością odwołań do teorii wyższego rzędu np. „niepowodzenie zwiększa nasilenie

⁷¹ Kody będą używane w dalszej części rozdziału do zilustrowania przykładów. W rozdziale 4 przedstawiam nieco zmodyfikowaną wersję tego podziału (zmodyfikowaną po analizie części literatury).

⁷² Można się o tym przekonać wpisując zapytanie „sex (lub gender) differences” i „meta-analysis”.

postaw religijnych” – wymaga to doprecyzowania pewnej liczby szczegółów, jak rozumiemy niepowodzenie, religię, czym powinno się objawiać nasilenie itd.; wydaje się jednak, że zdanie to samo w sobie powinno być intuicyjnie uchwytnie dla większości ludzi. W tym sensie stanowiłoby podstawową obserwację, choć pamiętajmy, że została wyciągnięta na podstawie eksperymentów o bardzo złożonej strukturze.

Można też obserwację i opis rozumieć jako dużą liczbę danych surowych dostępnych badaczom do dalszych analiz (np. nagrań wideo, nagrań wypowiedzi itd.).

Przykładowo, takimi danymi surowymi mogłyby być zanimizowane wypowiedzi badanych. Dane takie, jeśli zapisane w sensownym formacie, mogłyby służyć innym badaczom do testowania wielu nowych hipotez i w takim sensie trudno odmówić racji postulatowi większej liczby danych obserwacyjnych (szczególnie, że przy współczesnych technologiach informacyjnych, można się efektywnie takimi danymi dzielić).

Postulat Rozina można jeszcze rozumieć w trzeci sposób, mianowicie taki, że badacz indukuje jakąś sytuację i nie czyni z góry żadnych założeń, co będzie obserwował (np. nie rejestruje konkretnego, z góry założonego wycinka zmiennych zależnych), ale jest otwarty na wszelkie ciekawe zdarzenia, jakie będą miały miejsce. Ten postulat ma słuszną stronę i taką, do której podchodzę sceptycznie. Zapewne w wielu dziedzinach psychologii nie zawadziłyby nam badania, w których daje się ludziom możliwość wykazania spontanicznych reakcji, ale niestety w rejestrowaniu tych reakcji musi już być zawarty pewien ładunek „uteoretyzowania” obserwacji (nie muszą to być przy tym formalne teorie, mogą być czysto intuicyjne, ale tak czy owak). Wydaje się, że w procesie obserwacji zawsze przyjmujemy jakieś założenia, co jest istotne, warte uwagi, warte zarejestrowania i sprawozdania. W takim sensie zapewne sensie postulat Rozina jest chybiony, bo niemożliwa jest „czysta obserwacja”. Natomiast, nie wyklucza to tego, że obserwacja zachowania w naturalnych kontekstach, może czasami pozwolić badaczom sformułować dostrzec ważne, wcześniej pomijane aspekty zachowania lub sformułować nowatorskie hipotezy, na które by nigdy nie wpadł bez niej.

Jeśli więc dobrze rozumiem postulaty Rozina, wynikają z nich trzy zasadnicze postulaty, które na pewno nie zaszkodzą, a mogą pomóc:

- a) więcej danych surowych (szczególnie udostępnionych dla szerokiego grona badaczy)
- b) tworzenie sytuacji, gdzie badanym daje się możliwości wykazania spontanicznych reakcji
- c) obserwowanie naturalnie występującego zachowania, celem potencjalnych inspiracji.

Stopień spełnienia tych postulatów zostanie we właściwym badaniu przeanalizowany jakościowo.

Drugi kierunek to rozwój teorii, oraz pośrednio prowadzenie badań „kierowanych przez teorię” („theory-driven research” np.). Tak jak wcześniej pisałem, jestem przeciwnikiem dorabiania naciąganych teorii do prawidłowości niskiego poziomu, które powstały przypadkiem, lub, które trudno na tym etapie rozwoju sensownie wyjaśnić. Uważam też, że naszej nauce nie zaszkodziłby większe akcentowanie rozróżnienia w artykułach pomiędzy sferą faktów i (zawsze hipotetycznych) wyjaśnień. Zarazem jednak rozwój teorii (prawidłowości wysokiego poziomu) jest niezbędny dla rozwoju w miarę dojrzałej nauki. Dlatego rozwój teorii zoperacjonalizowałbym w sposób następujący:

- dane generują alternatywne wyjaśnienia teoretyczne, które są systematycznie testowane i dąży się do posiadania jednego najlepszego wyjaśnienia (choć nie narzuca się go na siłę, jeśli nie ma takiej możliwości) (testowanie alternatywnych teorii - **TeorAlter**)
- doprecyzowuje się teorie poprzez zbieranie nowych danych np. wskazując na ograniczenia jej stosowania, dodatkowe moderatory itd. (dodatkowe moderatory teorii - **TeorModer**⁷³)
- najlepszym sprawdzianem prawdziwości teorii, jest to, że generują nowe odkrycia, szczególnie niebanalne odkrycia wcześniej nie przewidziane (generowanie odkryć - **GenerOdkr**)

Powyższe odkrycia są możliwe do ilościowego zoperacjonalizowania na podstawie zawartości czasopism. Chciałbym również przyrzeć się rozwojowi teorii w sposób

⁷³ Ostatecznie po analizie części literatury różne rodzaje wyjaśnień teoretycznych zostały sklasyfikowane jako „zwiększanie precyzji teorii”

trudniejszy do uchwycenia, dlatego poniższe postulaty zostaną przeanalizowane jakościowo:

- różnorakie odkrycia, wcześniej niepowiązane, są ze sobą wiązane poprzez system teorii średniego zasięgu, które z kolei wiążą się ze sobą poprzez system metateorii
- teorie, początkowo niezależne, są ze sobą integrowane, szczególnie, jeśli dotyczą zbliżonego obszaru
- określony obszar korzysta z inspiracji z innych obszarów, teorii i dyscyplin

3.11. Analiza operacjonalizacji na przykładach

Zakończę ten rozdział analizą przykładów z rozdziału 2 pod kątem tego, jak w przypadku tych badań mogłaby przebiegać sensowna kumulacja wiedzy. Niektóre zalecenia, takie jak przetestowanie prawidłowości na osobach wywodzących się z różnych grup kulturowych, zreplicowanie badań przez niezależne zespoły itd. są oczywiste i nie będę tu ich każdorazowo wymieniał. Podobnie jak wcześniej, podane za chwilę przykłady bardziej służą zobrazowaniu wcześniejszych punktów, a nie dokładnemu teoretycznemu „prześwietleniu” opisywanych teorii.

- I. Badanie dokumentujące większą skłonność do pomagania osób z niższych klas społecznych (Piff, Kraus, Côté, Cheng, Keltner, 2010)

Wydaję się, że program badań powinien rozpocząć się od kilku klaryfikacji. W szczególności ważne byłoby stwierdzenie czy opisywana prawidłowość dotyczy osób znajdujących się na dole hierarchii społecznej, osób o niskim poziomie dochodu materialnego, a może tzw. wykluczonych (**ModSp**). Choć te trzy kategorie często się ze sobą łączą, możliwe są sytuacje, gdy osoba niekoniecznie zalicza się do wszystkich trzech (np. osoba uboga, ale mająca bogatą sieć wsparcia społecznego; relatywnie zamożny emeryt, który nie ma rodziny i ma niewielkie wsparcie społeczne; osoby o relatywnie niskich dochodach, ale wykonujące zawody spostrzegane jako ważne np. pielęgniarki, lub osoby wykonujące prace „wykluczające” w powszechnym odbiorze, pomimo tego, że mogą się wiązać z przyzwoitym dochodem np. przestępcy, prostytutki itp.). Niewykluczone, że zbadawszy różnicę, pomiędzy tymi trzema kategoriami, możliwe byłyby odkrycia, które rzuciłyby światło na samą naturę

zjawiska (**TeorModer**). Dalszym pytaniem byłoby to, czy zjawisko, to dotyczy całego społeczeństwa, czy może niewielkich grup społecznych również (np. jak zachowują się pracownicy różnych organizacji, którzy są na dole hierarchii, pomimo tego, że posiadają przyzwoity dochód itp.) (**ModSp**). Jeśli chodzi o samą zmienną zależną, też można by zadać kilka pytań. W badanych eksperymentach altruizm był mierzony samoopisowo, lub poprzez gotowość do udzielenia pomocy drugiemu badanemu. Jest otwartym pytaniem, co mierzy samoopisowa deklaracja skłonności do pomagania. Pomoc drugiemu badanemu jest natomiast rodzajem „pragmatycznego” altruizmu tj. takiego, gdzie druga osoba może się odwdziżyć. Interesująca byłaby odpowiedź na pytanie, czy takie osoby też będą skłonne pomagać, gdy altruizm nie będzie się wiązał z żadnymi potencjalnie wymiernymi korzyściami (np. czy będą skłonne dać datek dla ubogich w Bangladeszu). (**DaneBeh**)

Jeśli chodzi o samą prawidłowość, pytanie o jej naturę też narzuca kilka potencjalnych wyjaśnień, które mogą się uzupełniać, ale nie muszą. Intuicyjnie narzucają się trzy możliwe wyjaśnienia (**TeorAlter**):

- a) zjawisko jest rodzajem korelacji pozornej, gdyż istnieje trzecia zmienna, która predestynuje do biedy oraz wysokiego altruizmu (wykluczenia; niskiej pozycji społecznej, patrz punkt wyżej) np. bardzo wysoki poziom ugodowości
- b) zjawisko jest rodzajem strategii, adaptacyjnej w sytuacji niskiego sytuowania na drabinie społecznej
- c) bieda kształtuje altruizm/emпатиę (tu z kolei możliwe są dalsze wyjaśnienia, dlaczego tak się dzieje)

Pytanie o naturę zjawiska nie jest tylko ciekawostką, ale może w znaczącym stopniu kierunkować dalsze eksploracje. Jeśli stwierdzilibyśmy, że a) jest najlepszym wyjaśnieniem, wówczas wysiłek badawczy trzeba by skoncentrować na tym, dlaczego rzeczywista zmienna determinuje skłonność do pomagania i/lub biedę. W przypadku prawidłowości b) nacisk będzie położony na mechanizmy przetwarzania informacji, szczególnie co powoduje, że ludzie oceniają, że są na dole drabiny społecznej, czy w sytuacji, gdy altruizm nie wiąże się z potencjalnymi korzyściami, takie osoby też będą pomagały (co powoduje, że osoby postrzegają określoną sytuację, w której mogą pomóc jako korzystną itd.). Jeśli odpowiedź c) jest prawdopodobna, wtedy można sobie wyobrazić, że nacisk będzie położony na

różnego rodzaju badania korelacyjne i podłużne, mające na celu ustalić np. jak długi okres biedy sprzyja pomaganiu, czy różnią się od siebie, które od dzieciństwa wzrastały w biedzie, od tych, które weszły w nią w stosunkowo późnym wieku itd. Oczywiście, możliwe jest też, że ostateczne wyjaśnienie jest jakąś kombinacją trzech powyższych hipotez (lub innych, tutaj niewymienionych).

Cały program badawczy mógłby znacząco zaowocować, gdyby w pokrewnych obszarach zaproponowano np. jakąś solidną teorię rozwoju empatii/altruizmu, teorię dotyczącą strategii ingracjacji społecznych, teorię zachowania w sytuacji spostrzeganego zagrożenia lub wykluczenia społecznego itd. (niewykluczone, że istnieją już sensowne teorie i dane, które mogłyby wesprzeć ten program badawczy).

II. Badanie dokumentujące związek pomiędzy porażką a nasileniem postaw religijnych (McGregor, Nash, Prentice, 2010)

Jest to niewątpliwie problem badawczy, w którym wielkim ułatwieniem byłaby solidna teoria religii. Wg mojej wiedzy nie ma takiej teorii, która cieszyłaby się powszechnym konsensusem. Tym niemniej na tym etapie można zadawać wiele pytań, np. z jakimi komponentami religii efekt się wiąże (**ModSpec**; **TeorWykl**; **BehOgol**). Przykładowo, niektóre teorie wyróżniają takie jej aspekty jak duchowość (emocjonalny aspekt przeżywania religii), wewnętrzna religijność (osobisty związek z przedmiotem kultu), zewnętrzna religijność (uczestniczenie w praktykach religijnych i tym samym wysyłanie różnych sygnałów członkom własnej grupy społecznej). Szereg dalszych pytań mógłby dotyczyć tego, jakie formy niepowodzenia wywołują ten efekt (np. obiektywne/subiektywne), czy są sytuacje niwelujące ten efekt (**ModSp**). Teoria postuluje, że niepowodzenia wywołują stan lękowej niepewności, byłoby więc ważnym stwierdzić, jakie jeszcze sytuacje wywołują ten stan (**ModSp**). Oczywiście ważnym pytaniem byłoby też, czy ten efekt dotyczy wszystkich możliwych religii, gdyż jak wynika z innych danych, różne religie mogą oddziaływać na inne aspekty funkcjonowania psychologicznego (np. Cohen, Rozin, 2001) (**ModSp**). Badacze postulują, że religia jest formą „dążenia do abstrakcyjnych ideałów”), pytanie jest jakie jeszcze inne formy takiego dążenia mogą występować (np. działalność polityczna?, działalność społeczna?)

W szczególności na jeszcze bardziej abstrakcyjnym poziomie byłoby interesujące wykazanie związków (braku związków) tych zjawisk z fenomenami opisywanymi

przez teorię opanowywania trwogi, gdzie postuluje się, że np. poczucie własnej śmiertelności, (które można traktować jako abstrakcyjny rodzaj „porażki”), wywołuje u badanych większe przywiązanie do wartości kulturowo istotnych (np. flagi), w tym również religii (zbliżone badanie: Jonas, Fischer, 2006). Trudno nie odnieść wrażenia, że powyższe zjawiska mogą mieć wspólny mianownik, lub wręcz być innym przejawem tych samych prawidłowości.

III. Badanie dokumentujące inne strategie autoprezentacji u osób różnych ras w interakcji z osobą innej rasy (Bergsieker, Shelton, Richeson, 2010)

Badanie to zostało wykonane na studentach amerykańskich, którzy mogą mieć nieco inne postawy rasowe niż ogół społeczeństwa (źródło). Najbardziej pilnym postulatem wydaje się, więc przetestowanie w/w na innych grupach społecznych (**ModOgol**). Również inne grupy kulturowe byłyby bardzo potencjalnie ciekawe, np. czy zjawisko byłoby zbliżone w innych krajach, gdzie występują żyjące od dawna populacje czarnoskórych i białych (np. RPA), albo w krajach, gdzie nie ma dużych mniejszości rasowych (np. Polska). (**ModOgol**). Dalej interesującym byłoby sprawdzenie, jak interakcje kształtują się w innych parach osób różnych ras niż osoba czarna-osoba biała (**ModSp; TeorMod**).

Następnie, bardzo ciekawym sprawdzianem teorii byłoby badanie, w którym badano by zachowanie par osób, gdzie tradycyjne role rasowe są odwrócone np. czarnoskóry menedżer i biały podwładny (**ModSp; TeorMod**). Można sobie wyobrazić szereg moderatorów sytuacyjnych np. drugiego badanego nie lubimy, ale musimy mu się przymilać, lub *vice versa*.

Podstawowe pytanie dotyczy jednak tego, czy opisywany efekt jest w jakiś sposób szczególny, czy też jest przejawem działania bardzo ogólnej prawidłowości, którą sformułowałbym tak „czasami chcemy/musimy się przymilić innym osobom, tworzymy wtedy wyobrażenie, jakie aspekty naszego zachowania mogą doprowadzić do tego celu i na tej podstawie modyfikujemy zachowanie”. Nawet jeśli w opisywanych prawidłowościach „rasowych” nie ma nic szczególnego, tak czy owak jest to wartościowy dodatek do teorii. Oczywiście, jeśli jest to działanie ogólniejszej teorii, to wtedy aspekty międzyrasowe nie będą szczególnie istotne w programie badawczym i można sobie wyobrazić nieskończenie wiele badań różnych par osób z różnych grup społecznych (zarówno grup znanych badanych, np. kibole, Romowie,

poznaniacy, jak i utworzonych *ad hoc* w laboratorium, tak jak w badaniach Tajfela, 1970).

ROZDZIAŁ 4

METODOLOGIA BADANIA WŁASNEGO

4.1. Wprowadzenie

We wcześniejszych częściach zaprezentowałem problemy psychologii ze zbieraniem rzetelnych danych (rozdział 1), prawa do generalizowania wyników badań do szerszego spektrum sytuacji i osób (rozdział 2) oraz problemów związanych z kumulacją wiedzy (rozdział 3).

Zaprezentowałem też trzy przykłady stosunkowo niedawnych badań, pokazując, jak, w moim przekonaniu, mógłby następować proces kumulacji wiedzy w związku z tymi odkryciami. Badania te mają dopiero kilka lat, więc trudno byłoby oczekiwać, żeby w tym momencie duża wiedza została zebrana. Dlatego w empirycznej części tej pracy postanowiłem przyjrzeć się badaniom, które mają dobrze ponad dekadę i powinny już być dobrze zweryfikowane. Mam nadzieję, że pozwoli to rzucić nieco światła na to, jak wyglądają systematyczne praktyki badawcze w psychologii społecznej (co może też wskazywać na stan psychologii w ogóle). Podstawowe cele, jakie sobie postawiłem, to przede wszystkim odpowiedzieć (w sposób ilościowy) na pytania:

- a) jak często występują dokładne replikacje badań w psychologii (co jest najważniejszym sprawdzianem rzetelności danych empirycznych, mającym zapewniać ich wysoką rzetelność zarówno w przypadku artefaktów, nieuświadomianych błędów, jak i świadomego oszustwa)
- b) jak często badacze dowodząc istnienia określonych efektów psychologicznych starają się korzystać z różnych metod, nie tylko metod opartych na miarach pośrednich, ale też jak często wśród tych metod pojawia się zachowanie (zgodnie z definicją Furra, 2009)
- c) jak często replikują badania na członkach różnych grup społecznych (w szczególności innych niż studenci) oraz różnych grup kulturowych (w szczególności innych niż mieszkańcy zachodniego świata)

Odpowiedzi te są stosunkowo proste do uzyskania (choć oczywiście analiza setek publikacji jest wysoce czasochłonna). Dodatkowo wyszukiwałem badania, które rozwijają teoretycznie oryginalne hipotezy, są pośrednią konsekwencją tych teorii

(szczególnie tzw. replikacje konceptualne), lub testują alternatywne teorie wyjaśniające efekt.

W uzupełnieniu, w sposób już bardziej jakościowy, na podstawie analizy systematycznych programów badawczych, chciałem zobaczyć, jak kolejne badania przyczyniają się do rozwoju teorii i kumulacji sensownej wiedzy.

4.2. Wybór próby

Zgodnie z rankingiem Journal Citation Report instytutu Thompson Reuters z 2010 roku (podobnie w ostatniej edycji z 2014 roku), cztery najczęściej cytowane czasopisma w psychologii społecznej, publikujące artykuły empiryczne, to:

- *Journal of Personality and Social Psychology*
- *Journal of Experimental Social Psychology*
- *Personality and Social Psychology Bulletin*
- *European Journal of Social Psychology*

Wydaje się, że lista tych czterech czasopism pokrywa się z nieformalną listą najlepszych czasopism w obszarze ogólnej psychologii społecznej. W rankingu przygotowanym przez czasopismo *The Serials Librarian* trzy spośród powyższych czasopism (bez EJSP) zostały zakwalifikowane do grona czterech najważniejszych czasopism z obszaru psychologii społecznej/osobowości (na tej liście znalazł się też *Journal of Personality*) (Burgard, 2001).

Inne często cytowane czasopisma empiryczne z listy JCR w podkategorii „Psychology. Social” publikują artykuły przeglądowe (np. *Personality and Social Psychology Review*), lub obejmują relatywnie wąski obszar zainteresowań psychologów społecznych (np. *Cyberpsychology Behavior and Social Networking*, *Organizational Behavior and Human Decision Processes*, czy *Journal of Health and Social Behavior*). Pominąłem też czasopisma poświęcone wyłącznie psychologii osobowości (np. *Journal of Personality*), choć artykuły poświęcone osobowości i różnicom indywidualnym z czasopism z wybranej czwórki były analizowane, jeśli miały wyraźne implikacje behawioralne.

Wybrałem te cztery czasopisma zakładając, że odkrycia w nich przedstawiane powinny być najszerzej analizowane, replikowane i wyjaśniane itd. Zakładam intuicyjnie, że odkrycia prezentowane w tych czasopismach będą w miarę dobrze oddawały najlepsze kierunki badań, a sytuacja w czasopismach o niższych wskaźnikach cytowań nie będzie lepsza.

Z tych czterech czasopism wyodrębniłem pierwsze 10 artykułów z rocznika 2000, spełniających następujące kryteria:

- są artykułami empirycznymi;
- opisują jakiś efekt psychologiczny wywierający wpływ na zachowanie; tj. różnice w zachowaniu w zależności od określonych zmiennych dyspozycyjnych lub sytuacyjnych lub ich interakcji (wyłączyłem artykuły, w których opisywany efekt nie ma, lub trudno określić, jakie ma implikacje dla codziennych zachowań);
- opisują ten efekt jako pierwsze (tj. z analizy artykułu i bibliografii nie wynika, że jest on replikacją wcześniejszych badań analizujących ten sam efekt);
- wyłączyłem artykuły, które były badaniami podłużnymi, szczególnie szerokimi badaniami ankietowymi w modelu korelacyjnym (w Polsce „Diagnoza społeczna” jest przykładem tego typu badania). Uzasadniam to tym, że takie badania trudno ze względów logistycznych dokładnie replikować, dlatego ich uwzględnienie może zaniżać częstość replikacji.

Listę artykułów uwzględnionych i odrzuconych (wraz z uzasadnieniem) przedstawia Załącznik 1.

Na podstawie analizy tych 40 artykułów wyodrębniłem kluczowe efekty, które zostały odkryte w tych tekstach. Dalsze poszukiwania replikacji dokładnych, konceptualnych i prac rozwijających teoretycznie odbywały się pod kątem rzeczonych efektów. Te 40 prac będę dalej nazywał „oryginalnymi/wyjściowymi” badaniami lub efektami.

Na podstawie bazy SCOPUS wyodrębniłem artykuły, które cytowały artykuły uwzględnione do analizy (Załącznik 1) (cytowania pierwszego stopnia).

Postanowiłem ograniczyć się do analizy tylko artykułów bezpośrednio cytujących te artykuły, a nie szukałem dalszych tekstów, które cytowały cytowania pierwszego stopnia (czyli cytowań drugiego, potem trzeciego itd. stopnia). Przyczyn jest kilka:

- Analiza tylko cytowań pierwszego stopnia jest wysoce czasochłonna. Analizowane przeze mnie artykuły wyjściowe były cytowane łącznie około 2500 razy. Uwzględnienie cytowań drugiego stopnia zwielokrotniłoby tę liczbę w sposób geometryczny.
- Przeglądanie cytowań drugiego stopnia przynosi bardzo niewiele nowych badań nawiązujących wprost lub koncepcyjnie do oryginalnego badania. Żeby to sprawdzić przeanalizowałem bardziej szczegółowo cykl badań nad efektem opisanym przez Haselton i Bussa (2000). Jest on cytowany w bazie SCOPUS 157 razy (w latach 2000-2012). Z tej liczby 15 artykułów weszło do zasadniczej analizy. Te 15 tekstów łącznie było cytowane w bazie SCOPUS ponad 400 razy. Po przeanalizowaniu abstraktów tych kolejnych 400 tekstów znalazłem jeden tekst, który luźno nawiązywał do wyjściowego tematu (dotyczył zmiany percepcji zainteresowania seksualnego pod wpływem alkoholu: Farris, Treat, Viken, 2010). Dodatkowo przeanalizowałem też artykuł przeglądowy (Lindgren, Parkhill, George, Hendershot, 2008) i meta-analizę (La France, Henningsen, Oates, Shaw, 2009) poświęcone tematowi. Na ich podstawie udało mi się wyodrębnić w interesującym okresie dwa dodatkowe teksty będące pośrednią replikacją konceptualną wyjściowego efektu (Levesque, Nave, Lowe, 2006; Bartolucci, Zeichner, 2003). Tak więc nie mam wątpliwości, że w literaturze mogą być gdzieś teksty nawiązujące do wyjściowego efektu i rozszerzenie poszukiwań o analizę cytowań drugiego (i wyższego) stopnia oraz o inne heurystyki może przynieść kilka dodatkowych badań. Ale nie sądzę, żeby było to znacznie więcej badań. Uwzględnienie powyższych 3 dodatkowych artykułów w niczym nie zmieniłoby zasadniczych jakościowych wniosków na temat tego cyklu badań (patrz rdz. 6.2.4)
- Cytowania drugiego i wyższego stopnia mogą zawierać potencjalnie zbliżone konceptualnie odkrycia, ale odkrycia te mogą się nie łączyć ze sobą w sposób ułatwiający pragmatyczną analizę. Wydaje się, że badacze analizują zwykle artykuły cytowane i cytujące dany tekst, oraz posiłkują się poszukiwaniami na podstawie innych heurystyk (np. artykuły tego samego autora; artykuły z określonymi słowami kluczowymi). Jeśli artykuł X jest cytowany przez artykuł Y, a ten z kolei jest cytowany przez artykuł Z (a Z nie cytuje X), to jest mniejsza szansa, że na podstawie X badacz znajdzie Z (i *vice versa*). Więc, choć teksty X i

Z mogą być potencjalnie zbliżone, to jednak badacze mogą w niewielkim stopniu je ze sobą zestawiać, o ile nie posługują się dodatkowymi heurystykami.

Dodatkowo ograniczyłem się do artykułów tylko do końca 2012 roku. Artykuły wyjściowe ukazały się na początku 2000 roku, więc analizowany okres to prawie 13 lat. Wydaje się, że jest to wystarczający okres, żeby cykl badań pozwolił z grubsza odpowiedzieć na pytania o empiryczną zasadność efektu, jego podstawowe moderatory, zastosowania itd. Dodatkowo znaczenie miało to, że analiza odbywała się na przestrzeni wielu miesięcy i artykuły analizowane w późniejszym czasie miałyby potencjalnie bogatszą bazę cytowań. Żeby nie musieć ciągle wracać do już wcześniej analizowanych tekstów, musiałem sobie postawić jakąś cezurę czasową.

4.3. Właściwa analiza

Po wyodrębnieniu cytowań pierwszego stopnia przeanalizowałem wszystkie tytuły i abstrakty rzeczonych tekstów cytujących 40 analizowanych artykułów.

Automatycznie wyłączałem z dalszej analizy teksty w języku innym niż angielski, oraz teksty przeglądowe lub teoretyczne. W pozostałych tekstach szukałem informacji w abstrakcie, które wskazywałyby na to, że to badanie jest replikacją dokładną, konceptualną lub rozwinięciem teoretycznym oryginalnego badania. Teksty spełniające ten warunek były kwalifikowane do dalszego etapu. W przypadku niejasności artykuł był uznawany wstępnie jako spełniający ten warunek.

Teksty zakwalifikowane do następnego etapu były następnie przeze mnie ściągane w pełnej wersji. Korzystając z tekstów udostępnianych w internecie, baz danych artykułów akademickich, oraz – w ostateczności – kontaktując się z autorem z prośbą o przesłanie pełnej wersji tekstu.

Artykuły (cytowania pierwszego stopnia) były dzielone na poszczególne badania (chyba, że tylko wybrane badania nawiązywały do oryginalnego artykułu). Przy każdym badaniu kodowano jakościowo następujące podstawowe informacje:

- odkrytą prawidłowość psychologiczną (o ile różniła się jakoś od oryginalnego efektu);
- krótki opis procedury badania.

Dodatkowo kodowałem informacje „techniczne”, tj. możliwe do jednoznacznego określenia. Te informacje to:

- rok badania;
- czy badanie było wykonane przez niezależny zespół (za „niezależny zespół” uznawałem sytuację, gdy nie było żadnych wspólnych autorów oryginalnego badania i badania analizowanego);
- grupa badana (np. studenci, populacja ogólna, itp.);
- pochodzenie grupy badanej (narodowość);
- liczba osób badanych; przy czym nie brano pod uwagę liczby osób z ewentualnych badań pilotażowych, ani liczby osób, która brała udział w ewentualnych badaniach przesiewowych, z których wyłoniono finalną próbę;
- średnia liczba osób w kratce (grupie eksperymentalnej) (dotyczy tylko eksperymentów z manipulacją międzygrupową); przy czym przy liczbie grup eksperymentalnych uwzględniano tylko te, gdzie manipulacja była międzygrupowa. Płeć i inne zmienne „międzygrupowe”, którymi się nie manipuluje nie były uwzględniane, chyba, że były kluczowe z punktu widzenia badanej hipotezy.

Kodowałem też zmienną, która wymaga pewnej dozy jakościowej interpretacji:

- czy badanie potwierdza wyjściową hipotezę; na podstawie własnej oceny kodowałem jako „tak/nie/częściowe”. Ten ostatni kod dotyczył tych sytuacji, gdy niektóre aspekty hipotezy zostały potwierdzone, a inne nie. Dodatkowo badania mające na celu zwiększenie precyzji teorii i oryginalne kodowałem jako „n/d” („nie dotyczy”). W przypadku tych pierwszych nie może być mowy o potwierdzeniu lub falsyfikacji wyjściowej hipotezy. W przypadku tych drugich, choć jest logicznie możliwe, że badania w oryginalnym artykule zawierają falsyfikację oryginalnej hipotezy, to jednak nigdy się tak nie zdarzyło w analizowanej próbie.

Nadmienię w tym miejscu, że pierwotnie planowałem kodować wielkość efektów, aby sprawdzić czy pierwsze badania nie dają zawyżonych efektów, a więc czy nie występuje tzw. „declining effect”, czyli zmniejszanie się siły efektu w kolejnych badaniach (Ioannidis, 2008; Schooler, 2011). Ostatecznie jednak zrezygnowałem z rejestrowania tej zmiennej. Po pierwsze artykuły miały bardzo różny standard, jeśli

chodzi o podawanie wielkości efektów. Często siła efektów nie była podana, lub nie podano statystyk, które pozwalałyby obliczyć taką wielkość. Najważniejszy jednak powód jest taki, że, jak się z czasem okazało, dokładnych replikacji było niezwykle mało. Olbrzymią większość badań stanowiły replikacje konceptualne (a nie dokładne) i to raczej dosyć luźno związane z oryginalnym badaniem. Tutaj rejestrowanie siły efektów w moim przekonaniu niewiele wnosi. Jeśli siła efektów spada w kolejnych replikacjach konceptualnych, zasadniczo nie mówi to nic o oryginalnym efekcie. Może to wskazywać na to, że pierwsze raporty o sile efektu są przesadzone, ale może też świadczyć o tym, że w innych warunkach lub w innych grupach taki efekt jest po prostu słabszy, w żaden sposób nie kwestionując rzetelności oryginalnej wielkości. Po analizie kilku pierwszych oryginalnych badań okazało się, że góra 30-40 % artykułów podaje statystyki pozwalające na wyliczenie wielkości efektów, a i tak ze względu na niską liczbę replikacji dokładnych wnioski z takiej analizy byłyby niewiele warte.

Następnie badanie było klasyfikowane pod kątem 4 kluczowych zmiennych. Te zmienne to:

1. **Cel badania.** Na podstawie operacjonalizacji zarysowanej w rdz. 3.10, następnie nieco zmodyfikowanej w trakcie analizy, wyróżniłem następujące klasyfikacje:
 - a. Replikacja dokładna (badanie było zaliczane jako replikacja dokładna, jeśli wszystkie elementy procedury były takie same lub bardzo podobne jak w oryginalnym badaniu; dotyczy to również grupy badanej);
 - b. Replikacja w miarę dokładna z uwzględnieniem podstawowych moderatorów efektu, czyli grupy kulturowej, grupy społecznej (przy założeniu, że procedura była bardzo zbliżona) lub grupy wiekowej (jeżeli wyjściowe badanie odbywało się na dorosłych, to w przy badaniu dzieci, akceptowałem to, że procedura może mocno odbiegać od oryginalnego badania);
 - c. Replikacja konceptualna, lub replikacja z uwzględnieniem moderatorów specyficznych (przykłady takich badań podałem w rdz. 3.11);
 - d. Generowanie odkryć na podstawie teorii (lub jej testowanie w pośredni sposób) - badanie nie będące replikacją konceptualną oryginalnego efektu, ale bezpośrednio testujące teorię, która stała za wyjściowym efektem.

Przykładowo, w omawianym już badaniu Haselton i Buss (2000) pośrednim wskaźnikiem teorii jest to, że mężczyźni częściej deklarowali, że ich awanse miłosne zostały odrzucone przez kobiety, co może (przy dodatkowych założeniach) wskazywać na to, że częściej błędnie oceniali, że kobiety są nimi zainteresowane seksualnie.

- e. Badanie mające na celu zwiększenie precyzji teorii, np. rozstrzyganie między konkurencyjnymi teoriami wyjaśniającymi efekt, znalezienie dodatkowych moderatorów (ale bez próby konceptualnej replikacji) itd.;
- f. Oryginalne – tak zawsze kodowałem badania z wyjściowych artykułów.

Nadmienię, że czasami trudno było o jednoznaczną klasyfikację, szczególnie odróżnienie sytuacji c od d.

2. Rodzaj badania. Wyróżniłem następujące rodzaje:

- a. Eksperyment (gdzie miała miejsce manipulacja jakąś zmienną i randomizacja);
- b. Eksperyment terenowy (jw. ale badania w początkowym etapie nie wiedzieli, że uczestniczą w eksperymencie psychologicznym);
- c. Badanie korelacyjne;
- d. Dane behawioralne lub wytwory ludzkie, które powstały niezależnie od badania i zostały przez badacza przeanalizowane pod kątem wyjściowej hipotezy (np. tak jak Kuhle, 2011, który analizował zachowania ludzi w programie telewizyjnym);
- e. Dane opisowe, którego celem jest najogólniej stwierdzenie „jak jest”, bez badania związków z innymi zmiennymi (przykładem może być Dunbar, Marriott, Duncan, 1997, którzy przysłuchiwali się treści rozmów prowadzonych w miejscach publicznych); szczególną formą są „dane opisowe (otwarte)”, gdzie badany może zaprezentować spontaniczną reakcję (np. wypowiedź).

Rodzaj manipulacji eksperymentalnej:

- a. Realistyczna;
- b. Nierealistyczna;

- c. Nie dotyczy (w przypadku badań nieeksperymentalnych).

W rozdziale 2.3 przedstawiłem przykłady różnych sposobów manipulowania „motywacją romantyczną”. Idea realizmu manipulacji eksperymentalnej jest trudna do wyczerpującego zdefiniowania, dlatego zastrzegam, że jest to zmienna o naturze najmocniej chyba subiektywnej ze wszystkich tutaj opisanych zmiennych. Ogółem zakładam, że różne formy manipulacji polegające na pokazywaniu badanym określonych słów lub obrazków (w domyśle wzbudzania określonych nastawień), wyobrażania sobie określonych sytuacji, lub prezentowania hipotetycznych sytuacji są nierealistyczne. Przykłady. Jeśli grupa zachowuje się wrogo lub przyjaźnie wobec badanego mamy do czynienia z manipulacją realistyczną, jeśli natomiast badany jest pytany: gdyby „grupa zachowywała się wobec ciebie wrogo / przyjaźnie, jakbyś zareagował” mamy do czynienia z manipulacją nierealistyczną.

Czasami manipulowano więcej niż jedną zmienną i na różne sposoby (jedną realistycznie, drugą nierealistycznie). W takim wypadku kodowałem rodzaj dla tej manipulacji, która jest istotniejsza teoretycznie (co sporadycznie też nie było jednoznaczne do określenia).

Natura zmiennej zależnej:

- a. Zmienna biologiczna (w tym wyniki metod neuroobrazowania);
- b. Zachowanie (zgodnie z definicją Furra, 2007⁷⁴);
- c. Zachowanie w grach ekonomicznych (ten dodatkowy rodzaj zmiennej postanowiłem wyróżnić ze względu na to, że w moim przekonaniu nie do końca wpisuje się w a.);
- d. Zmienna związana ze sprawnością układu poznawczego (szczególnie pamięć), dalej zwana „zmienną poznawczą”;
- e. Zachowanie na bieżąco raportowane (np. dzienniczki aktywności);
- f. Zmienna utajona (np. wynik w teście IAT, w teście decyzji leksykalnych itd.);
- g. Zmienna samoopisowa (szeroko pojmowana).

⁷⁴ „Wypowiedzi (wyłączając raporty werbalne w kontekście testowania psychologicznego) lub ruchy, które są potencjalnie dostępne dla uważnego obserwatora korzystającego z normalnych procesów zmysłowych” (Furr, 2009, s. 372)

Sporadycznie zdarzało się, że w badaniu wykorzystywano dwie zmienne zależne o różnej naturze (np. wynik badania kwestionariuszowego i zachowanie w grze ekonomicznej). Każdorazowo taka sytuacja dotyczyła wykorzystania zmiennej samoopisowej i jakiejś dodatkowej zmiennej. W takim wypadku kodowałem tylko tą dodatkową zmienną.

4.4. Dane surowe badania

Zaznaczam, że o ile pewne dane bardzo łatwo zakodować (np. liczbę osób badanych), o tyle w przypadku niektórych informacji jest to często trudne i niejednoznaczne. Dotyczy to szczególnie celu badania i rodzaju manipulacji. Dlatego żeby ułatwić intersubiektywną kontrolę trafności moich analiz, w załączniku 2 dołączam wszystkie dane surowe, jakie wytworzyłem.

4.5. Podsumowanie kodowanych zmiennych

W Tabeli 4.1. przedstawiłem podsumowanie zmiennych wykorzystanych do analizy ilościowej (rozdział 5).

TABELA 4.1. Zmienne użyte do analizy ilościowej i sposób ich kodowania

Zmienne	Sposób kodowania
ROK	[Wartość liczbowa] (ostatnie 2 cyfry roku)
ZESPÓŁ (CZY NIEZALEŻNY)	TAK/NIE/N.D
POTWIERDZENIE WYJŚCIOWEJ HIPOTEZY	TAK/NIE/CZĘŚCIOWE/N.D.
GRUPA BADANA	[Tekstowo] (zgodnie ze stanem faktycznym)
NARODOWOŚĆ BADANYCH	[Tekstowo] (zgodnie ze stanem faktycznym)
LICZBA OSÓB BADANYCH	[Wartość liczbowa]
LICZBA OSÓB W KRATCE	[Wartość liczbowa]
CEL BADANIA	Replikacja dokładna/ replikacja konceptualna/ moderatory podstawowe/ generowanie odkryć / zwiększanie precyzji teorii
RODZAJ BADANIA	Eksperyment / Eksperyment terenowy / Badanie korelacyjne / Dane wytworzone poza badanie / Dane opisowe / Dane opisowe (otwarte)
REALIZM MANIPULACJI EKSPERYMENTALNEJ	Realistyczna / Nierealistyczna / N.D.
NATURA ZMIENNEJ WYJAŚNIANEJ	Zachowanie / Zachowanie na bieżąco raportowane / Zachowanie w grach ekonomicznych / Zmienna poznawcza / Zmienna utajona / Zmienna samoopisowa / Zmienna biologiczna

ROZDZIAŁ 5

ANALIZA ILOŚCIOWA WYNIKÓW

W tym rozdziale przedstawię podstawowe charakterystyki ilościowe analizowanej próby badań empirycznych. Interpretacja tych wyników będzie tutaj ograniczona do minimum. Integracja wyników ilościowych, w połączeniu z analizą jakościową (rozdział 6), zostanie przedstawiona w rozdziale 7.

5.1. Pozyskana próba artykułów do analizy

W Załączniku 1 znajduje się lista artykułów, które spełniały podane wcześniej kryteria (rdz. 4.2.) i zostały ostatecznie uwzględnione w analizie. Artykuły te zawierają łącznie 90 badań. Będę je dalej nazywał oryginalnymi lub wyjściowymi badaniami. W Załączniku 1 zawarłem również informacje o merytorycznych przyczynach odrzucenia artykułów, które kwalifikowałyby się tylko na podstawie pozycji.

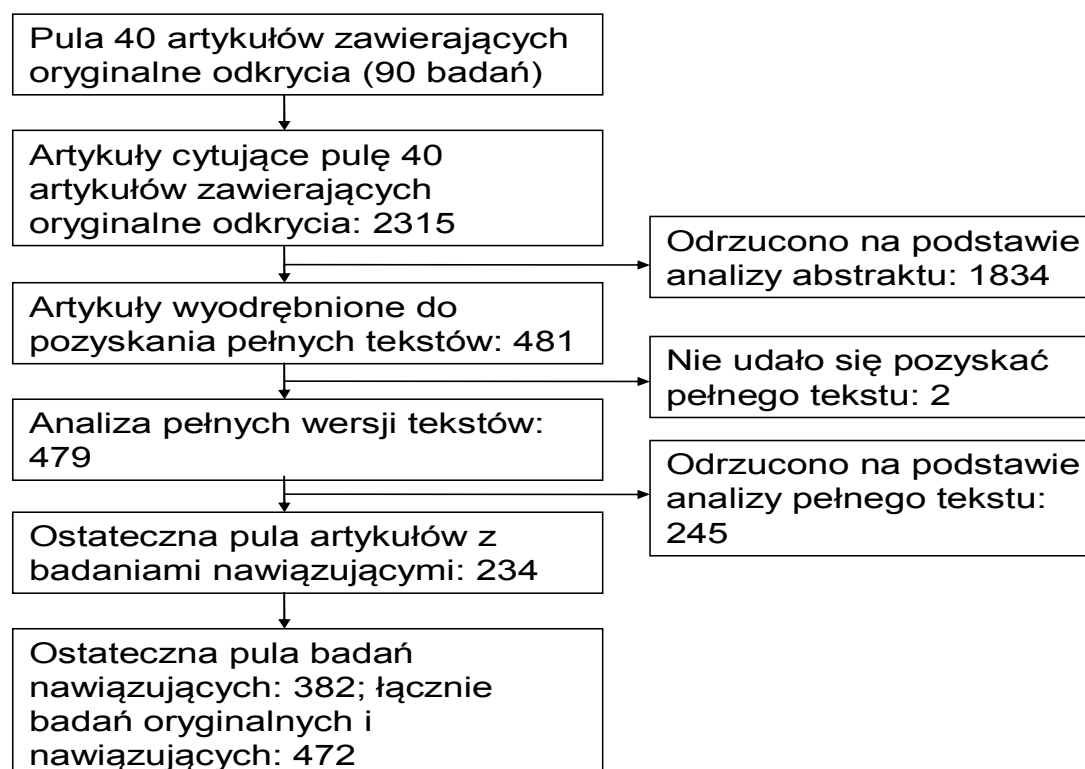
Po wyłonieniu 40 artykułów z oryginalnymi badaniami analizowałem dokładnie treści teoretyczne i materiał empirycznych w nich omówiony. Określiłem treść podstawowego odkrycia (lub odkryć), jakie zostały w nich opisane i empirycznie przetestowane. Jako że teksty często badają wiele hipotez, sporadycznie odrzucałem mniej ważne lub poboczne hipotezy (przykładowo, w badaniu JPSP4 taką odrzuconą hipotezą było twierdzenie, że efekt przeceniania zainteresowania seksualnego zanika, gdy badani mężczyźni analizują zachowanie swojego rodzeństwa). W Załączniku 2 przy każdym cyklu badań znajduje się opis każdego odkrycia.

Następnie wyodrębniałem pulę badań nawiązujących do oryginalnych badań poprzez fakt replikacji dokładnej, konceptualnej i teoretycznej, lub rozszerzanie rozumienia wyjściowego efektu. W dalszej części tekstu będę je nazywał pulą badań nawiązujących, a pulę badań oryginalnych i nawiązujących łącznie będę nazywał „cyklem badań”.

Najpierw w analizie wyodrębniłem wszystkie teksty cytujące te badania w bazie SCOPUS. Na podstawie analizy abstraktów wyodrębniłem artykuły do analizy pełnych wersji tekstów. Poprzez repozytoria artykułów, bazy danych lub kontakt z

badaczami pozyskiwałem pełne wersje artykułów (udało się pozyskać prawie wszystkie teksty⁷⁵). Analizowałem następnie pełne wersje tekstów, szczególnie dział metody i wyników badań i jeśli spełniały założone kryteria wchodziły ostatecznie w skład puli badań nawiązujących. Wszystkie etapy analizy przedstawia *Rycina 5.1*.

RYCINA 5.1. Etapy pozyskiwania puli badań empirycznych do analizy



5.2. Opis artykułów zawierających oryginalne badania

5.2.1. Liczba cytowań

W Tabeli 5.1. zawarłem informacje o częstotliwości cytowań artykułów, które zawierały pulę oryginalnych badań. Wszystkie wartości to liczba cytowań do końca 2012 roku, gdyż, jak zaznaczyłem (rdz. 4.2), do puli badań nawiązujących mogły wejść teksty opublikowane do 2012 roku włącznie. W dalszej części dla wygody i oszczędności miejsca nie będę powtarzał tytułu, tylko uwzględnił kod artykułu (pełny adres bibliograficzny tych artykułów znajduje się w Załączniku 2).

⁷⁵ Nie udało się pozyskać dwóch artykułów: Hatta, Ohbuchi (2003), Heim, Asting, Schliemann (2002)

TABELA 5.1. Częstość cytowań artykułów zawierających pulę badań oryginalnych⁷⁶

Kod artykułu	Tytuł artykułu	liczba cytowań (do końca 2012 roku)
JPSP1	How far do we go beyond the information given? The impact of knowledge activation on interpretation and inference	34
JPSP2	The relation between self-discrepancies and emotion: The moderating roles of self-guide importance, location relevance, and social self-domain centrality	31
JPSP3	When are we better than them and they worse than us? A closer look at social discrimination in positive and negative domains	75
JPSP4	Error management theory: A new perspective on biases in cross-sex mind reading	238
JPSP5	Decision-induced focusing in social dilemmas: Give-some, keep-some, take-some, and leave-some dilemmas	46
JPSP6	The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance	99
JPSP7	Varieties of groups and the perception of group entitativity	195
JPSP8	Motivated cultural cognition: The impact of implicit cultural theories on dispositional attribution varies as a function of need for closure	152
JPSP9	On inferring one's beliefs from one's attempt and consequences for subsequent compliance	7
JPSP10	Couples' shared participation in novel and arousing activities and experienced relationship quality	105
JESP1	Strategy and Fairness in Social Decision Making: Sometimes It Pays to Be Powerless	65
JESP2	Rapport in Conflict Resolution: Accounting for How Face-to-Face Contact Fosters Mutual Cooperation in Mixed-Motive Conflicts	115
JESP3	Perceived Variability and Stereotype Change	33
JESP4	Contents and Correlates of Whites' and Blacks' Racial Attitudes	55
JESP5	Moderation of Priming by Goals: Feeling Entitled to Judge Increases Judged Usability of Evaluative Primes	18
JESP6	An Experimental Test of the Role of Alcohol in Relationship Conflict	27
JESP7	Sources of Mental Contamination: Comparing the Effects of Self-Generated versus Externally Provided Primes	40
JESP8	Group-Based Dominance and Opposition to Equality as Independent Predictors of Self-Esteem, Ethnocentrism, and Social Policy Attitudes among African Americans and European Americans	138
JESP9	Converging Interracial Consequences of Exposure to Violent Rap Music on Stereotypical Attributions of Blacks	29

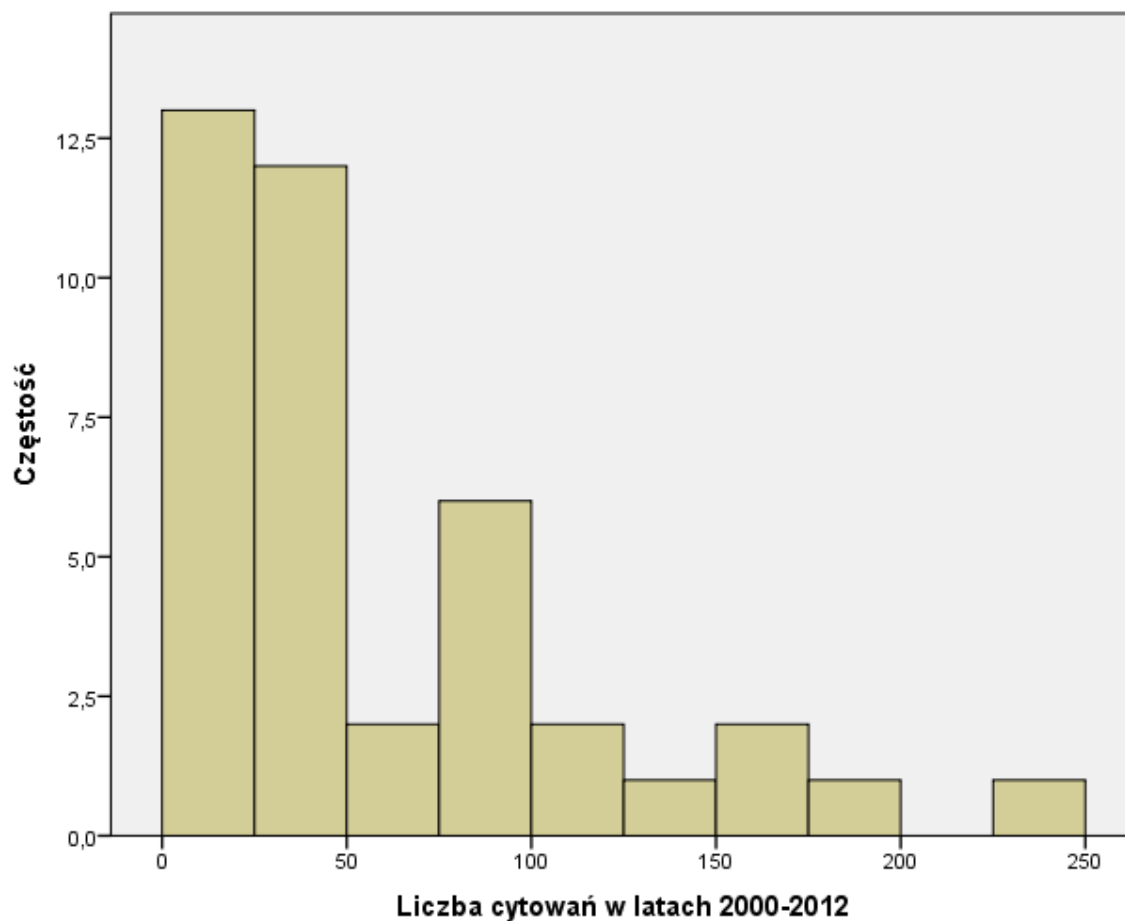
⁷⁶ We wszystkich tabelach w tym rozdziale źródłem danych jest opracowanie własne na podstawie analizy opisanej w rozdziale 4

JESP10	Distinguishing Gains from Nonlosses and Losses from Nongains: A Regulatory Focus Perspective on Hedonic Intensity	197
PSPB1	The roles of motivation and ability in controlling the consequences of stereotype suppression	41
PSPB2	On the encoding of stereotype-relevant information under cognitive load	41
PSPB3	Does personal morality exacerbate or restrain retaliation after being harmed?	13
PSPB4	Sandbagging as a self-presentational strategy: Claiming to be less than you are	19
PSPB5	Beyond self-presentation: Evidence for self-criticism among Japanese	93
PSPB6	When an adversary is caught telling the truth: Reciprocal cooperation versus self-interest in distributive bargaining	10
PSPB7	Proximal and distal defenses in response to reminders of one's mortality: Evidence of a temporal sequence	78
PSPB8	The role of internal and external sources of evaluation in motivating task performance	11
PSPB9	Detecting deception and judge's involvement: Lower task involvement leads to better lie detection	33
PSPB10	The distinctiveness effect in social interaction: Creation of a self-fulfilling prophecy	7
EJSP1	Social dominance orientation, anti-egalitarianism and the political psychology of gender: An extension and cross-cultural replication	77
EJSP2	Conversation as a resource for influence: Evidence for prototypical arguments and social identification processes	14
EJSP3	Learning to play golf under different goal conditions: Their effects on irrelevant thoughts and on subsequent control strategies	17
EJSP4	The reinstatement of dissonance and psychological discomfort following failed affirmations	26
EJSP5	Value priorities and subjective well-being: Direct relations and congruity effects	123
EJSP6	Communication context, explanation, and social judgment	6
EJSP7	The typical student as an in-group member: Eliminating optimistic bias by reducing social distance	24
EJSP8	Interdependence in negotiation: Effects of exit options and social motive on distributive and integrative negotiation	40
EJSP9	Counterfactual intensity	22
EJSP10	Acculturation discrepancies and well-being: The moderating role of conformity	42

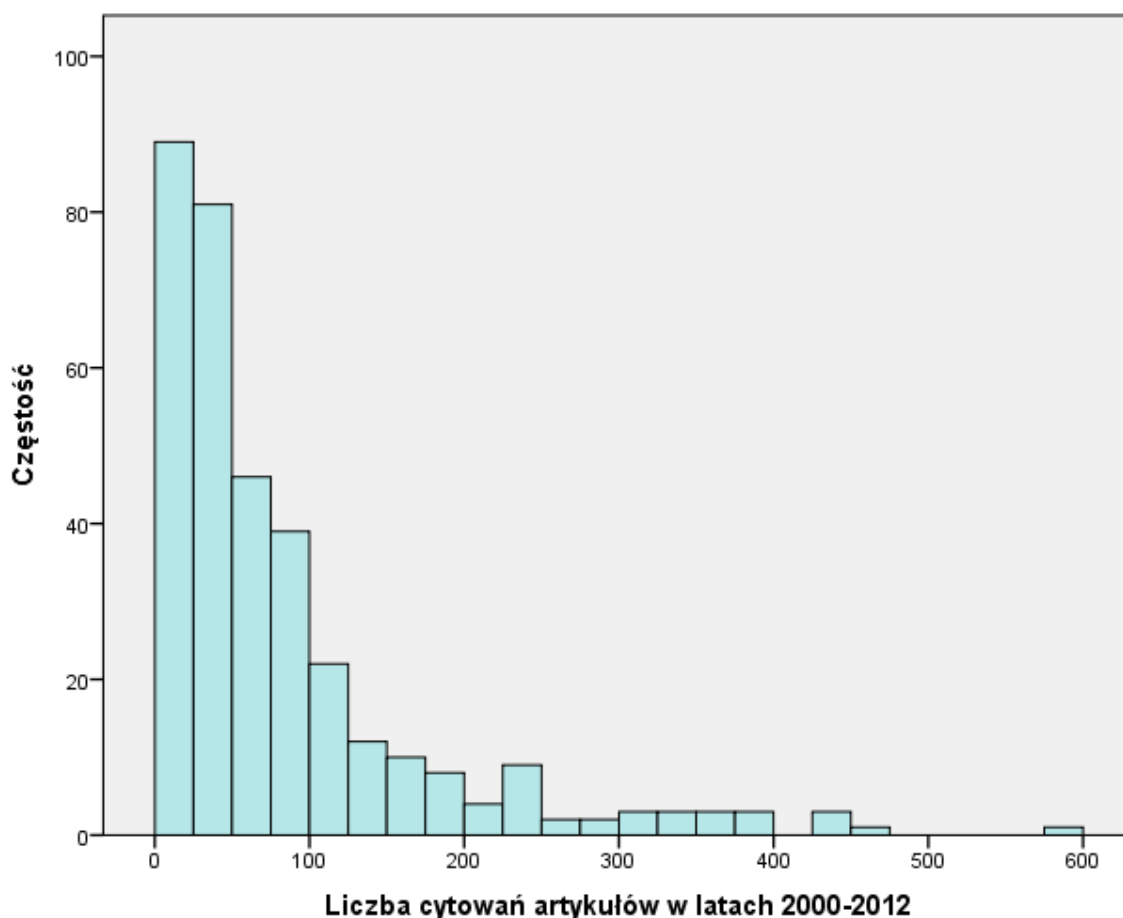
Rozkład częstości cytowań w analizowanej próbie artykułów jest silnie prawoskośny, co jest zgodne z tendencjami opisywanymi w literaturze (np. Opthof, Coronel, Piper,

2004). Jak widać na wykresie 5.2, częstość cytowań wszystkich artykułów z tych 4 czasopism z rocznika 2000 jest również silnie prawoskośna.

WYKRES 5.1. Rozkład częstości cytowań 40 artykułów z pulą badań oryginalnych



WYKRES 5.2. Rozkład częstości cytowań wszystkich artykułów z rocznika 2000 z 4 czasopism (JPSP, PSPB, JESP, EJSP).



W tabeli 5.2. przedstawiłem informacje o średniej częstości cytowań (do końca roku 2012) w zależności od czasopisma. Co nie jest zaskakujące, kolejność układa się w przybliżeniu zgodnie z wartością *impact factor* dla poszczególnych czasopism.

Tabela 5.2. zawiera również informacje o częstości cytowań pozostałych artykułów z rocznika 2000 z 4 czasopism, z których wyodrębniono pulę badań oryginalnych. Jako że rozkłady te odbiegają wyraźnie od rozkładu normalnego (Test Kołmogorova Smirnova $p < 0,001$), użyłem testu nieparametrycznego U Manna-Whitneya do porównania częstości cytowań artykułów z pulą badań oryginalnych i pozostałych artykułów z rocznika 2000. W przypadku żadnego czasopisma nie zanotowałem istotnych różnic. Podobnie, jeśli zbiorczo przeanalizować wszystkie artykuły, różnica również jest nieistotna statystycznie ($U = 5152$; $z = 1,48$; $p > 0,1$). Wskazuje to na to, że 40 artykułów z badaniami oryginalnymi jest względnie podobna do reszty badań z

tego rocznika, przynajmniej jeśli chodzi o zainteresowanie społeczności badawczej mierzone częstością cytowań.

TABELA 5.2. Średnia liczba cytowań artykułów w zależności od czasopisma

Czasopismo	Artykuły z pulą badań oryginalnych		Pozostałe artykuły z rocznika 2000	
	Liczba cytowań	n	Liczba cytowań	n
JPSP	98,2	10	134	132
JESP	59,6	10	39,6	22
PSPB	34,6	10	50,73	115
EJSP	39,1	10	38,4	32
WSZYSTKIE	57,8	40	65,6	301

5.2.2. Podstawowe charakterystyki cykli badań w zależności od czasopisma

W Tabeli 5.3. przedstawiłem liczbę artykułów z badaniami nawiązującymi oraz liczbę osób badanych w badaniach oryginalnych, badaniach nawiązujących oraz podgrupie badań nawiązujących z replikacjami⁷⁷.

TABELA 5.3. Podstawowe charakterystyki artykułów z pulą badań oryginalnych (liczba artykułów nawiązujących i liczba osób badanych)

Kod artykułu	Liczba artykułów z badaniami nawiązującymi	Liczba osób badanych w badaniach oryginalnych	Liczba osób badanych w badaniach nawiązujących	W tym liczba osób badanych w artykułach zawierających replikacje dokładne, konceptualne i moderatory podstawowe
JPSP1	0	521	0	0
JPSP2	0	619	0	0
JPSP3	0	149	0	0
JPSP4	15	506	3943	1531
JPSP5	0	307	0	0
JPSP6	8	455	1878	1525
JPSP7	38	395	10044	3249
JPSP8	6	249	1790	262
JPSP9	0	1100	0	0
JPSP10	0	444	0	0

⁷⁷Replikacje dokładne, konceptualne i replikacje badające podstawowe moderatory.

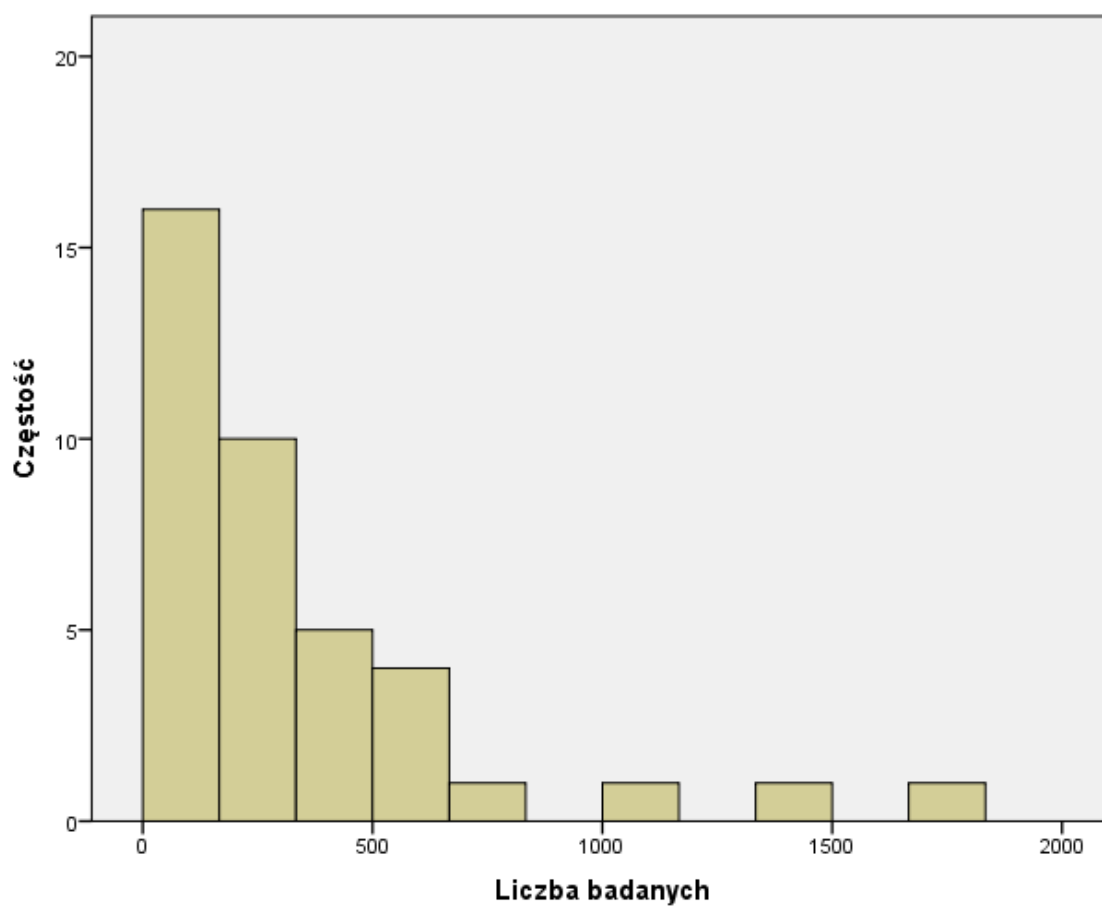
JESP1	7	187	1184	1184
JESP2	9	176	1129	1009
JESP3	4	132	1480	1378
JESP4	9	791	3059	2043
JESP5	2	209	308	108
JESP6	5	56	537	0
JESP7	1	113	64	0
JESP8	7	1675	18375	7520
JESP9	7	180	2980	772
JESP10	4	358	1434	1434
PSPB1	9	286	897	841
PSPB2	9	109	2070	2070
PSPB3	1	80	40	40
PSPB4	3	615	893	382
PSPB5	23	263	8808	1288
PSPB6	2	77	731	188
PSPB7	3	64	487	0
PSPB8	4	452	1098	565
PSPB9	3	66	631	631
PSPB10	1	78	26	26
EJSP1	20	6997	13716	9295
EJSP2	1	24	48	48
EJSP3	0	103	0	0
EJSP4	8	100	1552	427
EJSP5	14	1343	4023	3577
EJSP6	1	108	48	48
EJSP7	1	173	90	90
EJSP8	2	142	288	288
EJSP9	6	197	1120	883
EJSP10	1	100	220	0

Uwagi: pominięto cztery badania, w których przeanalizowano ponownie wyniki innych badań

W każdym przypadku rozkład jest wysoce prawoskośny. Artykuły przyciągają uwagę społeczności badawczej w bardzo różnym stopniu, oraz istnieje duże zróżnicowanie ilości badań nawiązujących.

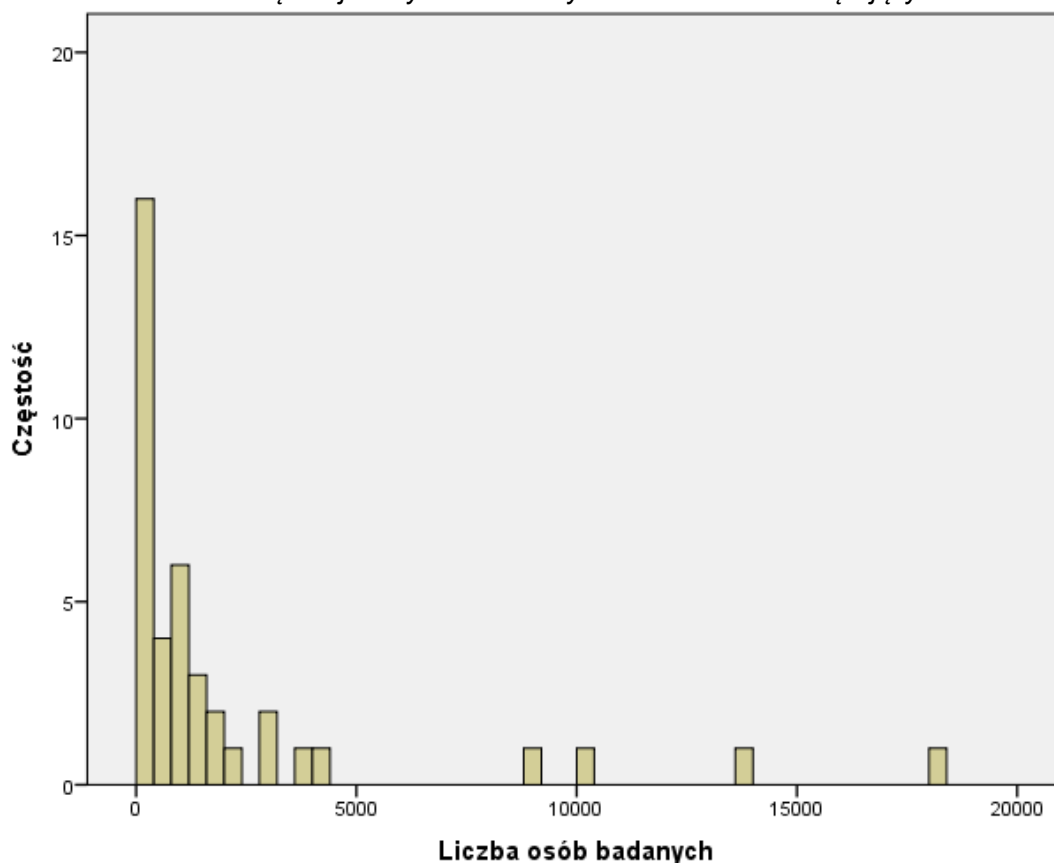
Na wykresie 5.3. przedstawiłem rozkład liczebności osób badanych w badaniach oryginalnych, a na wykresie 5.4. w kompletnym cyklu badań. Ponownie potwierdza się reguła wysokiej prawoskośności takiego rozkładu. Warto zwrócić uwagę, że pewna część badań (7 na 40) nie miała żadnych badań nawiązujących, więc najczęstszą wartością w tym przypadku jest 0. Jednak usunięcie wyników z zerową liczbą osób badanych nie zmieniłoby prawoskośnego charakteru danych.

WYKRES 5.3. Rozkład liczby osób badanych w badaniach oryginalnych



Uwaga: dla większej czytelności usunięto jedną skrajnie wysoką wartość (6997)

WYKRES 5.4. Rozkład łącznej liczby osób badanych w badaniach nawiązujących



W Tabeli 5.4. przedstawiłem te same dane z podziałem na 4 czasopisma. Okazuje się, co jest trochę zaskakujące, że liczba osób badanych zarówno w badaniach oryginalnych, jak i nawiązujących niekoniecznie musi być proporcjonalna do wysokości wskaźnika *impact factor*. Wyniki te przedstawiam tylko w celach poglądowych, gdyż wyciąganie z nich daleko idących wniosków wydaje się mało racjonalne. Wynik ten zależy przede wszystkim od rozkładu artykułów z wysokimi wynikami, a jak pokazałem w tabeli 5.3., rozrzut ten jest bardzo duży. Czasopisma więc swoje niskie lub wysokie wyniki zawdzięczają przede wszystkim artykułom z wysokimi liczbami osób badanych (lub braku takowych).

TABELA 5.4. Liczba artykułów z badaniami nawiązującymi oraz liczba osób badanych wg czasopisma

Czasopismo	Liczba artykułów z badaniami nawiązującymi	Liczba osób badanych w badaniach oryginalnych	Liczba osób badanych w badaniach rozwijających	Replikacje (patrz przypis 75)	SUMA WG CZASOPISMA
JPSP	67	4745	17655	6567	28967
JESP	55	3877	30550	15448	49875
PSPB	58	2090	15681	6031	23802
EJSP	54	9287	21105	14656	45048
SUMA	234	19999	84991	42702	

Na uwagę zasługuje też fakt, że w zarówno w badaniach oryginalnych w całym cyklu badań (Tabela 5.5.) występowało zróżnicowanie pochodzenia prób. W JPSP i JESP dominowały wyraźnie próby amerykańskie, natomiast w pozostałych dwóch czasopismach Amerykanie stanowili mniejszość.

TABELA 5.5. Odsetek prób badanych amerykańskich i nieamerykańskich w zależności od czasopisma

			Narodowość badanych	
			Amerykanin	nie-Amerykanin
CZASOPISMO	JPSP	Liczebność	101	54
		%	65,2%	34,8%
	JESP	Liczebność	71	41
		%	63,4%	36,6%
	PSPB	Liczebność	48	58
		%	45,3%	54,7%
	EJSP	Liczebność	41	57
		%	41,8%	58,2%

5.2.3. Klasyfikacja badania, rodzaj badań nawiązujących oraz rodzaj zmiennej zależnej w zależności od czasopisma

W tabelach 5.6., 5.7. i 5.8. przedstawiłem jakiego rodzaju występowały (w zależności od czasopismach): a) cele badań nawiązujących, b) rodzaj badania, c) rodzaje zmiennych zależnych. W moim przekonaniu ze względu na stosunkowo ograniczoną (w kategoriach statystycznych) próbę (10 artykułów z każdego czasopisma) obserwowane zróżnicowanie nie musi wynikać z systematycznych różnic, ale być wynikiem przypadku. Jak przedstawię dalej, rodzaj badań nawiązujących zależy w dużej mierze od charakteru badań oryginalnych. Odnotuję jednak, że w przypadku JPSP nadreprezentowane były artykuły, które zwiększały wyjaśnienie efektu (bez jego replikacji). Z kolei w przypadku JESP dosyć częste były replikacje konceptualne.

Replikacje dokładne i replikacje mające na celu sprawdzenie podstawowych moderatorów (grupy społecznej, kultury) były bardzo rzadkie w każdym przypadku. W PSPB częściej pojawiały się eksperymenty, a w EJSP badania korelacyjne. Z tego prawdopodobnie wynika, że zdecydowana większość zmiennych zależnych w tym ostatnim czasopiśmie to zmienne samoopisowe..

TABELA 5.6. Cel badań nawiązujących w zależności od czasopisma

		Klasyfikacja badania					
		oryginalne	replikacja dokładna	replikacja konceptualna / moderatorzy specyficzne	moderatorzy podstawowe	generowanie odkryć / testowanie teorii	zwiększanie precyzji teorii
JPSP	Liczebność	39	0	47	2	38	29
	%	25,2%	0,0%	30,3%	1,3%	24,5%	18,7%
JESP	Liczebność	21	4	56	5	15	11
	%	18,8%	3,6%	50,0%	4,5%	13,4%	9,8%
PSPB	Liczebność	15	1	45	3	32	10
	%	14,2%	,9%	42,5%	2,8%	30,2%	9,4%
EJSP	Liczebność	15	6	44	6	17	11
	%	15,2%	6,1%	44,4%	6,1%	17,2%	11,1%

Tabela 5.7. Rodzaj badania w zależności od czasopisma

			Rodzaj badania					
			eksperyment	eksperyment terenowy	badanie korelacyjne	dane opisowe	dane opisowe (otwarte)	dane wytworzone poza badaniem
CZASOPISMO	JPSP	Liczebność	72	7	55	20	0	1
		%	46,5%	4,5%	35,5%	12,9%	0,0%	,6%
	JESP	Liczebność	61	1	43	4	2	1
		%	54,5%	,9%	38,4%	3,6%	1,8%	,9%
	PSPB	Liczebność	79	0	23	1	1	2
		%	74,5%	0,0%	21,7%	,9%	,9%	1,9%
	EJSP	Liczebność	43	0	49	2	5	0
		%	43,4%	0,0%	49,5%	2,0%	5,1%	0,0%

TABELA 5.8 Rodzaj zmiennej zależnej w zależności od czasopisma

			Rodzaj zmiennej zależnej						
			zachowania	zmienna	zachowania	zachowanie	zmienna	zmienna	zmienna
			e	poznawcza	w grach	na bieżąco	samoopisowa	utajona	biologiczna
					ekonomicznych	raportowane			
CZASO PISMO	JPSP	Liczebność	11	0	4	0	132	8	0
		%	7,1%	0,0%	2,6%	0,0%	85,2%	5,2%	0,0%
	JESP	Liczebność	4	0	26	2	79	0	1
		%	3,6%	0,0%	23,2%	1,8%	70,5%	0,0%	,9%
	PSPB	Liczebność	10	25	6	0	57	8	0
		%	9,4%	23,6%	5,7%	0,0%	53,8%	7,5%	0,0%
	EJSP	Liczebność	1	0	4	0	93	1	0
		%	1,0%	0,0%	4,0%	0,0%	93,9%	1,0%	0,0%

5.2.4. Czasopismo a potwierdzenie oryginalnej hipotezy i niezależność zespołu badawczego

W tabeli 5.9. przedstawiłem częstości występowania niezależnych zespołów badawczych w badaniach nawiązujących, a w tabeli 5.10. częstość potwierdzenia wyjściowej hipotezy w tychże badaniach (dla przypomnienia: badania oryginalne oraz zwiększające rozumienie efektu były kwalifikowane jako nieweryfikujące prawdziwości wyjściowej teorii) w zależności od czasopisma. Nie ma zasadniczych różnic między trzema czasopismami, jedynie w JESP nieco częściej niezależne zespoły przystępowały do dalszych badań (alternatywnie można powiedzieć, że autorzy z wyjściowych badań rzadziej podejmowali weryfikacje) ($\chi^2 = 8,87$; $df = 3$; $p < 0,05$). W JPSP najrzadziej wyjściowa hipoteza była falsyfikowana (jeśli analizować tylko sytuacje, gdy hipoteza jest jednoznacznie potwierdzona lub niepotwierdzona, wówczas czasopismo różnicuje częstość falsyfikacji: $\chi^2 = 11,88$; $df = 3$; $p < 0,01$).

Różnice te nie muszą odzwierciedlać systematycznego trendu związanego z charakterystyką badań w danym czasopiśmie, ale podobnie jak w poprzednim przypadku wynikać z charakterystyk określonych cykli badawczych.

TABELA 5.9. Autorstwo badań z puli badań nawiązujących

			Niezależny zespół	
			NIE	TAK
CZASOPISMO	JPSP	Liczebność	40	76
		%	34,5%	65,5%
	JESP	Liczebność	17	74
		%	18,7%	81,3%
	PSPB	Liczebność	34	57
		%	37,4%	62,6%
	EJSP	Liczebność	26	58
		%	31,0%	69,0%

TABELA 5.10. Częstość potwierdzenia wyjściowej hipotezy w zależności od czasopisma

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
CZASOPISMO	JPSP	Liczebność	3	79	5
		%	3,4%	90,8%	5,7%
	JESP	Liczebność	13	48	19
		%	16,3%	60,0%	23,8%
	PSPB	Liczebność	11	46	24
		%	13,6%	56,8%	29,6%
	EJSP	Liczebność	11	47	14
		%	15,3%	65,3%	19,4%

5.3. Opis cykli badań – podstawowe charakterystyki

5.3.1. Niezależny zespół

30,6 % badań nawiązujących zostało wykonanych przez ten sam zespół (tj. przynajmniej jeden ze współautorów badania oryginalnego był współautorem badania nawiązującego). Jeśli jednak uwzględnić cały cykl badań (tj. łącznie z badaniami oryginalnymi) w takim wypadku odsetek wzrasta do 44 %. Tak więc nieznacznie więcej niż połowa badań w danym cyklu to niezależne weryfikacje lub rozszerzenia efektu. Ten odsetek może być nawet niższy, gdyż nie można wykluczyć, że część badań jest publikowanych przez osoby związane z autorami oryginalnych badań (np. współpracownicy, doktoranci itp.) i w takim przypadku trudno mówić o całkowicie niezależnej replikacji. Oczywiście, sam fakt relatywnie wysokiej częstości badań nad efektem przez te same osoby nie jest jeszcze oznaką patologii. Określony temat badawczy w sposób naturalny może interesować szczególnie tych badaczy, którzy już się nim wcześniej zajmowali.

5.3.2. Potwierdzenie wyjściowej hipotezy

W 220 (68,8 %) przypadkach wyjściowa hipoteza została potwierdzona, w 38 (11,9 %) przypadkach nie została potwierdzona, a w 62 (19,4 %) nie jest to jednoznaczne (tzn. niektóre elementy wyjściowej hipotezy zostały potwierdzone, inne nie).

Pozostałe badania nie były klasyfikowane ze względu na potwierdzenie oryginalnej hipotezy, gdyż były to albo badania oryginalne – gdzie na marginesie, zawsze wyjściowa hipoteza została potwierdzona, lub badania zwiększające rozumienie efektu i tutaj nie może być mowy o potwierdzeniu lub falsyfikacji. (Na marginesie: brak potwierdzenia hipotezy z wyjściowego badania nie oznacza, że nie została potwierdzona żadna hipoteza badaczy artykułu z badaniem nawiązującym. Często zakładano, że uzyska się jakieś wyniki istotne statystycznie, ale o kierunku innym niż wyjściowa hipoteza.)

5.3.3. Rok badania

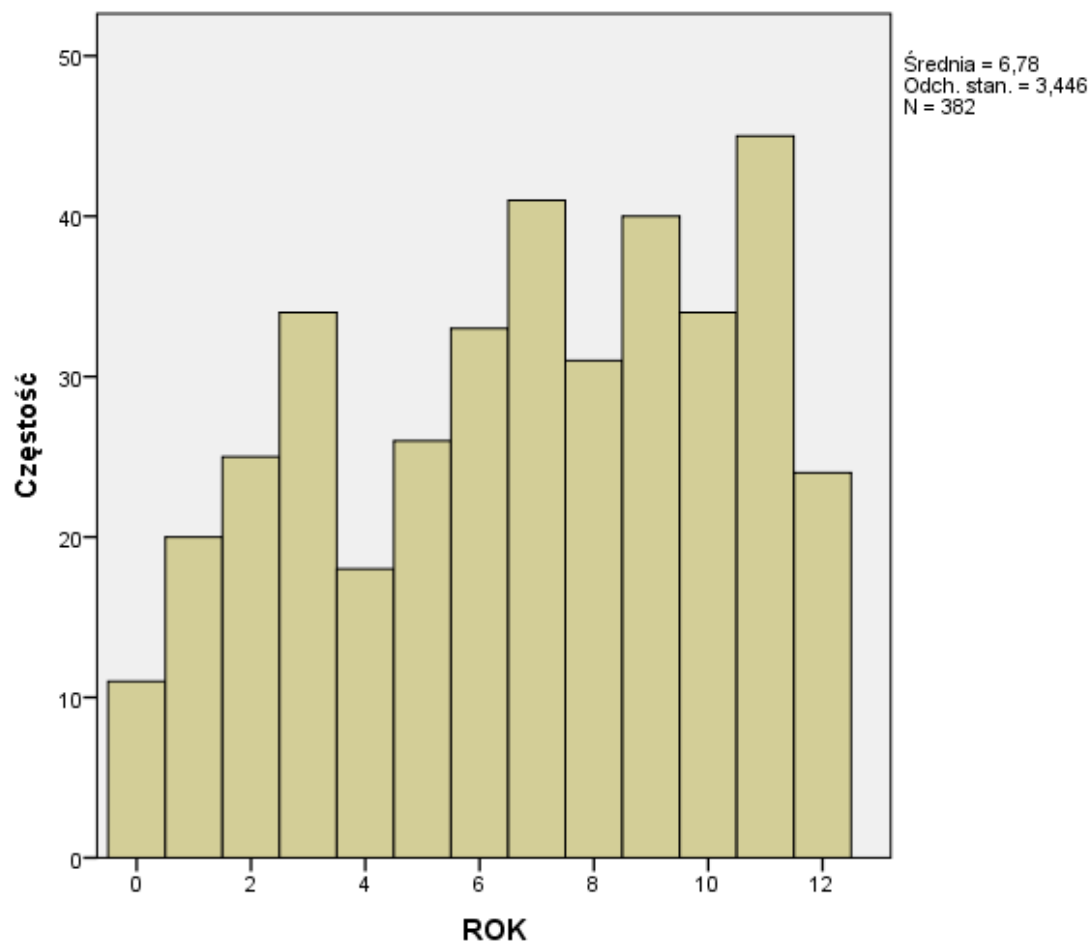
Uwzględniam tutaj tylko badania nawiązujące, gdyż badania oryginalne zawsze ukazywały się w pierwszym roku. Rok zapisałem jako ostatnią cyfrę roku (lub dwie), w którym ukazało się badanie, a więc jako że pierwsze badania miały miejsce w roku 2000, jest to zarazem wskaźnik liczby lat, jaka upłynęła między publikacjami oryginalnego i nawiązującego badania.

Rok badania na pierwszy rzut oka może się wydawać zmienną mało istotną. Jednak warto go przeanalizować, gdyż rozkład czasu badań może dostarczać istotnych wskazówek, m.in. na temat tego, czy z czasem spada zainteresowanie tematyką badawczą, kiedy są replikowane hipotezy itd. Przykładowo, jak wynika z analizy przeprowadzonej przez Makel, Plucker i Hegarty (2012), im więcej czasu upływa od badania, tym mniejsza szansa, że ktoś je zreplikuje (która tak czy owak jest bardzo mała).

Na wykresie 5.5. przedstawiłem częstość badań w zależności od roku badania. Badań, które ukazały się w tym samym roku co badanie oryginalne, było stosunkowo niewiele i prawie zawsze (10/11 przypadków) były autorstwa tego samego zespołu. Występuje zresztą reguła, że w pierwszych latach po publikacji, to najczęściej autorzy badań oryginalnych publikują badania nawiązujące, z czasem dopiero

przybywa zespołów niezależnych (po 4 roku jest ich zdecydowanie więcej) (Wykres 5.6.). Różnica ta jest istotna $t(380) = 9,18$, $p < 0,001$ (Tabela 5.11).

WYKRES 5.5. Odsetek badań w zależności od roku badania



WYKRES 5.6. Odsetek badań nawiązujących wykonanych przez niezależne zespoły

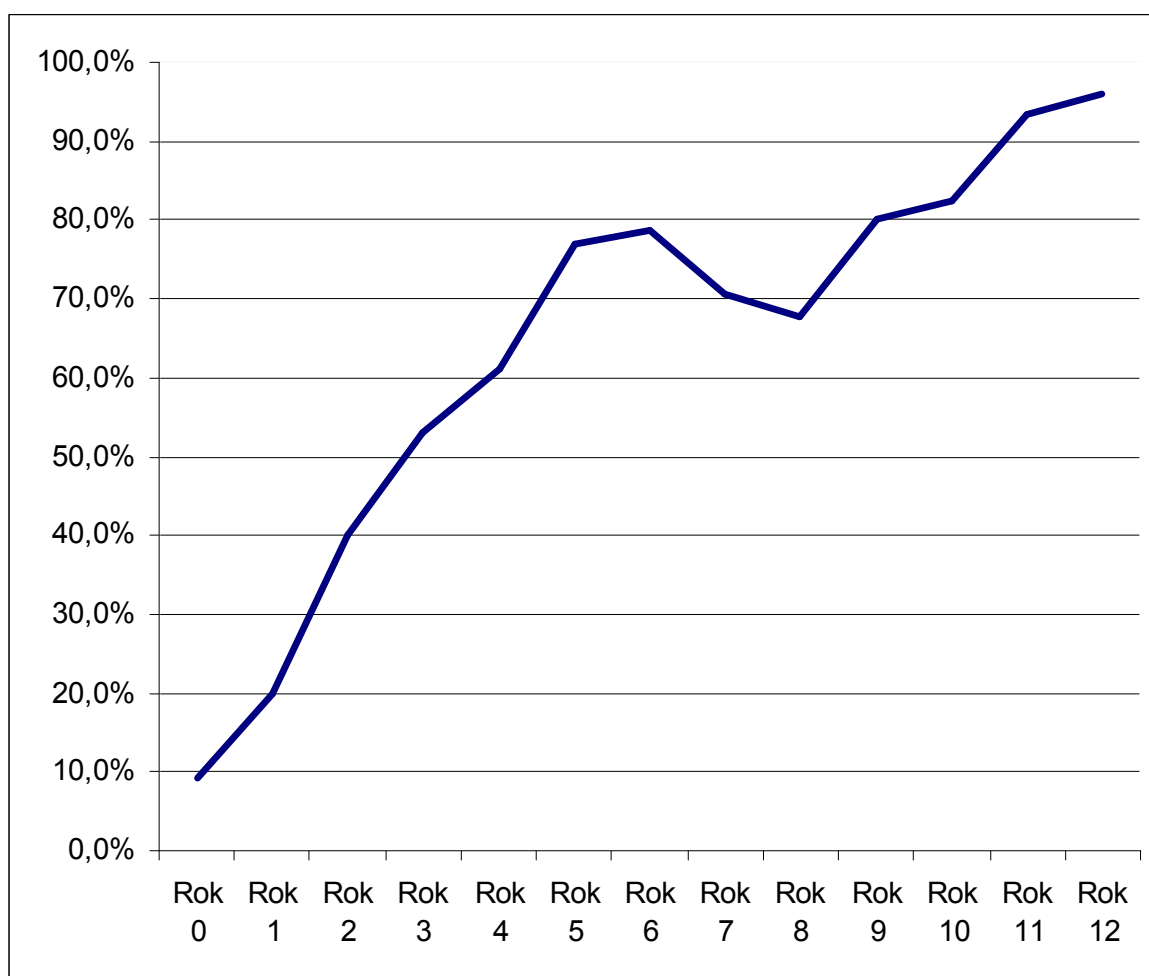


TABELA 5.11. Średni rok publikacji badań nawiązujących w zależności od tego, czy przeprowadził go niezależny zespół

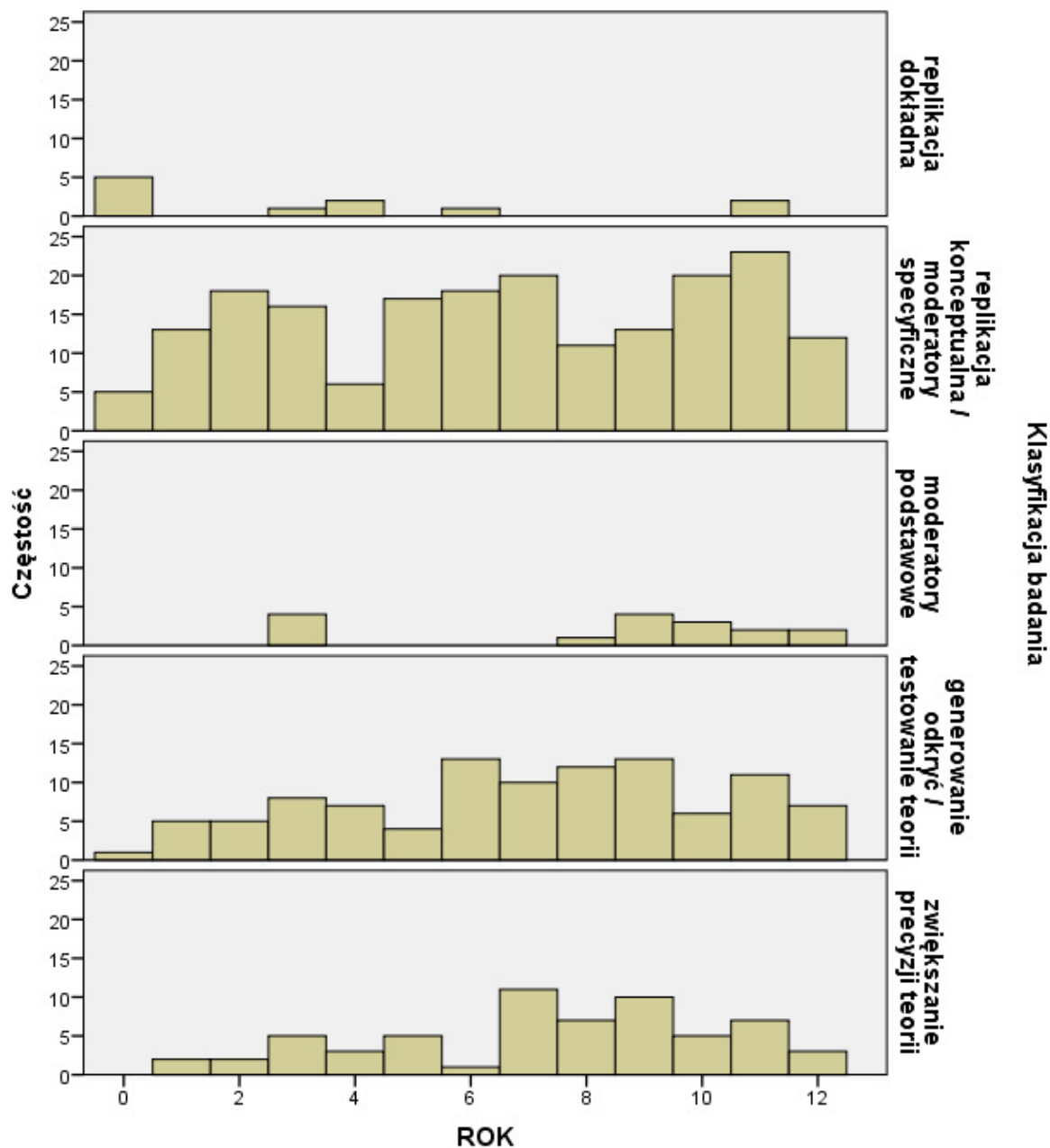
Niezależny zespół	Średnia	N	Odchylenie standardowe
NIE	4,57	117	3,310
TAK	7,75	265	3,036
Ogółem	6,78	382	3,446

Nie potwierdza się teza, że w miarę upływu czasu maleje liczba badań nawiązujących do oryginalnego badania (przynajmniej w badanym przedziale; nie mam wątpliwości, że w jakimś dłuższym przedziale liczba ta w końcu musi zacząć spadać). Wręcz przeciwnie, liczba ta rośnie (korelacja między rokiem a liczbą badań nawiązujących: $r = 0,65$, $p < 0,05$).

Postanowiłem sprawdzić również czy określonych badań przybywa w miarę upływu czasu. Wykres 5.7. przedstawia taki rozkład w zależności od klasyfikacji badania.

Tylko w przypadku badań mających na celu generowanie odkryć i testowanie teorii zachodzi istotna zależność między rokiem a ilością badań ($r = 0,58$, $p < 0,05$).

WYKRES 5.7. Częstość badań w zależności od roku i celu badania



5.3.4. Liczba osób badanych

Liczba osób badanych w pojedynczym badaniu była zmienną wysoce zróżnicowaną. W tabeli 5.12. przedstawiłem podstawowe statystyki. W pierwszej kolumnie przedstawiłem wyniki dla całego zbioru danych (liczba jest mniejsza o 3 od ogólnej liczby badań, gdyż w 2 badaniach nie podano liczby osób badanych, a w jednym „badanymi” nie byli ludzie, tylko ich wytwory; patrz PSPB5_17). W drugiej kolumnie

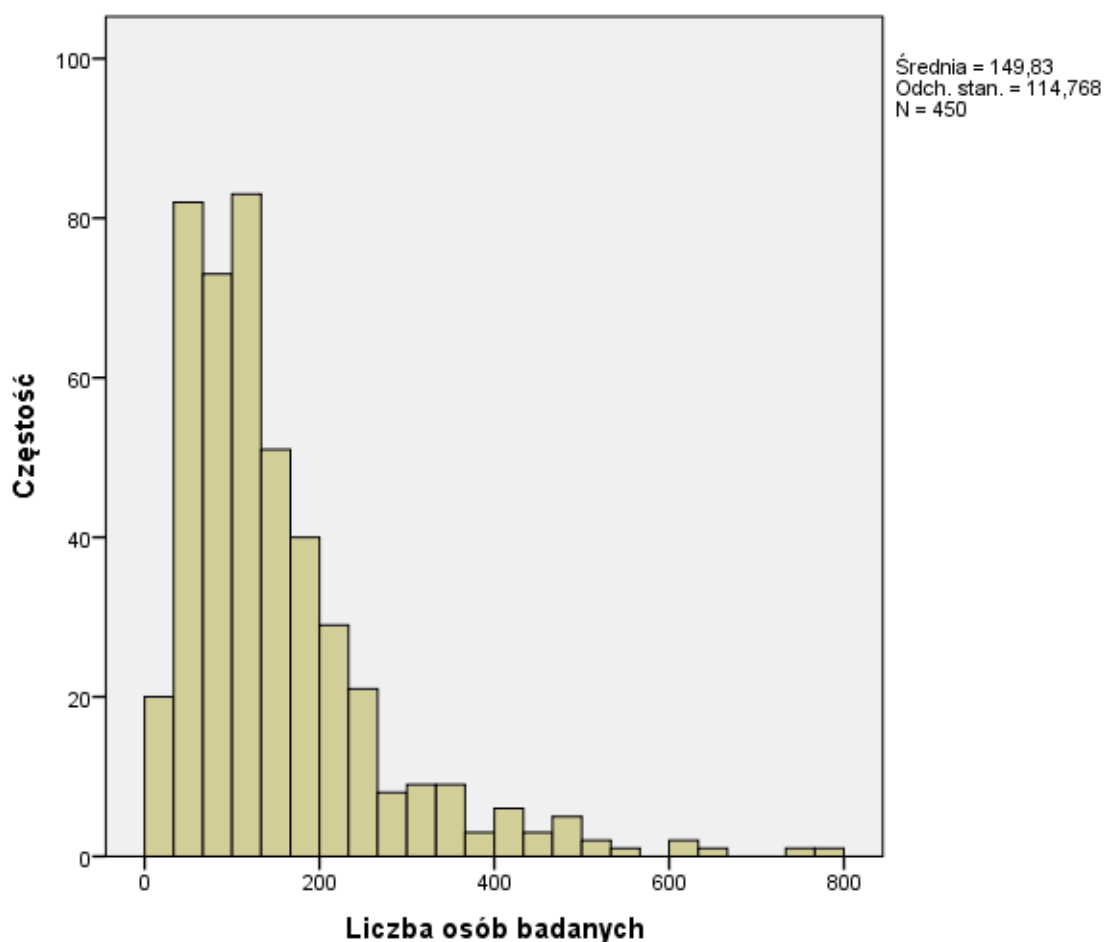
przedstawiłem wyniki bez czterech badań, gdzie bardzo duże liczby osób badanych (ponad 8 tys.) nie zostały rekrutowane przez zespół badaczy publikujących artykuł, ale były wynikiem reanalizy niezależnie wytworzonych dużych zbiorów danych (np. rejestrów armii amerykańskiej dotyczących zachowań żołnierzy). Te przypadki wyłączyłem z dalszych analiz w rdz. 5.4.2., gdzie liczba osób badanych jest zmienną zależną.

TABELA 5.12. Podstawowe charakterystyki zmiennej liczba osób badanych

Statystyka	Cała próba badań	z wyłączeniem dużych prób nie zebranych przez badaczy publikujących artykuł
Średnia	382	224
Odchylenie standardowe	2075	607
Minimum	15	15
Maksimum	33048	9537
10. centyl	48	48
25. centyl	74	73
Mediana	123	123
75. centyl	200	199
90. centyl	348	337
N	469	465

Dla lepszego zobrazowania wysokiego zróżnicowania liczby osób badanych na wykresie 5.8. przedstawiłem rozkład częstości stosowania określonych wielkości prób badanych. Dla czytelności rozkład kończy się na wartości 800 osób (około 96. centyla). Charakteryzuje go wysoki stopień prawoskośności.

WYKRES 5.8. Częstość występowania liczby osób badanych (tylko do wartości 800 osób)



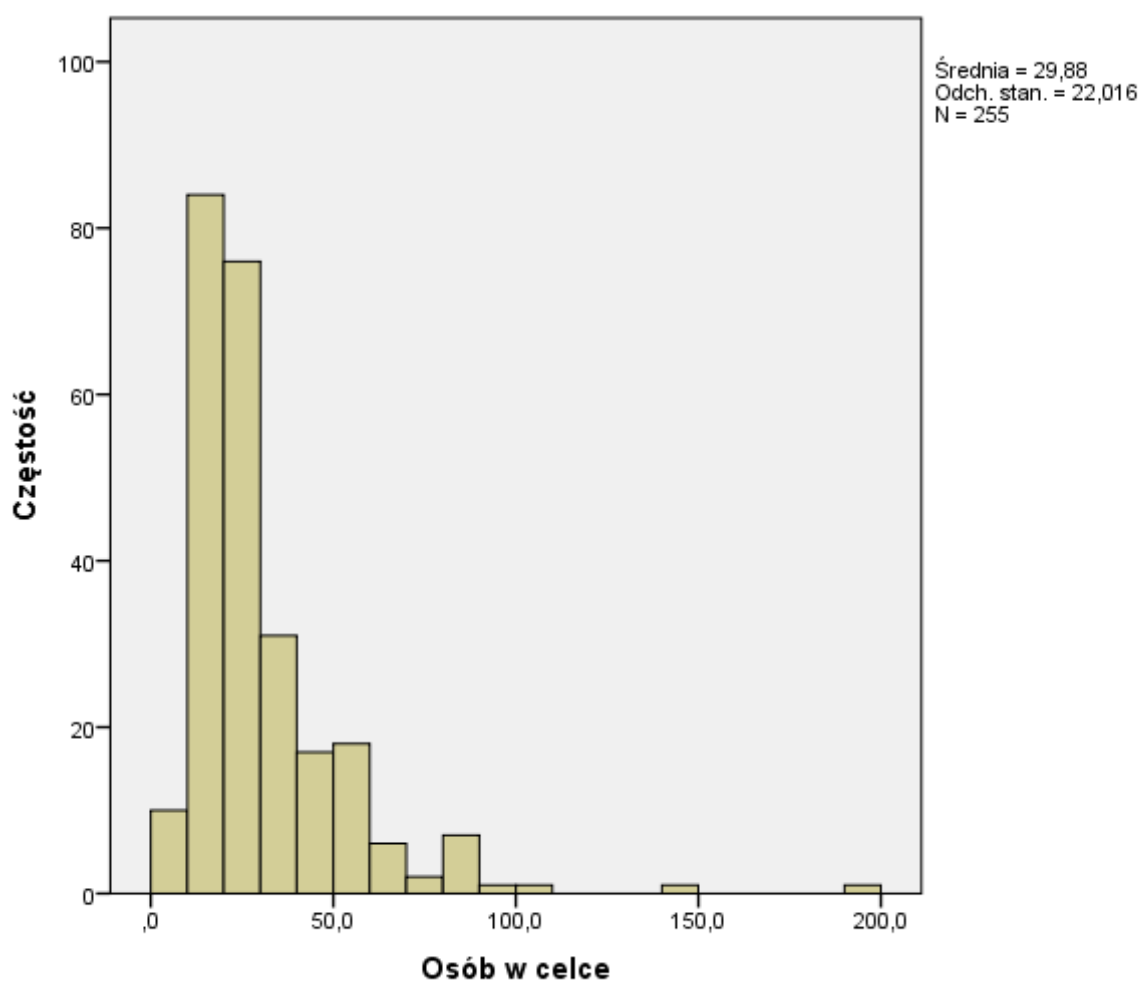
Ogólna reguła wskazuje na to, że wyższa liczba osób badanych jest bardziej pożądana, gdyż zwiększa moc badania. Jednak w przypadku badań eksperymentalnych ważna jest również liczba osób badanych w jednej kratce⁷⁸ (czyli grupie eksperymentalnej). Badanie eksperymentalne z setką osób w dwóch kratkach ma dużo wyższą wartość poznawczą niż badanie z taką samą liczbą osób w 8 grupach eksperymentalnych. Dlatego w tabeli 5.13. i na Wykresie 5.9. przedstawiam analogiczne dane, ale tylko dla wielkości kratek.

⁷⁸ W literaturze polskiej używa się też określenia „celka”.

TABELA 5.13. Podstawowe charakterystyki zmiennej liczba osób w kratce

		Wartość
N		255
Średnia		29,884
Mediana		24,000
Minimum		5,8
Maksimum		192,0
Percentyle	10	12,150
	25	16,000
	50	24,000
	75	34,750
	90	56,400

WYKRES 5.9. Rozkład częstości liczby osób w kratce (tylko badania eksperymentalne)



Warto zauważyć, że w 40 % przypadków kratka miała wielkość 20 lub mniejszą. Tak niska wielkość powoduje, że moc jest bardzo niska, oraz daje to możliwość nieuczciwym badaczom na poszukiwanie „odkryć” drogą psychologii fałszywie pozytywnej. Również Simmons, Nelson i Simonsohn (2011) wskazują, że liczba 20 osób jest absolutnym „minimum przyzwoitości”.

W 80 % przypadków wielkość kratki była mniejsza niż 40. Warto zauważyć, że aby uzyskać wysoką moc (90 %) dla dosyć słabego efektu ($d = 0,3$) potrzeba ponad 200 osób w jednej kratce (przy prostym porównaniu dwóch grup eksperymentalnych). Takie wartości były niespotykane w mojej analizie.

5.3.5. Grupa badana

W tabeli 5.14. przedstawiłem wykorzystywane próby do badań. Potwierdzają się wnioski o dominacji studentów jako osób badanych (por. rdz. 2.6). Dodatkowo, gdy badano więcej niż jedną próbę, w 6 (na 17) przypadkach studenci stanowili jedną z tych badanych prób.

Tabela 5.14. Częstość badania określonych grup

Próba	Częstość	Procent
studenci	374	79,2
Internet	5	1,1
populacja ogólna	38	8,1
pracownicy (konkretnych firm)	5	1,1
dzieci ⁷⁹	7	1,5
adolescenci	4	,8
studenci MBA	8	1,7
grupa kliniczna	3	,6
osoby starsze	1	,2
inna	10	2,1
więcej niż jedna	17	3,6
Ogółem	472	100,0

Ze względu na niedużą liczbę prób badanych grup nie-studentów w dalszej części (5.4.) przy testowaniu szczegółowych hipotez będę stosował rozróżnienie tylko na grupę studentów i nie-studentów.

5.3.6. Narodowość badanych

W 48 badaniach badano więcej niż jedną próbę, dlatego w tym miejscu analizuję wyniki przy założeniu, że każda próba z innego kraju stanowi jedno osobne badanie.

⁷⁹ Arbitralnie przyjąłem podział: do lat 12 – dzieci; wiek 13-17 – adolescenci.

W tabeli 5.15. przedstawiłem częstość określonego pochodzenia grup badanych. Badani z 9 najczęściej badanych krajów stanowili aż 89 % prób.

TABELA 5.15 Liczba prób w zależności od kraju (uszeregowano od najczęstszych)

Liczba badań	Kraj pochodzenia
296	Stany Zjednoczone
39	Chiny (w tym Hongkong)
27	Holandia
24	Japonia
21	Izrael
18	Niemcy
17	Wielka Brytania
15	Australia, Kanada
9	Francja, Polska
8	Włochy
6	Nowa Zelandia, Szwecja
2	Belgia, Iran, Korea, Meksyk, Tajwan
1	Bośnia, Czechy, Filipiny, Rosja, Singapur, Turcja
6	nie podano

W Tabeli 5.16. przygotowałem to samo zestawienie, ale z rozbiem na grupy kulturowe. Zdecydowaną większość badanych stanowiły osoby z rozwiniętych krajów zachodniego świata

TABELA 5.16. Liczba i odsetek prób badanych z różnych grup kulturowych

Kraj	Liczba badań	Odsetek badań
Stany Zjednoczone, Wielka Brytania, Australia, Belgia, Francja, Holandia, Kanada, Niemcy, Nowa Zelandia, Szwecja, Włochy	419	78,46%
(w tym tylko USA)	296	56,27%
Czechy, Polska, Bośnia	11	2,06%
Japonia	24	4,49%
Chiny, Hongkong, Singapur, Tajwan	42	7,87%
Izrael	21	3,93%
inne (Filipiny, Iran, Korea, Meksyk, Rosja, Turcja)	9	1,69%

Warto zauważyć, że w dwóch badaniach (JPSP8 i PSPB5) testowano hipotezy dotyczące ściśle różnic międzykulturowych. Siłą rzeczy te badania powinny były uwzględniać zróżnicowane próby (i tak się działo). Dlatego w tabeli 5.17. przedstawiłem analogiczne wyniki, ale bez tych dwóch badań.

TABELA 5.17. Liczba i odsetek prób badanych z różnych grup kulturowych z wyłączeniem dwóch cykli badań (JPSP8, PSPB5)

Kraj	Liczba badań	Odsetek badań
Stany Zjednoczone, Wielka Brytania, Australia, Belgia, Francja, Holandia, Kanada, Niemcy, Nowa Zelandia, Szwecja, Włochy	380	86,56%
(w tym tylko USA)	262	59,68%
Czechy, Polska, Bośnia	10	2,28%
Japonia	5	1,14%
Chiny, Hongkong, Singapur, Tajwan	17	3,87%
Izrael	20	4,56%
inne	7	1,59%

Dwa dalsze badania (EJPS1 i EJSP5) z jakichś względów były wyjątkowo zróżnicowane kulturowo. Wyniki (tabela 5.18.) po odjęciu tych dwóch cykli wskazują na jeszcze mniejsze zróżnicowanie osób badanych.

Takie przedstawienie danych może sprawiać wrażenie „wybierania wisienek”, żeby podtrzymać daną tezę (tu: o niskim zróżnicowaniu pochodzenia badanych). W rzeczywistości jednak ma na celu pokazanie, że w zdecydowanej większości badań zróżnicowanie kulturowe jest niewielkie. Owszem, sporadycznie występują badania z efektami przetestowanymi na wielu kulturach. Problem w tym, że typowy efekt psychologiczny ma nikłe szanse na takie potraktowanie.

TABELA 5.18. Liczba i odsetek prób badanych z różnych grup kulturowych z wyłączeniem czterech cykli badań (JPSP8, PSPB5, EJPS1, EJSP5)

Kraj	Liczba badań	Odsetek badań
Stany Zjednoczone, Wielka Brytania, Australia, Belgia, Francja, Holandia, Kanada, Niemcy, Nowa Zelandia, Szwecja, Włochy	345	89,15%
(w tym tylko USA)	253	65,37%
Czechy, Polska, Bośnia	10	2,58%
Japonia	5	1,29%
Chiny, Hongkong, Singapur, Tajwan	10	2,58%
Izrael	14	3,62%
inne	3	0,78%

Bardziej jakościowa ocena tych wyników zostanie przeprowadzona w rozdziale 7. Tutaj warto zauważyć tylko, że nie została zanotowana żadna próba pochodząca z Indii (ponad 1,2 mld populacji), Afryki (ponad 1 mld populacji), zdecydowanej

większości krajów azjatyckich oraz krajów Ameryki Łacińskiej i Środkowej (z wyjątkiem Meksyku). Również nie wykorzystano żadnej próby, która wiedzie styl życia daleki od stylu życia mieszkańców krajów uprzemysłowionych (np. społeczności łowiecko-zbierackich). Tezy Arnetta (2008) o „porzuconych 95 %” znajdują pełne potwierdzenie.

Ze względu na niewielką liczbę badań z kultur innych niż amerykańska (i większy margines błędu) w dalszej części analiz ilościowych będę stosował rozróżnienie na grupę Amerykanów i nie-Amerykanów.

5.3.7. Cel badania, rodzaj badania, sposób manipulacji zmiennymi oraz rodzaj zmiennej zależnej

W tabelach 5.19., 5.20., 5.21., 5.22. przedstawiłem podstawowe statystyki częstości odpowiednio różnych rodzajów, celów badań, rodzajów badań, sposobów manipulacji zmienną niezależną (tylko badania eksperymentalne), oraz rodzajów zmiennej zależnej.

W części 5.4. replikacje dokładne, replikacje konceptualne / moderatory specyficzne oraz moderatory podstawowe będę klasyfikował razem jako „replikacje” (ze względu na niską liczbę replikacji dokładnych i moderatorów podstawowych, a więc potencjalnie większy błąd pomiaru, przedstawianie wyników tylko dla tych danych nie ma dużo sensu). Analogicznie będę grupował razem eksperymenty i eksperymenty terenowe, oraz zmienne zachowanie, zmienna poznawcza, zachowanie w grach ekonomicznych i zachowanie na bieżąco raportowane jako zmienne „(około)behawioralne”

TABELA 5.19. Częstość celów badania w całej próbie badań

	Częstość	Procent
oryginalne	90	19,1
replikacja dokładna	11	2,3
replikacja konceptualna / moderatory specyficzne	192	40,7
moderatory podstawowe	16	3,4
generowanie odkryć / testowanie teorii	102	21,6
zwiększanie precyzji teorii	61	12,9

TABELA 5.20. Częstość rodzajów badań w całej próbie badań

	Częstość	Procent
eksperyment	255	54,0
eksperyment terenowy	8	1,7
badanie korelacyjne	170	36,0
dane opisowe	27	5,7
dane opisowe (otwarte)	8	1,7
dane wytworzone poza badaniem	4	,8

TABELA 5.21. Częstość rodzajów manipulacji (dotyczy tylko eksperymentów)

	Częstość	Procent
nierealistyczna	141	29,9
realistyczna	135	28,6
N/D	196	41,5
Ogółem	472	100,0

TABELA 5.22. Częstość stosowania zmiennych zależnych

	Częstość	Procent
zachowanie	26	5,5
zmienna poznawcza	25	5,3
zachowanie w grach ekonomicznych	40	8,5
zachowanie na bieżąco raportowane	2	,4
zmienna samoopisowa	361	76,5
zmienna utajona	17	3,6
zmienna biologiczna	1	,2
Ogółem	472	100,0

Dominacja zmiennych samoopisowych jest pokłosiem dominacji badań korelacyjnych, gdzie praktycznie zawsze występują takie zmienne (tabela 5.23.). Dlatego w tabeli 5.24. przedstawiłem dodatkowo dane częstość występowania określonego rodzaju zmiennej zależnej tylko dla badań eksperymentalnych.

TABELA 5.23. Częstość występowanie określonej zmiennej zależnej (tylko badania korelacyjne)

	Częstość	Procent
zachowanie w grach ekonomicznych	5	2,9
zachowanie na bieżąco raportowane	2	1,2
zmienna samoopisowa	163	95,9

TABELA 5.24. Częstość występowanie określonej zmiennej zależnej (tylko eksperymenty)

	Częstość	Procent
zachowanie	23	8,7
zmienna poznawcza	25	9,5
zachowanie w grach ekonomicznych	35	13,3
zmienna samoopisowa	162	61,6
zmienna utajona	17	6,5
zmienna biologiczna	1	,4

5.4. Związki pomiędzy poszczególnymi zmiennymi

W tej części dla większej czytelności kolejne podrozdziały będą miały tytuł „co wpływa na...”. Zakładam, że ewentualna zależność przyczynowo-skutkowa między pewnymi zmiennymi może być tylko w jednym kierunku, a nie w drugim (np. niezależność zespołu może wpływać na potwierdzenie wyjściowej hipotezy, ale potwierdzenie wyjściowej hipotezy nie będzie wpływać na niezależność zespołu). Oczywiście nie wykluczam, że czasami zależność między niektórymi zmiennymi ma inny kierunek (o ile nie jest, co oczywiste, jakąś formą korelacji pozornej).

5.4.1. Co wpływa na potwierdzenie oryginalnej hipotezy badawczej?

Pierwsza hipoteza, jaka intuicyjnie się narzuca, jest taka, że wpływ musi wywierać to, czy hipoteza jest testowana przez niezależny zespół. Istotnie taka zmienna wywiera wpływ. Niezależne zespoły potwierdziły 65 % wyjściowych hipotez, natomiast gdy badanie przeprowadzał ten sam zespół wynik ten był o 13 % wyższy (tabela 5.25.). Jeśli uwzględnić badania niejednoznacznie potwierdzające hipotezę, wówczas efekt ten jest na granicy istotności $\chi^2 = 5,805$, $df = 2$, $p = 0,55$. Natomiast jeśli uwzględnić tylko te badania, które jednoznacznie potwierdzają lub falsyfikują wyjściową hipotezę, w takim wypadku efekt jest istotny $\chi^2 = 4,05$, $df = 1$, $p < 0,05$. (Zaznaczę tylko, że niepotwierdzenie wyjściowej hipotezy nie oznacza artykułu bez istotnych efektów).

TABELA 5.25. Potwierdzenie wyjściowej hipotezy w zależności od tego, czy badanie przeprowadził niezależny zespół

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
Niezależny zespół	NIE	Liczebność	7	77	15
		%	7,1%	77,8%	15,2%
	TAK	Liczebność	31	143	47
		%	14,0%	64,7%	21,3%

Efekt ten był jeszcze wyraźniejszy, jeśli uwzględnić tylko replikacje (tabela 5.26.) ($\chi^2 = 14,64$, $df = 2$, $p < 0,001$).

TABELA 5.26. Potwierdzenie w zależności od niezależności zespołu (tylko replikacje)

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
Niezależny zespół	NIE	Liczebność	7	58	9
		%	9,5%	78,4%	12,2%
	TAK	Liczebność	28	75	42
		%	19,3%	51,7%	29,0%

Okazuje się też, że w badaniach, w których falsyfikuje się oryginalną hipotezę, występują większe liczby osób badanych (tabela 5.27.). Ten efekt jest jednak przypadkowy. Usunięcie dwóch skrajnie wysokich wyników (badania EJSP1_18 i EJSP5_09) powoduje znaczący spadek średniej liczby osób badanych w grupie badań nie potwierdzających hipotezy (średnia wynosi wówczas 172 osoby, a nie 224).

TABELA 5.27. Średnia liczba osób badanych w zależności od potwierdzenia wyjściowej hipotezy

Potwierdzenie oryginalnej hipotezy?	Średnia	N	Odchylenie standardowe	Mediana
NIE	224,72	36	260,499	133,50
TAK	196,76	215	346,317	120,00
Częściowe	172,24	62	248,386	120,00

Kolejna zmienna, która wpływa na potwierdzenie oryginalnej hipotezy, to pochodzenie badanych. Na próbach nie-amerykańskich hipotezy były częściej falsyfikowane ($\chi^2 = 14,32$, $df = 2$, $p < 0,01$). Może to wynikać z tego, że duża część badań oryginalnych (52 %) jest przeprowadzana na Amerykanach, więc – ze względów różnic kulturowych – efekty po prostu się nie replikują na innych grupach. Może też świadczyć o tym, że próby amerykańskie siłą rzeczą pochodzą od amerykańskich badaczy, a jak już wspominałem, istnieją pewne dane, że amerykański system rywalizacji akademickiej wymusza częstsze publikowanie danych istotnych (Fanelli, 2010b).

TABELA 5.28. Potwierdzenie oryginalnej hipotezy w zależności od pochodzenia badanych

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
Narodowość badanych	Amerykanin	Liczebność	12	131	37
		%	6,7%	72,8%	20,6%
	nie-Amerykanin	Liczebność	22	59	19
		%	22,0%	59,6%	19,0%

Zauważyłem również, że w grupach nie-studenckich (Tabela 5.29.) efekt był nieco częściej falsyfikowany, choć nie była to zależność istotna ($p > 0,1$). Trend (choć ciągle nieistotny) utrzymuje się także, gdy kontrolować pochodzenie.

TABELA 5.29. Potwierdzenie wyjściowej hipotezy w zależności od grupy badanej

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
Grupa badana	student	Liczebność	27	181	48
		%	10,5%	70,7%	18,8%
	nie-student	Liczebność	11	39	14
		%	17,2%	60,9%	21,9%

Zaobserwowałem również ciekawy efekt polegający na tym, że częściej hipoteza była falsyfikowana w replikacjach (konceptualnych) niż badaniach, których celem było przetestowanie teorii, i te ostatnie prawie zawsze współgrały z wyjściową hipotezą (pomijam tu przypadki niejednoznaczne) (tabela 5.30.). Efekt jest statystycznie istotny ($\chi^2 = 14,29$; $df = 1$; $p < 0,001$). Jeśli efekt ten nie jest przypadkowy, to pierwsza jego najbardziej intuicyjna interpretacja jest taka, że w przypadku badania zbyt daleko odbiegającego od oryginalnej hipotezy, badacze muszą się wykazać dodatkowymi „atutami” (istotnością statystyczną), żeby badanie zostało dopuszczone do druku. Należy jednak zauważyć, że olbrzymia większość badań mających na celu przetestowanie teorii nie przedstawiała się jako robiące to za sprawą badania wyjściowego. Nadanie im takiej etykiety wynika z mojej interpretacji treści tych badań. Alternatywnym wyjaśnieniem jest więc to, że o ile badania pośrednio testujące z sukcesem wyjściową hipotezę są łatwo dostrzegalne wśród artykułów cytujących, to także badania falsyfikujące są albo mniej widoczne, albo nie odwołują się do konkretnego oryginalnego artykułu. Nie jestem w stanie rozstrzygnąć, która interpretacja może być poprawna.

TABELA 5.30. Potwierdzenie oryginalnej hipotezy w zależności od celu badania

			Potwierdzenie oryginalnej hipotezy?	
			NIE	TAK
Cel badania	replikacje	Liczebność	35	133
		%	20,8%	79,2%
	generowanie odkryć / testowanie teorii	Liczebność	3	87
		%	3,3%	96,7%

Okazuje się również (tabela 5.31.), że trudniej było potwierdzić wyjściową hipotezę, jeśli natura zmiennej zależnej była behawioralna lub okółobehawioralna ($\chi^2 = 20,17$ $df = 4$, $p < 0,001$). Jest to odkrycie wysoce zgodne z intuicją, wydaje się bowiem, że w przypadku zmiennych behawioralnych istnieją mniejsze możliwości „wybierania wisienek z tortu”. Zmiennych behawioralnych można również zwykle przetestować zwykle znacznie mniej niż zmiennych samoopisowych. Efekt ten zachodzi również dla tylko dla podgrupy eksperymentów (tabela 5.31a.): $\chi^2 = 11,24$ $df = 2$, $p < 0,001$.

TABELA 5.31. Potwierdzenie wyjściowej hipotezy w zależności od rodzaju zmiennej zależnej

			Potwierdzenie oryginalnej hipotezy?		
			NIE	TAK	częściowe
Rodzaj zmiennej zależnej	(około)behawioralne	Liczebność	17	33	16
		%	25,8%	50,0%	24,2%
	samoopisowe	Liczebność	21	175	42
		%	8,8%	73,5%	17,6%
	inne (biologiczne i utajone)	Liczebność	0	12	4
		%	0,0%	75,0%	25,0%

TABELA 5.31a. Potwierdzenie wyjściowej hipotezy w zależności od natury zmiennej zależnej (tylko eksperymenty)

			Potwierdzenie oryginalnej hipotezy?	
			NIE	TAK
Natura zmiennej zależnej	(około)behawioralne	Liczebność	12	29
		%	29,3%	70,7%
	samoopisowe	Liczebność	7	87
		%	7,4%	92,6%

Sam rodzaj badania (szczególnie porównanie eksperymentów i badań korelacyjnych) nie miał wpływu na potwierdzenie wyjściowej hipotezy ($p > 0,2$).

5.4.2. Co wpływa na liczbę osób badanych?

Ze względu na obserwowany brak normalności w rozkładzie liczby osób badanych do testowania szczegółowych hipotez używam testów nieparametrycznych.

Zasadniczy wpływ na wielkość próby badanej ma rodzaj badania. Badania korelacyjne są dużo większe niż badania eksperymentalne (Tabela 5.32.)

(porównanie eksperymentów i badań korelacyjnych: $U = 11406$, $z = 7,95$, $p < 0,001$).

Efekt ten utrzymuje się nawet, gdy usunąć skrajnie wysokie wyniki (tj. badania powyżej 1000 osób, które prawie zawsze są korelacyjne).

TABELA 5.32. Średnia liczba i mediana osób badanych w zależności od rodzaju badania

Rodzaj badania	Średnia	N	Odchylenie standardowe	Mediana
eksperyment	123,19	254	118,980	103,00
eksperyment terenowy	143,50	8	50,094	150,00
badanie korelacyjne	403,39	166	980,606	183,00
dane opisowe	105,41	27	69,806	97,00
dane opisowe (otwarte)	187,25	8	257,410	86,50
dane wytworzone poza badaniem	355,50	2	204,354	355,50
Ogółem	224,63	465	607,592	123,00

Żeby to lepiej zobrazować, w tabeli 5.33. przedstawiłem odsetek badań korelacyjnych, eksperymentalnych (eksperymenty i eksperymenty terenowe) oraz innych w zależności od próby badanej.

Ogólnie ten fakt nie jest zaskakujący. Badania korelacyjne to najczęściej badania kwestionariuszowe, które są logistycznie łatwiejsze, przynajmniej jeśli chodzi o etap zebrania odpowiedzi (a już w szczególności, gdy ma się dostęp do dużych grup studentów). Zapewne też inne są nieformalne oczekiwania redakcji, co do wielkości badania eksperymentalnego i korelacyjnego.

TABELA 5.33. Odsetek różnych rodzajów badań w zależności od wielkości próby

Liczba osób badanych	Odsetek badań eksperymentalnych	Odsetek badań korelacyjnych	Odsetek pozostałych badań	N
Do 100 osób	67,7 %	20,8 %	11,4 %	183
100-200 osób	62,9 %	31,5 %	5,6 %	178
200-500 osób	34 %	60,6 %	5,3 %	94
500-1000 osób	9,1 %	72,7 %	18,2 %	11
ponad 1000 osób	6,3 %	87,5 %	6,3 %	16

Również w przypadku badania prób nie-studenckich liczba osób badanych jest większa ($U = 11469$; $z = 4,92$; $p < 0,001$). W przypadku bardzo dużych prób (ponad 1,5 tys. osób) badanymi są częściej nie-studenci (8 na 13 przypadków), ale nawet gdy usunąć te dane z analizy, istotna różnica utrzymuje się. Wydawać to się może paradoksalne, gdyż na pozór studenci są łatwiej dostępni. Przypuszczam jednak, że różnica ta wynika z relatywnej rzadkości sięgania po innych niż studenci badanych. Zapewne, gdy już badacze to robią, posiadają środki z grantów i są zobowiązani do

publikowania badań. Niewykluczone również, że w przypadku badań studentów badacze mogą być bardziej nastawieni na wybiórcze publikowanie, tj. przeprowadzanie wielu względnie niedużych badań, z których wybiorą sobie tylko te z istotnymi wynikami.

TABELA 5.34. Średnia i mediana osób badanych w zależności od grupy

Grupa badana	Średnia	N	Odchylenie standardowe	Mediana
student	181,29	373	412,796	115,00
nie-student	400,35	92	1070,912	162,50
Ogółem	224,63	465	607,592	123,00

Również, gdy badań podejmowały się niezależne zespoły, zaobserwowałem wyższe próby (Tabela 5.35.) ($U = 12144$, $z = 2,96$, $p < 0,001$). Efekt wynika z tego, że niezależne zespoły częściej podejmowały badania korelacyjne i efekt zanika, gdy kontrolować tę zmienną.

TABELA 5.35. Średnia liczba osób badanych w zależności od tego, czy badanie powtarzał niezależny zespół (tylko badania nawiązujące)

Niezależny zespół	Średnia	Odchylenie standardowe	Mediana	N
NIE	164,01	244,282	114,00	116
TAK	252,63	667,254	140,00	259
Ogółem	224,63	607,592	123,00	465

Okazuje się również, że próby nie-amerykańskie są średnio nieco większe (tabela 5.36), choć to nie jest istotna różnica. Próba badana jest prawie zawsze pochodną pochodzenia badacza, więc być może w przypadku badaczy i/lub prób nie-amerykańskich próby muszą być nieco większe, żeby dodatkowo zachęcić redaktorów do publikacji takich badań.

TABELA 5.36 Średnia liczba i mediana osób badanych w zależności od pochodzenia osób badanych

Narodowość	Średnia	Odchylenie standardowe	Mediana	N
Amerykanin	154,2	221,4	105,00	259
nie-Amerykanin	189,3	232,1	137,00	155

Nie ma zasadniczego wpływu na wielkość prób cel badania, rok badania, rodzaj manipulacji oraz fakt potwierdzenia (lub nie) wyjściowej hipotezy. Natura zmiennej

zależnej ma wpływ, ale gdy kontrolować rodzaj badania efekt zanika (przykładowo w Tabeli 5.37. wielkość prób badanych w zależności od natury zmiennej zależnej, ale tylko dla badań eksperymentalnych)

TABELA 5.37 Wielkość prób badanych w zależności od rodzaju zmiennej zależnej (tylko badania eksperymentalne)

Natura zmiennej zależnej	Średnia	Mediana	N	Odchylenie standardowe
(około)behawioralne	124,84	120,00	83	84,580
samoopisowe	123,61	102,00	161	135,274
inne (biologiczne i utajone)	120,78	124,50	18	71,792
Ogółem	123,81	104,50	262	117,482

5.4.3. Co wpływa na wybór grupy badanej (student vs. nie-student)?

Na wybór grupy badanej zasadniczy wpływ ma pochodzenie badanych. W przypadku prób amerykański wyraźnie częściej występują próby studenckie (Tabela 5.38.): $\chi^2 = 14,74$, $df = 1$; $p < 0,001$).

TABELA 5.38. Grupa badana w zależności od pochodzenia badanych

			Grupa badana	
			student	nie-student
Narodowość dwa	Amerykanin	Liczebność	224	37
		%	85,8%	14,2%
	nie-Amerykanin	Liczebność	150	60
		%	71,4%	28,6%

Na podobnej zasadzie niezależne zespoły częściej wykorzystywały próby nie-studenckie (Tabela 5.39.): $\chi^2 = 7,09$, $df = 1$, $p < 0,01$.

TABELA 5.39. Grupa badana w zależności od tego, czy badanie zostało przeprowadzone przez niezależny zespół

			Grupa badana	
			student	nie-student
Niezależny zespół	NIE	Liczebność	102	15
		%	87,2%	12,8%
	TAK	Liczebność	199	66
		%	75,1%	24,9%

Kolejny wyraźny efekt polegał na tym, że w badaniach eksperymentalnych dużo chętniej korzystano z prób studenckich niż w badaniach korelacyjnych (Tabela 5.40.; $\chi^2 = 31,56$; $df = 2$; $p < 0,001$). Efekt ten nie jest zaskakujący, szczególnie biorąc pod uwagę to, że większość badań korelacyjnych (szczególnie kwestionariuszowych) można zrealizować w miejscu zamieszkania badanego, lub nawet na ulicy, czego nie można zawsze powiedzieć o eksperymentach.

TABELA 5.40. Grupa badana w zależności od rodzaju badania

			Grupa badana	
			student	nie-student
Rodzaj badania	eksperymenty	Liczebność	230	33
		%	87,5%	12,5%
	badania korelacyjne	Liczebność	111	59
		%	65,3%	34,7%
	inne	Liczebność	33	6
		%	84,6%	15,4%

Jako że w przypadku badań korelacyjnych natura zmiennej zależnej jest prawie zawsze samoopisowa, więc konsekwencją tego jest fakt, że studenci są częściej badani, gdy zmienna ma taką właśnie naturę (Tabela 5.41.): $\chi^2 = 8,84$, $df = 2$; $p < 0,05$.

TABELA 5.41 Grupa badana w zależności od natury zmiennej zależnej

			Grupa badana	
			student	nie-student
Natura zmiennej zależnej	(około)behawioralne	Liczebność	66	27
		%	71,0%	29,0%
	samoopisowe	Liczebność	290	71
		%	80,3%	19,7%
	inne (biologiczne i utajone)	Liczebność	18	0
		%	100,0%	0,0%

Nie wpływa na dobór próby cel badania i rodzaj manipulacji.

5.4.4 Co wpływa na wielkość kratek w badaniach eksperymentalnych?

Występują pewne nieduże różnice w wielkości kratek w eksperymentach w zależności od celu badania (Tabela 5.42.) jednak nie są to różnice istotne.

TABELA 5.42. Średnia i mediana osób w kratce w zależności od celu badania

Klasyfikacja badania	Średnia	Odchylenie standardowe	Mediana	N
oryginalne	27,944	23,6102	19,750	64
replikacja dokładna	31,500	8,8459	33,000	3
replikacja koncepcyjna / moderatory specyficzne	28,701	22,7596	24,000	113
moderatory podstawowe	25,000	12,9679	20,250	4
generowanie odkryć / testowanie teorii	32,323	17,8932	27,750	55
zwiększanie precyzji teorii	38,529	25,8209	29,750	16
Ogółem	29,884	22,0156	24,000	255

Pewne różnice można też zaobserwować w zależności od natury zmiennej zależnej (Tabela 5.43.). W przypadku wszystkich zmiennych okołobehawioralnych mediana wynosi 24, a w przypadku zmiennych samoopisowych 22,8. Jest to głównie spowodowane serią badań o dużych wartościach (50 osób) w jednym tylko artykule (JPSP9). Po wyłączeniu tych danych wszelkie różnice między zachowaniem, a innymi zmiennymi zanikają po wyłączeniu tego konkretnego artykułu z analizy. Tak czy owak jest to nieco zaskakujące, gdyż wydawać by się mogło, że w badaniach eksperymentalnych w których natura zmiennej zależnej jest nie-behawioralna, a więc badania są logistycznie prostsze, badacze powinni stosować wyższe próby. Okazuje się jednak, że nie ma takich oczekiwań.

TABELA 5.43. Wielkość kratek (grup eksperymentalnych) w zależności od rodzaju zmiennej zależnej

Rodzaj zmiennej zależnej	Średnia	Odchylenie standardowe	Mediana	N
zachowanie	33,018	15,7480	32,000	23
zmienna poznawcza	21,548	8,6418	20,000	25
zachowanie w grach ekonomicznych	30,605	16,4718	24,000	34
zmienna samoopisowa	30,103	24,6431	22,875	155
zmienna utajona	35,876	25,2568	27,000	17
zmienna biologiczna	5,750	.	5,750	1
Ogółem	29,884	22,0156	24,000	255

Kolejnym paradoksem jest to, że nieco większe próby organizuje się, gdy badacze rekrutują nie-studentów do badania ($U = 1936$; $z = 3,58$, $p < 0,001$) (nie-studenci ogólnie stanowili tylko nieco ponad 13 % prób w badaniach eksperymentalnych. W dużej mierze ta różnica wynika z serii eksperymentów terenowych (JPSP9) (po usunięciu tych wyników średnia spada do około 31,5). Tak czy owak nawet ten brak

różnic jest zaskakujący, gdyż pokazuje, że w przypadku grup studenckich badacze również nie mają presji, żeby zwiększać liczbę osób badanych.

TABELA 5.44. Liczba osób w kratce w zależności od grupy badanej

Grupa badana	Średnia	Odchylenie standardowe	Mediana	N
student	29,0	22,7	22,0	226
nie-student	36,4	13,8	33,5	29
Ogółem	29,8	22,0	24,0	255

Nie miały wpływu na liczbę osób w kratkach pochodzenie badanych i realistyczność manipulacji.

5.4.5. Co wpływa cel badania?

Autorzy oryginalnego badania nieco częściej przeprowadzali replikacje, a nieco rzadziej przeprowadzali badania mające na celu (niebanalne) przetestowanie teorii (Tabela 5.45.). Efekt ten nie jest jednak istotny.

TABELA 5.45. Cel badania w zależności od tego, czy badanie było przeprowadzane przez niezależny zespół

			Cel badania		
			replikacje	generowanie odkryć / testowanie teorii	zwiększanie precyzji teorii
Niezależny zespół	NIE	Liczebność	74	25	18
		%	63,2%	21,4%	15,4%
	TAK	Liczebność	145	77	43
		%	54,7%	29,1%	16,2%

Konsekwencją tego faktu jest to, że replikacje były nieco wcześniej publikowane (Tabela 5.46.) niż inne kategorie badań. Efekt ten nie jest istotny i jest konsekwencją tego, że autorzy oryginalnego badania średnio szybciej publikowali artykuły nawiązujące.

TABELA 5.46. Czas publikacji badań w zależności od celu badania

Cel badania	Średnia	N	Odchylenie standardowe
replikacje	6,52	219	3,66
generowanie odkryć / testowanie teorii	6,99	102	3,21
zwiększanie precyzji teorii	7,38	61	2,94

Na cel podejmowanych badań nie miały wpływu zmienne związane z próbą, a więc grupa badana i pochodzenie badanych.

5.4.6. Co wpływa na rodzaj badania?

W przypadku prób nie-amerykańskich częściej były podejmowane badania nie-eksperymentalne (Tabela 5.47.). Efekt ten jest istotny ($\chi^2 = 15,39$; $df = 2$; $p < 0,001$). Prawdopodobnie nie jest to artefakt tzn. w badaniach testujących explicite hipotezy o różnicach międzykulturowych (szczególnie JPSP8 i PSPB5) używano często eksperymentów. Niewykluczone, że jakiś wpływ na obserwowany efekt mogą też mieć inne tradycje badawcze w USA lub większe negatywne nastawienie redakcji do eksperymentalnych badań autorów nie-amerykańskich. Bardziej prozaicznym powodem może być też to, że w przypadku badań korelacyjnych łatwiej zorganizować międzynarodowy zespół.

TABELA 5.47. Narodowość badanych a rodzaj badania

			Rodzaj badania		
			eksperymenty	badania korelacyjne	inne
Narodowość badanych	Amerykanin	Liczebność	156	76	32
		%	58,6%	29,1%	12,3%
	nie-Amerykanin	Liczebność	92	61	3
		%	59,0%	39,1%	1,9%

Również jeśli chodzi o cel badania zaobserwowałem spore zróżnicowanie (Tabela 5.48.) ($\chi^2 = 32,95$; $df = 6$; $p < 0,001$). Przede wszystkim w przypadku badań oryginalnych nadreprezentowane były eksperymenty (prawie $\frac{3}{4}$ wszystkich badań), z kolei w przypadku badań mających na celu zwiększenie precyzji teorii eksperymenty były niedoreprezentowane. Pierwszy efekt wynika prawdopodobnie z tego, że 4 analizowane czasopisma szczególnie często przyjmują eksperymenty relatywnie do innych rodzajów badań. Z kolei ostatni efekt najprawdopodobniej wynika z tego, że zwiększanie precyzji teorii dotyczyło najczęściej zmiennych dyspozycyjnych, a tymi nie można manipulować, można je tylko kontrolować.

TABELA 5.48. Rodzaj badania w zależności od celu badania

			Rodzaj badania		
			eksperymenty	badania korelacyjne	inne
Cel badania	oryginalne	Liczebność	66	18	6
		% z Cel badania	73,3%	20,0%	6,7%
	replikacje	Liczebność	122	76	21
		% z Cel badania	55,7%	34,7%	9,6%
	generowanie odkryć / testowanie teorii	Liczebność	56	37	9
		% z Cel badania	54,9%	36,3%	8,8%
	zwiększanie precyzji teorii	Liczebność	19	39	3
		% z Cel badania	31,1%	63,9%	4,9%

Prawdopodobnie konsekwencją powyższej prawidłowości jest to, że niezależne zespoły częściej podejmowały badania nie-eksperymentalne, jak pokazałem w Tabeli 5.49. ($\chi^2 = 14,67$; $df = 2$; $p < 0,01$). Być może autorzy oryginalnych badań tj. osoby publikujące w najbardziej prestiżowych czasopismach, siłą rzeczy częściej przeprowadzają eksperymenty, gdyż wiedzą, że takiego rodzaju dane są bardziej pożądane w takich czasopismach.

TABELA 5.49. Rodzaj badania w zależności od tego czy było przeprowadzane przez niezależny zespół

			Rodzaj badania		
			eksperymenty	badania korelacyjne	inne
Niezależny zespół	NIE	Liczebność	70	31	16
		%	59,8%	26,5%	13,7%
	TAK	Liczebność	127	121	17
		%	47,9%	45,7%	6,4%

Warto zauważyć jako ciekawostkę, że rodzaj badania różnicował czas ukazania się badania $F(2,465) = 15,31$, $p < 0,001$ (Tabela 5.50.). W początkowym okresie ukazywało się więcej eksperymentów. Efekt ten nie jest rodzajem artefaktu wynikającego z tego, że niezależne zespoły przeprowadzały mniej eksperymentów (i ogólnie niezależne zespoły publikowały badania później). Efekt ten występuje nawet w obrębie niezależnych zespołów (porównując eksperymenty i badania korelacyjne $t(246)=3,64$, $p < 0,001$. W przypadku tych samych autorów, co oryginalne badanie, nie było różnicy między rodzajem badania a datą publikacji.

TABELA 5.50. Średni rok badania w zależności od rodzaju badania

Rodzaj badania	Średnia	Mediana	N	Odchylenie standardowe
Eksperymenty	4,77	5,00	263	4,010
badania korelacyjne	6,86	8,00	170	3,926
Inne	4,33	3,00	39	3,807
Ogółem	5,49	6,00	472	4,089

5.4.7. Co wpływa na rodzaj manipulacji?

Nie znalazłem zmiennych, które miałyby wpływ na to, jakiego rodzaju manipulacje stosowano w badaniach.

5.4.8. Co wpływa na wybór zmiennej zależnej?

W przypadku badań korelacyjnych natura zmiennej zależnej jest prawie zawsze samoopisowa, w eksperymentach taka zmienna też dominuje (62 %) jednak pojawiają się też inne rodzaje (Tabele 5.22. i 5.23.).

Poza tym zaobserwowałem kilka prawidłowości. W przypadku badań oryginalnych oraz replikacji występuje stosunkowo najwięcej zmiennych okołobehawioralnych (Tabela 5.53.). Efekt ten jest na pograniczu istotności $\chi^2 = 7,81$; $df=3$; $p<0,05$

TABELA 5.51. Natura zmiennej zależnej w zależności od celu badania

			Natura zmiennej zależnej		
			(około)behawioralne	samoopisowe	inne (biologiczne i utajone)
Cel badania	oryginalne	Liczebność	22	68	0
		%	24,4%	75,6%	0,0%
	replikacje	Liczebność	49	159	11
		%	22,4%	72,6%	5,0%
	generowanie odkryć / testowanie teorii	Liczebność	17	80	5
		%	16,7%	78,4%	4,9%
	zwiększanie precyzji teorii	Liczebność	5	54	2
		%	8,2%	88,5%	3,3%

Częściej też przy badaniu nie-studentów wykorzystuje się zmienne behawioralne (Tabela 5.52.). Efekt jest istotny ($\chi^2 = 8,84$ $df = 2$; $p < 0,05$), jednak znika, gdy usunąć jeden cykl badań (JPSP9), w którym korzystano tylko z prób niestudenckich i zmienna zależna miała charakter okołobehawioralny. Wydaje się więc, że grupa badana nie ma zasadniczego wpływu na naturę zmiennej zależnej.

Tabela 5.52. Natura zmiennej zależnej w zależności od grupy badanej

			Natura zmiennej zależnej		
			(około)behawioralne	samoopisowe	inne (biologiczne i utajone)
Grupa badana	student	Liczebność	66	290	18
		%	17,6%	77,5%	4,8%
	nie-student	Liczebność	27	71	0
		%	27,6%	72,4%	0,0%

Okazuje się natomiast, że taki wpływ może mieć pochodzenie badanych (Tabela 5.53.). W przypadku prób nie-amerykańskich częściej korzystano z miar okołobehawioralnych, a w przypadku prób amerykańskich częściej z samoopisowych ($\chi^2 = 12,63$; $df = 2$; $p < 0,01$).

TABELA 5.53. Natura zmiennej zależnej w zależności od pochodzenia badanych

			Natura zmiennej zależnej		
			(około)behawioralne	samoopisowe	inne (biologiczne i utajone)
Narodowość	Amerykanin	Liczebność	38	209	14
		%	14,6%	80,1%	5,4%
	nie-Amerykanin	Liczebność	55	151	4
		%	26,2%	71,9%	1,9%

Wyraźny efekt zanotowałem również zestawiając częstość stosowania realistycznych miar manipulacji i natury zmiennej zależnej. Zgodnie z intuicją, jeśli manipulacja była realistyczna, częściej również zmienna zależna była okołobehawioralna. Efekt jest bardzo wyraźny ($\chi^2 = 54,26$; $df = 2$; $p < 0,001$)

TABELA 5.54. Natura zmiennej zależnej w zależności od rodzaju manipulacji

			Natura zmiennej zależnej		
			okołobehawioralne	samoopisowe	inne (biologiczne i utajone)
Rodzaj manipulacji	nierealistyczna	Liczebność	17	109	15
		%	12,1%	77,3%	10,6%
	realistyczna	Liczebność	70	63	2
		%	51,9%	46,7%	1,5%

Na wybór zmiennej zależnej nie miała natomiast wpływu niezależność zespołu badawczego.

5.5. Zróżnicowanie metod badawczych

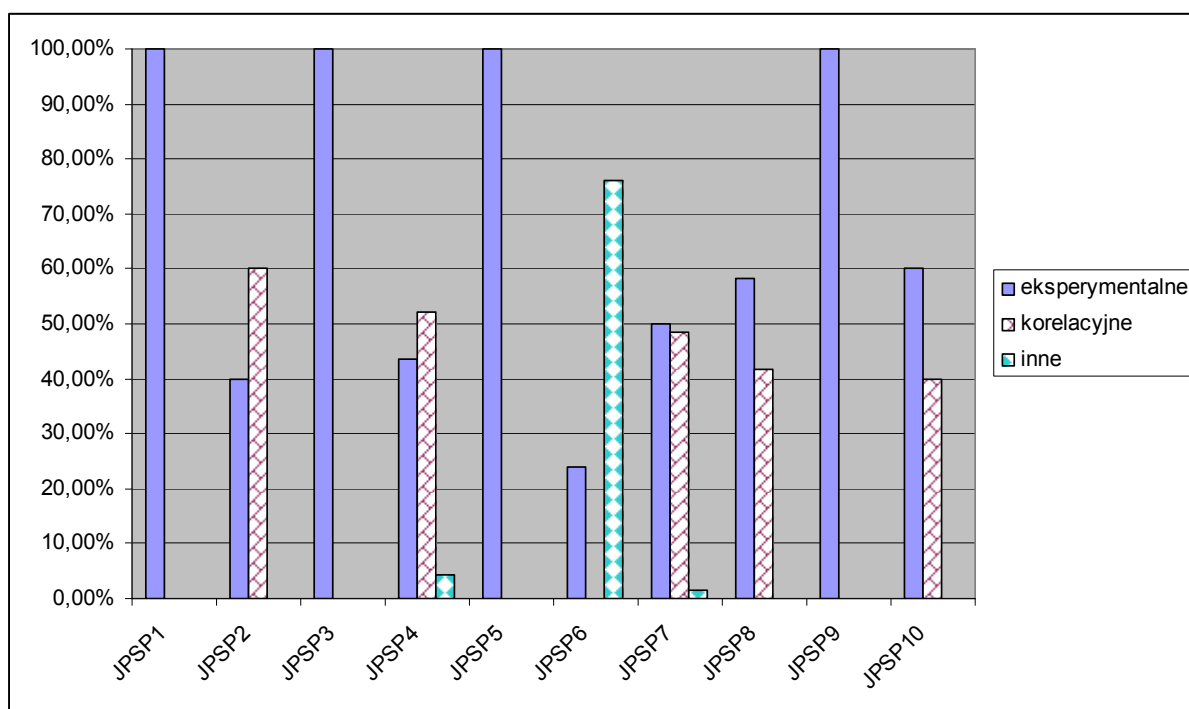
Interesujące z punktu widzenia celu tej pracy jest również zestawienie mające na celu pokazania zróżnicowania stosowanych metod badawczych. Jak wspominałem wcześniej (rdz. 2), jest wysoce zalecane, żeby określone hipotezy były weryfikowane przy pomocy wielu różnych metod i źródeł danych. Dzięki temu można zminimalizować częstość występowania artefaktów wynikających z ograniczeń danej metody badawczej.

Przedstawię tu podstawowe wskaźniki zróżnicowania (a raczej jego braku), przy czym dane te służą tylko ogólnej orientacji. Prawdziwe zróżnicowanie metod badawczych można poznać dopiero analizując jakościowo treść badań, procedurę, sposób manipulacji zmiennymi itd. Poświęcam temu zagadnieniu dużo miejsca w rozdziale 6.

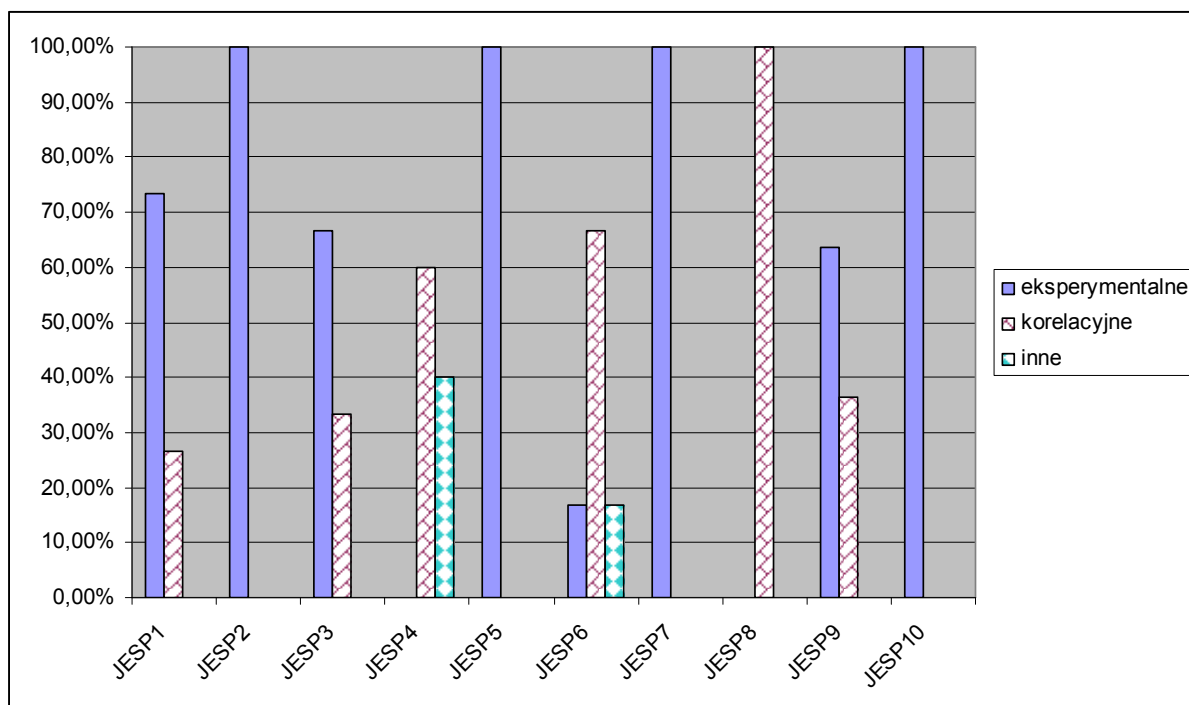
Mierząc zróżnicowanie przygotowałem dwa wskaźniki. Pierwszy to odsetek stosowania określonych rodzajów eksperymentów w zależności od cyklu badań. W 25 przypadkach jedna metoda wyraźnie zdominował cykl badań (tj. ponad 70 %)

Dodatkowo na wykresach 5.14 i 5.15 przedstawiłem analogiczne dane tylko, odpowiednio dla replikacji oraz dla badań polegających na testowaniu teorii i zwiększaniu jej precyzji (przy założeniu, że odbyło się przynajmniej 5 takich badań). Zróżnicowanie jest również niewielkie, w 12 na 17 przypadkach jedna z trzech metod stanowiła ponad 70 % badań, a w drugim przypadku odsetek ten wynosił 7 na 10.

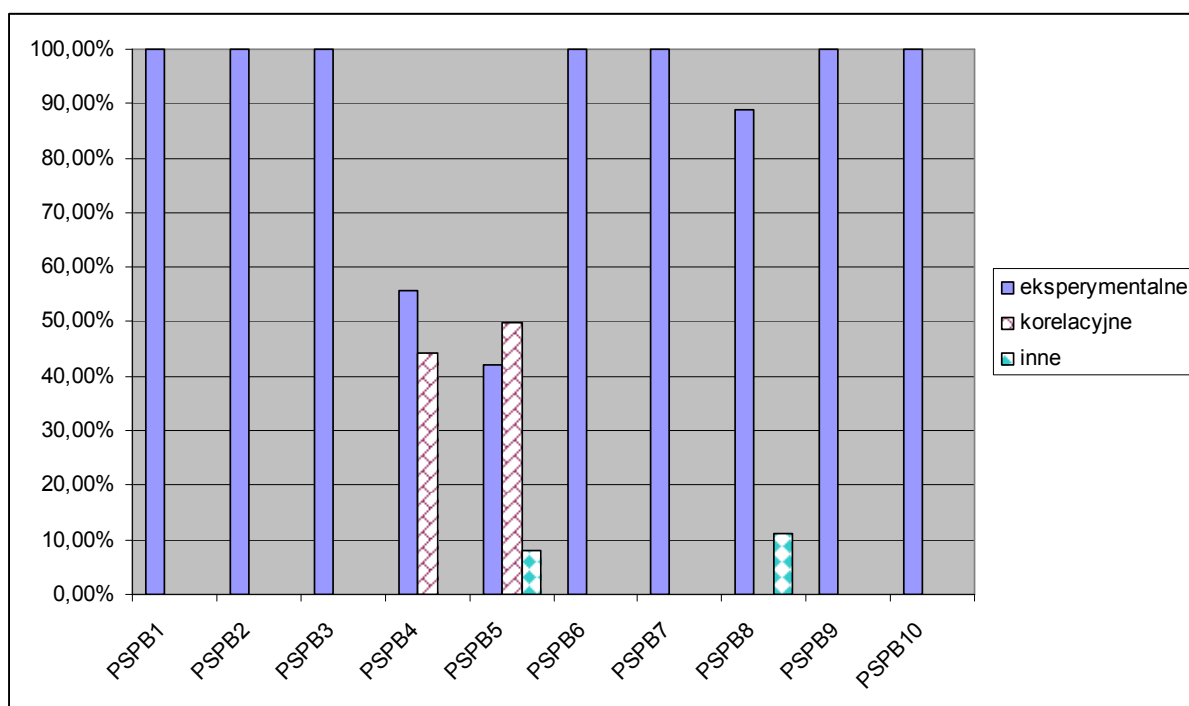
WYKRES 5.10. Częstość stosowania metod badawczych w zależności od cyklu badań (JPSP)



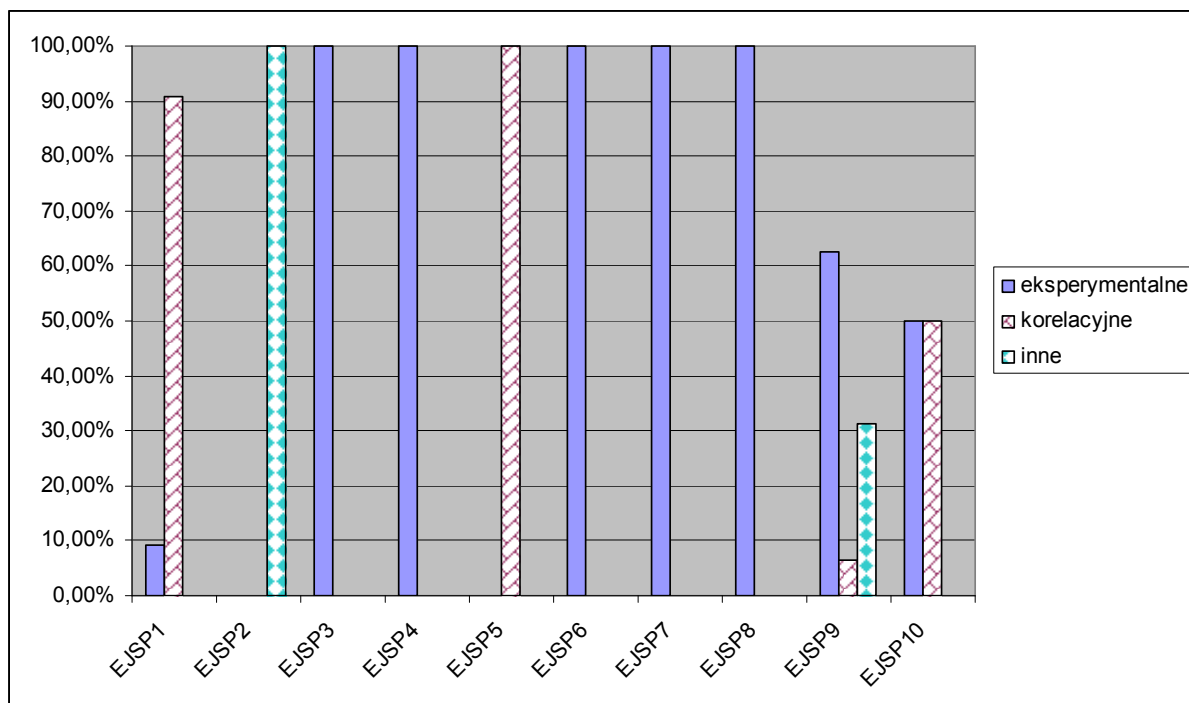
Wykres 5.11. Częstość stosowania metod badawczych w zależności od cyklu badań (JEJSP)



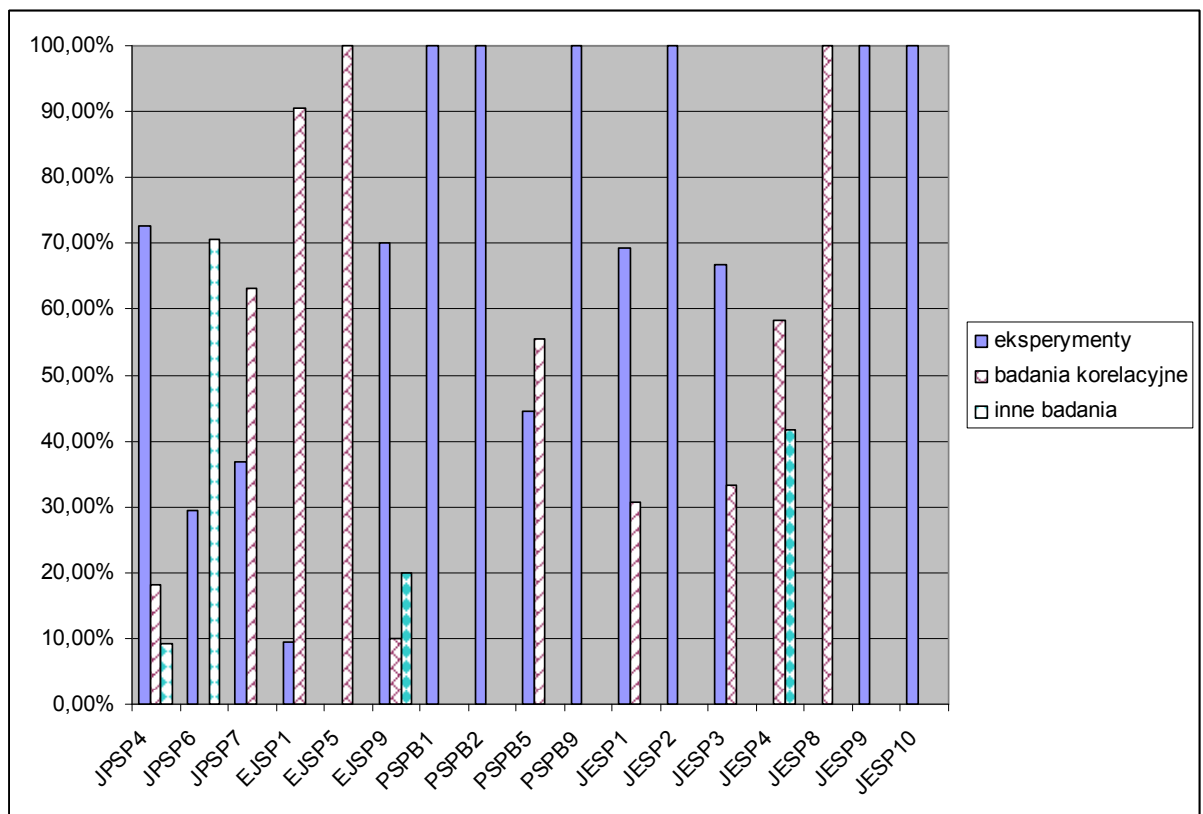
WYKRES 5.12. Częstość stosowania metod badawczych w zależności od cyklu badań (PSPB)



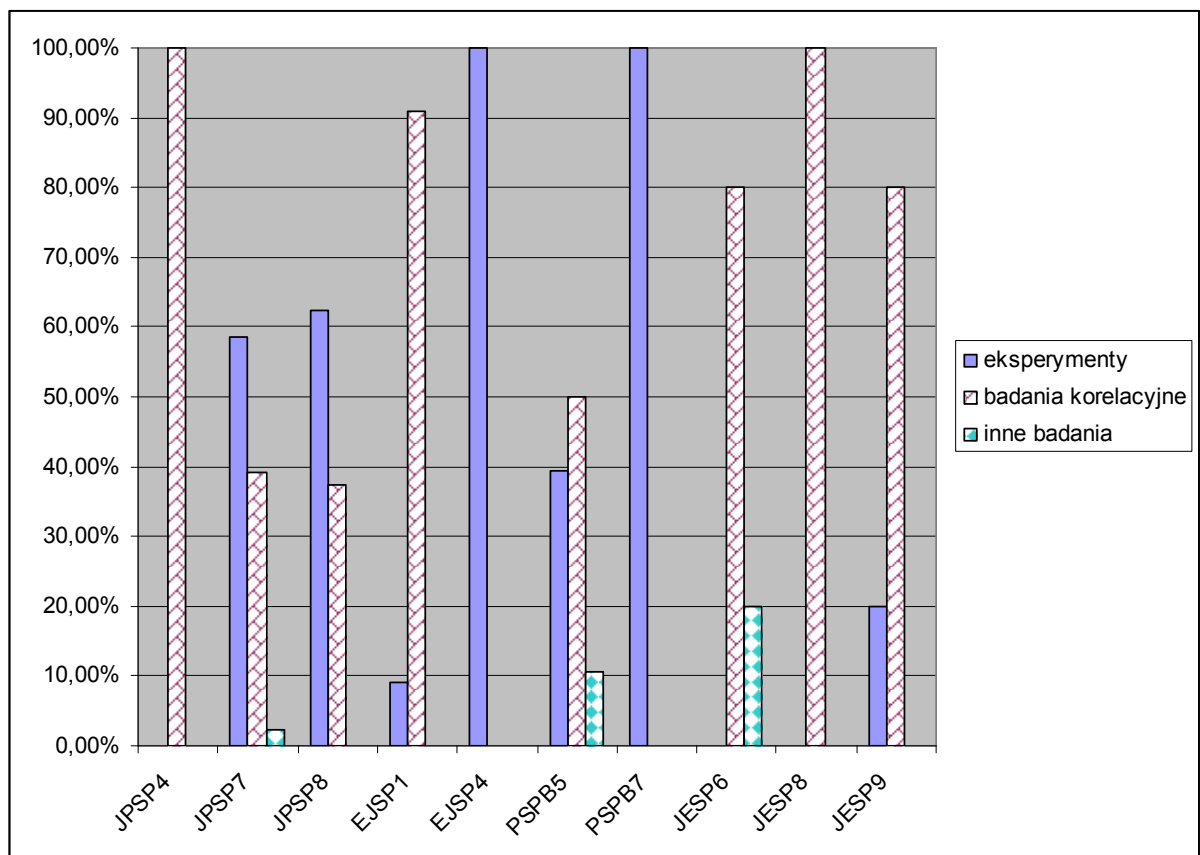
WYKRES 5.13. Częstość stosowania metod badawczych w zależności od cyklu badań (EJSP)



WYKRES 5.14. Częstość stosowania metod badawczych w zależności od cyklu badań (tylko replikacje; minimum 5 takich badań w cyklu)



WYKRES 5.15. Częstość stosowania metod badawczych w zależności od cyklu badań (tylko testowanie teorii i zwiększanie precyzji teorii; minimum 5 takich badań w cyklu)



Na koniec warto odnotować, że występowały wysokie korelacje między odsetkiem określonych badań w badaniach oryginalnych, a odsetkiem takich badań w cyklu badań nawiązujących. W przypadku eksperymentów korelacja taka wynosiła $r = 0,66$ ($p < 0,001$), w przypadku badań korelacyjnych ($r = 0,59$; $p < 0,001$). Odsetek badań oryginalnych ze zmienną zależną o charakterze samoopisowym przewidywał dobrze analogiczny odsetek w cyklu badań nawiązujących ($r = 0,69$; $p < 0,001$). Podobny efekt zachodził dla zmiennych behawioralnych ($r = 0,71$; $p < 0,001$). Jednak, jak już wcześniej zaznaczałem, interpretacja tych wyników jest niemożliwa w oderwaniu od bardziej jakościowej analizy treści samych badań (rozdział 6).

ROZDZIAŁ 6

ANALIZA JAKOŚCIOWA POSZCZEGÓLNYCH CYKLI BADAŃ

6.1. Uwagi wstępne

W tym rozdziale przeanalizuję poszczególne cykle badań. Najwięcej uwagi chcę poświęcić podstawowym pytaniom, które wynikają z wcześniejszych rozdziałów, a więc szczególnie:

- a) Czy w badaniach występują elementy wskazujące na stosowanie „psychologii fałszywie pozytywnej”, szczególnie małe próby, arbitralne decyzje dotyczące doboru zmiennych, czy fakt, że efekty działają tylko w interakcji z szeregiem dodatkowych zmiennych?
- b) Czy występują replikacje dokładne, które są podstawowym narzędziem kontroli rzetelności w nauce?
- c) W przypadku replikacji konceptualnych, czy badacze stosują pomysłowe procedury i niebanalne źródła danych, czy raczej ma miejsce częste powtarzanie tego samego schematu (jednak z na tyle znaczącymi zmianami, że nie można badania zakwalifikować jako replikacji dokładnej)?
- d) Czy stosuje się zmienne behawioralne jako zmienne zależne?
- e) Czy na podstawie wyjściowych teorii robi się niebanalne empirycznie testowalne przewidywania oraz czy próbuje się znaleźć zrozumienie efektu?
- f) Czy badane hipotezy są testowane na grupach innych niż studenckie?
- g) Czy badane hipotezy są testowane na osobach pochodzących z innych krajów niż USA (i inne rozwinięte kraje)?
- h) Czy cykl badań można opisać jako systematyczny program badawczy, czy raczej jako zbiór badań luźno powiązanych wyjściową hipotezą?

W przypadku każdego efektu staram się go opisać, oraz omówić podstawową metodę jego sprawdzenia w badaniu oryginalnym. Całościowa analiza jest zawarta w Załączniku 2, gdzie znajduje się wiele szczegółowych informacji. Tutaj nie powtarzam zdecydowanej większości z nich, a jedynie wskazuję na te dane, które uważam za szczególnie interesujące do omówienia czy to w kontekście rozwoju

wiedzy o wyjściowym efekcie, czy to w kontekście problemu wartości poznawczej praktyk badawczych w psychologii.

Poniższa analiza jest nierozzerwalnie związana z treściami przedstawionymi w Załączniku 2, dlatego odwołując się w tekście do poszczególnych badań w cyklu nie używam tradycyjnych odwołań („Kowalski, 2008” itp.), ale wskazuje numer badania w formacie [X], gdzie X to numer porządkowy artykułu w danym cyklu badań w pierwszej kolumnie załącznika, (jeśli zachodzi taka potrzeba to wskazuje też konkretny numer badania w tekście]. A więc przykładowo w podrozdziale JPSP4 numer [09] to badanie opisane w załączniku jako JPSP4_09, a [03:2], to badanie opisane jako drugie (np. „Study 2”) w artykule JPSP4_03. Taki sposób kodowania ma też ułatwić Czytelnikowi odnajdywanie odpowiednich badań w tabeli w załączniku 2. Jeśli w podrozdziale odwołuje się do badania z innego cyklu wówczas używam pełnego kodu (np. [JPSP6_03]).

6.2. Szczegółowa analiza jakościowa poszczególnych cykli badań.

6.2.1. Stapel, D. A., Koomen, W. (2000). How far do we go beyond the information given? The impact of knowledge activation on interpretation and inference. *Journal of Personality and Social Psychology*, 78(1), 19-37. (JPSP1)

Przy lekturze tego artykułu jako pierwsze narzuca się nazwisko autora, niesławnego Dederika Stapela. Komisja badająca działalność Stapela nie znalazła w nim jednak niczego, co sugerowałoby oszustwo (por. Levelt, 2012, s. 97).

W 5 badaniach eksperymentalnych wskazywano, że przymowanie słowami oznaczającymi cechy ekstremalne (np. „okrutny”) oraz ogólne (np. „dobry”) będzie wywierało większy wpływ na następujące potem sądy społeczne niż przymowanie słowami oznaczającymi cechy umiarkowane (np. „zły”) lub wąskie (np. „pomocny”). Manipulacja eksperymentalna dokonywała się najczęściej poprzez układanie zdań z rozproszonych słów (gdzie były ukryte słowa o takich charakterystykach). Badani następnie czytali teksty opisujące zachowania hipotetycznej osoby, których nie można było jednoznacznie zinterpretować (np. mogły zarówno wskazywać na wytrwałość jak i niezdrowy upór), a następnie musieli dokonywać oceny tej osoby m.in. pod kątem tego, na ile jest uparta lub wytrwała. Sama procedura nie jest więc

zbyt wyszukana i mogłaby posłużyć jako artykuł ilustrujący tezę o odejściu od metod behawioralnych i realistycznych manipulacji eksperymentalnych.

Liczba osób w kratkach jest stosunkowo nieduża (standardowo od 15 do 19, w jednym badaniu 24). Badanymi są każdorazowo studenci holenderscy. Artykuł ten, jak na JPSP, był stosunkowo rzadko cytowany (34 razy na przestrzeni 13 lat). Na podstawie analizy abstraktów oraz szczegółowej lektury dwóch artykułów nie znalazłem tekstów, które w jakiś sposób weryfikowałyby powyższą hipotezę. Oczywiście, jest bardziej niż pewne, że hipoteza ta została zweryfikowana w innych badaniach, gdyż wydaje się nieprawdopodobne, żeby np. stopień ekstremalności prymy nie był szerzej badany.

W ewentualnych przyszłych (mało prawdopodobnych) replikacjach można by się skupić na bardziej realistycznej manipulacji (np. określone cechy są prymowane poprzez zachowanie pomocnika eksperymentatora) oraz ocenach bardziej realistycznych bodźców (np. zachowania partnerów interakcji).

Ze względu na to, że cykl badań jest stosunkowo skromny, ograniczony do jednego artykułu, siłą rzeczy ocena jego wartości poznawczej jest niska.

6.2.2. Boldero, J., Francis, J. (2000). The relation between self-discrepancies and emotion: the moderating roles of self-guide. *Journal of Personality and Social Psychology*, 78(1), 38-52. (JPSP2)

W tym artykule wykazano moderatory efektu polegającego na tym, że rozbieżności między ja „idealnym” (jaki ktoś chciałby być) a ja „realnym” są skorelowane z przygnębieniem, a rozbieżności między ja „realnym”, a ja „powinnościowym” (jaki ktoś powinien być wg rodziców czy społeczeństwa) wiążą się z niepokojem. Sama ta hipoteza wymaga wielkiej ostrożności, gdyż zarówno rozbieżności, jak i poziom emocji są oceniane przez samych badanych, a więc łatwo tu o efekt wpływania odpowiedzi na siebie nawzajem (tak jak w badaniu Strack, Martin, Schwarz, 1988, por. rdz. 2.4).

W tym konkretnym tekście wskazywano na to, że efekt ma dodatkowe moderatory. Jeden z nich wydaje się być banalny: gdy kontrolować ważność określonej domeny, efekt ulega zmniejszeniu lub zanika. Drugi moderator jest już bardziej zaskakujący: jeśli badanie przeprowadzić w korespondującej lokalizacji (np. badani oceniają siebie

jako studenta na uczelni, lub siebie jako członka rodziny w swoim domu), wówczas korelacja „ja realne”-„ja powinnościowe” a niepokój była wyraźniejsza, ale tylko dla ważnych domen.

Ogólnie badanie wydaje się być mało poznawczo inspirujące, a i sam tekst jest bardzo nieprzystępnie napisany. Jak można się domyślić, zarówno zmienne zależne, jak i niezależne mają charakter samoopisowy (z wyjątkiem lokalizacji, którą manipulowano w jednym z pięciu badań). Warto też zauważyć, że dwa pierwsze (kwestionariuszowe korelacyjne) badania mają bardzo małe liczby badanych (46 i 47), dopiero ostatnie badanie ma dużą liczbę (334) (dwa pośrednie badania eksperymentalne mają zadowalającą liczbę osób badanych tj. około 40-50 w kratce). Wadą tego rodzaju badań wydaje się być to, że sam fakt oceny rozbieżności między różnymi rodzajami „ja” może automatycznie wpływać na następującą ocenę przeżywanych emocji. Być może warto byłoby wprowadzić większe przerwy między pomiarem emocji a rozbieżnościami w różnych wersjach „ja”. Alternatywnie badani mogliby prowadzić przez jakiś czas dzienniczki, w których rejestrowaliby emocje i owe rozbieżności (ale w innych porach dnia).

Są to propozycje hipotetycznych badań, gdyż podobnie jak w przypadku badania poprzedniego, nie było żadnych replikacji konceptualnych, ani badań choćby nawiązujących teoretycznie do wyjściowego efektu.

Osobiście nie ukrywam, że fakt, że znalazłem taki artykuł w JPSP, był dla mnie zaskoczeniem. Mówiąc dosadnie, powodem tego zaskoczenia, jest jego teoretyczna miałość. Nie czuję się ekspertem od badań nad ja realnym czy idealnym, ale nie wydaje się, żeby to był bardzo ważny dział psychologii. A badanie jest dosyć wydumaną wariacją wokół mało istotnego efektu. Jest oczywiście technicznie poprawne i być może zasługuje na publikację, ale w czasopiśmie dużo niższej rangi. Podobnie jak w przypadku wcześniejszego badania ocena wartości poznawczej tego cyklu badań musi być niska.

6.2.3. Reynolds, K. J., Turner, J. C., Haslam, S. (2000). When are we better than them and they worse than us? A closer look at social discrimination in positive and negative domains. *Journal of Personality and Social Psychology*, 78(1), 64-80. (JPSP3)

Badanie to jest próbą wskazania moderatorów ważnego efektu nadmiernie pozytywnego oceniania grupy własnej (i negatywnego oceniania grupy obcej). Zdaniem badaczy efekt ten dotyczy tylko cech, które są stereotypowo charakterystyczne dla danej grupy. I tak, studenci psychologii oceniali, że studenci uniwersytetu (czyli ich grupa własna) w porównaniu do studentów politechniki mają wyższe natężenie cech „intelektualny”, czy „ambitny” (a więc cech stereotypowych dla tej grupy), ale z kolei studenci politechniki są bardziej praktyczni i twardo stąpający po ziemi. Analogicznie było z cechami negatywnymi. W jednym z trzech badań studenci porównywali grupy osób młodych i starszych.

Warto zaznaczyć, że pomimo kwestionariuszowego (choć eksperymentalnego) charakteru badania, próby były bardzo niewielkie (około 30-70 osób), a liczba osób w kratce wręcz na bardzo niskim poziomie (7-17 osób).

Zaletą artykułu jest niewątpliwie ciekawa, intuicyjnie łatwo zrozumiała hipoteza. Ale zarówno pod względem liczby osób badanych, jak i ograniczonej próby (studenci australijscy) badanie można zaliczyć do kategorii „po linii najmniejszego oporu”.

Wśród 75 tekstów bezpośrednio cytujących ten artykuł nie ma żadnego, który by się okazał po bliższej inspekcji replikacją lub testowaniem teorii, więc niestety na tym omawianie tego interesującego odkrycia musi się zakończyć. Efekt ten jest przywoływany dosyć często, ale nie jest badany.

6.2.4. Haselton, M. G., Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, 78(1), 81-91. (JPSP4)

Badanie to dokumentuje efekt, który wcześniej (rozdział 3) posłużył mi do zilustrowania kilku idei. Dla przypomnienia: stwierdzono, że mężczyźni doszukują się we względnie neutralnych zachowaniach kobiet zainteresowania seksualnego, w większym stopniu niż (wg deklaracji samych kobiet) te zachowania są rzeczywistą oznaką takiego zainteresowania. Badanie to zainicjowało teorię zarządzania błędami (ang. *error management theory*), która cieszy się dużą estymą w środowisku psychologów zorientowanych ewolucyjnie (np. Haselton, Nettle, 2006). W tym artykule pokazano też „odwrotną stronę medalu” tj. niedocenianie przez kobiety

poziomu zainteresowania romantycznego u mężczyzn, ale zdecydowałem skoncentrować się tylko na tej wcześniejszej hipotezie.

Samo badanie nie jest proceduralnie zbyt wyszukane. Badani (studenci) oceniali na ile hipotetyczna osoba jest zainteresowana seksualnie na podstawie opisanych zachowań.

Badanie to było często replikowane konceptualnie (15 artykułów i 21 badań) i na szczęście ta procedura została nieco udoskonalona w dalszych badaniach:

- efekt badano też, gdy mężczyźni oceniali nagrane interakcje kobiet i mężczyzn w trakcie sesji szybkich randek [11]; tutaj jednak nie potwierdzono wyjściowej hipotezy;
 - mężczyźni oceniali stopień zainteresowania seksualnego na podstawie zdjęć (wyrażających różne emocje) [09]; tutaj hipotezę potwierdzono;
 - prezentowano w formie nagrania wideo, samego audio, lub transkryptu samoopisy osób z rzekomego portalu randkowego [13]; tutaj potwierdzono hipotezę;
 - badani angażowali się w sesje „szybkich randek”, po których oceniano na ile osoby płci przeciwnej były zainteresowane nimi seksualnie ([15]; zasadniczo potwierdzono hipotezę). Jednak w badaniu [14], w którym dwie osoby płci przeciwnej wykonywały wspólnie zadanie nie stwierdzono wyjściowego efektu.
- Dodatkowo badacze zgromadzili sporo pośrednich danych wskazujących na to, że efekt może być prawdziwy. W szczególności, w kilku badaniach, w których badani przywoływali wspomnienia ze swojego życia, stwierdzono że:
- mężczyźni częściej deklarowali, że źle zinterpretowali „intencje” przyjaciółek [01], lub po prostu osób płci przeciwnej [12], od których następnie „dostali kosza”;
 - kobiety częściej deklarowały, że ich przyjacielskie zachowanie było odbierane jako oznaka zainteresowania seksualnego [02];
 - mężczyźni zawyżali rzekome zainteresowanie seksualne ich osobą ze strony ich przyjaciółek (owe przyjaciółki niezależnie pytano o rzeczywisty poziom zainteresowania) [07];
 - kobiety częściej deklarowały wykorzystywanie atrakcyjności fizycznej i swojego rzekomego zainteresowania seksualnego do pozyskiwania korzyści, a mężczyźni częściej deklarowali, że padli "ofiara" takich taktyk [10].

Kilka dalszych badań próbowało zidentyfikować moderatory i mediatory tego efektu.

Oryginalne badanie zasadniczo nie miało dokładnej replikacji, choć jedno badanie [04] jest dosyć podobne. W prawie wszystkich (22 na 23) badaniach wykorzystywano tylko próby studenckie (w jednym przypadku [10] badanymi były osoby, które wypełniły ankietę w Internecie). Można sobie wyobrazić, że studenci są grupą szczególnie zaktywizowaną pod względem seksualnym i w innych grupach nie stwierdzono by tego efektu (osobiście w to wątpię, ale mimo wszystko szkoda, że tego nie sprawdzono). We wszystkich próbach badani pochodzili z USA lub Wielkiej Brytanii. I tu można się zastanawiać, czy w bardziej „powściągliwych” kulturach (np. Japonii) podobny efekt byłby zaobserwowany.

Podsumowując, wydaje się, że cykl badań dobrze dokumentuje występowanie wyjściowego efektu. Samo wyjściowe badanie nie zostało bezpośrednio zreplikowane, ale występuje sporo replikacji konceptualnych, oraz sporo badań, które można uznać za niebanalne testowanie teorii (najczęściej udane). Zmiennymi zależnymi były prawie zawsze zmienne samoopisowe (i sporadycznie utajone [05]). W żadnym badaniu nie pokuszono się o próbę znalezienia behawioralnej miary potwierdzającej hipotezę. W przypadku tego efektu nie jest to być może aż takie ważne, a ewentualne badania mogłyby być problematyczne pod względem praktycznym i etycznym, ale można sobie wyobrazić, że choćby próba umówienia się na randkę przez telefon (tak jak w klasycznym badaniu Duttona i Arona, 1974) mogłaby być sensownym wyborem zmiennej zależnej. Łatwo też sobie wyobrazić, na czym mogłaby polegać manipulacja eksperymentalna: przykładowo kobieta, pomocniczka badacza, uśmiecha się i wykazuje serdeczne (ale nie seksualne) gesty w kierunku badanego, lub zachowuje się naturalnie. Badani następnie oceniają jej poziom zainteresowania seksualnego lub mają możliwość podjąć próbę umówienia się na randkę.

Podsumowując, hipoteza, na tle wielu innych analizowanych tutaj hipotez, jest nieźle udokumentowana. Przebadano prawie 4,5 tys. osób; 12 na 15 badań potwierdziło bezpośrednio efekt⁸⁰. Największą wadą tego cyklu badań są znaczące niedostatki, jeśli chodzi o próbę międzykulturowego i międzygeneracyjnego weryfikowania odkrycia. Stąd wartość poznawcza tego zbioru badań jest umiarkowana.

⁸⁰ Pozostałe badania dotyczyły moderatorów lub mediatorów.

Przypomnę, że niezależny przegląd literatury i cytowań drugiego stopnia (por. rdz. 4.2) wskazał mi, że w literaturze nie ma dużo więcej danych ponad te, które zostały tu uwzględnione.

6.2.5. van Dijk, E., Wilke, H. (2000). Decision-induced focusing in social dilemmas: give-some, keep-some, take-some, and leave-some dilemmas. *Journal of Personality and Social Psychology*, 78(1), 92-104. (JPSP5)

Badacze aranżowali kilkuosobowe gry ekonomiczne, których istota polegała na tym, że środki we „wspólnej” puli zasobów były pomnażane przez badacza (i rozdzielane równo między badanych), a badani decydowali, ile ich własnych środków powinno się w tej puli znaleźć. Wykazano, że zachodzą różnice w zależności od tego, czy środki są najpierw u badanego, czy w ogólnej puli (i badany może zdecydować ile ich zabrać dla siebie), oraz w zależności od tego to jak obramuje się zachowanie (czy przekazanie środków X to „dawanie X”, czy może „zostawianie sobie 1-X”).

Manipulowano też dodatkowymi zmiennymi, jednak tutaj szczegóły nie są istotne.

Badanie przeprowadzono na grupie studentów holenderskich. Wielkość $\square 2$ jest na pograniczu przyzwoitości (nieco poniżej 20 osób; por. Simmons, Nelson, Simonsohn, 2011).

Tak czy owak, niestety, poza oryginalnym badaniem, sam efekt nie doczekał się żadnych replikacji (tekst był cytowany 46 razy). Odpowiednikiem „życiowym” tego eksperymentu może być sytuacja prowadzenia firmy przez kilku wspólników, z których każdy może zainwestować własne środki (nie tylko finansowe), żeby potencjalnie je pomnożyć, a wspólnicy czerpią zyski w równym stopniu. Trudniej wskazać odpowiednik dotyczący codziennego życia i szerszego odsetka populacji. Być może dlatego to badanie nie spotkało się z późniejszym zainteresowaniem społeczności badawczej.

6.2.6. Gilovich, T., Medvec, V., Savitsky, K. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, 78(2), 211-222. (JPSP6)

Artykuł ten opisuje badanie niezwykle atrakcyjne w treści i formie. Stwierdzono, że ludzie mają tendencję do przeceniania stopnia, w jakim inni zwracają uwagę na ich cechy i zachowania (szczególnie negatywne, ale też pozytywne). (Badacze nazwali ten efekt „spotlight effect”, co przetłumaczyłbym jako „efekt-bycia-w-blasku-flesza”). Żeby to wykazać przeprowadzono 3 pomysłowe eksperymenty (oraz 2 dodatkowe próbujące znaleźć więcej uwarunkowań efektu). W jednym badaniu zakładali koszulkę z niepopularnym wśród studentów piosenkarzem (polskim odpowiednikiem byłby zapewne jakiś niewyszukany zespół disco polo), wchodzili do grupy (też rzeczywistych badanych), przebywali w niej pewien czas i po jej opuszczeniu szacowali, ile osób zwróciło uwagę na ich koszulkę. Natomiast osoby w grupie pytano czy zwróciły uwagę na coś szczególnego u tamtego badanego. Okazało się, że szacunki badanych dotyczących tego, ile osób zwróciło uwagę na ich koszulkę były prawie dwukrotnie zawyżone. W badaniu drugim zastosowano identyczną procedurę, z tą różnicą, że studenci zakładali koszulki z popularną postacią wśród studentów (np. Bobem Marleyem). Również tym uczestnikom wydawało się, że więcej osób zwróciło na nich uwagę niż miało to miejsce w rzeczywistości. W trzecim eksperymencie badani uczestniczyli w dyskusji grupowej. Przeceniali liczbę swoich pomyłek językowych, agresywnych wypowiedzi czy wypowiedzi odebranych krytycznie, które dostrzegły inne osoby (z kolei nie doceniali liczby znaczących komentarzy, jakie w oczach innych wygłosili).

Badanie to miało adekwatny do swojej atrakcyjności wskaźnik cytowań (99 razy). Z tej liczby 8 artykułów (z 20 badaniami) spełnia kryteria mojej analizy. Wśród nich 5 artykułów (16 badań) zostało przeprowadzone przez tych samych badaczy co badanie wyjściowe (i nie będzie zapewne zaskoczeniem, że każdorazowo przedstawiono wyniki potwierdzające wyjściowe hipotezy). Te badania omówię w pierwszej kolejności.

W dużej mierze (11 razy) były to realistyczne i pomysłowe eksperymenty, wykraczające poza standardowe procedury „papier i ołówek” (lub „ekran i klawisze”). Były też zróżnicowane pod względem tematycznym, choć niestety nie występowały żadne replikacje dokładne). Wśród zbadanych prawidłowości warto wspomnieć następujące:

- badani przeszacowują ile osób zwróci rzeczywiście uwagę na ich udany i mało udany wygląd (w jednym i drugim wypadku mniej osób zwraca uwagę) ([02]; w tym tekście analogiczne efekty w obszarze gier sportowych i gier komputerowych);
- badani w trakcie zadania oceniają, że inni oceniają niżej ich funkcjonowanie intelektualne niż to ma miejsce w rzeczywistości [01];
- badani przeszacowują, ile osób zauważy usunięcie ich osoby z grupy, oraz jak grupa będzie funkcjonowała bez nich [05].

Dwa dalsze badania miały mniej realistyczny charakter (np. szacowanie częstości gaf i błędów: [03]), a trzy testowały możliwe wytłumaczenie rzeczzonego efektu [04] (ludzie sądzą, że inni popełniają większy podstawowy błąd atrybucyjny niż oni sami).

Cztery dodatkowe badania (w trzech artykułach) zostały przeprowadzone przez niezależne zespoły. Dwa [07] miały bardzo podobną formę jak badanie oryginalne, ale dotyczyły oceny częstości zauważenia i zapamiętania określonego napisu na koszulce (a nie samej koszulki). W kolejnym badaniu [06] badani w symulowanych interakcjach przeszacowywali stopień zagrożenia, jaki stanowią dla drugiej strony. W ostatnim badaniu z tego cyklu [08] badani (uczestnicy kursu publicznego przemawiania) przeszacowywali poziom lęku społecznego, jaki dostrzegą inne osoby w ich wystąpieniu.

Cały cykl 25 badań miał kilka cech charakterystycznych. Wszystkie one odbyły się na Amerykanach (tylko w [04:3] przy okazji zaangażowano grupę Japończyków). W 23 badaniach byli to studenci (dodatkowo w jednym przypadku studenci MBA, oraz wspomnieni już uczestnicy kursu publicznego przemawiania). Cykl ten mógłby więc idealnie posłużyć jako ilustracja do postulatu Arnetta (2008) o przywrócenie psychologii „porzuconych 95 %” (alternatywnie „neglected” można tłumaczyć jako „zaniedbanych”). To, że wszystkie badania bez wyjątku potwierdzają wyjściową hipotezę, również jest trochę niepokojące. Efekt jest silny, a grupy były dosyć liczne, tym niemniej warto byłoby przeprowadzić szczegółowe analizy statystyczne, czy taki odsetek sukcesów jest realistyczny bez uciekania się choćby do wybiórczego publikowania.

Na plus natomiast wyróżnia się dosyć szerokie zróżnicowanie tematyczne i to, że na ogół wykorzystywano eksperymenty o realistycznej procedurze. Nie miałbym w tej kwestii większych zastrzeżeń. Zmienna zależna miała charakter szacowania przez

badanych, ale w przypadku takiej hipotezy trudno o jakieś inne (behawioralne) zmienne.

Choć pod względem liczby badań ten cykl wyróżnia się na tle innych tutaj omawianych, to badania są na ogół do siebie mocno niepodobne, co potwierdza tezę o niekorzystnej w psychologii „kulturze [fałszywego] nowatorstwa” (a wysoki wskaźnik sukcesów potwierdza tezę o „kulturze [rzekomego] sukcesu”). Do tego wątku jeszcze powrócę przy okazji podsumowania analizy jakościowej.

6.2.7. Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., Uhles, A. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, 78(2), 223-246. (JPSP7)

Artykuł dokumentuje występowanie właściwości grup społecznych, jaką jest entytatywność. W przypadku tego artykułu miałem dylemat, czy włączyć go do analizy, gdyż badanie samo w sobie nie opisuje nowego fenomenu. Pojęcie entytatywności grupy było już wcześniej opisywane w literaturze (np. Yzerbyt, Rogier, Fiske, 1998). Ostatecznie zważył fakt, że badanie testuje występowanie tego zjawiska w sposób dosyć nowatorski (nie znalazłem podobnego we wcześniejszych tekstach), dlatego postanowiłem je uwzględnić. Używam tutaj dosłownego polskiego tłumaczenia „entytatywności” (ang. *entitativity*), gdyż nie jest mi znany polski odpowiednik tego pojęcia. W angielskim jest to pojęcie techniczne, nie pochodzące z ogólnego słownika, więc takie dosłowne przetłumaczenie nie wydaje mi się błędem.

W entytatywnych grupach członkowie dzielą wspólne cele, są do siebie bardziej podobni i częściej wchodzi w interakcje. Przykładem grupy wysoce entytatywnej jest rodzina lub grupa bliskich przyjaciół, a przykładem grupy mało entytatywnej jest zbiór osób oczekujących na przystanku na autobus. Entytatywność grupy koreluje z subiektywnym poczuciem ważności przynależności do niej, dodatkowo entytatywność grupy jest najlepszym predyktorem poczucia ważności przynależności do niej.

Przedstawiono w wyjściowym artykule 3 badania. Dwa pierwsze były proceduralnie prawie identyczne, różniła je tylko grupa badana (pierwszą byli studenci amerykańscy, a drugą studenci, tu ciekawostka, polscy). Badani otrzymywali listę 40 grup społecznych i mieli je ocenić pod kątem szeregu właściwości np. tego, jak

często osoby wchodzą ze sobą w interakcje, jak członkowie tych grup są do siebie podobni itd. W trzecim badaniu osoby badane (tym razem studenci tylko amerykańscy) generowali listę grup, do których należą, a następnie oceniali te grupy pod kątem szeregu właściwości (m.in. jak ważna jest dla nich osobiście przynależność do tych grup).

Badanie to było często cytowane (195 razy). Podczas analizy artykułów cytujących pojawił się drugi problem dotyczący tego, które artykuły uwzględniać. Szybko okazało się, że zdecydowana większość badań nie jest proceduralnie zbliżona do wyjściowego badania, oraz nie testuje samego faktu występowania fenomenu entytatywności, tylko zakłada go *a priori*. Mówiąc inaczej, większość badań dotyczy tego, na jakie aspekty poznania społecznego postrzegana entytatywność grupy wpływa lub co wpływa na postrzeganą entytatywność, ale zakładając, że fenomen entytatywności jest faktem. Natomiast wyjściowe badanie w pierwszym rzędzie stara się wykazać sam fakt występowania. Ostatecznie postanowiłem uwzględnić wszystkie takie badania, żeby analiza tego konkretnego artykułu pozwoliła prześledzić, jak określony fenomen jest badany w najszerszy możliwy sposób. Jednak podkreślę, że zdecydowana większość badań jest konceptualnie dosyć odległa od problemu występowania entytatywności i sposobu jego badania z wyjściowego badania.

Ostatecznie do analizy tego badania zostało zaliczonych 38 artykułów (zawierających 65 badań). Jest to największa liczba ze wszystkich analizowanych tu oryginalnych badań. W tej liczbie 30 to badania korelacyjne, a 34 badania eksperymentalne (oraz jedno badanie oparte na analizie danych wytworzonych poza badaniem).

Mimo tak dużej liczby badań, żadne z nich nie jest proceduralnie bliską replikacją (pod względem proceduralnym) badania wyjściowego. Występuje natomiast sporo badań, które dokumentują to, jakie cechy sytuacji lub członków grupy wpływają na ich postrzeganą entytatywność (w badaniu oryginalnym, dla przypomnienia, korelowano oceny entytatywności z ocenami właściwości grupy). Jeśli chodzi o badania eksperymentalne, w których manipulowano określonymi cechami grup, stwierdzono, że zwiększają postrzeganą entytatywność następujące właściwości/sytuacje:

- zgadzanie się z decyzją grupy [09],
- poziom interakcji i zależności członków grupy (utworzonej w laboratorium [10] oraz prezentowanej badanym [21],
- podobieństwo fizyczne i wspólny cel członków grupy [12].

Natomiast w badaniach korelacyjnych stwierdzono, że wpływ na postrzeganie grup jako entytatywnych ma:

- wyższy poziom uprzedzeń względem określonej grupy [22],
- wielkość grupy (negatywnie) [33:3].

Dosyć często była replikowana hipoteza (oryginalne badanie: study 3) mówiąca o tym, że wyższa entytatywność grupy koreluje z wyższym poczuciem przynależności do tej grupy (np. [24], [25], [27], [34], [35], [38]). Grupy do których się przynależało były też postrzegane jako bardziej entytatywne [31]. Prawie wszystkie badania opisane w tym akapicie miały charakter korelacyjny, za wyjątkiem badania [25:3]. Tu jednak manipulacja eksperymentalna nie wydaje się zbyt realistyczna (badani wyobrażali sobie, że są mieszkańcami kraju o wysokiej lub niskiej entytatywności np. poprzez informacje o tym, że kraj jest jednolity społecznie).

Pośrednim wnioskiem z tej hipotezy może być to, że zagrożenie lub odrzucenie zwiększa poczucie entytatywności grupy własnej. Istotnie, badacze wykazali takie efekty w kilku badaniach ([02], [16], [20]), gdzie manipulowano eksperymentalnie poziomem odrzucenia, poziomem niepewności lub poziomem zagrożenia (odrzuceniem poprzez polecenia napisania historii o sytuacji, gdy zostało się odrzuconym; poczuciem niepewności poprzez myślenie o sytuacjach, gdy badani tak się czuli w życiu; poczuciem zagrożenia poprzez manipulacje znane z badań wokół teorii opanowywania trwogi, por. np. PSPB7). W innym badaniu [26] wyższa spostrzegana entytatywność grupy własnej była skorelowana z wyższym poziomem spostrzeganej różności z grupą obcą (ale w interakcji z poczuciem zagrożenia). Grupy własne, które miały większe możliwości determinacji swojego losu, były też postrzegane jako bardziej entytatywne [33]).

Istnieje wreszcie duża grupa badań, które dokumentują to, w jaki sposób postrzegana (lub manipulowana) entytatywność grupy wpływa na procesy poznania społecznego. Przykładowo, stwierdzono, że zmienna ta ma wpływ na:

- poziom zagrożenia ze strony grupy obcej [03],
- rodzaj poszukiwanych informacji o zachowaniach członków grupy [04],
- postrzeganą spójność grupy, pomimo różnych zachowań w różnych sytuacjach [05],
- gotowość do porównywania członków grupy między sobą ([01] i [06]),
- ocenę argumentów członka takiej grupy ([07], [23]),
- rodzaj zasad, którymi grupy się kierują [11],
- postrzeganą odpowiedzialność członków grupy za przewinienia osób z tej grupy ([13], [36]) i gotowość zemsty za takie zachowania ([18], [19]),
- sposób opisywania zachowań członków grupy ([14], [29]),
- tendencję do używania stereotypów w ocenie zachowań członków grupy ([15], [35]),
- większą gotowość do używania teorii spiskowych dotyczących działania członków danej grupy [17],
- ocenę „naturalności” grupy [30],
- wysoka entytatywność grupy ma też zmieniać ocenę niskich osiągnięć członków grupy własnej i likwidować klasyczny efekt preferencyjnego ich oceniania [28].

Z tej grupy badania [11], [13], [15], [17], [18], [23:1a], [23: 1b], [30], [35], [36] mają charakter korelacyjny, pozostałe eksperymentalny.

Udokumentowano również pewne różnice kulturowe mające wpływ na postrzeganie zachowań i cech wysoce entytatywnych grup [08] (w badaniu wzięło udział ponad 800 osób z 8 krajów). W badaniu [37] sprawdzano, jakie są różnice w przekonaniach Amerykanów i Japończyków na temat wpływu cech esencjalnych (np. podobieństwo) i cech dynamicznych (np. podobne cele) na postrzeganą entytatywność. Z kolei w badaniu [32] sprawdzano w jaki sposób dzieci 10-letnie spostrzegają entytatywność grup.

W tym cyklu, co widać powyżej, zostało przetestowanych wiele hipotez, wszystkie związane z ideą entytatywności. Jeśli chodzi o metody, wspomnę o kilku badaniach, które wyróżniają się oryginalnością na tle innych. Zdecydowana większość badań niestety nie wyróżniała się niczym szczególnym, tj. polegały albo na korelowaniu ze sobą zestawu skal kwestionariuszowych (w przypadku badań korelacyjnych), albo na

wyświetlaniu obrazków lub podawaniu w formie werbalnej informacji o grupach i pomiarze zmiennych zależnych (których natura też była samoopisowa).

W badaniu [01] zmienną zależną była tendencja do ulegania złudzeniu optycznemu, które jest odmianą złudzenia Ebbinghausa (tj. małe koło na tle dużych wydaje się większe i *vice versa*). Zaliczyłem taką zmienną do zmiennych utajonych. Ma to wskazywać na to, że osoby z grup wysoce entytatywnych są do siebie bardziej porównywane.

W 4 badaniach w artykule [10] aranżowano sytuację w taki sposób, że badani współpracowali ze sobą (lub nie), byli od siebie zależni, oceniali atrakcyjność własnej grupy, po czym mierzono postrzeganą entytatywność. Badanie to wyróżnia się realizmem w manipulacji zmienną niezależną na tle zdecydowanej większości innych badań.

W badaniach [14] i [29] badani wypowiadali się swobodnie na zadane tematy, a zmienną zależną był rodzaj używanego języka do opisywania grup w zależności od ich entytatywności, m.in. stopień abstrakcyjności lub konkretności wypowiedzi.

W badaniu [19] badani byli rzekomo odrzucani przez jednego członka trzyosobowej grupy. Grupa była wysoce (wszyscy nosili koszulki tej samej drużyny sportowej) lub nisko entytatywna (osoby były różnie ubrane). Badany miał możliwość "zemsty" poprzez zaaplikowanie sesji białego szumu.

W badaniu [26:1] wykorzystano niezwykle oryginalną próbę badaną: osoby uczestniczące w strajku. Stwierdzono, że wyższa postrzegana entytatywność grupy własnej jest skorelowana z wyższym poziomem spostrzeganej różności od grupy obcej (w interakcji z poczuciem zagrożenia). Taka „grupa” badana nie jest oczywiście dostępna na zawołanie, a badanie pozytywnie świadczy o pewnej czujności badaczy w gotowości do wykorzystania szczególnej sytuacji do celów badawczych. W drugim badaniu w tym artykule [26:2] badano podobne zależności, ale tym razem na populacji ogólnej (zrekrutowanej do badania w centrum handlowym), co także nie zdarza się często.

Wreszcie warto wspomnieć o badaniu [09:1], w którym testowano hipotezę o tym, że zgadzanie się z określoną decyzją grupową będzie powodowało ocenę grupy ją wydającą jako bardziej entytatywną. Badacze przeanalizowali artykuły prasowe dotyczące wyroków amerykańskiego Sądu Najwyższego. W zależności od tego czy

autor artykułu się zgadzał z decyzją czy nie, miało to wpływ na opisywany poziom entytatywności sądu, jaki był zawarty w tekście (jeśli się nie zgadzał, była ona niższa). Ocen entytatywności w artykułach dokonywali sędziowie kompetentni. Jest to jeden z nielicznych przykładów w tej analizie poszukiwania danych pozalaboratoryjnych, dzięki którym można sprawdzić hipotezę. W dwóch dalszych badaniach w tym artykule, dane te potwierdzono przy pomocy badań na studentach (jeśli się nie zgadzali z decyzją sądu oceniali go jako mniej entytatywny).

Pomimo dużego zróżnicowania tematycznego badań, takiego dużego zróżnicowania nie odnotowałem, jeśli chodzi o narodowość badanych. I tak na 68 badań w 49 (72 %) badana próba była wyłącznie amerykańska (dodatkowo w dwóch dalszych badaniach Amerykanie stanowili jedną z kilku prób). Dodatkowo 6 prób było złożonych z osób z podobnego kręgu kulturowego (4 z Wielkiej Brytanii i 2 z Australii). Kolejne 5 prób pochodziło z Włoch i dwie z Polski. Z niezachodniego kręgu kulturowego pochodziły 4 próby chińskie (wszystkie z jednego artykułu [12]), jedna mieszana japońsko-amerykańska i jedna wielonarodowa [08].

Podobnie rzecz wygląda, jeśli chodzi o zróżnicowanie grupowe. W skład aż 57 (83 %) prób badanych wchodziło tylko studenci (dodatkowo w dwóch dalszych badaniach studenci byli jedną z dwóch grup badanych). W dwóch przypadkach badanych rekrutowano przez Internet, 3 próby badane to „populacja ogólna”, w 3 przypadkach to uczniowie szkoły podstawowej lub średniej. Poza tym w skład trzech prób (po jednej) wchodziły osoby cierpiące na zaburzenia psychiczne, osoby strajkujące i autorzy artykułów w prasie.

Podsumowując więc ten cykl badań, ogólnie wskazuje on na to, że entytatywność jest konstruktem występującym w ludzkim umyśle (abstrahując od tego, na ile występuje „realnie”) i wywiera różnoraki wpływ na poznanie społeczne. Jest skrajnie mało prawdopodobne, żeby tak duża ilość danych była fundamentalnie błędna.

Zarazem do słabości tego cyklu można zaliczyć to, że nie było wiele replikacji zbliżonych do wyjściowego badania. Entytatywność jest w większości badań przyjmowana jako pewnik. Choć ogółem badań jest wiele, to wiele jest też szczegółowych hipotez. Każdą z nich zwykle potwierdza jedno lub dwa badania, więc trudno uznać, żeby były szeroko przetestowane (niewykluczone, że hipotezy te są

dalej potwierdzone w artykułach cytujących drugiego stopnia; o ograniczeniu mojej metody będę jeszcze wspominał w podsumowaniu).

Niepokojące jest również to, że praktycznie wszystkie badania obalają hipotezę zerową – entytatywność na coś wywiera wpływ, lub jakieś czynniki na nią wpływają (istotnie statystycznie). Nawet, jeśli entytatywność jest czynnikiem mającym znaczący wpływ na poznanie społeczne, taki jednoznaczny wzór wyników wydaje się niemożliwy ze względu na problem mocy badań i ogólny problem ograniczeń metod badawczych w psychologii. Niewykluczone więc, że zaprezentowany w czasopiśmie zbiór badań jest tylko wybiórczą „próbą” z większej całości. Minimalne różnicowanie prób badanych (zarówno narodowościowo, jak i grupowo) nie wymaga chyba dalszego komentarza.

Wydaje się więc, że z jednej strony ten zbiór badań dostarcza nam niemałej liczby informacji o ciekawym fenomenie z obszaru poznania społecznego, a zarazem jest doskonałym świadectwem przenikających psychologię społeczną „kultury sukcesu” i „kultury nowatorstwa”.

6.2.8. Chiu, C. Y., Morris, M. W., Hong, Y. Y., Menon, T. (2000). Motivated cultural cognition: the impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. *Journal of Personality and Social Psychology*, 78(2), 247-259. (JPSP8)

W tym artykule udokumentowano ciekawy efekt wskazujący na występowanie złożonych interakcji między kulturą a osobowością. Okazuje się, że osoby mające wysokie natężenie pewnej zmiennej dyspozycyjnej (potrzeba domknięcia poznawczego) inaczej dokonują atrybucji zachowania (przejaw wpływu grupy lub sytuacji vs. dyspozycyjnie) w zależności od tego, jaka jest zasadnicza narracja takich wyjaśnień w ich kraju. I tak studenci amerykańscy o wysokim natężeniu potrzeby domknięcia poznawczego wyjaśniali zachowania bardziej dyspozycyjnie, a studenci chińscy jako przejaw wpływu grupy (sytuacyjnie). W artykule przedstawiono dwa badania (stosunkowo niedużo jak na JPSP). Jedną ze zmiennych niezależnych zawsze było pochodzenie badanych. W pierwszym badaniu poziom potrzeby domknięcia mierzono kwestionariuszowo (formalnie było to więc badanie

korelacyjne), a w drugim przeprowadzono eksperyment, w którym potrzebę domknięcia poznawczego niejako wymuszano poprzez krótki czas na odpowiedź.

Spośród 152 tekstów cytujących, tylko 6 spełniało założone przeze mnie kryteria (łącznie opisywały 10 badań). 2 artykuły można uznać za zbliżone replikacje konceptualne ([01] i [03]), przy czym w jednym [01] wprowadzono jeszcze inną manipulację mającą na celu wymuszenie domknięcia poznawczego (obciążenie pamięci roboczej), a w drugim [03] porównywano tylko studentów amerykańskich o wysokim i niskim poziomie potrzeby domknięcia poznawczego (ci pierwsi byli bardziej podatni na podstawowy błąd atrybucji).

Dwa badania ([04] i [06]) można uznać za generowanie empirycznie weryfikowalnych hipotez na podstawie wyjściowego efektu. W jednym [04] stwierdzono, że zachodzi interakcja między kulturą, natężeniem potrzeby domknięcia poznawczego a sposobem rozdzielania nagród i kar (choć badani nie rozdzielali ich, ale wyobrażali sobie pewien scenariusz i deklarowali, jak by zareagowali). W drugim [06] stwierdzono, że wysokie natężenie potrzeby domknięcia poznawczego utrudnia badanym negocjacje z członkami innej kultury.

Wreszcie dwa badania próbowały wyjaśnić wyjściowy efekt, albo poprzez to, że ludzie o wysokim natężeniu potrzeby domknięcia poznawczego dążą do pewności usankcjonowanej kulturowym konsensusem [02], albo poprzez to, że takie osoby wykazują większą gotowość do zaszczepienia swojemu dziecku wartości swojej kultury [05].

Można więc powiedzieć, że cykl badań charakteryzował się idealnym „wyważeniem” między trzema rodzajami badań: replikującymi efekt, będącymi konsekwencją teorii, będącymi próbami rozszerzenia rozumienia efektu. Tym niemniej w każdym wypadku mamy do czynienia tylko z dwoma artykułami, co nie jest liczbą bardzo wysoką.

Badania te cechują się za to dosyć częstym porównywaniem wyników osób z dwóch kultur, amerykańskiej i chińskiej (i tylko takich, nie uwzględniono żadnych innych porównań). Nie jest to jednak w tym kontekście jakaś wielka zaleta, gdyż istotą oryginalnego efektu było występowanie różnic międzykulturowych i trudno je badać w inny sposób niż rzeczywiście porównując członków dwóch kultur.

Zróznicowanie grupowe pozostawia wiele do życzenia: spośród 12 badań w 10 wykorzystywano tylko próby studenckie. W jednym przypadku [01:1] rekrutowano

osoby na lotniskach, a w drugim [05] badanymi były matki uczących się dzieci w szkole.

Badania mniej więcej w połowie miały charakter korelacyjny, a w połowie eksperymentalny. Zmienna zależna miała zawsze charakter samoopisowy, z wyjątkiem badania [06], w którym sprawdzano zachowania w grze ekonomicznej.

Ogółem, podsumowując, wydaje się, że cykl badań stanowi niezły prognostyk wskazujący na fakt występowania oryginalnego efektu. Ilość badań (10) jest większa niż w wielu analizowanych tutaj artykułach, choć w kategoriach absolutnych jest minimalna. Nie przeanalizowano dziesiątków sensownych moderatorów, analizowano tylko osoby z Chin i USA, nie testowano hipotez przy pomocy bardziej behawioralnych metod, nie przeprowadzono replikacji dokładnych itd.

Uwaga na marginesie: jeśli opisywane prawidłowości są rzeczywiście prawdziwe, jest to kolejny argument za tym, żeby prawidłowości psychologiczne każdorazowo testować w różnych kulturach (rdz. 2.5).

6.2.9. Dolinski, D. (2000). On inferring one's beliefs from one's attempt and consequences for subsequent compliance. *Journal of Personality and Social Psychology*, 78(2), 260-272. (JPSP9)

W badaniu tym na podstawowym poziomie wskazywano na efekt polegający na tym, że nieudana próba pomocy innej osobie będzie skutkować większą gotowością do udzielenia pomocy innej osobie w innym zadaniu. W pewnym sensie jest to więc wariant efektu „stopy w drzwi”, samo badanie jest jednak oparte na teorii samospostrzegania Bema i stanowi jej rozwinięcie.

Wszystkie 7 badań to eksperymenty terenowe. Pod tym względem jest to jedyny taki artykuł, zarówno wśród tekstów oryginalnych, jak i cytujących. Badanymi były osoby z ogólnej populacji rekrutowane na ulicach miast (badani nie wiedzieli, że uczestniczyli w eksperymencie). Pierwsze badanie weryfikuje efekt w czystej postaci, a następne sześć badań ma na celu wyeliminowanie alternatywnych wyjaśnień i poszerzenie jego rozumienia. We wszystkich badaniach zmienną zależną jest zachowanie (gotowość do udzielenia pomocy), jedynie w ostatnim zmienna zależna ma charakter samoopisowy (deklarowane posiadanie określonych cech charakterów). Grupy badane są stosunkowo duże (50 osób w kratce).

Pomimo tego, że badanie jest kwintesencją postulatów Baumeistera, Vohs i Fundera (2007), jest ono niestety stosunkowo rzadko cytowane (7 razy), a wśród tekstów je cytujących nie ma takiego, które można zakwalifikować jako replikację.

6.2.10. Aron, A., Norman, C. C., Aron, E. N., McKenna, C., Heyman, R. E. (2000). Couples' shared participation in novel and arousing activities and experienced relationship quality. *Journal of Personality and Social Psychology*, 78(2), 273-284. (JPSP10)

W tym badaniu wykazywano, że wspólne ekscytujące przeżycia sprzyjają zadowoleniu par romantycznych ze związku. Serię oryginalnych badań zaliczyłbym do modelowych. Przeprowadzono 5 badań, dwa pierwsze miały charakter ankietowy (siłą rzeczą korelacyjny), jednak próbę pozyskano z ogólnej populacji. Dwa następne badania miały charakter eksperymentalny, gdzie w sposób wysoce realistyczny manipulowano poziomem „ekscytacji” wspólnego przeżycia (badani wykonywali nudne lub nienudne ćwiczenia fizyczne), po czym badani określali samoopisowo poziom zadowolenia ze związku (jedna próba była studencka, ale już kolejna to małżeństwa z ogólnej populacji). Jednak w ostatnim eksperymencie badacze wyszli poza ograniczenia miar samoopisowych i nagrywali interakcje par romantycznych po manipulacji eksperymentalnej, a sędziowie kompetentni na podstawie odpowiedniego systemu kodowania określali np. poziom wykazywanej akceptacji lub wrogości.

Nietrudno wskazać możliwe próby replikacji np. podłużne badanie, w którym kontrolowano by poziom „ekscytujących” przeżyć (zarówno odczuwanych subiektywnie, jak i poprzez wskazywanie częstości obiektywnych wydarzeń np. wspólnych wakacji, wyjść do kina itd.), a zmienną zależną, poza oczywistym deklarowanym zadowoleniem ze związku, mógłby być np. wskaźnik rozwodów. Oczywiście, łatwo sobie też wyobrazić niezliczone możliwości eksperymentalnego przetestowania hipotez.

Oryginalne dane wydają się wyraźnie wskazywać na to, że zakładany efekt może występować. Dodatkowo badanie to jest najprawdopodobniej najbardziej „medialnym” ze wszystkich oryginalnych badań tutaj analizowanych. Istota odkrycia jest łatwo, intuicyjnie zrozumiała, a przy tym dotyczy obszaru interesującego dla

większości ludzi. Bardzo się więc zdziwiłem, gdy okazało się, że wśród niemałej liczby 105 cytujących tekstów, nie było żadnego (!), które można zaklasyfikować jako replikację (choćby konceptualną). Teksty cytujące w olbrzymiej większości dotyczyły przemocy i kłótni w związku, co przypomina znany fakt, że psychologia jest bardziej skoncentrowana na tym, co negatywne niż pozytywne (por. Krueger, Funder, 2004).

Temat badania jest na tyle ogólny, że prawdopodobnie w całej literaturze psychologicznej uda się znaleźć teksty, które można by zaklasyfikować jako replikację konceptualną. Hipoteza sama w sobie też nie jest jakoś „odkrywcza” i prawdopodobnie niejedynemu badacz niezależnie doszedłby do takich wniosków. Tym niemniej w obrębie zaproponowanej tu metodologii badania postępu wiedzy trzeba stwierdzić, że po bardzo efektownym początku rozumienie efektu nie posunęło się do przodu ani o krok.

6.2.11. Sidanius, J., Levin, S., Liu, J., Pratto, F. (2000). Social dominance orientation, anti - egalitarianism and the political psychology of gender: an extension and cross - cultural replication. *European Journal of Social Psychology*, 30(1), 41-67. (EJSP1)

W badaniu tym analizowano różnice międzypłciowe i międzygrupowe w natężeniu hipotetycznego konstruktu psychologicznego orientacji na społeczną dominację (ang. *social domination orientation*; SDO), który objawia się m.in. przekonaniem, że pewne grupy społeczne mają prawo dominować nad innymi i nie należy podejmować wysiłków, żeby zmniejszać różnice w sile politycznej lub ekonomicznej mniej i bardziej uprzywilejowanych.

Zasadnicza hipoteza badania mówi, że różnice międzypłciowe w SDO są wrodzone i względnie niezależne od uwarunkowań kulturowych, gdyż mają swoje źródło w historii naturalnej. Natomiast różnice międzygrupowe (np. między białymi i Afroamerykanami w USA) wynikają ze względnej pozycji grup w hierarchii społecznej. Im jest ona wyższa, tym wyższe jest natężenie SDO (w tym celu, oprócz wyliczenia różnic międzypłciowych, badacze skorelowali różnice w natężeniu SDO między różnymi grupami etnicznymi w USA i Izraelu, a ich pozycją w hierarchii społecznej; względną pozycję tych grup w hierarchii wyznaczono na podstawie oceny innych badanych). Badanie oryginalne wyróżnia się zdecydowanie na tle

większości analizowanych tu badań dlatego, że przebadano prawie 7 tys. osób pochodzących z 6 krajów.

Hipotezy te były dalej weryfikowane w przynajmniej 20 artykułach (32 badaniach), przy czym w olbrzymiej większości były to weryfikacje hipotezy o różnicach międzypłciowych, natomiast hipoteza o różnicach międzygrupowych była weryfikowana stosunkowo rzadko (w [03], [05] i [07]).

Prawie wszystkie badania miały charakter kwestionariuszowy (co sprzyjało dużym grupom; we wszystkich badaniach wzięło łącznie udział ponad 36 tys. osób⁸¹ i jest to rekord wśród omawianych tu badań). Olbrzymia większość miała też charakter korelacyjny, z wyjątkiem trzech. W jednym [05] studentom mówiono, że ich uniwersytet otrzymuje dużo lub mało środków ze specjalnego funduszu stanowego i tym samym manipulowano ich poczuciem uprzywilejowania. Była to udana manipulacja, studenci „uprzywilejowani” wykazywali następnie istotnie wyższe wyniki SDO. W kolejnym badaniu [19] zamysł był podobny, ale inaczej manipulowano „statusem” (badanym dawano informacje zwrotne o bardzo wysokim lub względnie niskim wyniku w teście poznawczym). W ostatnim badaniu eksperymentalnym [08] podkreślano koncept „płci” lub „etniczności” poprzez to, że badani wypełniali kwestionariusz dotyczący relacji między płciami lub (alternatywnie) relacji między grupami etnicznymi. Jeśli podkreślano koncept „etniczności” zanikały różnice płciowe, a jeśli podkreślano koncept „płci” zanikały różnice między grupami etnicznymi.

Jednak, tak jak wspominałem, olbrzymia większość badań miała charakter kwestionariuszowy. Na pozytywne podkreślenie zasługuje fakt, że przebadano szerokie spektrum narodowości m.in. Nowozelandczyków, Izraelczyków, Chińczyków, Niemców, czy Szwedów (łącznie obywateli 11 krajów). Dodatkowo w 9 próbach (na 33) badanymi nie byli studenci, co na pierwszy rzut oka nie wygląda imponująco, ale jak się okazuje, jest wynikiem naprawdę niezłym na tle innych cykli badań. Jako ciekawostkę warto zauważyć, że w jednym z tekstów badano różnice płciowe u osób grających w sieciową grę „World of Warcraft” [12] (obok uczestników strajku JPSP7_26 jest to najoryginalniejsza próba w 472 badaniach).

⁸¹ Przy czym duża część tej próby pochodziła z jednego dużego badania internetowego.

Kilka badań zaliczyłem w tym cyklu do replikacji dokładnych, jednak czasami można się nad tą dokładnością zastanawiać np. stosowano bardzo podobne kwestionariusze w połączeniu z dodatkowymi miarami.

Jeśli chodzi o to, czy hipoteza została potwierdzona, daje się zauważyć ciekawe zjawiska, mówiące coś o przenikaniu się nauki i wartości ideologicznych. Autorzy oryginalnego badania w kilku następnych dużych badaniach ([01], [09]) potwierdzili wyjściową hipotezę (i wsparło ich kilku niezależnych badaczy). Jednak szereg innych badaczy przedstawił badania, w których zasadniczo nie kwestionowali samego faktu występowania różnic międzypłciowych, ale wskazywali, że efekt ten jest rodzajem zależności pozornej i zanika, gdy kontrolować dodatkowe zmienne. W szczególności stwierdzono, że różnice mają zanikać, gdy kontrolować:

- identyfikację z własną płcią i rolami płciowymi ([06], [14]),
- poziom seksizmu [15],
- atrybucyjną złożoność [16],
- tendencję do zwiększania/zmniejszania hierarchii [19],
- wartości [11].

Wskazywano również, że różnice płciowe nie występują u studentów psychologii [02], w kulturze szwedzkiej [19] i niektórych organizacjach [13].

Nie chcę tutaj szerzej omawiać tego sporu teoretycznego, trudno jednak nie było czasami odnieść wrażenia, że niektórzy badacze postawili sobie za cel zakwestionowanie biologicznych uwarunkowań różnic międzypłciowych, szczególnie w przypadku artykułów w czasopismach znanych z niechęci do takiego myślenia (jak *Sex Roles* czy *Psychology of Women Quarterly*). Wydaje się też, że statystycznie jest niemożliwe, żeby wszystkie w/w zmienne faktycznie niwelowały różnice. Muszą raczej być wyrazem jakiegoś wspólnego podstawowego konstruktu leżącego u ich podłoża lub też sytuacja konieczności określenia swojej identyfikacji z płcią wywołuje zmianę w odpowiadaniu na pytania skal SDO.

W serii badań nie użyto żadnych zmiennych, które miałyby charakter behawioralny i niewątpliwie przyszłe uzupełnienie badań o takie miary byłoby wartościowe (np. poprzez aranżowanie gier ekonomicznych, w których badani mogą wykazać gotowość do rezygnacji z przywilejów na rzecz grup dyskryminowanych).

Ogółem wydaje się, że choć spór o źródła różnic międzypłciowych i międzygrupowych nie został ostatecznie rozstrzygnięty w omawianej tu serii badań, to jednak serię tę samą w sobie należy ocenić dosyć pozytywnie. W badaniach tych uczestniczyły bardzo duże próby z większej liczby krajów niż „standardowe” USA i kraje europejskie, (choć i tak całe połacie świata, jak Ameryka Południowa, Afryka i większość Azji były pominięte). Zaproponowano kilka alternatywnych wyjaśnień efektu, w dodatku pokazano, że może on być niwelowany przez pomysłowe manipulacje eksperymentalne. Można zauważyć, że temat budzi zainteresowanie społeczności badawczej. Jedyne, czego można żałować, to że w badaniach tych bada się deklarowane preferencje, a nie rzeczywiste zachowania.

6.2.12. Reid, S. A., Ng, S. H. (2000). Conversation as a resource for influence: evidence for prototypical arguments and social identification processes. *European Journal of Social Psychology*, 30(1), 83-100. (EJSP2)

W tym artykule omówiono ciekawe badanie dotyczące wpływu w grupie. Badani spotykali się w 8-osobowych grupach, które dyskutowały kontrowersyjną kwestię (konkretnie dopuszczalność kary śmierci). Okazało się, że osoby, które więcej mówią, częściej zabierają głos i częściej przerywają innym, są postrzegane jako bardziej wpływowe, zarówno przez oponentów, jak i zwolenników dyskutowanej kwestii.

Rozmowy toczyły się 30 minut, tworzono z ich zapisu dokładny stenogram, który poddawano szczegółowej lingwistycznej analizie. Jest to zapewne bardzo pracochłonna procedura, więc nie dziwi specjalnie, że przebadano tylko 3 grupy (łącznie 24 osoby). Wpływowość osób w grupie mierzono rzecz jasna poprzez oceny uczestników badania.

Badanie to niewątpliwie wyróżnia się na tle olbrzymiej większości badań dużym realizmem. Choć zmienna zależna ma naturę wysoce samoopisową, to jednak samo badanie zaklasyfikowałbym jako wysoce „behawioralne” i, jak sądzę, realizujące postulat Rozina o większym znaczeniu obserwacji. Wada, która jest dla mnie najwyraźniejsza, to brak niezależnej oceny wpływowości osób. Mogę sobie wyobrazić, że niezależni obserwatorzy mogliby oglądać dyskusje i dokonywać samodzielnych ocen. Przy już obecnej pracochłonności badania, ten dodatkowy

wysiłek nie byłby duży. Zastanawia też fakt, że badacze zaprezentowali wyniki tylko jednej zmiennej zależnej (wpływowość członków grupy), dodam, że statystycznie istotnej. Trudno uwierzyć, że przy tak pracochłonnej procedurze badacze nie sprawdzali szeregu innych zmiennych.

Niestety, pomimo tych zalet badanie to nie spotkało się z jakimś szerszym zainteresowaniem społeczności badaczy. W interesującym mnie okresie było cytowane tylko 14 razy. Jedno z tych cytowań to badanie o bardzo podobnej procedurze przeprowadzone przez ten sam zespół. Więcej replikacji lub badań wyjaśniających efekt nie zidentyfikowałem.

Ogółem więc, badanie to wysoce wyróżnia się na tle innych badań tutaj omawianych. Niestety, nie spotkało się z szerszym zainteresowaniem w społeczności badawczej (co być może jest znamienne) i ilość wiedzy o wyjściowym efekcie jest stosunkowo nieduża.

6.2.13. Thill, E. E., Cury, F. (2000). Learning to play golf under different goal conditions: their effects on irrelevant thoughts and on subsequent control strategies. *European Journal of Social Psychology*, 30(1), 101-122. (EJSP3)

W tym tekście przedstawiono badanie, w którym sprawdzano skuteczność różnych technik doskonalenia umiejętności (konkretnie tutaj: gry w golfa) i ich wpływ na mierzone samoopisowo m.in. poziom lęku i strategie kontroli.

Badani byli przydzieleni do jednej z 4 grup eksperymentalnych. Przykładowo w jednej grupie badani wielokrotnie powtarzali podstawowe czynności, w innej rywalizowali bezpośrednio z innym badanym, w jeszcze innej przyznawano im na bieżąco punkty w zależności od jakości uderzenia itp. Zaletą tego badania jest jego wysoka realistyczność i prawdopodobnie dużo większy wysiłek logistyczny niż w typowym badaniu psychologii społecznej. W grupach eksperymentalnych było średnio 25 osób, więcej niż w wielu badaniach eksperymentalnych o dużo prostszej procedurze. Badanymi nie byli studenci pierwszego roku psychologii, ale osoby rzeczywiście uczące grać się w golfa.

Po sesji nauki badani podchodzili do kilku kwestionariuszy mierzących zmienne zależne. Badacze przedstawili szereg wyników, praktycznie wszystkie, które omówili, były istotne statystycznie. Trudno nie mieć poczucia, że badacze nie przedstawili

podstawowej zmiennej, a mianowicie zmiany poziomu gry, tym bardziej że golf jest taką grą, w której poziom ten można dosyć łatwo zoperacjonalizować. Trudno uwierzyć, że badacze wysilali się z manipulacją eksperymentalną, a nie zadali sobie odrobiny trudu, żeby sprawdzić tą zmienną. Niewykluczone więc, że artykuł jest po prostu wyborem zmiennych działających. Statystycznie jest skrajnie mało prawdopodobne, żeby przedstawione wyniki (praktycznie wszystkie istotne) były jedynymi danymi. Przy tym z tytułu i abstraktu wynika, że badanie było poświęcone mierzeniu tylko zmiennych, które potem okazały się różnić pod wpływem manipulacji (i/lub płci).

Badanie to wyróżnia się na tle innych tu przedstawionych wysokim poziomem realizmu, przynajmniej na poziomie manipulacji eksperymentalnej. Gdy przyszło do pomiaru zmiennych zależnych badacze dziwnym trafem zrezygnowali z najciekawszej zmiennej i ograniczyli się do samoopisów.

Badanie to było stosunkowo rzadko cytowane (17 razy). Wśród tekstów cytujących nie znalazłem żadnych badań, które byłyby replikacją czy rozwinięciem głównych idei.

6.2.14. Galinsky, A. D., Stone, J., Cooper, J. (2000). The reinstatement of dissonance and psychological discomfort following failed affirmations. *European Journal of Social Psychology*, 30(1), 123-147. (EJSP4)

Badanie opisuje efekt polegający na tym, że afirmacja ważnych dla osoby wartości może zredukować dysonans poznawczy. Badani wypowiadali się w eseju na niepopularny temat (zakazanie lubianej przez studentów imprezy sportowej), manipulowano stopniem swobody w wyborze takiego tematu (podobnie jak w klasycznych eksperymentach z obszaru dysonansu poznawczego). Manipulacja eksperymentalna polegała też na tym, że badani rozwiązywali kwestionariusz wartości. Udzielano informacji zwrotnej wskazującej na to, które wartości są dla badanego ważne (lub w drugiej grupie udzielano informacji fałszywej wskazując, że rzeczywiście ważne dla badanych wartości nie są dla nich ważne). Tego rodzaju afirmacja w pierwszym badaniu redukowała dysonans w postaci psychologicznego dyskomfortu (ale nie wpływała na zmianę postawy), a w drugim badaniu powodowała

zmianę postawy (w kierunku tej, która jest bliższa większości), ale nie wpływała na komfort.

Te niejednoznaczne wyniki mogą brać się z tego, że badanie ma bardzo niską moc. W pierwszym eksperymencie w grupie eksperymentalnej uczestniczyło średnio 10 osób, w drugim 12. W taki przypadku nawet silny efekt ($d=0,8$) ma nieduże szanse na bycie wykrytym (odpowiednio: 43 % i 50 %).

Hipoteza sama w sobie wydaje się być interesująca poznawczo i, na szczęście, ukazało się trochę prac próbujących zreplikować konceptualnie to badanie, lub wygenerować niebanalne przewidywania na ich podstawie (8 artykułów i 17 badań). Niektóre z nich są w miarę zbliżone do oryginalnej procedury (np. [01], [02], [03]). W jednym z tekstów [04] połączono tę procedurę z paradygmatem teorii opanowywania tworgi. Okazało się, że pod wpływem manipulacji polegającej na wzbudzaniu poczucia własnej śmiertelności oraz afirmacji ważnych wartości badani następnie bardziej przychylnie oceniają esej krytycznie wypowiadający się o ich kraju. Zbliżone efekty uzyskano w kolejnym badaniu [05]. Samoafirmacja powodowała, że badani Amerykanie byli bardziej skłonni zaakceptować raport krytycznie wypowiadający się o polityce amerykańskiej (choć, efekt ten działał tylko w interakcji z dodatkową manipulacją: poleceniem, żeby „myśleć racjonalnie”). W tym samym artykule samoafirmacja powodowała, że badani zwolennicy i przeciwnicy aborcji nie oceniali tak bardzo krytycznie drugiej strony sporu, jak bez tego oddziaływania.

Zróznicowanie tematyczne w dalszych badaniach jest również duże:

- badani, którzy mają okazję dokonać samoafirmacji ważnych dla siebie wartości, ulegają następnie w mniejszym stopniu efektowi "za dużo zainwestowałem, żeby się wycofać" (polegającemu na wkładaniu zasobów w inwestycje, w które się dużo zainwestowało, choć najrozsądniej byłoby się wycofać) [06];
- osoby, które przeszły procedurę samoafirmacji wykazywały wyższą gotowość do profilaktycznych badań [07];
- przymowanie standardów osobistych zwiększało poczucie winy za krzywdy doznawane przez inne grupy spowodowane działaniem grupy własnej. Ten efekt przetestowano na Izraelczykach i bośniackich Serbach [08].

Mamy tu więc do czynienia z ciekawym zróznicowaniem tematycznym, ciągle jednak związanym z testowaniem mniej więcej tej samej hipotezy (oczywiście, czasami

trzeba przyjmować dodatkowe założenia teoretyczne np. że sytuacja, gdy grupa własna popełnia zbrodnie wojenne wywołuje dysonans; nie są to jednak założenia, które kłóciłyby się z elementarną logiką). Pod tym względem, ten cykl badań wydaje mi się wyjątkowo wart naśladowania. Niestety, podobnego zróżnicowania nie wykazano, jeśli chodzi o narodowość (wszystkie próby, z wyjątkiem ostatniego badania, pochodziły z USA), oraz zróżnicowanie grupowe (poza jednym badaniem, wszystkie grupy były studenckie). Powodem do niepokoju są też małe kratki (10 prób na 19 ma poniżej 20). Wszystkie badania potwierdzały wyjściową hipotezę, co jest statystycznie patrząc mało prawdopodobne.

Ogólnie więc oceniam ten cykl badań umiarkowanie pozytywnie. Z jednej strony mamy do czynienia z podstawowymi „grzechami”, takimi jak niezróżnicowanie grupowe i narodowe badanych oraz małe kratki, z drugiej jednak hipotezę rozwinięto i przetestowano w kilka niebanalnych, mocno odbiegających od wyjściowej procedury, sposobów.

6.2.15. Sagiv, L., Schwartz, S. H. (2000). Value priorities and subjective well-being: direct relations and congruity effects. *European Journal of Social Psychology*, 30(2), 177-198. (EJSP5)

W kolejnym artykule inny zespół badaczy pozostał przy temacie wartości, tym razem badano związek wyznawanych wartości z poczuciem subiektywnego dobrostanu. Mierzone tu wartości wywodziły się z koncepcji 10 wartości wg Schwartz (m.in. tradycja, stymulacja, bezpieczeństwo, autonomia, osiągnięcia, władza itd.). W 4 niemieckich i dwóch izraelskich próbach stwierdzono występowanie istotnych zależności (szczegółowe omówienie wyników w załączniku). Warto jednak zaznaczyć, że olbrzymia większość zaobserwowanych korelacji była w przedziale $0,1 < r < 0,2$, bardzo sporadycznie $r > 0,2$ i w żadnym wypadku nie więcej niż 0,3. Badanie miało oczywiście charakter kwestionariuszowy i korelacyjny.

Drugim odkryciem, które zaobserwowano, jest to, że zgodność między wartościami wyznawanymi przez grupę a własnymi jest skorelowana z poczuciem dobrostanu. Porównywano tu konkretnie studentów zarządzania i psychologii. I tak przykładowo u tych pierwszych zaobserwowano, że im wyższa dla badanych wartość władzy, tym

wyższy subiektywny dobrostan. Natomiast u studentów psychologii efekt był odwrotny.

O ile w pierwszym badaniu przebadano niemałą grupę osób (ponad 1200), o tyle w drugim liczba ta była bardzo skromna (82), jak na badanie kwestionariuszowo-korelacyjne (można powiedzieć, że zaledwie po dwie grupy ćwiczeniowe z każdego kierunku).

Tezy te były następnie weryfikowane w 15 badaniach (omówionych w 14 artykułach). Wszystkie miały charakter samoopisowy i korelacyjny. Nie będę tutaj omawiał szczegółowo wzoru wyników w każdym badaniu (odsyłam do załącznika 2). Czasami potwierdzano oryginalne wyniki, czasami uzyskiwano nieco odmienne. Natomiast nadmienię, że prawie zawsze efekty miały niewielką siłę, porównywalną do tych z wyjściowego badania.

W 15 badaniach w 8 przypadkach wykorzystywano próby studenckie, a w 7 inne, najczęściej z populacji ogólnej (w jednym badaniu przebadano tylko policjantów, a w innym emerytów). Jeśli chodzi o kraje, występowało tu ponadprzeciętnie duże zróżnicowanie (choć i tak niezbyt duże patrząc bardziej obiektywnie). Poza osobami z USA, UK, Izraela i kilku krajów europejskich przebadano osoby z Turcji, Iranu, Chin i Meksyku. Próby były umiarkowanie duże jak na tego typu badanie. W jednym z badań [12] wykorzystano dane z ogólnoeuropejskiego badania i liczba osób badanych była bardzo wysoka (ponad 33 tys.). Poza tym jednym odstającym przypadkiem średnio przebadano nieco ponad 300 osób.

Najczęściej testowano hipotezę o związkach niektórych wartości z dobrostanem. Druga hipoteza o związkach wartości grupowych z dobrostanem była testowana dużo rzadziej ([04], [07], [08]), przy czym w dwóch ([04], [07]) poziom rozbieżności między swoimi wartościami a społecznymi był określany przez samych badanych. W badaniu [08] obserwacjami były kraje, a korelowano poziom rozwoju z poziomem korelacji między sekularyzmem a dobrostanem.

Ogółem więc cykl badań cechował się przyzwoitym poziomem zróżnicowania międzynarodowego i grupowego. Nie jestem ekspertem od badania ludzkich wartości i dobrostanu, jednak w moim przekonaniu ogólna wartość poznawcza takich badań jest niska. Uzyskuje się zróżnicowany wzór wyników i nie wiadomo, czy należy go przypisać ograniczeniom metody, czy systematycznym różnicom między badanymi

grupami. Efekty są poza tym słabe i bardzo słabe, ale załapują się na istotność dzięki temu, że bada się dosyć duże grupy (przy czym, jeśli badania te polegają na rozdaniu pliku kwestionariuszy swoim grupom studenckim, to są to badania logistycznie relatywnie mało pracochłonne, więc nie jest to w tym kontekście jakaś wielka zaleta). Ze względu na samoopisowy charakter badań te niewielkie korelacje mogą odzwierciedlać jakieś subtelne artefakty, a nie obiektywną „rzeczywistość”. Można też zauważyć dużą mieszaninę konstruktów (subiektywny dobrostan, pozytywny/negatywny afekt, dobrostan indywidualny, dobrostan grupowy, zadowolenie z życia), które dodatkowo zaciemniają obraz. Natomiast najciekawsza hipoteza (rozbieżność między wartościami własnymi i grupowymi a dobrostan) była przetestowana stosunkowo rzadko.

6.2.16. Todorov, A., Lalljee, M., Hirst, W. (2000). Communication context, explanation, and social judgment. *European Journal of Social Psychology*, 30(2), 199-209. (EJSP6)

W badaniu tym stwierdzono, że jeśli ludzie mają wyjaśniać innej osobie zachowanie innego człowieka, to w zależności od nastawienia adresata komunikatu robią to bardziej sytuacyjnie (jeśli adresat ma nastawienie surowe), lub dyspozycyjnie (jeśli adresat ma nastawienie łagodne). Oczywiście, sam fakt tego, że ludzie dostosowują treść komunikatu do adresata jest banalny, jednak, jak wynika z treści artykułu, odkrycie o wyjaśnianiu sytuacyjnym lub dyspozycyjnym w zależności od nastawienia jest nowe.

Tak czy owak, abstrahując od hipotez, samo badanie nie należy do zbyt wyszukanych. Badani wcielali się w rolę „pedagoga”, czytali opis zachowania młodego człowieka i mieli następnie przekazać wnioski do osoby, o której wiedzieli, że jest łagodna lub surowa. Można sobie łatwo wyobrazić nieco bardziej realistyczne warunki (np. obserwowanie czyichś rzeczywistych zachowań lub nagranych na wideo) lub takie zaaranżowanie sytuacji eksperymentalnej, żeby badany miał poczucie, że faktycznie przekazuje eksperymentatorowi dane o czymś rzeczywistym zachowaniu (np. drugiego badanego), i te informacje mogą mieć wpływ na jakieś istotne kwestie (np. jak eksperymentator oceni drugiego badanego). Nic takiego jednak nie miało miejsca.

Stwierdzono też w drugim badaniu, że gdy badacz informował w sposób natarczywy o nastawieniu adresata, wówczas efekt był słabszy. Dodatkowo w przypadku tego badania mamy do czynienia z małą liczbą 10 osób w kratce.

Wśród nielicznych tekstów cytujących (6 cytowań) jest jedno przeprowadzone przez tego samego autora [01], w którym testowano zbliżoną hipotezę i dodatkowo potwierdzono efekt natarczywości/subtelności wskazówek (przy równie małych liczebnościach grup, jak w drugim badaniu).

Ogólnie można powiedzieć, że ilość wiedzy na temat tego efektu jest minimalna i badanie co najwyżej stanowi sensowny wstęp do szerszego projektu badawczego. Prawdopodobnie jednak on nie nastąpi, gdyż opisywany efekt nie wydaje się mieć jakichś ważnych implikacji teoretycznych czy praktycznych.

6.2.17. Harris, P., Middleton, W., Joiner, R. (2000). The typical student as an in-group member: eliminating optimistic bias by reducing social distance. *European Journal of Social Psychology*, 30(2), 235-253. (EJSP7)

Badanie to jest wariantem tzw. złudzenia optymisty (ang. *optimistic bias*), który polega na tym, że ludzie przeszacowują prawdopodobieństwo przyjemnych rzeczy, które mogą ich w życiu spotkać (np. wygrana na loterii), a nie doszacowują prawdopodobieństwa rzeczy nieprzyjemnych (np. nowotwór). Tutaj stwierdzono, że efekt ten można zmniejszyć, jeśli badanym kazać najpierw oceniać szanse pozytywnych i negatywnych wydarzeń członków grupy obcej (studenci innego uniwersytetu), a następnie dopiero członków grupy własnej. Szanse wydarzeń pozytywnych i negatywnych dla samego siebie badani oceniali na samym początku, więc nie była to zmienna zależna, ale punkt odniesienia do następnych ocen.

Natura zmiennej zależnej jest oczywiście samoopisowa, ale nie wydaje się, żeby w przypadku badań nad złudzeniem optymisty miało to aż takie wielkie znaczenie. Grupą badaną byli studenci na brytyjskim uniwersytecie.

Badanie to zasadniczo nie było bezpośrednio replikowane, choć jedno [01] badanie można uznać za ciekawą replikację conceptualną. Nie stwierdzono tu efektu kolejności ocen, ale za to stwierdzono, że jeśli oceny grupy własnej i obcej są wykonywane w tym samym oknie komputera, to efekt ulega zmniejszeniu w sytuacji, gdy są one wykonywane na niezależnych oknach następujących po sobie.

Dodatkowo, manipulacja ta powodowała, że badani deklarowali większe „zmartwienie” możliwymi problemami zdrowotnymi.

Poza tym efekt zanotowany w tym tekście nie był w żaden sposób replikowany. Trudno mi odnieść się do jego znaczenia teoretycznego, ale wydaje się, że redukcja złudzenia optymizmu jest pod względem praktycznym ważną rzeczą (m.in. może zachęcać ludzi do większej profilaktyki). Natomiast badanie oryginalne ze względu na mało realistyczną manipulację i ograniczoną próbę może być uznane tylko za wstępny prognostyk, że efekt występuje.

6.2.18. Giebels, E., De Dreu, C. W., Van de Vliert, E. (2000). Interdependence in negotiation: effects of exit options and social motive on distributive and integrative negotiation. *European Journal of Social Psychology*, 30(2), 255-272. (EJSP8)

Badacze stwierdzili, że negocjatorzy osiągają najgorsze wyniki, jeśli jedna strona ma możliwość wyjścia z negocjacji i rozpoczęcia nowych negocjacji (z kimś innym). Natomiast, jeśli obydwie strony mają możliwość wyjścia lub żadna strona nie ma takiej możliwości, wówczas efekty negocjacji są lepsze (nie ma różnicy między tymi dwoma warunkami). Dodatkowo ustalono, że efekt ten występował tylko, gdy wzbudzano "motyw egoistyczny" (przy „motywie prospołecznym” tj. informacji, żeby uwzględnić interesy drugiej strony, nie stwierdzono różnic między grupami). (Takie dodatkowe, dosyć banalne moderatory, wzbudzają we mnie od razu podejrzenie o występowanie zjawiska psychologii fałszywie pozytywnej).

Badanie miało formę dosyć realistycznego eksperymentu. Zostało następnie zreplikowane konceptualnie przez 2 niezależne zespoły (3 badania), jedno [01] pod względem procedury zbliżone do oryginalnego badania, drugie [02] uznałem też za replikację konceptualną, chociaż procedura jest odległa. Wyniki były dosyć podobne do tych oryginalnym badaniu, choć nie było to jednoznaczne (odsyłam do załącznika po szczegóły).

Na tym niestety cykl badań się kończy, jest on więc stosunkowo skromny. Każdorazowo badano studentów, wyjątkowo nie byli to ani razu studenci amerykańscy (ale japońscy, holenderscy i włoscy). To zróżnicowanie nie jest zabiegiem celowym, ale wynika po prostu z pochodzenia badaczy.

Ze względu na niewielką ilość badań siłą rzeczy wartość poznawcza tego cyklu jest niewielka. Niewątpliwą zaletą jest natomiast dosyć wysoki poziom realizmu.

6.2.19. Sanna, L. J., Turley-Ames, K. J. (2000). Counterfactual intensity. *European Journal of Social Psychology*, 30(2), 273-296. (EJSP9)

W artykule tym testowano wpływ myśli kontrfaktycznych (np. „gdybym mniej chodził na imprezy, wtedy zaliczyłbym semestr”) na poczucie przygotowania do przyszłych podobnych zdarzeń, oraz nastroj. Stwierdzono, że myśli o gorszym alternatywnym biegu wypadków poprawiają nastrój, ale zmniejszają poczucie przygotowania, a myśli o lepszym biegu wypadków obniżają nastrój, ale zwiększają poczucie przygotowania.

W wyjściowym badaniu przeprowadzono trzy eksperymenty. W pierwszym badaniu prowadzili dzienniczki, w których mieli zapisywać myśli kontrfaktyczne do bieżących zdarzeń (o lepszym lub gorszym biegu) oraz mierzono w/w zmienne zależne. W drugim badaniu badani otrzymywali (pozytywne lub negatywne) informacje zwrotne o poziomie wykonania w teście mierzącym kreatywność. Sprawdzano, z którymi myślami kontrfaktycznymi się utożsamiali (korelowano ten wynik następnie ze zmiennymi mierzącymi poziom przygotowania i nastroju). W trzecim eksperymencie zastosowano podobną procedurę, ale dodatkowo indukowano pozytywny lub negatywny nastrój (poprzez projekcje odpowiedniego filmu).

Nadmienię przy okazji, że pierwszym autorem badania jest Lawrence Sanna, który kilka miesięcy po aferze Stapela, również zakończył karierę akademicką na kanwie podejrzeń o upiększanie danych (Simonsohn, 2013; rdz. 1.1). Samo w sobie to nie dyskwalifikuje to badania, warto jednak odnotować, potencjalnie sprzyjające psychologii fałszywie pozytywnej, małe kratki w badaniach 1 i 3 (odpowiednio 13 i 19 osób).

Badanie to nie było bardzo często cytowane (22 razy), jednak wśród cytowań można znaleźć relatywnie dużo replikacji conceptualnych (6 artykułów z 13 badaniami), z czego 7 badań (w 3 artykułach) jest również autorstwa Sanny. Z ciekawszych badań, warto wspomnieć następujące:

- sprawdzano myśli kontrfaktyczne u studentów 3 tygodnie przed egzaminem i po egzaminie [01:1] oraz ich wpływ na poziom przygotowania i nastrój. Mimo

interesującej procedury nie zapytano o oceny z egzaminu (byłoby interesującym poznać, czy myśli kontrfaktyczne wywierają jakiś wpływ na rzeczywiste przygotowanie objawiające się wynikiem egzaminu);

- z kolei w badaniu [03:2] badano studentów, którzy właśnie dowiedzieli się, że dostali złą lub dobrą ocenę z egzaminu.

Inne badania można zaliczyć do replikacji konceptualnych, przy czym żadnej z tych replikacji nie można uznać za dokładną.

Warto nadmienić, że aż 8 eksperymentów (spośród 15) miało ekstremalnie małe wielkości kratek (mniej niż 15 osób). Przy tak niskiej mocy (i równoczesnym potwierdzaniu hipotez) musi to zmuszać do zadania pytania o możliwe wybiórcze publikowanie lub podkręcanie wyników. Również wielką wadą jest to, że wszystkie próby badane składały się ze studentów (w jednym przypadku ze studentów MBA) oraz mieszkańców USA.

Ogółem cykl badań jest interesujący poznawczo, występuje pewne zróżnicowanie stosowanych procedur badawczych, mamy zarówno badania eksperymentalne, jak i dane bardziej opisowe. Wydaje się jednak, że ze względu na niską moc, oraz wysoce jednorodną próbę w całym cyklu, potrzeba znacznie więcej badań, żeby uznać efekt za wysoce pewny, o testowaniu licznych szczegółowych hipotez nie wspominając.

6.2.20. Roccas, S., Horenczyk, G., Schwartz, S. H. (2000). Acculturation discrepancies and well - being: the moderating role of conformity. *European Journal of Social Psychology*, 30(3), 323-334. (EJSP10)

Stwierdzono tutaj, że u imigrantów rozbieżność pomiędzy chęcią asymilacji a spostrzeganym oczekiwaniem asymilacji ze strony miejscowej ludności jest negatywnie skorelowana z poczuciem dobrostanu. Jednak efekt ten dotyczy tylko osób, które cenią konformizm jako wartość w życiu.

Badanie to miało charakter wyłącznie kwestionariuszowy, zostało przeprowadzone na izraelskich studentach, ale tylko takich, którzy byli imigrantami z Rosji. Wielkość próby jak na badanie korelacyjne jest stosunkowo niewysoka (100 osób), choć można przypuszczać, że nie jest łatwo pozyskać dużo większą próbę osób o takiej

charakterystyce. Poza tym badanie miało wyłącznie charakter samoopisowy, więc niesie ze sobą wszystkie wady tego rodzaju pomiarów.

Wśród 42 cytujących tekstów znalazłem jeden, który można uznać za próbę wyjaśnienia tego zjawiska. W eksperymencie przedstawiano opisy imigrantów i m.in. manipulowano chęcią ich integracji z nowym społeczeństwem. Badani oceniali niżej imigrantów, którzy nie chcieli się integrować.

Poza tym w tekstach cytujących nie można wyodrębnić tekstu, w którym w bezpośredni sposób nawiązywano by do tych hipotez. Wartość poznawczą oryginalnego badania uważam za niską, przede wszystkim ze względu na jej samoopisowy charakter. Poza tym efekt działa tylko w interakcji z dodatkowymi moderatorami, które z jednej strony są banalne (cenienie konformizmu w życiu), z drugiej mogą wskazywać na występowanie artefaktów (np. badani, którzy zaznaczają, że cenią konformizm w życiu oraz występuje rozbieżność między ich chęcią integracji a presją społeczną, mogą niejako „przy okazji” uświadomić sobie, że wywołuje to w nich negatywne odczucia).

6.2.21. Wyer, N. A., Sherman, J. W., Stroessner, S. J. (2000). The roles of motivation and ability in controlling the consequences of stereotype suppression. *Personality and Social Psychology Bulletin*, 26(1), 13-25. (PSPB1)

Artykuł opisuje efekt polegający na tym, że powstrzymywanie się od myślenia stereotypowego zwiększa następnie ocenianie niejednoznacznych zachowań w zgodzie z jakimś stereotypem.

Jednak, jak to nierzadko bywa w analizowanych tutaj artykułach, efekt ten „działa” tylko w interakcji z dodatkowymi moderatorami. Po pierwsze zwiększa go obciążenie poznawcze. Po drugie zachodzi tylko, jeśli osoba oceniana jest nieopisana pod względem kategorii (np. rasy), której dane stereotypy dotyczą.

Pod innymi względami badanie też jest dosyć „typowe”. W uproszczeniu wyglądało mniej więcej tak: badani opisywali osobę na podstawie dostarczonego zdjęcia (osoba ta była Azjatą lub Afroamerykaninem), które miało aktywizować określone stereotypy (np. agresywności w przypadku oglądania Afroamerykanina). Następnie, w rzekomo

innym badaniu, osoby badane oceniały człowieka opisanego w sposób niejednoznaczny (np. w jego zachowaniu są elementy agresywności, ale też pasywności). Główną zmienną zależną była miara „stereotypowości” opisu (przy czym w zgodzie z wcześniej prezentowaną osobą, a nie osobą, którą oceniali później).

Efekty te były relatywnie często replikowane konceptualnie (9 artykułów z 11 badaniami). Niestety, większość tych replikacji jest w ogólnej procedurze bardzo zbliżona do badania oryginalnego, tzn. badanym prezentowano bodźce w postaci zdjęć lub opisów (które miały aktywizować określone stereotypy), udzielano instrukcji (lub nie) o powstrzymaniu się od myślenia stereotypowego, następnie badani oceniali niejednoznaczne zachowania (lub po prostu osoby z określonych grup) i była to miara stereotypizowania. Niestety, olbrzymia większość tych badań jest mało inspirująca. Zwykle w badaniach sprawdzano też dodatkowe moderatory (np. poziom „władzy”, poziom lęku społecznego itd.). W niektórych badaniach ([01] i [03]) jako zmiennej zależnej użyto dostępności poznawczej słów (mierzonej testem decyzji leksykalnych) związanych z określonym stereotypem.

Wyróżniającym się w tym cyklu wyjątkiem jest badanie [09], które polegało na tym, że badane kobiety współpracowały z rzekomym drugim badanym (w rzeczywistości pomocnikiem badacza) w zadaniu, które wymagało kreatywności. Otrzymywały wcześniej instrukcję, żeby powstrzymać się od myślenia stereotypowego o kobietach jako osobach mniej kreatywnych od mężczyzn. Zmiennymi zależnymi były zarówno samoocena pewności siebie kobiet w trakcie badania, jak również ich zachowanie (m.in. poziom pewności siebie), co oceniał zarówno rzekomy drugi badany, jak i sędziowie kompetentni na podstawie nagrań interakcji.

W całym cyklu 13 badań zróżnicowanie grupowe było zerowe (same grupy studenckie), podobnie jak zróżnicowanie kulturowe (3 grupy holenderskie, po 1 australijskiej i brytyjskiej, reszta amerykańska).

Ogólnie, cykl tych badań wydaje się mieć raczej niską wartość poznawczą. Przede wszystkim badacze powtarzali standardową procedurę opisaną powyżej, jednak wprowadzali na tyle dużo zmian, że trudno uznać te badania za replikacje dokładne. Z wyjątkiem jednego opisanego badania nie szukali ani sposobów urealistycznienia procedury, ani wykorzystania jako zmiennych zależnych bardziej behawioralnych

wskaźników. Sporo badań miało małe kratki (7/13 mniejsze niż 20), oraz często efekty były istotne w interakcji z jakimiś dodatkowymi moderatorami. Wszystko to budzi skojarzenia z psychologią fałszywie pozytywną. Obrazu nieszczęścia dopełniają niefortunny brak zróżnicowania grupowego, o zróżnicowaniu narodowościowym nawet nie wspominając. Jak już wspomniano, nie warto wyciągać ogólnych wniosków w obszarze uprzedzeń tylko na grupach studenckich (Henry, 2008).

6.2.22. Sherman, J. W., Frost, L. A. (2000). On the encoding of stereotype-relevant information under cognitive load. *Personality and Social Psychology Bulletin*, 26(1), 26-34. (PSPB2)

W badaniu tym sprawdzano trwałość pamięci w zależności od tego, czy informacje są zgodne ze stereotypem czy nie, oraz w zależności od dodatkowych czynników. Stwierdzono, że pod wpływem obciążenia poznawczego badani pamiętają więcej niezgodnych ze stereotypem informacji (społecznych), ale tylko w przypadku rozpoznawania, bowiem w przypadku odtwarzania (swobodnego) badani przywołują więcej zgodnych ze stereotypem informacji. Dodatkowo stwierdzono, że niskie obciążenie poznawcze ułatwia odtwarzanie informacji zgodnych ze stereotypem, ale utrudnia ich rozpoznawanie. Wysokie obciążenie nie różnicuje zasadniczo jakości dwóch rodzajów odpamiętania. Stwierdzono jednak, że w przypadku rozpoznawania wysokie obciążenie poznawcze paradoksalnie podnosi poziom wykonania.

Zanim przejdę do dalszej analizy zauważę, że już opis odkrytych prawidłowości wskazuje na to, że możemy mieć do czynienia z „psychologią fałszywie pozytywną”. Efekt nie jest prosty, intuicyjnie zrozumiały i nie wydaje się też jakiś szczególnie uzasadniony teoretycznie (choć badacze jakieś uzasadnienie do niego dorobili). Łącznie występowało 8 grup eksperymentalnych (międzygrupowo manipulowano 3 zmiennymi, plus jedną wewnątrzgrupową), co oznacza, że liczba badanych w kratce wynosiła ok. 14 osób.

Procedura polegała na tym, że OB czytały informacje o osobie, która była opisana jako skinhead lub jako ksiądz. Część informacji była zgodna, część niezgodna, część neutralna względem stereotypu. Manipulowano dodatkowo obciążeniem

poznawczym (poprzez powtarzanie w pamięci zbioru liczb), manipulowano również rodzajem odpamiętania (rozpoznawanie vs. odtwarzanie swobodne). Już choćby ze względu na wielość porównań, mogą zachodzić przesłanki do odkrywania przypadkowych zależności.

Przyznam szczerze, że w przypadku tego badania jedyną jego „zaletą” wydaje się być to, że niektóre różnice są statystycznie istotne. Poza tym badane prawidłowości nie są poznawczo interesujące, grupy są względnie małe, a procedura jest banalna. Dodatkowo badanymi są, jakżeby inaczej, studenci amerykańscy.

Badanie zostało jednak opublikowane i doczekało się 41 cytowań, z których relatywnie dużo (9 artykułów z 14 badaniami) zostało przeze mnie uznanych za spełniające kryteria analizy. Wzór wyników w tych replikacjach jest również niejednoznaczny, co pośrednio może wskazywać na jakość wyjściowych badań. W 4 przypadkach wyjściowe prawidłowości zostały z grubsza potwierdzone, w 4 przypadkach nie zostały, a w kolejnych 6 jest to trudne do jednoznacznej klasyfikacji. Przy czym brak potwierdzenia wyjściowej hipotezy nie oznacza braku istotnych statystycznie zależności, ale na ogół oznacza po prostu (istotne) zależności w przeciwnym kierunku.

Z 14 badań replikujących/rozszerzających trzy ([01], [02], [04]) zostały przeprowadzone przez ten sam zespół i były proceduralnie dosyć zbliżone (oraz jedno badanie [07], gdzie testowano wyjściowe prawidłowości w interakcji z dodatkowymi moderatorami: poziomem uprzedzeń). Co ciekawe prawie zawsze wykorzystywano te same bodźce („skinhead i ksiądz”). 7 badań można zakwalifikować jako replikacje efektu, choć procedury były dosyć różne. Z niebanalnych procedur badawczych warto wymienić badanie [09], gdzie testowano, czy osoby pamiętały m.in. informacje o wykształceniu w hipotetycznym procesie rekrutacji (oczywiście, brano pod uwagę, czy określone grupy mają stereotypowo wysokie wykształcenie czy nie). W innym badaniu [08] „pamięć” do informacji testowano poprzez rodzaj miary utajonej (szybkość reakcji w zależności od zgodności lub niezgodności ze stereotypem) (podobnie w [04]). Na wspomnienie zasługuje też badanie [03], gdzie sprawdzano, czy opisywane zależności zachodzą na dzieciach w wieku 7-10 lat (procedura była mocno odległa, jednak zakwalifikowałem to badanie jako testujące moderatory podstawowe).

Nie licząc wspomnianych badań na dzieciach, pozostałe badania bez wyjątku testowały prawidłowości na studentach. Zróżnicowanie kulturowe również było minimalne, 7 prób pochodziło z USA, 3 z Niemiec i 1 z Izraela.

Ogółem cykl takich badań moim zdaniem przynosi niedużo pewnej wiedzy. Mamy tu do czynienia z typową dla psychologii „pogonią za nowatorstwem”, tzn. niechęcią do dokładnych lub w miarę dokładnych replikacji, nawet jeśli badania są konceptualnie podobne, to muszą być „wymyślone na nowo”. Wadą takich badań jest też minimalny realizm polegający na serwowaniu badanym opisów w rodzaju „Jan (skinhead) lubi chodzić do kawiarni z przyjaciółmi”. Wątpię, czy tego rodzaju „bodźce” angażują dużo uwagi badanych, czy może raczej podchodzą do badania z postawą „skończę to i dostanę dodatkowe punkty do zaliczenia”. Warto też wspomnieć, że 7 spośród 15 badań ma małe wielkości kratek (mniej niż 20 osób). To w połączeniu z wielością dokonywanych porównań (i być może wybiórczym raportowaniem) powoduje, że w moim przekonaniu wysokie jest prawdopodobieństwo psychologii fałszywie pozytywnej.

6.2.23. Batson, C., Bowers, M. J., Leonard, E. A., Smith, E. C. (2000). Does personal mortality exacerbate or restrain retaliation after being harmed?. *Personality and Social Psychology Bulletin*, 26(1), 35-45. (PSPB3)

Interesujący efekt polegający na tym, że stwierdzono dwa rodzaje motywacji do przywracania sprawiedliwości: wyrównanie niesprawiedliwej nierówności i zemsta za spowodowanie niesprawiedliwej nierówności. Badani uczestniczyli w rodzaju gry ekonomicznej, w której z rzekomym drugim badanym transferowali żetony między sobą. Jeśli doszło do nieuzasadnionego przetransferowania żetonów, badani przywracali sprawiedliwy podział. Jeśli z kontekstu wynikało, że do takiej sytuacji doszło przypadkiem, nie karali drugiego badanego. Jeśli jednak wynikało, że niesprawiedliwy transfer był wynikiem celowego działania, w takim wypadku dodatkowo karali współgracza (w rzeczywistości nie było drugiego badanego). Wysokie poczucie moralnej odpowiedzialności (mierzone samoopisowo) nie wpływało na pierwszy rodzaj motywacji, ale powodowało zmniejszenie drugiego.

Eksperyment jest niewątpliwie interesujący na tle innych tu omawianych badań, gdzie trzeba było zareagować na kilka zdań opisów jakichś sytuacji (np. EJSP6). Niestety, nie był zbyt często powtarzany. W jednym badaniu [01], które można uznać za replikację konceptualną i zwiększanie precyzji teorii stwierdzono, że efekt „wyrównywania niesprawiedliwości” nie dotyczy sytuacji, gdy ofiarą jest trzecia osoba. Niezależnie od tego, czy osoba zrobiła krzywdę badanemu, czy trzeciej osobie, badani mogli byli się „zemścić” poprzez odjęcie badanemu wirtualnych dolarów, które on mógł wydać w czasie wirtualnej wycieczki. Dodatkowe zmienne zależne w tym badaniu to między innymi złość na rzekomego badanego, oraz jego ocena. Wybór (?) tych zmiennych jest dosyć znamieny, gdyż znowu można sobie wyobrazić, że skoro badacze zadali sobie trud przeprowadzenia takiej dosyć złożonej procedury, mogli się bardziej postarać i wybrać jakąś bardziej realistyczną zmienną zależną np. wymierzenie „kary” w postaci wykonywania nudnego zadania. Chyba, że badacze faktycznie mierzyli również inne zmienne zależne, ale akurat wyszły im tylko te, które zaraportowali.

Tak więc podsumowując: nasza wiedza na temat tego efektu jest nieduża. Prawdziwość efektu oczywiście bardzo przemawia do intuicji, ale nie było żadnych replikacji dokładnych, ale jedno badanie, które można zaliczyć do replikacji konceptualnych, nie sprawdzało podstawowych moderatorów, ani bardziej behawioralnych zmiennych zależnych. Mimo przebadania interesujących zależności, badanie oryginalne było rzadko cytowane (13 razy).

6.2.24. Gibson, B., Sachau, D. (2000). Sandbagging as a self-presentational strategy: claiming to be less than you are. *Personality and Social Psychology Bulletin*, 26(1), 56-70. (PSPB4)

Artykuł wskazuje na występowanie strategii autoprezentacji polegającą na tym, że osoby przedstawiają swoje potencjalne osiągnięcia w gorszym świetle niż tego można realistycznie oczekiwać (ang. *sand-bagging*, dalej S-B; można przetłumaczyć dosłownie jako „branie worka na plecy”). Stwierdzono przy tym, że tendencja do stosowania tej strategii, choć podlega uwarunkowaniom sytuacyjnym, jest zasadniczo zmienną dyspozycyjną.

Dlatego w pierwszym badaniu wyjściowego artykułu badacze przebadali psychometryczne właściwości takiej skali. W drugim i trzecim badaniu (eksperymentalnym) osoby o wysokim natężeniu S-B pod wpływem presji (oczekiwanie wysokich wyników lub publika oglądająca wyniki) twierdziły, że osiągną wyniki niższe niż wynikało to z ich wcześniejszych osiągnięć (badani grali w grę komputerową). Niestety, zwracają uwagę stosunkowo małe grupy badane 15-17 osób w warunku eksperymentalnym.

Zespół badaczy przedstawił jeszcze 4 badania w dwóch artykułach, które replikowały konceptualnie efekt oraz znajdowały dodatkowe moderatory efektu. W każdym wypadku w badaniu uczestniczyli amerykańscy studenci.

Dodatkowo niezależny zespół badaczy z Niemiec [03] w dwóch badaniach korelacyjnych (w tym jedno podłużne) stwierdził, że przedstawianie siebie w skromnym świetle w połączeniu z „wysokimi umiejętnościami politycznymi” sprzyja wyższym osiągnięciom. Należy jednak zaznaczyć, że wszystkie zmienne były tutaj dostarczane przez osoby badane. Dodam też, że można dyskutować, czy faktycznie użyte tu zmienne do pomiaru skromności są identyczne ze skalą S-B.

Ogółem program badań dotyczących *sensu stricte* S-B jest stosunkowo skromny (6 badań jednego zespołu), całkowicie niezróżnicowany etnicznie i grupowo oraz stosunkowo jednorodny pod względem metody, choć trzeba przyznać, że manipulacje eksperymentalne są zwykle realistyczne (samoopisowy charakter zmiennych zależnych nie wydaje się dużym problemem, bo istotą tych badań jest strategia samoprezentacji – czy ta strategia prowadzi do jakichś wyników behawioralnych jest osobnym, ale w tym kontekście marginalnym pytaniem). Choć jest to więcej badań niż w niektórych innych analizowanych tu przypadkach, trzeba jednak stwierdzić, że zakres wiedzy o S-B przez 12 lat zwiększył się w niewielkim stopniu.

6.2.25. Heine, S. J., Takata, T., Lehman, D. R. (2000). Beyond self-presentation: evidence for self-criticism among Japanese. *Personality and Social Psychology Bulletin*, 26(1), 71-78. (PSPB5)

Badanie opisuje interesujący efekt z obszaru różnic międzykulturowych. Japońscy i kanadyjscy studenci wykonywali zadanie uwagowo-percepcyjne. Otrzymywali

następnie „informacje zwrotne” o poziomie wykonania (część badanych losowo otrzymywała informację o wykonaniu poniżej, a część powyżej przeciętnej). Pytani o to Japończycy częściej niż Kanadyjczycy uważali, że wypadli w zadaniu poniżej przeciętnej. Potrzebowali też więcej czasu i więcej danych, żeby stwierdzić, że wypadają powyżej przeciętnej. „Wypadając” poniżej przeciętnej, wyżej oceniali trafność narzędzia mierzącego ich rzekome cechy poznawcze w porównaniu do Kanadyjczyków. Dodam, że badanie miało bardzo przyzwoitą moc (średnio około 60 osób w kratce).

Efekt ten i samo badanie spotkało się z dosyć dużym zainteresowaniem (93 cytowania). Z tej liczby wyodrębniłem 23 teksty (zawierające 37 badań), które konceptualnie nawiązywały do oryginalnego efektu. Badania te można podzielić na przede wszystkim na badania eksperymentalne i korelacyjne. W pierwszej grupie wyróżniłem 12 badań. Przy czym tylko jedno z nich [03] jest podobne proceduralnie do wyjściowego eksperymentu, a w drugim [07] wprowadzono dodatkową modyfikację polegającą na tym, że badani ze sobą rywalizowali (w takim wypadku wzór zaniżania samooceny u Azjatów zanikał).

Do badań eksperymentalnych testujących wprost wyjściową hipotezę zaliczyłem:

- [04:4] – odbywały się „negocjacje”, gdzie manipulowano informacją zwrotną o poziomie wykonania badanego, zmienną zależną było m.in. poczucie winy za taki wynik;
- [20] (trzy badania) – procedura polegała na tym, że badani zadawali sobie nawzajem pytania z wiedzy ogólnej (tak jak w klasycznym eksperymencie Ross, Amabile, Steinmetz, 1977).
- [01:1-3] – tutaj zasadniczą zmienną zależną była wytrwałość w wykonywaniu zadania po pozytywnych lub negatywnych informacjach zwrotnych.

Dwa eksperymenty ([16] i [22]) badały trafność potencjalnego wyjaśnienia obserwowanych różnic (osoby przechwalające się w kulturach kolektywistycznych są oceniane bardziej negatywnie).

Wreszcie w ostatnim z badań eksperymentalnych (PSPB5_05) zastosowano niezwykle pomysłową procedurę, którą warto tu omówić, pokazuje bowiem, że badacze bywają czasami bardzo kreatywni w poszukiwaniu niebanalnych testów teorii. Badani zaliczali się do jednej z trzech grup: mieszkających w Kanadzie

chińskich imigrantów, Kanadyjczyków pochodzenia chińskiego (mieszkających od urodzenia w Kanadzie) oraz Kanadyjczyków pochodzenia europejskiego. Badani odpowiadali na pytania dotyczące m.in. samooceny. Chińscy imigranci robili to (losowo) albo po chińsku, albo po angielsku. W przypadku odpowiadania po chińsku wzór wyników był inny niż gdy odpowiadali po angielsku i w tym ostatnim przypadku był zbliżony do wzoru wyników pozostałych dwóch grup. Oczywistą wadą takiego rodzaju badania jest to, że badanie to nie jest replikacją konceptualną bez przyjęcia dużej liczby dodatkowych założeń. Trudno jednak nie odmówić badaczom pomysłowości w testowaniu wyjściowej hipotezy.

Dalej występuje duża grupa badań korelacyjnych (22 badania w 16 artykułach). Z tych badań bardzo duża grupa jest proceduralnie banalna i bada po prostu różnice kulturowe w mierzonych kwestionariuszowo zmiennych (szczególnie związanych z samoocenną) (szczególnie [09], [10], [11], [12], [14], [15]; podobnie w [08], ale tutaj badanymi byli uczniowie szkół średnich; w [04] badano samoocenianą uczciwość; w [19] oprócz jawnej samooceny mierzono samoocenę ukrytą przy pomocy miary utajonej).

W kilku badaniach korelacyjnych podejmowano próby zbadania przyczyn efektów. Wśród zaproponowanych można wymienić teorie wskazujące, że obserwowane różnice wynikają z:

- [18:1] tendencji osób z kulturach azjatyckich do wywieraniu dobrego wrażenia (ang. *impression management*) w przeciwieństwie do osób z kultur zachodnich, które są bardziej skoncentrowane na nieszczerym przechwalaniu (ang. *self-deceptive enhancement*). W [18:2] badano członków tylko jednej kultury (USA), ale poziom kolektywizmu badanych kontrolowano kwestionariuszowo;
- większego przekonania Amerykanów (w porównaniu do Japończyków), że sukces to kwestia wybitnych zdolności, a nie wysiłku i pracy [01] (stąd dla japońskich studentów porażka w zadaniu poznawczym może być mniej zagrażająca, gdyż potencjalnie mogą ją zmienić większą pracą);
- wśród badanych Izraelczyków (mieszkańców miast i kibuców, świeckich i religijnych) stwierdzono, że stopień, w jakim w badaniu pomniejszają swoje zasługi i wykazują psychologiczną skromność, nie zależy od poziomu

indywidualizmu/kolektywizmu, ale od wymagań kulturowych, tutaj poziomu przestrzegania zasad religijnych [02].

W dwóch badaniach ([06] i [21]) porównywano samooceny badanych z ocenami dokonywanymi przez ich rówieśników.

Dodatkowo jedno badanie korelacyjne przeprowadzono na dzieciach. W [23] japońskie 7-11-latki (w porównaniu do amerykańskich rówieśników) oceniają przyznawanie się do dobrych uczynków (szczególnie publicznie) bardziej negatywnie.

Pewnych wskazówek, co do przyczyn efektu może też dostarczać badanie [13], gdzie stwierdzono, że chińskie dzieci (11 lat) twierdziły, że są w mniejszym stopniu chwalone przez ich rodziców, niż twierdziły to dzieci amerykańskie [13:1]. Dodatkowo w tym badaniu obserwowano interakcję matek i dzieci w trakcie wykonywania zadania poznawczego i zauważono podobny wzór wyników [13:2].

Do osobnej kategorii można zaliczyć badanie [22:1], gdzie OB miały przypomnieć sobie sytuację, gdy rozmawiały z kolegą na temat odniesionego przez nich sukcesu, a następnie spisać dokładnie tę rozmowę. Skrypty były analizowane jakościowo i ilościowo, i stwierdzono, że badani chińscy studenci opisując swoje sukcesy najczęściej wyrażali skromność przypisując sukces czynnikom zewnętrznym (szczęściu, pracy zespołowej itp.), bardzo rzadko wyrażali samozachwyt. Badanie to wyróżnia się na tle innych w tym cyklu ze względu na swój opisowy charakter oraz na to, że badacze byli potencjalnie otwarci na różnorakie reakcje, a nie definiowali z góry zbioru zmiennych zależnych.

Natomiast najbardziej niebanalna replikacja konceptualna wyjściowego efektu została przedstawiona w [17] przez zespół badaczy z Australii i Hongkongu. Badano nie ludzi, ale ludzkie wytwory, mianowicie oficjalne oświadczenia spółek giełdowych dotyczących przyczyn ich osiągnięć ekonomicznych. Porównywano spółki z USA, Japonii i Singapuru. Nie stwierdzono zasadniczych różnic w stylu atrybucji sukcesów i porażek. W tej analizie jest to jeden z nielicznych przykładów danych „behawioralnych” wytworzonych poza badaniem i istotnych dla analizowanej hipotezy.

Jeśli chodzi o zróżnicowanie narodowościowe, nie będzie zaskakujące, że było one bardzo duże na tle innych cykli badań (wyjściowa hipoteza wszak dotyczyła

zróznicowania kulturowego). Ogółem badań było 38, z czego w 29 porównywano osoby badane z dwóch lub większej liczby krajów. Zwykle porównywano dwa kraje zachodnie (USA – 23 razy; lub Kanadę 30 razy) z dwoma krajami azjatyckimi (Japonią – 17 razy i Chinami 14 razy). Sporadycznie pojawiły się próby z jeszcze innych krajów (po jednej z Izraela, Korei, Meksyku, Filipin i Szwecji). Niewątpliwie to względnie duże zróznicowanie jest pozytywne, choć zwróć uwagę, że olbrzymia większość krajów azjatyckich (czy szerzej: kolektywistycznych) i zachodnich (czy szerzej: indywidualistycznych) nie była reprezentowana.

Natomiast jeśli chodzi o zróznicowanie grupowe, wszystko odbywało się niestety po tradycyjnych „torach”. Dzieci lub adolescenci wchodzili w skład 5 prób (przy czym w jednym wypadku były to pary dziecko i matka). Z 32 dorosłych prób 31 stanowiły próby tylko studencie, a w jednym przypadku [16] była to próba studencka i pochodząca z populacji ogólnej. Ten cykl (i większość innych) pokazuje dobitnie, że nawet gdy określona hipoteza jest testowana przez międzynarodowe zespoły, tak czy owak idą one po linii najmniejszego oporu w wyszukiwaniu osób badanych.

Podsumowując ten cykl badań, trzeba stwierdzić, że wyjściowa hipoteza jest dosyć szeroko przetestowana, zarówno w badaniach korelacyjnych, jak i eksperymentalnych. Ilość danych wskazuje, w moim przekonaniu, na to, że jest to coś więcej niż hipoteza podparta wstępnymi danymi. Wzór wyników jest przy tym dosyć spójny i zdecydowana większość badań potwierdza ją. Wskazano też na dodatkowe możliwe przyczyny efektu, na przykład na różnicę w ocenie do osiągnięć ze strony zachodnich i wschodnich matek (ale tylko w jednym badaniu).

Pomimo tego, że są tu szeroko testowane różnice międzykulturowe wybór „reprezentantów” kultur jest niezbyt szeroki (po jednej stronie USA i Kanada, po drugiej Japonia i Chiny). Ograniczanie się tylko do studentów nie wymaga dalszego komentarza.

Nawet jednak jeśli zostawić na boku ewidentne ograniczenia trafności zewnętrznej wynikające z ograniczonych prób oraz małej liczby danych terenowych (choć większość laboratoryjnych badań eksperymentalnych wydaje się całkiem realistyczna), trzeba zauważyć, że jest to jedna z najszerzej przetestowanych hipotez spośród wszystkich tutaj analizowanych. Czego tym badaniom brakuje, podobnie zresztą jak praktycznie wszystkim tu opisanym, to fakt, że nie występuje coś, co

opisałbym jako systematyczny program badawczy, a badania są raczej zbiorem badań mniej lub bardziej luźno połączonych wyjściową hipotezą. Wątek ten na razie zostawię, gdyż szerzej rozwinę go w rozdziale 7.

6.2.26. Paese, P. W., Gilin, D. A. (2000). When an adversary is caught telling the truth: reciprocal cooperation versus self-interest in distributive bargaining. *Personality and Social Psychology Bulletin*, 26(1), 79-90. (PSPB6)

Artykuł opisuje efekt polegający na tym, że odsłanianie się przed rywalami (np. w negocjacjach) poprzez ujawnianie wiarygodnych informacji na temat swojej pozycji negocjacyjnej będzie prowadziło paradoksalnie do większych ustępstw ze strony osób, które te informacje otrzymały. Wbrew pierwszej intuicji badani nie wykorzystują tych informacji, żeby poprawić sobie pozycję negocjacyjną (od razu wspomnę, że interesująca byłaby replikacja, w której badani poznają takie informacje z trzecich źródeł; takiej replikacji jednak nie było wśród badań cytujących). Zmienną zależną był m.in. poziom stawianych żądań oraz ocena jakości interakcji.

Badanie to było konceptualnie zreplikowane przez tego samego badacza [01]. Tym razem jednak okazało się, że efekt występuje tylko w negocjacjach drogą mailową lub telefoniczną, ale już nie twarzą w twarz. Kolejne badanie [02] zaliczam do kategorii testowania teorii (i generowania odkryć). Tutaj okazało się, że jeśli badanych poinformować, że druga strona negocjacji jest nastawiona konfrontacyjnie, wówczas ocena i wyniki będą niższe. W tym samym artykule w dodatkowym badaniu kwestionariuszowym na ogólnej populacji (rekrutowanej na ulicy) stwierdzono, że ludzie deklarują mniejszą gotowość do współpracy z konfrontacyjnie nastawionym partnerem negocjacji kupna-sprzedaży (zmienną zależną była m.in. deklarowana gotowość udzielania prawdziwych informacji o swojej pozycji negocjacyjnej). To badanie jest pewną odmianą, gdyż we wszystkich pozostałych badaniach w tym cyklu badanymi byli studenci (z dwóch krajów: USA i Holandii).

Ogólnie więc badania te są stosunkowo mało zróżnicowane jeśli chodzi o metody i źródła danych. Efekt jest poznawczo i życiowo interesujący i warto, by był dalej badany.

6.2.27. Greenberg, J., Arndt, J., Simon, L., Pyszczynski, T., Solomon, S. (2000). Proximal and distal defenses in response to reminders of one's mortality: evidence of a temporal sequence. *Personality and Social Psychology Bulletin*, 26(1), 91-99. (PSPB7)

Badanie jest wariantem teorii opanowywania trwogi. Stwierdzono, że „klasyczna” procedura wzbudzania poczucia własnej śmiertelności (ang. *mortality salience*; przykład takiej procedury to polecenie badanym, żeby opisali, jakie myśli, emocje wywołuje w nich myślenie o śmierci) ma dwojakie skutki: po pierwsze wywołuje „obronę bezpośrednią” (ang. *proximal defence*) polegającą na zaprzeczaniu własnej śmiertelności, po drugie „obronę pośrednią” (ang. *distal defence*) polegającą na faworyzowaniu własnej grupy kulturowej. Przy czym, jeśli po wzbudzeniu poczucia własnej śmiertelności badanych poddać dystrakcji, siła obrony pośredniej ulega zwiększeniu. Natomiast badani bezpośrednio po procedurze wzbudzania własnej śmiertelności wykazują wyższy poziom obrony bezpośredniej.

Badanie to jest dosyć często cytowane (78 razy), jednak jego replikacje/rozszerzenia są nieliczne (5 badań w 3 artykułach). Wszystkie zostały przeprowadzone przez osoby, które były współautorami wyjściowego artykułu. Wszędzie wykorzystano zbliżoną procedurę jak w badaniu wyjściowym, przynajmniej jeśli chodzi o sposób manipulacji poczuciem własnej śmiertelności. Różniły się natomiast zmienne zależne i w badaniach koncepcyjnie replikujących mierzono m.in. deklarowaną gotowość do uprawiania ćwiczeń, zachowania prozdrowotne oraz ocenianie znaczenia wzniosłych i przyziemnych celów w życiu. Choć badania te na pozór różnią się niewiele („tylko” inny rodzaj zmiennej zależnej) to kwalifikuję je jako testowanie teorii, a nie replikacje konceptualne, gdyż żadne z tych badań nie replikowało w sposób niewątpliwy wyjściowej hipotezy, ale z „nakładką” pewnej dodatkowej teorii (np. że rodzajem „obrony bezpośredniej” jest gotowość do uprawiania ćwiczeń fizycznych itd.).

Badanymi byli studenci amerykańscy (jeden wyjątek zanotowano w badaniu [02:2], gdzie badane rekrutowano na lokalnym targu spożywczym). W dwóch badaniach (i badaniu oryginalnym) grupy eksperymentalne też miały bardzo małe kratki (około 12 osób).

Wydaje się, że ten stosunkowo nieduży cykl badań jest dobrą ilustracją tezy, że czasami badacze zamiast w miarę sumiennie przetestować jedną hipotezę,

wybiegają do przodu testując hipotezy zbliżone, jednak wymagające dodatkowych założeń (które też nie są ponad wszelką wątpliwość pewne). Takie podejście jest zrozumiałe z punktu widzenia kultury „nowatorstwa”, utrudnia jednak wypracowanie w miarę pewnej wiedzy o konkretnych efektach. W omawianym przykładzie może być tak, że badacze nietrafnie teoretycznie interpretują określone zmienne i tak naprawdę testują zupełnie nowe hipotezy.

Ocena wartości poznawczej jest więc raczej niska. Dodatkowo wykorzystanie tylko miar samoopisowych jako zmiennych zależnych wzmacnia to przekonanie.

6.2.28. Harkins, S. G., White, P. H., Utman, C. H. (2000). The role of internal and external sources of evaluation in motivating task performance. *Personality and Social Psychology Bulletin*, 26(1), 100-117. (PSPB8)

Badanie to jest nowatorskim rozszerzeniem innego efektu. Wcześniej stwierdzono, że wykonanie zadania jest lepsze, gdy badacz określi poziom wykonania, niż gdy poda instrukcję "postaraj się najbardziej jak tylko możesz" (ang. *do your best*). Efekt ten był szczególnie wyraźny, gdy badacz zapowiedział, że oceni poziom wykonania. Tutaj okazuje się, że gdy badani mogą porównać swój poziom wykonania do podobnej osoby (lub osoby osiągającej wysokie wyniki), wykonują zadanie na wysokim poziomie niezależnie od tego, czy badacz sprawdza poziom wykonania. W wyjściowym badaniu zadaniem było wygenerowanie maksymalnej liczby zastosowań określonego przedmiotu (konkretnie tutaj: noża).

Ogółem 4 artykuły (z 6 badaniami) zakwalifikowałem do dalszej analizy (na 11 cytujących). Ten sam zespół uzyskał zbliżone wyniki w trzech replikacjach konceptualnych ([01], [03]). Dodatkowo w miarę zbliżony efekt uzyskał niezależny badacz [02]. W całym cyklu badań średnie wielkości $d \approx 0.2$ były na pograniczu przyzwoitości (20 osób).

Za to szczególną uwagę chciałbym poświęcić badaniu przedstawionemu w [04]. Dla ścisłości, badacze nie analizowali identycznej hipotezy, ale konceptualnie zbliżoną (badani, gdy wykonują zadanie równolegle z ekspertem lub jakąś osobą o wysokim poziomie wykonania, osiągają lepsze wyniki niż, gdy wykonują zadanie równolegle z osobą, która ma słabe wyniki). Żeby przetestować tę hipotezę badacze przeanalizowali osiągi zawodowych golfistów na oficjalnych turniejach, relatywnie do

poziomu ich przeciwników (drugie z badań w tym artykule było normalnym eksperymentem). Przy tym nie porównywano wyników gry zawodników z wynikami ich rywali w tym samym dniu, ale z wcześniejszych dni turnieju (a więc niejako w domyśle poziomem oczekiwań).

Niezależnie od możliwych potencjalnych artefaktów wynikających z nieeksperymentalnego charakteru takiego badania, wydaje się, że zasługuje ono na szczególne wyróżnienie (i naśladowanie przez społeczność badaczy). Jest to jedno z nielicznych tego rodzaju badań w całym cyklu. Zebrano dane o zachowaniu, które w oczywisty sposób nie podlega zmianie pod wpływem oddziaływania badacza i sytuacji badawczej. „Osoby badane” były również zmotywowane do osiągnięcia wysokiego wyniku, gdyż uczestniczyły w turniejach o wysoką stawkę.

Przykład ten pokazuje, że przy odrobinie inwencji i wysiłku badacza możliwe jest wyjście poza standardowe procedury i szukanie danych behawioralnych w rzeczywistych zarchiwizowanych danych.

Ogólnie jednak cały cykl badań liczy sobie niewiele badań, badanymi byli każdorazowo amerykańscy studenci (z wyjątkiem badania golfistów), dlatego jego wartość poznawcza jest niewielka.

6.2.29. Forrest, J. A., Feldman, R. S. (2000). Detecting deception and judge's involvement: lower task involvement leads to better lie detection. *Personality and Social Psychology Bulletin*, 26(1), 118-125. (PSPB9)

Badanie dotyczy trafności wykrywania kłamstwa. Stwierdzono, że mniejsze zaangażowanie w to zadanie podnosi poziom skuteczności w wykrywaniu kłamstw. Manipulacja eksperymentalna polegała na podawaniu informacji dla badanych, że udane wykonywanie tego zadania świadczy o „inteligencji i wysokich umiejętnościach społecznych” (lub nie podawano tej informacji w warunku kontrolnym). Zmienną zależną była ocena wiarygodności osób na nagranych wcześniej wideo (które prawdziwie lub fałszywie przedstawiały swoje poglądy na wybrane tematy).

Pięć badań (w 3 artykułach) można uznać za replikacje konceptualne. Zarówno badanie oryginalne, jak i badania replikujące miały jedną cechę wspólną: badanymi byli studenci. Jedna próba była amerykańska, druga kanadyjska, a cztery niemieckie. W badaniu oryginalnym i 2 dwóch badaniach replikujących ([01] i [02]) bodźce były

dosyć realistyczne: proszono wcześniej inną grupę studentów o nagranie swoich wypowiedzi, z których część była prawdziwa, a w części mieli za zadanie kłamać. Jedynie w badaniu [03] zastosowano chyba mniej sensowne bodźce: transkrypty wypowiedzi, do których załączano zdjęcia rzekomych osób wypowiadających się. Manipulacja poziomem zaangażowania polegała na podbijaniu ambicji badanych informacjami, że „tylko nieliczni potrafią dobrze wykrywać kłamstwa”, lub „udane wykrywanie kłamstw jest wskaźnikiem inteligencji”. W jednym z badań motywację miała zwiększać informacja, że badanie jest bardzo istotne dla nauki.

Zmienną zależną była ocena wiarygodności i/lub prawdomówności. Mam pewne kłopoty z tym, do jakiej kategorii zaklasyfikować taką zmienną. Zaklasyfikowałem ją ostatecznie do kategorii „zachowanie/zmienna poznawcza”, gdyż – tak jak w zadaniach poznawczych – trudno jest wykazać wyższy poziom niż pozwala na to aktualna sprawność, można co najwyżej wykazać poziom niższy (np. rozmyślnie odpowiadając błędnie). Oczywiście powierzchownie taka zmienna ma więcej wspólnego ze „zmienną samoopisową”. Tak czy owak w przypadku tego rodzaju badań trudno spodziewać się jakichś innych zmiennych zależnych, więc brak ich zróżnicowania nie jest tutaj żadną dużą wadą.

W jednym przypadku [01] potwierdzono oryginalny efekt, w drugim [02] nie potwierdzono. W trzecim badaniu [03] procedura była tak dalece różna, że niepotwierdzenie efektu nie jest tu jakimś szczególnie istotnym prognostykiem.

Podsumowując, oryginalny efekt jest poznawczo interesujący i niebanalny. Trudno mieć jakieś zarzuty zarówno do manipulacji eksperymentalnych, czy natury zmiennych zależnych. Nie licząc badania [03] (dosyć odległego konceptualnie) przebadano 337 osób i jest to liczba duża na tle wielu innych badań, choć mała, żeby móc mówić o bardzo mocnym potwierdzeniu efektu. Po raz kolejny natomiast powtarza się całkowity brak zróżnicowania grupowego osób badanych (tylko studenci), oraz niewielkie zróżnicowanie narodowościowe (tylko 3 kraje zachodniego świata).

6.2.30. Nelson, L. J., Klutas, K. (2000). The distinctiveness effect in social interaction: creation of a self-fulfilling prophecy. *Personality and Social Psychology Bulletin*, 26(1), 126-135. (PSPB10)

W tym eksperymencie badano poziom oczekiwań na podstawie dystynktywnych (występujących rzadko) i niedystynktywnych cech. Badanych przydzielano w parach do jednego z dwóch zadań: przeprowadzającego wywiad, lub udzielającego wywiadu. A więc dalej „właściwymi” badanymi jest ta pierwsza grupa. Badani dowiadywali się, że osoba, z którą będą przeprowadzali wywiad, ma określoną właściwość opisaną w hermetyczny sposób (typu "osoby myśląca konwergencyjnie") wraz z informacją, z czym owa właściwość się wiąże (szereg pozytywnych lub negatywnych cech). Właściwość była opisana jako dystynktywna lub nie (występująca u kilku procent lub ponad 60 procent populacji). Nagrywano wywiady. Zmiennymi zależnymi były m.in. oczekiwania badanych względem drugiej osoby oraz oceniane przez sędziów kompetentnych elementy interakcji np. swoboda uczestników i ich poziom zainteresowania rozmową. (Występował też warunek kontrolny, w którym drugi badany miał pozytywne i negatywne cechy niedystynktywne).

Badanie to wyróżnia się więc tym, że główna zmienna zależna ma charakter behawioralny. Również manipulacja eksperymentalna wydaje się być wysoce realistyczna. Należy jednak zaznaczyć, że w każdej grupie eksperymentalnej było tylko 13 osób (ściślej biorąc 26, gdyż osoba, z którą przeprowadzano wywiad była też badanym, ale oczywiście liczba 13 jest właściwa dla porównań statystycznych). Stąd więc stwierdzono, że efekt dotyczy tylko kobiet (jednak badacze uczciwie zaznaczyli, że mężczyzn było bardzo niewiele). Biorąc pod uwagę, że poza tym spośród 12 zmiennych zależnych 11 było istotnych, musi to wzbudzać wątpliwości, co do wybiórczego raportowania. Wyliczyłem⁸², że przy takiej liczbie osób badanych efekt musi wynosić około $d = 0,65$, żeby w połowie przypadków został on wykryty w formie istotnych statystycznie wyników, jeśli założymy moc na poziomie 90 %, w takim razie wielkość efektu musi wynosić, aż $d = 1,15$.

Wśród cytujących je badań tylko jedno można zaliczyć z grubsza do replikacji konceptualnej. Przy tym było to badanie dużo bardziej „standardowe”, m.in. manipulowano oczekiwaniami względem hipotetycznych, opisanych w tekście osób, oraz zmienną zależną były opinie badanych na temat posiadanych przez tą osobę cech (dodatkowo badanie to cechowało się równie niewielką mocą, gdyż brało w nim udział 26 osób w dwóch warunkach eksperymentalnych manipulowanych

⁸² Przy pomocy kalkulatora G-Power.

wewnątrzgrupowo). Więcej badań nawiązujących do tego efektu wśród 7 cytujących tekstów nie występowało.

Ogółem więc badanie oryginalne wyróżnia się pozytywnie pod względem metody, ale bardzo traci na niskiej mocy, która może wskazywać na wybiórcze publikowanie lub inne zabiegi ze szkoły „psychologii fałszywie pozytywnej”.

6.2.31. van Dijk, E., Vermunt, R. (2000). Strategy and fairness in social decision making: sometimes it pays to be powerless. *Journal of Experimental Social Psychology*, 36(1), 1-25. (JESP1)

Badanie opisuje interesujący, nieco paradoksalny efekt. Porównywano dwa rodzaje gier ekonomicznych: w „ultimatum” i „dyktatora” (w pierwszym rodzaju jedna osoba proponuje podział jakiejś kwoty, którą druga strona musi go zaakceptować, jeśli go odrzuci, nikt nie dostaje nic; z kolei w grze w „dyktatora” jeden z badanych deklaruje określony podział środków i decyzja ta jest ostateczna). W przypadku gry w „ultimatum” badani wykazują większe nastawienie egoistyczne (zwane też „strategicznym”) tj. w większym stopniu promują swoje interesy. Jednakże w przypadku gry w „dyktatora” badani paradoksalnie dbają o bardziej sprawiedliwy podział środków i okazują się bardziej hojni (dodatkowo manipulowano również względną siłą żetonów, które były rozdzielane i tym, czy badany o tym wiedział, dzięki czemu można było sprawdzać siłę nastawienia strategicznego badanych). Zarówno temu, jak i przyszłym badaniom tego rodzaju nadawano nieco realizmu informując, że żetony, o które grali będzie można przeliczyć na rzeczywiste pieniądze, które otrzymają od badacza.

7 artykułów (z 13 badaniami) replikowało efekt, lub próbowało znaleźć do niego dodatkowe moderatory. Jedno z badań [01] jest na tyle podobne do badania wyjściowego, że kwalifikuję je jako replikację dokładną. Badania [02] i [05] są pod względem procedury konceptualnie podobne jak wyjściowe. Interesującą modyfikację wprowadzono w badaniu [06]. Manipulowano tu poziomem władzy, poprzez to, że badani albo wyrażali „preferencję”, albo „podejmowali decyzję”, ile środków ma dostać drugi badany. W trzech eksperymentach większa władza prowadziła do większej hojności.

Efekt ten próbowano powtórzyć na dzieciach ([03], [07]) oraz na dzieciach i adolescentach [04]. Tylko w jednym przypadku [03] efekty się potwierdziły, ale i to przy współdziałaniu dodatkowych moderatorów. Natomiast w przypadku replikacji na próbach dorosłych (czyli zawsze studenckich) wyniki wyjściowe potwierdziły się w 6 badaniach, a nie potwierdziły w dwóch.

Ogólnie cały cykl badań oceniam dosyć pozytywnie na tle olbrzymiej większości innych omawianych tutaj cykli badań. Miała miejsce replikacja dokładna (to wielka rzadkość), i parę razy wysoce zbliżone procedury, oraz procedury odmienne, choć badające ten sam efekt. Rzadkością jest też obserwowana tu próba replikacji na osobach młodszych. Występuje też pewne nieduże, ale jednak zróżnicowanie międzynarodowe, ewidentnie w tego rodzaju badaniach wyspecjalizowali się badacze z Holandii (8 badań), pięć dalszych miało miejsce w Izraelu, oraz po jednym w USA i Chinach. Wadą jest ograniczenie się tylko do laboratoryjnych badań eksperymentalnych i brak prób weryfikacji efektu poza laboratorium, czy to poprzez eksperymenty terenowe, czy to poprzez analizę jakichś archiwalnych danych (około)behawioralnych.

6.2.32. Drolet, A. L., Morris, M. W. (2000). Rapport in conflict resolution: accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1), 26-50. (JESP2)

W tym artykule testowano, czy negocjacje (lub inne sytuacje konfliktowe, np. „dylemat więźnia”) twarzą w twarz mają przewagę w stosunku do innych form kontaktu szczególnie przez telefon lub przez e-mail. W badaniu oryginalnym stwierdzono, zgodnie z intuicją, przewagę wzajemnego kontaktu wzrokowego. Od razu jednak zaznaczę, że w badaniu pierwszym wyjściowego artykułu nie był to jednoznaczny wzór wyników, gdyż istotne różnice uzyskano tylko dla jednej z dwóch zmiennych, jak się wydaje mniej ważnej (badani wcielali się w przedstawiciela związku zawodowego lub przedstawiciela zarządu i negocjowali dwie kluczowe kwestie: wielkość płac w ich zakładzie i długość strajku; istotne wyniki uzyskano tylko dla tej drugiej zmiennej). W drugim badaniu dodatkowo stwierdzono, że poziom

negocjacji twarzą w twarz sprzyja dobremu emocjonalnemu współgraniu i pozytywniejszym ocenom partnera interakcji.

Następnie 13 badań miało być konceptualnie zbliżonych do badania oryginalnego. We wszystkich badani angażowali się w jakąś formę hipotetycznych negocjacji między sobą (lub w „dylemat więźnia”), sporadycznie badanych dodatkowo motywowano np. oferując nagrodę za najwyższy wynik negocjacji.

W badaniach tego rodzaju niewątpliwą zaletą jest fakt, że badacze wykazują się większym niż minimalny wysiłkiem organizacyjnym (mogliby wszak ograniczyć się do rozdania kwestionariuszy, w których badani deklarowaliby, jak hipotetycznie zareagowaliby na określone sytuacje negocjacyjne, tak jak w PSPB2_02:4). Zmienną zależną był w 11 sytuacjach wynik negocjacji (lub innej sytuacji konfliktowej, jak „dylemat więźnia”), często także oceny jakości interakcji, oceny partnera negocjacji. W jednym badaniu [07] nie ujawniono wyników negocjacji, tylko ocenianą jakość zaufania, a w innym badaniu sędziowie kompetentni na podstawie transkryptów oceniali częstość okłamywania się przez badanych (i zaklasyfikowałem tą zmienną jako zachowanie).

Dla równowagi poruszę te aspekty, które sprawiają, że przysłowiowa szklanka jest do połowy pusta. Widzę potencjalnie wiele możliwości bardziej realistycznych eksperymentów, z których badacze nie skorzystali. W szczególności można sobie wyobrazić rozmaite eksperymenty terenowe np. takie, w których badacze oferują w realistyczny sposób sprzedaż określonych produktów i manipulują formą negocjacji. Można też sobie wyobrazić, że badacze mogliby użyć danych archiwalnych gromadzonych przez działy sprzedaży, lub wraz z działami sprzedaży sprawdzać czy modalność kontaktu ma wpływ na gotowość do ustępstw.

W całym cyklu badań niestety występuje bardzo niewielkie zróżnicowanie narodowościowe i grupowe. Na 13 prób tylko jedna nie była amerykańska (izraelska). Wszystkie próby to próby studenckie, choć 6 z tych prób to próby na studentach MBA (nieco starsza i bardziej doświadczona zawodowo grupa niż „zwykli” studenci uniwersyteccy). Domyślam się, że wynika to prędkiej z tego, że tacy akurat studenci byli dostępni dla badaczy (dodam, że duża część badań była opublikowana nie w czasopismach *stricte* psychologicznych, ale z obszaru nauki o zarządzaniu i innych okołoeconomicznych).

Również wzór wyników jest niejednoznaczny. Z 11 badań replikujących, tylko 5 jednoznacznie potwierdza hipotezę, w 3 przypadkach nie znaleziono różnic, a w pozostałych przypadkach wyniki zostały potwierdzone tylko częściowo. Brak zgodności wyników nie jest fundamentalnie wadą, wydaje się jednak, że spore zróżnicowanie tematycznego może odpowiadać za taki stan rzeczy. Badania miały wspólne tylko dwa mianowniki: negocjacje i manipulacja rodzajem kontaktu, poza tym przekrój tematów i szczegółowych rozwiązań był dosyć szeroki. Dodam, że w tego rodzaju badaniach widzę duże możliwości stosowania psychologii fałszywie pozytywnej, szczególnie, że często negocjuje się wiele aspektów końcowego porozumienia.

Po raz kolejny nie odnotowałem też żadnych replikacji dokładnych. Wartość poznawcza całego cyklu badań jest więc dosyć niska.

6.2.33. Hewstone, M., Hamberger, J. (2000). Perceived variability and stereotype change. *Journal of Experimental Social Psychology*, 36(2), 103-124. (JESP3)

W tym badaniu eksperymentalnym stwierdzono, że gdy osoby sądzą, że członkowie określonej grupy społecznej są mało zróżnicowani pod kątem jakiejś cechy, wówczas dostrzegane większe zróżnicowanie powoduje mniejszą tendencję do stereotypizowania, a dostrzegane mniejsze zróżnicowanie większą tendencję do stereotypizowania. W sytuacji, gdy osoby myślą, że członkowie jakiejś grupy są bardziej zróżnicowani, wówczas spostrzegane zróżnicowanie nie ma dużego wpływu na stereotypizowanie.

W oryginalnym badaniu wzbudzano w badanych przekonanie, że określona grupa jest zróżnicowana (lub nie jest), a następnie przedstawiano rzekomych członków tej grupy z opisami. Manipulowano tym, czy była to grupa dosyć jednorodna (mieli podobne cechy) czy nie. Badani oceniali cechy członków tej grupy i ich „typowość” jako członka danej grupy.

Bardzo zbliżone badanie [01] przeprowadził ten sam zespół i uzyskał podobne wyniki. Poza tym 3 dalsze artykuły (6 badań), jak się wydaje, pośrednio testują wyjściową hipotezę. W pierwszym stwierdzono, że większe zróżnicowanie jakości kontaktów z dziadkami powoduje bardziej negatywne postawy względem osób starszych [01] (byłoby to więc zaprzeczenie oryginalnej hipotezy). Wyniki drugiego

[02] nie były jednoznaczne, w pierwszym badaniu okazało się, spostrzegane różnicowanie grup nie wpływa na poziom uprzedzeń, w drugim wyniki były odwrotne. Różnice te mogły wynikać, że testowano za pierwszym razem populację studentów, za drugim razem populacją ogólną.

W trzecim badaniu [03] stwierdzono, że eksperymentalnie manipulowane spostrzegane różnicowanie grup wpływa na poziom deklarowanych uprzedzeń (w dwóch pierwszych badaniach). W trzecim badaniu zmienna zależna miała charakter *strictly* behawioralny i to badanie zasługuje na dokładniejsze omówienie. Po skończonej procedurze manipulacji poczuciem różnicowania grupy (tutaj: grupy Arabów; badanymi byli francuscy studenci) badani opuszczali laboratorium. Spotykali na swojej drodze arabską studentkę, której rozsypały się książki. Zmienną zależną była decyzja o udzieleniu pomocy. Jest to jeden z nielicznych eksperymentów „terenowych” w omawianym tu cyklu. (Jedynie wątpliwości budzi fakt, że badacze określili, że badany student musiał zacząć udzielać pomocy w ciągu 20 sekund od dostrzeżenia sytuacji. Załóżmy, że badacze mają prawa do takiej arbitralnej decyzji, ale powinni jeszcze określić, czy po takim czasie ktoś jeszcze pomagał i jak by to ewentualnie zmieniło wyniki).

Oceniając całościowo cykl 9 badań można powiedzieć, że liczba badań nie była duża, ale były one względnie zróżnicowane. A więc występowały zarówno eksperymenty, jak i badania korelacyjne, oraz jeden eksperyment terenowy (jedyne wśród wszystkich badań z wyjątkiem JPSP9). Wyniki nie były jednoznaczne, ale oczywiście nie jest to żadna wada. Prawie zawsze (z wyjątkiem jednej próby) badano studentów, na szczęście nie byli to tylko studenci amerykańscy (ściślej, Amerykanów rekrutowano tylko do jednego badania, poza tym badanymi byli studenci francuscy, brytyjscy i niemieccy). Zaznaczę jednak, że badań, które bezpośrednio testowały hipotezę, poza oryginalnym, było tylko jedno. Pozostałe testowały tylko niektóre jej aspekty (tj. różnicowanie, zaindukowane eksperymentalnie lub mierzone kwestionariuszowo, ale już nie spostrzegane różnicowanie tu i teraz).

Wartość poznawcza cyklu jest więc umiarkowana, przede wszystkim ze względu na nieliczne próby dokładnych lub wysoce zbliżonych replikacji. Szkoda też, że badacze nie wykorzystywali częściej zmiennych behawioralnych jako zmiennych zależnych. Pomysł z operacjonalizacją poziomu uprzedzeń jako gotowością do udzielenia

pomocy osoby na korytarzu jest niebanalny, a przy tym, jak się wydaje, nie jest ani skomplikowany praktycznie, ani bardzo problematyczny pod względem etycznym.

6.2.34. Monteith, M. J., Spicer, C. V. (2000). Contents and correlates of whites' and blacks' racial attitudes. *Journal of Experimental Social Psychology*, 36(2), 125-154. (JESP4)

Badanie to dotyczy treści współczesnych postaw rasistowskich u osób białych i czarnych w USA. Jest ono szczególnie interesujące, z tego względu, że zdaje się realizować postulat Rozina (2001) o wprowadzenie większej ilości danych opisowo-obszernych. Procedura polegała na tym, że badani pisali esej zaczynający się od słów "Moje nastawienie do osób czarnych [białych] jest generalnie pozytywne/negatywne, ponieważ...". Eseje analizowano następnie jakościowo i klasyfikowano główne wątki. Zasadniczy wniosek z analizy był taki, że treścią współczesnych postaw rasistowskich białych jest przekonanie o większej przestępczości czarnych i ich postawie roszczeniowej, natomiast treścią postaw rasistowskich czarnych jest przekonanie, że biali utrudniają rozwój czarnych.

Dużą zaletą jest ponadprzeciętnie duża próba (ponad 700 osób), szczególnie biorąc pod uwagę pracochłonność tego typu badań (niestety, badanymi byli tylko studenci, a jak już wspominałem forma uprzedzeń w tej grupie niekoniecznie odzwierciedla uprzedzenia ogółu populacji por. Henry, 2008).

Największą zaletą takiego rodzaju badania jest jego otwartość, tj. badani mogą wskazać wątki, o których badacze mogliby nie pomyśleć, gdyby badanie miało typową formę ustosunkowania się do zamkniętego zestawu pytań kwestionariuszowych. Ograniczeniem takiej techniki są ograniczenia ludzkiej zdolności do samooceny własnych postaw (por. rdz. 2.4), jednak wydaje się, że treść postaw rasistowskich ogółu populacji trudno zbadać w inny sposób ze względów praktycznych i etycznych.

Spośród 9 artykułów (zawierających 15 badań) w jednym [07] pozwolono badanym wygenerować własne odpowiedzi, jakie są postawy innych ras względem własnej; w innym [02] pytano o przyczyny dyskryminacji. W kolejnym badaniu [01:pilotaż] pula odpowiedzi posłużyła do skonstruowania narzędzia kwestionariuszowego użytego w dalszej części (jednak wyniki tego pilotażu nie są omówione, więc nie jest on

uwzględniony w załączniku 2). W pozostałych przypadkach badani odpowiadali na zdefiniowane wcześniej przez badacza pytania dotyczące własnych uprzedzeń oraz postaw i zachowań innych grup rasowych (gdziekolwiek poszukiwano korelatów określonych postaw i w takim wypadku kwalifikowałem badanie jako korelacyjne).

Omówienie wzoru wyników nie jest w tym miejscu istotne (dosyć dokładnie są omówione w załączniku). W tym miejscu warto wspomnieć, że w całym cyklu (tj. badanie oryginalne + 15 badań replikujących) badani byli prawie zawsze Amerykanie (z wyjątkiem jednego badania z Kanadyjczykami). Badanych rekrutowano najczęściej wśród studentów (10 na 16), w pozostałych przypadkach wyszukiwano ich w populacji ogólnej.

Ogólnie więc, przeprowadzono dosyć dużo badań z wykorzystaniem dużej grupy (ponad 3 tys. osób). Nie uważam jednak, żeby wartość całego cyklu była wysoka. Przede wszystkim korzystano najczęściej z odpowiedzi studentów. Po drugie, badani prawie zawsze odpowiadali na zdefiniowany zbiór pytań. Po trzecie, poza ogólnym wątkiem (treść postaw rasistowskich) badania te są ze sobą luźno powiązane, szczególnie należy odnotować, że występuje dosyć duże zróżnicowanie typów używanych narzędzi psychometrycznych.

Tak jak wspominałem, w tym temacie trudno ze względów etycznych i praktycznych pozyskiwać inne dane niż samoopisowe, szczególnie behawioralne. Być może jednak warte rozważenia byłoby poszukanie danych też samoopisowych, ale nie wytworzonych w kontekście badania psychologicznego. Najbardziej przychodzą na myśl rozmaite wypowiedzi, jakie ludzie tworzą na forach, w komentarzach pod artykułami lub na portalach społecznościowych. Uzupełnieniem mogłyby być treści wytwarzane przez publicystów lub działaczy politycznych. Oczywiście, niekoniecznie muszą odzwierciedlać poglądy ogólnej populacji, jednak wydaje się, że takie grupy (szczególnie politycy) mówią rzeczy, które są bliskie ich słuchaczom.

6.2.35. Croizet, J., Fiske, S. T. (2000). Moderation of priming by goals: feeling entitled to judge increases judged usability of evaluative primes. *Journal of Experimental Social Psychology*, 36(2), 155-181. (JESP5)

Opisuje paradoksalny efekt związany z procedurą prymowania (podobnie jak JESP7). Polega na tym, że poczucie kompetencji („uprawomocnienia”) do

wydawania jakichś sądów zwiększa wpływ prymowania na późniejsze wydawanie sądów. Paradoks polega tu na tym, że kompetencja sama w sobie (i jej trafne rozeznanie) powinno zmniejszać wpływ niemerytorycznych czynników (w tym właśnie prymowania). Poczuciem kompetencji manipulowano w ten sposób, że badanych informowano, że otrzymali wysokie wyniki w teście „zdolności percepcji osobowości” (Study 1) lub poprzez nieprawdziwą instrukcję, że badani rzekomo otrzymali istotne do podjęcia sądu informacje podprogowo (w ignorowanym kanale w procedurze słuchania dwuosobowego; Study 2).

W badaniu oryginalnym uczestniczyli studenci, odpowiednio amerykańscy i francuscy. Kratki były bardzo małe (odpowiednio ok. 10 i 16), co budzi pewien niepokój o występowanie psychologii fałszywie pozytywnej.

Miały miejsce dwie replikacje konceptualne tego efektu. W pierwszej [01] badanym najpierw eksponowano same twarze osób. Badani w drugiej fazie mieli następnie większe tendencje do stereotypizowania (tj. opisywania osoby określonej daną kategorią w zgodzie ze stereotypami dotyczącymi tej kategorii) w przypadku osób, których twarze wcześniej oglądali. Można założyć, że być może jest to rodzaj poczucia uprawomocnienia (poprzez przekonanie badanych, że lepiej znają osoby, których twarze widzieli). Nadmienię, że współautorem tego badania był Stapel, choć komitet nie wykazał tu fałszerstwa (por. Levelt, 2012, s. 69).

W drugim [02] stwierdzono najpierw, że wiedza na temat książki zwiększa tendencję do oceniania jej jako ważną. Wiedzą manipulowano poprzez prezentowanie tylko przedniej lub tylnej strony okładki (na tylnej były dodatkowe informacje o książce i autorze). Następnie w kolejnym badaniu w tym samym artykule stwierdzono, że deklarowana znajomość książki powoduje, że dokładanie odważnika do książki (i zwiększanie jej wagi) wywołuje ocenienie książki jako ważnej.

Zaznaczę też, że można dyskutować, czy opisane tu replikacje konceptualne rzeczywiście badają ten sam efekt.

Ogółem więc cały cykl liczył 5 badań eksperymentalnych i wzięło w nim udział 517 osób. Był więc on stosunkowo nieduży, choć plusem jest to, że hipotezę weryfikowano w konceptualnie bardzo inny sposób. Wadą takiego podejścia jest niewiedza, czy ciągle była to ta sama hipoteza. Poza tą różnorodnością badania nie

wyróżniają się niczym szczególnym, każdorazowo używano studentów jako osób badanych, z czego raz byli to studenci holenderscy.

Ogólnie więc, pomimo interesującej procedury, wiedza jest mała.

6.2.36. MacDonald, G., Zanna, M. P. (2000). An experimental test of the role of alcohol in relationship conflict. *Journal of Experimental Social Psychology*, 36(2), 182-193. (JESP6)

Badanie z pogranicza psychologii społecznej i klinicznej. Stwierdzono, że wypicie alkoholu zmienia ocenę konfliktów z partnerem romantycznym, w szczególności poprzez większe obwinianie partnera, ocenę emocji partnera jako bardziej negatywnych i większą niepewność co do uczuć partnera. Zarazem stwierdzono, że efekty te występują tylko u osób o niskiej samoocenie (co jest logiczne, choć miejmy nadzieję, że badacze nie sprawdzali samooceny przy pomocy pięciu skal, żeby wybrać tę która daje potwierdzenie efektu).

Sam eksperyment miał wysoce realistyczną manipulację: badani rzeczywiście pili alkohol (lub placebo w grupie kontrolnej; w trzeciej grupie nic nie pili). Zmienne zależne są już bardziej „typowe” (tj. ocena hipotetycznych zdarzeń; wszystko metodą kwestionariuszową). W artykule omówiono tylko jedno badanie, ogółem próba liczyła 56 osób (nieco mniej niż 19 w jednej grupie), badanymi byli studenci kanadyjscy.

Hipotezy o wpływie związku alkoholu i oceny konfliktów w związkach romantycznych testowano następnie w 5 artykułach (samo badanie było cytowane 27 razy).

Wszystkie one mają charakter korelacyjny. Występuje w nich zerowe zróżnicowanie międzykulturowe (wszystkie zrobione w USA), ale za to spore zróżnicowanie grupowe. I tak, badanymi byli alkoholicy, osoby nadużywający alkoholu, żołnierze armii amerykańskiej, oraz populacja ogólna. Tylko w jednym tekście badani byli studenci.

Znaleziono trochę danych wskazujących pośrednio na to, że wyjściowa hipoteza jest prawdziwa. M.in. odkryto, że:

- alkoholicy przed konfliktem z użyciem przemocy częściej spożywali alkohol i w większych ilościach (niż przy konfliktach bez przemocy) [01];

- żołnierze amerykańscy, którzy byli sprawcami przemocy, częściej też nadużywali alkoholu [02]; wykorzystano tutaj dane z rejestrów armii stąd bardzo duża liczba osób badanych (ponad 25 tys.);

- osoby o niskiej samoocenie piły więcej w dni, w które doświadczały negatywnych wydarzeń związanych z partnerem romantycznym [03];

- w badaniu podłużnym stwierdzono mianowicie, że picie, szczególnie picie męża, przewidywało mierzone rok później problemy małżeńskie, szczególnie agresję werbalną i fizyczną oraz wycofywanie się małżonków z bliskich kontaktów [04].

Powyższe badania potwierdzały hipotezę. Jedno badanie dało niejednoznaczne wyniki [05], stwierdzono, że intensywne picie zmniejszało poziom bliskości, ale już nie umiarkowane, oraz że wspólne picie paradoksalnie podnosiło poziom bliskości (badani przez kilka miesięcy prowadzili dzienniczki, w których notowali istotne zmienne). Warto jednak zaznaczyć, że w tym badaniu uczestniczyli studenci, o których nie wspomniano, że mieli problemy z piciem, stąd wyniki mogą odbiegać od wcześniejszych.

Ogólnie więc cykl badań cechował się dosyć dużym zróżnicowaniem źródeł danych (i grup badanych), choć liczba badań jest niestety dosyć mała. Poza oryginalnym badaniem nie testowano tych zależności w sposób eksperymentalny. Szczególnie ważne byłoby przeprowadzenie badania o dużo większej mocy. Z kolei szczególnie interesujące mogłoby być zaproszenie do badania jako ochotników par romantycznych, które po spożytym alkoholu (lub placebo) mogłyby omawiać swoje problemy w związku. Wydaje się, że w przypadku tej hipotezy istnieje wielka różnorodność możliwych metod weryfikacji. Społeczność badawcza decydowała się tylko na metody korelacyjne, co ma oczywiście takie ograniczenia, że nie pozwala na wnioskowanie z dużą dozą pewności o zależnościach przyczynowo-skutkowych.

6.2.37. Mussweiler, T., Neumann, R. (2000). Sources of mental contamination: comparing the effects of self-generated versus externally provided primes. *Journal of Experimental Social Psychology*, 36(2), 194-206. (JESP7)

Artykuł opisuje interesujący, choć mocno techniczny, efekt polegający na tym, że prymy wygenerowane przez samych badanych wywierają większy wpływ na późniejsze oceny (społeczne) niż prymy podane z zewnątrz (np. przez

eksperymentatora). Na wyższym poziomie więc badanie to ma dowodzić, że badani są świadomi wpływu zewnętrznych prym na oceny, korygują te oceny o znany sobie wpływ. Generowanie prym polegało na podawaniu antonimów do podanych słów, lub rozwiązywaniu anagramów. Dalsza część była już mocno standardowa (ocena niejednoznacznych zachowań opisanej osoby). Również pod kątem próby badanej (studenci z zachodniego państwa) nie było tu żadnej oryginalności. Uwagę zwraca to, że w pierwszym badaniu wielkość kratki wynosiła tylko 8 osób! W drugim eksperymencie było to już osób 20.

Można więc powiedzieć, że jest to typowe badanie psychologii społecznej. Badanie to nie miało żadnej bezpośredniej replikacji. Wśród tekstów cytujących znalazłem natomiast badanie [01] konceptualnie nawiązujące do hipotez z oryginalnego badania, które przy okazji pokazuje „życiowe” zastosowanie omawianego efektu. W tym eksperymencie osoby badane, tj. studenci będący palaczami, generowały listę argumentów przeciwko paleniu tytoniu. Natomiast grupa kontrolna miała podawany przez badacza zestaw argumentów. Badani czekali na drugą część eksperymentu w poczekalni, w której mieli możliwość zapalić. Zmienną zależną była decyzja o zapaleniu papierosa w trakcie 30 minut oczekiwania. Okazało się, że osoby, które generowały samodzielnie listę dłużej (lub w ogóle) powstrzymywały się od zapalenia. Stwierdzono także, że manipulacja nie miała wpływu na deklarowaną postawę względem palenia.

Osobiście odnoszę się bardzo pozytywnie do tego badania, gdyż przetestowano tu typową zmienną behawioralną. Przy okazji autorzy pokazali, że taka zmienna może być naprawdę bardzo prosta. Mimo wszystko, jak pokazuje moja całościowa analiza, takie zmienne są rzadko spotykane. Obraz całości psuje nieco fakt, że moc najprawdopodobniej nie jest zbyt wysoka (w dwóch badaniach odpowiednio 12 i 20 osób w kratkach).

Więc poza tą pozytywną „wisienką” (pod względem metody) cykl tych badań dostarczył minimalnej wiedzy o wyjściowym efekcie.

6.2.38. Jost, J. T., Thompson, E. P. (2000). Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and

European Americans. *Journal of Experimental Social Psychology*, 36(3), 209-232. (JESP8)

Kolejny artykuł (po EJSP1), w którym główną uwagę skupiono na skali „Social Dominance Orientation” (skala dominacji społecznej). Stwierdzono tutaj, że skala ta nie jest jednorodna i lepiej opisują ją dwa czynniki: sprzeciw wobec równości (grup społecznych: SWR) i prawo grup do dominowania (PGDD). Wskazano też na różnice w ich nasileniu i związkach z innymi zmiennymi u różnych grup. Np. sprzeciw wobec równości był negatywnie skorelowany z poczuciem własnej wartości i etnocentryzmem u osób czarnych, a pozytywnie u białych. Prawo grup do dominowania było pozytywnie skorelowane z neurotyzmem u osób białych, ale negatywnie u osób czarnych itd.

W oryginalnym badaniu efekt pokazano w 4 niezależnych próbach (choć wszystkie były studenckie), łącznie przebadano ponad 1600 osób. Następnie w 7 badaniach (opisanych w 4 artykułach) badacze replikowali tę hipotezę. Ten sam zespół na dużej próbie (ponad 3,5 tys. osób) zreplikował swoje wyniki [07]. Zespół niezależnych badaczy uzyskał zbliżone wyniki [06], natomiast inny zespół zaproponował strukturę 3-czynnikowej skali [02]. Wreszcie jeden zespół [03] wskazał na możliwość występowania artefaktów, mianowicie, że dwuczynnikowy model skali orientacji na dominację społeczną może wynikać ze specyficznych właściwości pozycji testowych wchodzących w skład poszczególnych podskal: PGDD ma tylko zdania twierdzące, a SWR tylko zdania przeczące.

W dalszych 8 badaniach (w 3 artykułach) badacze próbowali szerzej zrozumieć nowe skale, niejako zakładając *a priori* hipotezę o występowaniu dwóch czynników (np. znajdować dodatkowe korelaty tych dwóch zmiennych). Nie ma tu potrzeby wyliczać listy znalezionych korelatów, odsyłam do załącznika z bardziej szczegółowymi informacjami.

Choć w oryginalnym badaniu występowały próby tylko studenckie, to już w replikacjach/rozwinęciach dobór badanych był bardziej zróżnicowany. Próby studenckie występowały w 7 badaniach, 3 próby zrekrutowano wyłącznie w Internecie, 3 stanowili pracownicy jednej korporacji (wszystkie próby w [02]), oraz w badaniu [01] badanych rekrutowano z populacji ogólnej na lotnisku. Dodatkowo w badaniu [07] było kilka prób (m.in. studenci, populacja ogólna oraz badani

rekrutowani w Internecie). Zróżnicowanie narodowościowe było niestety dużo mniejsze. W trzech próbach (z jednego artykułu [02]) badanymi byli obywatele Państwa Środka, a w kolejnym tekście [03] była to jedna z dwóch grup (obok obywateli USA). Poza tym wszyscy badani wywodzili się z USA, lub, w przypadku badań internetowych można się domyślić, że stamtąd (i ewentualnie innych krajów anglojęzycznych).

Zróżnicowanie metodologiczne również jest zerowe. Wszystkie badania miały charakter kwestionariuszowy i korelacyjny. Próby były dosyć duże, jak to często bywa w tego typu badaniach. Całkowita liczba badanych była bardzo duża, ponad 20 tys. w całym cyklu badań, z czego prawie połowa pochodziła tylko z jednego dużego badania internetowego [04].

Cykl badań wydaje się więc nie odbiegać jakoś zasadniczo od typowych badań psychometrycznych. Warto jednak zauważyć, że w żadnym wypadku nie próbowano zweryfikować hipotez inną drogą niż badania kwestionariuszowe. Jeżeli zakładany efekt ma miejsce i faktycznie dwa rodzaje orientacji na dominację społeczną są od siebie względnie niezależne, powinno to objawiać się w obserwowalnych zachowaniach. Jednym ze sposobów nie-kwestionariuszowej weryfikacji mogłaby być analiza programów partii politycznych pod kątem tych dwóch zmiennych. Wytworów ludzkich do analizy pod kątem tej hipotezy może być zresztą bardzo wiele (np. wypowiedzi na forach internetowych, artykuły prasowe itp.). Również interesującym sposobem weryfikacji tych hipotez byłoby stworzenie odpowiednich tekstów, w których uwypukla się jeden aspekt np. sprzeciw wobec (polityki) równości, a równocześnie odrzuca drugi (czyli prawo grup do dominowania). Jeśli są to faktycznie odrębne byty w ludzkich głowach, w takim razie takie teksty powinny budzić względnie mniejszy dysonans i zgoda z tymi dwoma rodzajami przekazu powinna być tylko umiarkowana skorelowana.

6.2.39. Johnson, J. D., Trawalter, S., Dovidio, J. F. (2000). Converging interracial consequences of exposure to violent rap music on stereotypical attributions of blacks. *Journal of Experimental Social Psychology*, 36(3), 233-251. (JESP9)

Badanie opisuje interesujący efekt z obszaru uprzedzeń rasowych. Polega on na tym, że słuchanie agresywnego czarnego rapu prowadzi następnie u (białych)

badanych do oceny niejednoznacznych zachowań osób czarnych jako agresywnych lub mało inteligentnych. W warunku kontrolnym badani słuchali również muzyki z gatunku rap, ale bez przekleństw i odniesień do przemocy (była również druga grupa kontrolna, która nie słuchała żadnej muzyki). W doborze zmiennych zależnych również użyto swego rodzaju „zmiennej kontrolnej”, nie zanotowano istotnych różnic w ocenie zdolności do wykonywania operacji przestrzennych przez opisywaną czarną osobę. Analogicznie taka manipulacja eksperymentalna nie miała wpływu na ocenę inteligencji i agresywności u osoby białej.

Poza ciekawą hipotezą oryginalne badanie nie miało jakichś szczególnych wyróżników: badanymi byli amerykańscy studenci i grupy badane nie były zbyt liczne (15 osób w jednej kratce).

Kilka badań było replikacjami konceptualnymi efektu, choć żadna nie badała efektu *sensu stricte* (tj. wpływu słuchania rapu na ocenę inteligencji i agresywności). I tak w replikacjach wykazano, że:

- słuchanie seksistowskiego rapu nie miało zasadniczego wpływu na deklarowany poziom seksizmu [01];
- słuchanie rapu o wysokim natężeniu treści seksualnych powoduje następnie ocenianie niejednoznacznych zachowań czarnej kobiety jako rozwiązłych seksualnie [02];
- słuchanie czarnego rapu, który odwołuje się do rozwiązłych zachowań seksualnych, wpływa na postawy badanych dotyczącego tego, jaka powinna być wysokość zasiłków przyznawanych czarnym (niższa) ([04] i [06]).

Stwierdzono też, przy wykorzystaniu fMRI, że u badanych, którzy wcześniej wysłuchali czarnego rapu, pod wpływem ekspozycji twarzy czarnych osób następuje większa aktywacja ciała migdałowatego (czyli tzw. mózgu emocjonalnego). Jest to jedyne badanie w całej mojej analizie, w której wykorzystano zmienne *stricte* biologiczne [07].

W jakiś sposób jako konsekwencje teorii można uznać odkrycia w badaniach, w których stwierdzono, że młodzi czarni adolescenci, którzy słuchają czarnego rapu mają bardziej negatywne postawy względem relacji damsko-męskich [03]. Jako luźne nawiązanie do oryginalnej hipotezy uznałem też badanie, w którym stwierdzono, że u

osób białych negatywna postawa względem rapu jest skorelowana z uprzedzeniami rasowymi [05].

Prawie wszystkie badania rozwijające wyjściową hipotezę miały miejsce w USA (z wyjątkiem [05:3], które odbyło się w Wielkiej Brytanii), oraz z wykorzystaniem prób studenckich (z wyjątkiem [03] i [05]).

Tak więc mamy tu do czynienia z dosyć częstymi replikacjami konceptualnymi, dotyczącymi innych „tematów”, ale z brakiem replikacji w miarę zbliżonej, która badałaby wpływ słuchania rapu na ocenę agresywności i inteligencji. Replikacje konceptualne, jak już wspominałem wcześniej, pozwalają wzmocnić wyjściową hipotezę, ale nie prowadzą do jej falsyfikacji. Porażka takiej replikacji (tak jak w [01]) nie pozwala stwierdzić, czy oryginalny efekt nie zachodzi, czy po prostu oryginalny efekt nie dotyczy tych warunków, jakie przeprowadzono w replikacji.

Cykl badań jest niewątpliwie niezłym wskaźnikiem, że istnieje jakiś wpływ słuchania niektórych gatunków muzyki na percepcję społeczną. Słabo wyjaśnione jest kluczowe pytanie, czy taki wpływ wynika z tylko werbalnej treści czy jest jeszcze potrzebny określony rodzaj muzyki (a więc czy np. agresywne wypowiedzi w połączeniu z muzyką poważną przyniosłyby zbliżony efekt). Nie jest też jasne, na ile efekt ten jest warunkowy kulturowo, stąd byłoby interesujące przebadanie go na członkach kultur, które nie mają zasadniczej styczności z muzyką rap.

Z tych względów oceniam więc wartość poznawczą całego cyklu badań jako relatywnie niską.

6.2.40. Idson, L., Liberman, N., Higgins, E. (2000). Distinguishing gains from nonlosses and losses from nongains: a regulatory focus perspective on hedonic intensity. *Journal of Experimental Social Psychology*, 36(3), 252-274. (JESP10)

Tekst opisuje interesujący efekt polegający na tym, że pozytywne i negatywne zdarzenia mogą być różnie obramowane i w związku z tym będą wywoływać inne emocje. I tak, negatywne wydarzenie można obramować jako stratę lub brak zysku (pozytywne jako zysk lub brak straty). Przykładowo, jeśli chcemy zakupić książkę za 65 zł i przy kasie dowiadujemy się, że możemy uzyskać 5 zł rabatu za zapłatę gotówką, ale gotówki akurat nie mamy, to mamy do czynienia z „brakiem zysku”.

Analogicznie, jeśli książka kosztuje 60 zł, ale przy kasie dowiadujemy się, że musimy dopłacić 5 zł za płatność kartą, w takim wypadku mamy do czynienia ze „stratą”. Pod względem finansowym są to identyczne sytuacje, ale budzą emocje o różnym nasileniu. W tym badaniu pokazano, że zysk budzi intensywniejsze pozytywne emocje niż brak straty, a strata intensywniejsze emocje negatywne niż brak zysku (w chwili, gdy piszę te słowa w Irlandii odbywają się demonstracje przeciwko wprowadzeniu opłat za wodę, która wcześniej była darmowa; jednak w Polsce nikt nie protestuje, żeby rząd wprowadził darmową wodę; jest to bardzo życiowe potwierdzenie rzeczonych prawidłowości). Efekt oczywiście ma daleko idące implikacje, szczególnie w obszarze handlu (czy sprzedawca będzie „dawał rabat” za jakieś zachowanie, czy raczej „wymagał dopłaty” za brak jakiegoś zachowania), ale nie tylko, czego dowodzi przykład darmowej wody w Irlandii, albo omawiane za chwilę badanie dotyczące komunikatów prozdrowotnych.

Oryginalny artykuł opisywał 4 eksperymenty. W 3 z nich manipulacja była mało realistyczna (badani wyobrażali sobie określone sytuacje i deklarowali swoje reakcje na nie), w jednym manipulacja miała bardziej realistyczny charakter (badani grali w grę,- w której albo odejmowano im punkty z wyjściowej puli, albo zaczynali od zera i dodawano im punkty). Każdorazowo zmienna zależna miała charakter samoopisowy.

Autorzy tego badania opublikowali następnie 2 teksty z 5 badaniami, które były replikacjami konceptualnymi i 1 badaniem, które było replikacją dokładną (co, jak już wiemy, występuje niezmiernie rzadko). Dodatkowe 2 artykuły (z 5 replikacjami konceptualnymi) opublikowały niezależne zespoły badawcze. Każde opublikowane badanie potwierdza wyjściowe hipotezy. Na ogół badania te mają mało realistyczny charakter (opis określonej sytuacji plus polecenie ustosunkowania się do nich przez badanych), i dotyczą prawie zawsze sytuacji okołofinansowych (kupno-sprzedaż, lub negocjacje). Wyjątkiem jest badanie [04], w którym zmienną zależną były intencje behawioralne w kierunku działań prozdrowotnych, gdzie stwierdzono, że najbardziej na nie wpływają wiadomości przedstawione w formie zysków, nieco mniej w formie strat, a najmniej w formie braku zysków lub braku strat.

Niestety, w całym cyklu piętnastu badań (łącznie z oryginalnym) jako osoby badane służyli studenci. W zdecydowanej większości byli to studenci amerykańscy. W dwóch badaniach [04] byli to studenci holenderscy.

Ogółem więc, sam oryginalny efekt jest dosyć szeroko przetestowany. Nie ma też oczywistych wskaźników psychologii fałszywie pozytywnej (np. średnia kratka w tym cyklu badań jest dosyć wysoka ok. 27 osób). Niestety, ale wartość poznawczą tego cyklu badań ograniczają praktycznie zerowe zróżnicowanie grupowe i narodowościowe grupy badanej, oraz wykorzystywanie prawie zawsze nierealistycznych manipulacji eksperymentalnych w połączeniu ze zmiennymi o naturze samoopisowej. Tym bardziej, że jak się wydaje, istnieje potencjalnie wiele możliwości bardziej realistycznych badań. Najprostsze, jakie przychodzi na myśl, to sytuacja, w której badani „dostają”, jakiś upominek od badacza, ale potem „tracą” go wskutek jakiegoś niepowodzenia (elementu procedury eksperymentalnej). Żal ze straty (i np. gotowość odkupienia) powinny być wyższe niż w sytuacji polegającej na tym, że badani upominku nie otrzymali na początku, ale też go „nie wygrali”.

6.3. Podsumowanie

W Tabeli 6.1. podsumowałem ten rozdział. Całościową ocenę praktyk badawczych w psychologii społecznej przedstawiam w rozdziale 7.

TABELA 6.1. Ocena poszczególnych cykli badawczych

	Stosowanie "psychologii fałszywie pozytywnej"?	zróżnicowanie grupowe	zróżnicowanie narodowościowe	ilość badań i danych	zróżnicowanie metod i źródeł danych	pewność wiedzy
JSP1	pewne wskaźniki	zerowe	zerowe	niska	bardzo niskie	bardzo niska
JSP2	pewne wskaźniki	zerowe	zerowe	bardzo niska	niskie	bardzo niska
JSP3	pewne wskaźniki	zerowe	zerowe	niska	niskie	bardzo niska
JSP4	mało prawdopodobna	niskie	zerowe	dosyć duża	dosyć duże	średnia
JSP5	pewne wskaźniki	zerowe	zerowe	bardzo niska	niskie	bardzo niska
JSP6	mało prawdopodobna	niskie	zerowe	dosyć duża	dosyć duże	średnia
JSP7	mało prawdopodobna	niskie	niskie	duża	średnie	średnia
JSP8	pewne wskaźniki	niskie	niskie	dosyć niska	średnie	niska
JSP9	mało prawdopodobna	średnie	zerowe	niski	niskie	niska
JSP10	mało prawdopodobna	duże	zerowe	niska	duże	niska
JESP1	mało prawdopodobna	średnie	średnie	średnio	dosyć niskie	dosyć niska/średnia
JESP2	pewne wskaźniki	niskie	niskie	średnio	niskie	dosyć niska
JESP3	pewne wskaźniki	niskie	dosyć niskie	niska	średnie	niska
JESP4	mało prawdopodobna	średnie	niskie	dosyć duża	niskie	dosyć niska
JESP5	pewne wskaźniki	zerowe	niskie	niska	średnie	niska
JESP6	pewne wskaźniki	średnie	niskie	niska	średnie	niska
JESP7	pewne wskaźniki	zerowe	niskie	niska	dosyć niskie	niska
JESP8	mało prawdopodobna	średnie	niskie	duża	niskie	średnia
JESP9	pewne wskaźniki	niskie	zerowe	dosyć niska	niska	niska
JESP10	mało prawdopodobna	zerowe	niskie	średnio	niskie	niska
PSPB1	pewne wskaźniki	zerowe	niskie	dosyć niska	niskie	niska
PSPB2	pewne wskaźniki	niskie	niskie	średnio	dosyć niskie	niska
PSPB3	pewne wskaźniki	zerowe	zerowe	niska	niskie	niska
PSPB4	pewne wskaźniki	zerowe	niskie	dosyć niska	niska	niska
PSPB5	mało prawdopodobna	niskie	średnie	duża	średnie	średnia
PSPB6	mało prawdopodobna	niskie	niskie	dosyć niska	niska	niska
PSPB7	pewne wskaźniki	niskie	zerowe	niska	niskie	niska
PSPB8	pewne wskaźniki	zerowe	zerowe	niska	dosyć niskie	niska

PSPB9	mało prawdopodobna	zerowe	niskie	dosyć niska	niskie	niska
PSPB10	pewne wskaźniki	zerowe	zerowe	niska	dosyć niskie	niska
EJSP1	mało prawdopodobna	średnie	średnie	duża	niska	średnia
EJSP2	pewne wskaźniki	zerowe	zerowe	niskie	niskie	niska
EJSP3	prawdopodobna	zerowe	zerowe	niski	niskie	bardzo niska
EJSP4	pewne wskaźniki	niskie	niskie	dosyć duża	duże	średnia
EJSP5	pewne wskaźniki	średnie	średnie	dosyć duża	niskie	niska
EJSP6	pewne wskaźniki	zerowe	zerowe	niska	niskie	niska
EJSP7	mało prawdopodobna	zerowe	niskie	niska	niskie	niska
EJSP8	pewne wskaźniki	zerowe	średnie	niska	niskie	niska
EJSP9	pewne wskaźniki	zerowe	zerowe	średnio	średnie	niska
EJSP10	pewne wskaźniki	niskie	niskie	niska	niska	niska

Zastosowane skale:

- Stosowanie "psychologii fałszywie pozytywnej"?; mało prawdopodobna; pewne wskaźniki
- „Zróżnicowanie grupowe” i „zróżnicowanie narodowościowe”: zerowe; niskie; dosyć niskie; średnie; dosyć duże; duże
- „Ilość badań i danych” oraz „zróżnicowanie metod i źródeł danych”: niska, dosyć niska; średnia; dosyć duża; duża
- Pewność wiedzy: bardzo niska; niska; dosyć niska; średnia; dosyć wysoka; wysoka

ROZDZIAŁ 7

PODSUMOWANIE ANALIZY ORAZ WNIOSKI NA PRZYSZŁOŚĆ

7.1. Wprowadzenie

Żeby lepiej zilustrować pewne tezy w dalszej części, na początku tego rozdziału przedstawię w uproszczeniu, jak rozwijała się część klimatologii poświęcona deszczowi, szczególnie przewidywaniu jego wystąpienia (w pewnej mierze opierałem się na Anderson, 2005 i Barnett, 2015).

Na początku rozwoju nauki badacze dysponują sporą liczbą danych anegdotycznych o przyczynach (a może tylko czynnikach poprzedzających?) deszczu. Dane te pozwalają zaprojektować badania korelacyjne, w trakcie których mierzy się te zmienne i sprawdza potem ich związek z opadami. Przy okazji mierzy się też sporo innych dodatkowych zmiennych, żeby sprawdzić, czy może coś nie koreluje systematycznie z interesującym nas zjawiskiem. Po pewnym czasie badacze dysponują już sporą liczbą nieźle udokumentowanych twierdzeń w rodzaju „w 84 % przed opadami deszczu występowało X, a gdy występowało X w 95 % przypadków padało”, albo „natężenie zmiennej X jest skorelowane z wielkością opadów na poziomie $r = 0,63$ ”. Jednak zbiór korelatów ich nie zadowala i chcą pogłębić rozumienie o mechanizmie fizyko-chemicznym, który powoduje opady. Przeprowadzają w ograniczonej skali eksperymenty w laboratorium, żeby zasymulować elementarne procesy (m.in. manipulując poziomem zmiennych, o których wiadomo, że są z deszczem skorelowane). Mając świadomość, że eksperymenty takie mogą być mało „trafne zewnętrznie” sprawdzają następnie czy wiedza tak uzyskana pozwala z większą precyzją przewidywać rzeczywiste opady. Przy okazji, dzięki osiągnięciom innych nauk, rozwijają się coraz precyzyjniejsze narzędzia pomiarowe, albo takie narzędzia, które pozwalają uzyskiwać dane zupełnie nowego rodzaju (np. zdjęcia z satelitów). Rozwój dziedziny matematycznego modelowania w połączeniu z coraz mocniejszymi komputerami pozwala na modelowanie interesujących zjawisk. Symulacje te są na początku wysoce niedoskonałe, ale są ulepszane dzięki temu, że są ciągle porównywane z rzeczywistością. Rozwój teorii pozwala na dostrzeżenie, które z wcześniejszych korelacji są pozorne, zarazem pozwala na szukanie związku deszczu ze zmiennymi,

o których wcześniej się nie myślało (pomaga niewątpliwie to, że zjawiska pogodowe są dosyć skrupulatnie rejestrowane na świecie). Z czasem prognozy są coraz trafniejsze i coraz precyzyjniejsze (z „w ciągu dnia będzie padać”, na „od godziny 8 do 9, spadnie X mm deszczu +/- 2 %”). Oczywiście, zjawiska pogodowe są systemem chaotycznym i niewielkie zmiany warunków początkowych znacząco zmieniają wydarzenia długookresowe. Ale rozwój nauki tutaj polega również na coraz większej świadomości błędu i wprowadzaniu go do modelu.

Ogólnie „nauka o deszczu” rozwija się dwoma ściśle powiązаныmi torami: z jednej strony mamy coraz więcej danych, coraz lepszej jakości, z drugiej coraz lepsze teorie wyjaśniające podstawowe mechanizmy. A wszystko to jest na bieżąco konfrontowane z rzeczywistością. Trafność i precyzja przewidywań są ostatecznym sprawdzianem rozwoju nauki.

Nauka o ludzkim zachowaniu ma w moim przekonaniu wiele cech podobnych. Zachowanie jest kształtowane przez dynamiczne systemy, w których wiele podsystemów i zmiennych wchodzi ze sobą w niezwykle złożone interakcje. Trudno je przewidywać, szczególnie w dłuższej perspektywie czasowej, nie tylko nie zna się wszystkich kluczowych zmiennych, ale też niewielkie różnice w warunkach początkowych radykalnie zmieniają wynik końcowy. Podobnie jak wielkich zjawisk atmosferycznych nie można dosłownie badać w laboratorium przy pomocy eksperymentów, tak samo nie można dosłownie badać zachowania. Nie tyle ze względu na skalę, ile ze względu na to, że nie można oddzielić czynników sytuacyjnych (którymi się manipuluje) od czynników dyspozycyjnych (które powinny, zgodnie ze sztuką eksperymentalną zawsze pozostać stałe, a tak się nie dzieje), albo wpływu samej sytuacji badania empirycznego na wynik końcowy.

Puszczę teraz wodze fantazji i przedstawię podobną hipotetyczną (nieco wyidealizowaną) historię rozumienia jakiegoś efektu psychologicznego. Załóżmy, że chcemy wyjaśnić, co wpływa na zadowolenie ze związku romantycznego. Dajmy na to, że badacze na podstawie anegdotycznych informacji dowiedzieli się, że wpływ mają wspólne ekscytujące przeżycia. W tym celu przeprowadzają serię eksperymentów i badań korelacyjnych (takich jak w JPSP10). W badaniu tym są duże próby i nie ma żadnych oczywistych cech, wskazujących na to, że badanie zostało przeprowadzone w sposób nieprawidłowy lub jest wynikiem przypadku. Wiemy jednak, że w psychologii czasami trafiają się różne artefakty, bywa też, że

przypadek może wypaczyć wyniki, a i sporadycznie niestety trafiają się oszuści. Dlatego 2 zespoły przeprowadzają niezależne replikacje na podobnych próbach. Są to replikacje o bardzo wysokiej mocy, bowiem badacze Ci wiedzą, że ich wyniki będą opublikowane ze względu na ich wysiłek (przypominam, że to historia „idealna”). Wyniki są potwierdzone w każdym przypadku. A więc 3 niezależne zespoły potwierdziły te wyniki na dużej próbie badanej. Jest to naprawdę mocny wskaźnik, że w efekcie jest coś na rzeczy.

Idąc dalej: wiadomo jednak, że występują w niektórych obszarach zachowania znaczące różnice międzykulturowe. Idealnie byłoby przetestować efekt na ludziach ze wszystkich krajów świata, ale jesteśmy realistami. Dlatego udane przeprowadzenie dużej międzykulturowej replikacji na członkach trzech, dosyć zróżnicowanych, kultur daje nam wysoki stopień pewności, że efekt ten nie jest rodzajem artefaktu występującego tylko u Amerykanów.

Efekt jest rozwijany dalej, próbuje się innych sposobów manipulacji „ekscytacją”, nie tylko poprzez wspólne ćwiczenia fizyczne. Manipuluje się poziomem natężenia „ekscytacji”, żeby sprawdzić, czy przypadkiem ekstremalnie wysokie jej poziomy nie powodują znaczącego spadku zadowolenia. Poszukuje się dalszych korelatów sytuacyjnych i dyspozycyjnych tego efektu. Liczba potencjalnych takich korelatów jest oczywiście nieskończona, dlatego dzięki wywiadam i obserwacjom par żyjących w ekscytujących (i zadowolonych) związkach tworzy się potencjalne hipotezy do empirycznego przetestowania (przy okazji sprawdza się też czy najlepiej przebadane cechy dyspozycyjne mają wpływ na rzeczony efekt). W miarę możliwości odkrycia powtarza się poprzez niezależne replikacje, oraz w innych grupach kulturowych i społecznych. Część z tych replikacji jest nieudana, ale na szczęście są one opublikowane, więc inni badacze nie będą w przyszłości brnęli niepotrzebnie w ślepe zaułki.

Równolegle badacze starają się uzyskać dane „pozalaboratoryjne”, które będą rzetelnym wskaźnikiem, że efekt ten zachodzi w codziennym życiu. Czy przypadkiem nie jest tak, że małżeństwa, które wykonują wspólny ekscytujący zawód rzadziej się rozwodzą? A może okazało się, że po otwarciu parku rozrywki wśród okolicznych mieszkańców wzrósł odsetek nowych małżeństw? (być może ten ostatni przykład jest zbyt wydumany i bardziej kreatywni badacze znajdą lepsze dowody).

Z czasem rozwijają się w pokrewnych dyscyplinach teorie dotyczące wpływu pobudzenia na emocje, które pozwalają wygenerować dalsze hipotezy. W obszarze nauk o badaniu testowym z kolei poprawia się rozumienie tego, co wpływa na sposób odpowiadania, dzięki czemu zmniejsza się wariancja błędu wynikająca z tego czynnika. W subdyscyplinie integrującej psychologię pozytywną z neuronauką uzyskuje się sposób mierzenia zadowolenia ze związku, który jest w miarę obiektywny bo jest oparty na wskaźnikach fizjologicznych, których nie można zafałszować. Badacze uzyskują więc jeszcze precyzyjniejsze narzędzie do pomiaru. Dzięki temu też, badacze nie muszą już ograniczać się do relatywnie mało precyzyjnych twierdzeń w rodzaju „u grupy z ekscytującymi przeżyciami poziom zadowolenia ze związku średnio poprawił się”, a przejść do „u pary Jan Kowalski i Anna Nowak na podstawie znanych zmiennych wiemy, że poziom zadowolenia ze związku pod wpływem X poprawi się o $Y \pm 3\%$ ”. W rozwoju nauki poza chęcią poznania mają też znaczenie aplikacje praktyczne, więc jesteśmy w stanie projektować oddziaływania, które sprawiają, że w społeczeństwie znacząco spada odsetek rozwodów...

Wybiegając nieco dalej przyszłość, posiadamy wyrafinowane modele funkcjonowania mózgu, dzięki czemu możemy symulować, jak określone bodźce wpłyną na poziom zadowolenia, a potem porównujemy wyniki z rzeczywistością...

Być może zbyt mocno popuściłem wodze fantazji. Tym niemniej, wydaję mi się, że pokazałem tu pożądany model rozwoju nauki. Nie chodzi przy tym o konkretne przykłady, które mogą się okazać zupełnie inne, ale o ogólne zasady.

Dobór przykładu nie jest przypadkowy. Jak można było sprawdzić w poprzednim rozdziale, to akurat badanie nie doczekało się żadnych replikacji lub badań rozwijających i to pomimo niezwykle atrakcyjnej treści i mocnych dowodów początkowych w wyjściowym artykule.

Ale w przypadku pozostałych badań ujętych w tej analizie jest niewiele lepiej. Nigdzie nie dostrzegam systematycznie planowanego szerokiego programu badań, który choćby nieznacznie przybliżał się do zarysowanego powyżej (mówiąc o „planowaniu” niekoniecznie mam na myśli odgórne zarządzanie serii badań przez jakieś nadrzędne instancje; wyobrażam sobie, że „planowany” program badań może wyniknąć też z autonomicznych decyzji indywidualnych badaczy, którzy mają poczucie, że istnieją luki w wiedzy, które powinni wypełnić). Nie znaczy to, że w

niektórych przypadkach nie ma dużej liczby badań i dużej liczby danych. Jednak w olbrzymiej większości polega to na skakaniu od jednej do drugiej hipotezy, często dosyć luźno powiązanych z hipotezą wyjściową. By użyć nieco militarystycznego porównania: badacze nie próbują zdobyć „przyczółka” wysoce pewnej wiedzy, który pozwoli później rozszerzać rozumienie efektu. Zamiast utrzymać taki przyczółek, prą dalej naprzód, a zdobywając nowe tereny, zostawiają za sobą zdobycze, które są na tyle słabo zabezpieczone, że wracają z powrotem w ręce wroga (czyli odchodzą do lamusa historii i nie stają się kanonem wysoce pewnej wiedzy). Uważam, że jest to najważniejszy wniosek z mojej analizy. W dalszej części tego rozdziału postaram się go uzasadnić. Zanim jednak to uczynię, chciałbym zrobić kilka zastrzeżeń dotyczących ograniczeń obranej przeze mnie metody.

7.2. Ograniczenia metody

Omawiam ograniczenia przed zasadniczym podsumowaniem wyników, gdyż dalsze wnioski muszą być rozpatrywane z punktu widzenia tychże ograniczeń.

Metoda, którą obrałem, jest według mojej wiedzy pierwszy raz stosowana. Istnieje dużo analiz, w których sprawdza się częstość określonych cechy badań (np. częstość replikacji: Makel, Plucker Hegarty, 2012; rozkład wartości p : Masicampo, Lalande, 2012; poprawność raportowania danych o istotności statystycznej: Bakker, Wicherts, 2011; moc badań: Francis, 2014; formalne analizy mocy badania: Vankov, Bowers, Munafò, 2014; częstość stosowania metod behawioralnych: Baumeister, Vohs, Funder, 2007 i Furr, 2009; częstość raportowania wielkości efektów: Sun, Pan, Wang, 2010; wielkość efektów względem wielkości próby: Kuhberger, Fritz, Scherndl, 2014; grupa badana: Sears, 1986; pochodzenie próby badanej: Arnett, 2008). We wszystkich tych analizach dobiera się próbę artykułów albo losowo, albo poprzez analizę wszystkich artykułów z danego numeru (lub rocznika) czasopisma. Nie jest mi natomiast znana w literaturze analiza polegająca na systematycznym sprawdzaniu kilkudziesięciu cykli badań i to pod kątem tak wielu charakterystyk (pojedyncze cykle badań są często analizowane, ale nie są to artykuły metodologiczne, tylko poświęcone konkretnemu efektowi).

Jest to metoda wysoce pracochłonna, i kosztowała mnie znacznie więcej czasu niż pierwotnie zakładałem. Do celów tej pracy musiałem z dużą uwagą przeanalizować

ponad 2,5 tys. abstraktów i prawie pół tysiąca pełnych tekstów artykułów. Pomimo tego, liczba przeanalizowanych cykli badawczych jest w kategoriach statystycznych relatywnie niewielka. Istnieją bardzo duże różnice pomiędzy poszczególnymi cyklami badań, szczególnie w liczbie cytowań oryginalnego artykułu, liczbie osób badanych, częstości stosowania określonych metod badawczych czy poziomu „umiędzynarodowienia” próby badanej. Jeżeli takie zróżnicowanie cechuje też całe uniwersum wszystkich cykli badań, to może się zdarzyć tak, że analizowana próba z powodu czystego przypadku posiada nadreprezentacje badań (lub ich cech) określonego typu (i niedostateczną reprezentację innego typu).

Zaznaczę przy tym, że nie sędzę, żeby dobór próby wyjściowych artykułów polegający na uwzględnieniu pierwszych badań z określonego czasu spełniających założone kryteria, a nie poprzez losowanie, powodował sam w sobie zaburzenia reprezentatywności. Jedynie taka sytuacja mogłaby mieć miejsce, gdyby artykuły z początku roku różniły się od artykułów z dalszej części rocznika, ale trudno sobie wyobrazić, dlaczego taka sytuacja miałaby zachodzić. Problemem może być co najwyżej relatywnie niewielka próba, czyli mówiąc w kategoriach statystycznych – niska moc.

Nie mogę więc wykluczyć, że gdyby przeanalizować znacznie więcej cykli badawczych zapoczątkowanych artykułami z tych czterech czasopism, to wyniki byłyby inne pod wieloma względami. Uważam jednak, że jest mało prawdopodobne, że inne cykle badawcze posiadały dużo więcej badań, przede wszystkim z tego względu, że inne artykuły z tego rocznika były podobnie często cytowane.

Metoda ta może mieć bardziej fundamentalne ograniczenie. W określonym cyklu badań wokół jednej hipotezy określony artykuł („wyjściowy”) niekoniecznie musi być standardowo cytowany przez inne artykuły jednak testujące tą hipotezę i artykuły te są „niewidoczne” w mojej analizie. Może się więc zdarzyć tak, że cykl badawczy jest znacznie szerszy. Zarazem zaznaczę, że, choć tego nie wykluczam, to uważam taką opcję za mało prawdopodobną. W przypadku cyklu badań opisanego w JPSP4 przegląd cytowań drugiego stopnia, oraz niezależny przegląd baz danych nie pozwolił mi odkryć zasadniczo większej liczby danych, która mogłaby zmienić ostateczny obraz. Podobnie w przypadku JPSP7 prawie wszystkie teksty po 2000 używające słowa „entitativity” cytują rzeczony artykuł.

Ograniczeniem mojej metody jest też to, że w centrum analizy stoi jakiś efekt psychologiczny („empiryczny”). Ktoś może argumentować, że nauka rozwija się bardzo dobrze poprzez rozwój teorii. To jest spośród wielu badań (niekoniecznie cytujących się nawzajem) kreatywni badacze są w stanie ułożyć spójne teorie. Niekoniecznie muszą korzystać z badań nad podobnymi efektami, ale z zestawienia badań nad (powierzchnownie) różnymi efektami są w stanie wyciągnąć niebanalne wnioski. Jednak nawet w takim wypadku pozostaje problem rzetelności i trafności tych badań, a przypuszczam, że moja metoda analizy powinna w miarę sprawnie odpowiadać na te pytania. Trudno oczekiwać, że replikacje dokładne nie cytują artykułu z efektem, który replikują (szczególnie trudno sobie wyobrazić sytuację, że ktoś zupełnie przypadkowo wykona niezwykle dokładną, nie konceptualną, replikację jakiegoś badania).

Wreszcie fundamentalnym ograniczeniem mojej metody może być to, że nie pozwala dobrze prześledzić drogi rozwoju teorii wykraczającej znacząco poza efekt wyjściowych (czyli prześledzić czegoś, co w rozdziale 3 określiłem jako prawidłowości wysokiego poziomu). W rozdziale 3 sygnalizowałem gotowość przeanalizowania tego, jak wyjściowy efekt pozwala rozwijać wyższego rzędu twierdzenia teoretyczne, i jak te z kolei generują empiryczne weryfikacje. Ograniczenia czasowo-logistyczne kazały mi ostatecznie pozostać na poziomie cytowań pierwszego stopnia i analiza takich cytowań w moim przekonaniu nie pozwala na uważne śledzenie rozwoju teorii wyższego poziomu. Tym niemniej jestem przekonany, że analiza cytowań pierwszego stopnia jest wystarczającą metodą, żeby odpowiedzieć na pytania o wartość poznawczą cyklu badawczego nad konkretnym efektem przynajmniej na poziomie jego intersubiektywnej sprawdzalności i walidacji z wykorzystaniem innych metod i innych grup.

7.3. Wnioski – częstość replikacji

Wniosek z mojej analizy, który uważam za najważniejszy jest taki, że w psychologii występują silne tendencje do badań „szerokich” (tj. luźno nawiązujących do wyjściowego efektu), a nie „głębokich” (tj. próbujących zreplikować efekt poprzez replikacje dokładne i maksymalnie zrozumieć uwarunkowania wyjściowego efektu poprzez zbliżone replikacje konceptualne). Kolejne badania nawiązujące są zwykle na tyle różne od badania wyjściowego, że można by je zakwalifikować jako nowe

„badania oryginalne” i prawie nigdy nie można ich uznać za replikację dokładną. Nie pozwalają więc na falsyfikację wyjściowego efektu, co najwyżej na jego wzmocnienie (ma miejsce w nauce coś w rodzaju „błędu konfirmacji”). Zwykle kolejne badania nawiązujące muszą zawierać, przynajmniej na poziomie narracji, jakieś „odkrycie”, a więc element nowy, który pozwoli im wyraźnie odróżniać się od badań wcześniejszych i nie być odrzuconym przez redaktorów z powodu tego, że „to już dobrze wiemy”.

Najdobitniejszym tego wskaźnikiem jest częstość replikacji dokładnych. Na 382 badania nawiązujące tylko 11 (2,8 %) zostało zaliczonych do tej kategorii. Ale i tutaj należy poczynić pewne zastrzeżenie. Z tej liczby 8 badań było replikacjami wcześniejszych badań kwestionariuszowych (dotyczyło to tylko dwóch cykli badań: EJSP1 i JESP8). Innymi słowy, replikacja polegała na rozdaniu tego samego pliku kwestionariuszy. W niektórych przypadkach można by kwestionować czy są to rzeczywiście sensu stricte replikacja, bo np. wykorzystano tą samą skalę, ale jej krótszą wersję, względnie dołożono do badania kilka kwestionariuszy. Nawet jeśli uznać takie badania za replikacje dokładne, to powiedzmy szczerze, że replikacja polegająca na ponownym wypełnieniu tych samych kwestionariuszy nie należy do najbardziej wyszukanych.

Z 3 pozostałych replikacji dokładnych 2 zostały przeprowadzone przez te same zespoły (PSPB6_01 i JESP10_01). A więc wśród wszystkich badań (niekwestionariuszowych) miała miejsce tylko jedna sytuacja (0,26 %) kiedy przeprowadzono replikację przez niezależny zespół, a więc replikację, którą uznaje się za najbardziej wartościową (dokładnie chodzi o JESP1_01). Na marginesie mówiąc, akurat ta (jedyna) replikacja niezależnego zespołu, nie potwierdziła wyjściowych wyników.

Niezwykle wymowne jest też to, że nawet w tych nielicznych badaniach, które ewidentnie na podstawie analizy działu metoda były replikacjami dokładnymi, autorzy nie przedstawiali ich jako takich, i najczęściej w artykułach nie padało nawet słowo replikacja! Oczywiście w ostatecznym rozrachunku liczy się to, jakie badacze robią badania, a nie jak je przedstawiają, tym niemniej to odcinanie się od replikacji jest bardzo wymowne (na podobnej zasadzie nie było artykułu w którym stwierdzano – „my tu chcemy po prostu sprawdzić czy efekt powtarza się w grupie X” – do każdego

badania musiała być dorobione dosyć bogate uzasadnienie teoretyczne, po to, żeby mu przydać, jak sądzę, nimbu oryginalności).

Poza tym, że badania nie są replikowane dokładnie, istnieje też spora grupa badań wyjściowych, które w ogóle są nie rozwijane (lub w minimalnym stopniu), ani poprzez fakt przeprowadzania replikacji konceptualnych czy innego rodzaju badań nawiązujących. W niektórych przypadkach trudno mieć o to żal, bo wyjściowe efekty są banalne (np. JPSP1, JPSP2), teoretycznie i proceduralnie trywialne (EJSP5), albo wydają się być oparte na technikach psychologii fałszywie pozytywnej (np. efekty działające tylko w interakcji EJSP10, PSPB1, PSPB2; lub zachodzi mocne podejrzenie prezentowania tylko wyników pozytywnych: EJSP3). Dla równowagi zauważę jednak, że istnieje kilka ciekawych (np. JPSP5, EJSP7, PSPB3) czy wręcz niezwykle atrakcyjnych (np. JPSP9, JPSP10, PSPB3) badań, które też nie doczekały się żadnych lub prawie żadnych badań nawiązujących.

Publikowane replikacje w mojej analizie były niezwykle rzadkie. Na podstawie własnych kontaktów z literaturą nie sądzę, żeby reszta nauki była zasadniczo inna niż ten jej wycinek. Owszem, sporadycznie replikacje występują, szczególnie w przypadku badań niezwykle kontrowersyjnych (np. replikacje prekognicji Bema: Ritchie, Wiseman, French, 2012; Galak, LeBoeuf, Nelson i Simmons, 2012), lub sławnych (np. replikacje eksperymentów Miligrama nad posłuszeństwem względem autorytetu; Burger, 2009). Jednak, statystyczny efekt w psychologii ma znikome szanse, żeby doczekać się opublikowanych replikacji. Kultura nowatorstwa prowadzi nas do tego, że w pogoni za ciągłymi nowościami, nie budujemy fundamentu wiedzy tak pewnej jak to tylko możliwe w obrębie metody naukowej. Na długo przed Stapelem wiadomo, że w psychologii występuje problem z niską mocą badań (Maxwell, 2004), całą gamą artefaktów badawczych (Rosenthal, Rosnow, 2009), oraz wybiórczym publikowaniem (Cooper, DeNeve, Charlton, 1997). Dokładne replikacje są potrzebne nie tylko po to, żeby wykrywać potencjalne oszustwa, ale choćby po to, żeby ograniczać skalę takich zjawisk.

7.4. Wnioski – trafność zewnętrzna (grupa badana)

Założyłem (rozdział 3), że plan minimum, to oprócz dokładnego zreplikowania efektu przez niezależne zespoły, również sprawdzenie występowania efektu pod kątem

trzech podstawowych moderatorów, które mają znany wpływ na zachowanie, a więc grupy społecznej, wieku i grupy kulturowej.

Zaznaczę na początku, że nie chcę tu promować jakichś nierealistycznych standardów. Nie ma zasobów, żeby sprawdzić każdy efekt na każdym mieszkańcu każdego zakątka świata (w tym każdej możliwej grupie społecznej i zawodowej). Tym niemniej plan minimum polegający na sprawdzeniu, czy efekt utrzymuje się u ludzi z kilku odmiennych grup społecznych i kulturowych wydaje się być potencjalnie w zasięgu możliwości.

Pod tym względem niestety moja analiza również zmusza do niepokoju. 80 % prób badanych stanowili studenci. W wielu dosyć obszernych cyklach badawczych studenci byli jedynymi badanymi. Sporadycznie występowały inne grupy, ale trudno nie odnieść wrażenia, że często były to grupy najłatwiej dostępne. Stąd dużo częściej pojawiały się dzieci (które można zebrać w szkole) niż osoby starsze (tylko jedno badanie z taką próbą); do tego wystąpiło kilka innych prób łatwo dostępnych, a więc próby zebrane przez badanie internetowe, studenci MBA itd.

Jeśli chodzi o zróżnicowanie narodowościowe wnioski również muszą być równie negatywne. Nie ma co prawda jednej absolutnie dominującej grupy (Amerykanie to „tylko” nieco ponad 56 %), ale kilka najczęściej badanych narodowości stanowi olbrzymią większość prób. Próby amerykańskie plus próby z trzech krajów anglojęzycznych (Australia, Kanada, Wielka Brytania) to już 65 % prób. Jeśli do tego dołożymy próby z trzech innych wysoce rozwiniętych krajów zachodnich (Niemcy, Holandia i Francja) to odsetek urasta do 75 %. Jeśli dalej uwzględnimy badanych z Izraela odsetek urasta do 80 %. Z pozostałych, niezachodnich krajów, wyróżniają się liczebnością Chiny i Japonia. Stanowią łącznie 12 % prób badanych. Istnieje wiele bardzo dużych krajów, z których nie pochodziła żadna próba (Brazylia, Argentyna, Indie, RPA, Indonezja, Nigeria, Pakistan, Bangladesz) lub pochodziły pojedyncze próby (np. Rosja, Turcja, Meksyk).

Przedstawiając te wyniki inaczej można powiedzieć, że relatywnie do udziału w światowej populacji, próby amerykańskie są nadreprezentowane 12 krotnie, próby holenderskie 22 razy, a próby izraelskie aż 33 razy. Z kolei próba chińska jest niedoreprezentowana prawie 2,5 krotnie itd. Wśród 20 najludniejszych krajów świata

dziesięć⁸³ nie było reprezentowanych w żadnych badaniach, a dalsze pięć⁸⁴ miało reprezentacje czysto symboliczną.

Tak jak zaznaczałem w rozdziale 5, zróżnicowanie kulturowe jest jeszcze niższe, gdy nie uwzględniam kilku pojedynczych cykli wysoce umiędzynarodowionych badań. A te relatywnie niskie zróżnicowanie osób badanych można ująć w szerszym kontekście: nawet, gdy badane były różne próby, były to zwykle replikacje konceptualne lub inne formy badań, ale nie replikacje dokładne. Oznacza to dodatkowe trudności w interpretacji: gdy nie można potwierdzić wyjściowej hipotezy, nie wiadomo, czy wynika to ze zróżnicowania grupowego, czy może z innej procedury (przy okazji nieco innej hipotezy).

Ogółem 16, a więc 4,5 % badań nawiązujących zaliczyłem do kategorii badań testujących podstawowe moderatory. Z tej liczby 10 badań testowało je na dzieciach (lub adolescentach), a dalsze trzy na innych grupach społecznych (po jednym na emerytach, policjantach i graczach konkretnej gry wideo). Tylko w 4 przypadkach badanie nawiązujące testowało identyczne hipotezy w bardzo zbliżony sposób na osobach z innych krajów niż badanie wyjściowe.

Trudno więc o inny wniosek na podsumowanie, że w psychologii bada się ludzką populację w sposób bardzo wybiórczy.

7.5. Wnioski - trafność zewnętrzna (metody badawcze)

Z jednej strony pożądane jest, żeby hipotezy były testowane przy pomocy różnych metod badawczych i źródeł danych, żeby zmniejszać ograniczenia jednej metody (a każda metoda ma swoje ograniczenia). Z drugiej strony powinno dążyć się do tego, żeby ostatecznym sprawdzianem hipotez „behawioralnych” (tj. takich, które mają bezpośrednie implikacje dla codziennych zachowań) były dane tegoż rodzaju, a więc szczególnie rejestry zachowań (lub wytworów ludzkich) oraz eksperymenty terenowe. W ostateczności używanie jakiegś formy zachowania jako zmiennej zależnej w połączeniu z realistycznymi manipulacjami też jest pożądane.

⁸³ Od największych: Indie, Indonezja, Brazylia, Pakistan, Nigeria, Bangladesz, Wietnam, Etiopia, Egipt, Kongo

⁸⁴ Rosja, Meksyk, Filipiny, Iran, Turcja

Analiza ta od początku była nastawiona na hipotezy mające implikacje behawioralne i tylko badania, które mają takie potencjalne czytelne implikacje były uwzględniane w wyjściowej puli efektów.

Eksperymentów terenowych było bardzo niewiele (8), z czego prawie wszystkie z wyjątkiem jednego pochodziły z jednego artykułu (JPSP9). Eksperymenty terenowe uważam za bardziej pożądaną formę badania niż eksperymenty „zwykłe” z tego względu, że badany w dobrze zaprojektowanym eksperymencie nie powinien modyfikować swojego zachowania poprzez świadomość tego, że uczestniczy w badaniu.

W trzech przypadkach pojawiły się analizy niezależnie od badania zarejestrowanych ludzkich zachowań i wytworów. Dla przypomnienia były to: wyniki golfistów (PSPB8_04), sprawozdania zarządów spółek giełdowych (PSPB5_17) i treść artykułów prasowych (JPSP7_09). Dane te są o tyle pożądane, że trudno zakładać tutaj występowanie jakichkolwiek artefaktów wynikających z interakcji badany-badacz (oczywiście, może się natomiast zdarzyć, że z jakichś względów, choćby przypadku, dane takie nie są reprezentatywne). Tak czy owak te przykłady pokazują, że badacze są w stanie sięgać po tego rodzaju dane, ale niestety robią to niezwykle rzadko. Do wartościowych pomysłów na reanalizę danych można też zaliczyć badanie, gdzie skorzystano z dużych rejestrów armii amerykańskiej i testowano hipotezy o związku picia alkoholu z przemocą w związku (JESP6_02). W przeciwieństwie jednak do wcześniejszych badań część danych dostarczali sami badani, tym niemniej wielką zaletą tego badania jest jego duża moc.

Warto też dodać, że reanaliza danych już istniejących na ogół jest związana z dużo mniejszą liczbą dylematów etycznych. Przede wszystkim pozwalają na niebanalne przetestowanie teorii za pomocą rejestrów rzeczywistych zachowań. Wydaje się, że potencjał do wykorzystywania takich danych jest ogromny. Wymaga to jednak wyjścia poza standardowe schematy badań, i nierzadko pomysłowości ze strony badacza. Nie bez znaczenie jest też to, że kodowanie takich danych może być w wielu przypadkach wysoce pracochłonne, i pochłaniać dużo więcej czasu badacza niż przeprowadzenie jakiegoś niewyszukanego eksperymentu na grupie studentów. Niestety, poza przytoczeniem tych interesujących przykładów wykorzystania danych pochodzących „z życia”, trzeba stwierdzić, że takich danych się nie poszukuje, gdyż stanowiły tylko około 1 % wszystkich badań nawiązujących.

Zachowanie jako zmienna zależna występowało na tyle rzadko, że mogę sobie pozwolić tutaj na przywołanie wszystkich sytuacji, gdy wystąpiło:

- reakcje i zachowania badanego w zaaranżowanej sytuacji eksperymentalnej nagrywane i oceniane przez sędziów kompetentnych i/lub przez pomocników badacza (PSPB1_09; PSPB10; JESP2_09; JPSP10; w tym jedno badanie dotyczyło interakcji matek z ich dziećmi PSPB5_13);
- elementy lingwistyczne swobodnej wypowiedzi, np. tendencja do opisywania jakiejś grupy w abstrakcyjny sposób (JPSP7_14; JPSP7_29; PSPB1_08);
- decyzję o dokonaniu „zemsty” na członkach zaaranżowanej grupy (JPSP7_19);
- udzielenie pomocy obcej osobie np. przypilnowanie walizki (6 badań w JPSP9);
- wzięcie udziału w badaniu diagnostycznym wykrywającym nowotwór (EJSP4_07);
- decyzję o wyborze określonego typu zadania do rzeczywistego wykonywania (PSPB4_02);
- wytrwałość w rozwiązywaniu problemów w czasie rzekomej „przerwy” w eksperymencie (3 badania w PSPB5);
- zapalenie papierosa w czasie „przerwy” w eksperymencie (2 badania w JESP7_01);
- udzielenie pomocy (lub nie) osobie, której rozsypały się książki na korytarzu (JESP3_04).

Jeśli chodzi o samo zachowanie jako zmienną zależną (wyjaśnianą), to warto nadmienić, że było one używane jako zmienna tylko w eksperymentach, choć nie ma żadnych formalnych przeszkód, żeby je wykorzystywać w badania korelacyjnych. Siłą więc rzeczy, gdy badano wpływ określonych czynników na zachowanie to owymi czynnikami były tylko takie, którymi możemy manipulować, a więc zmienne sytuacyjne, bo, co oczywiste, nie możemy manipulować eksperymentalnie natężeniem cech dyspozycyjnych. Sporadycznie występowały sytuacje, gdy wpływ takich zmiennych dodatkowo kontrolowano poprzez ich pomiar w eksperymentach.

Ogółem zachowanie pojawiło się w 23 eksperymentach (4,9 % wszystkich badań). Osobną kategorię stanowią badania, gdzie zmienną zależną ma naturę poznawczą, szczególnie pamiętanie określonych informacji (głównie związane z badaniami nad

stereotypami) oraz badania, w których badani uczestniczyli w rozmaitych grach ekonomicznych (szczególnie negocjacjach) i zmienną zależną były określone decyzje. Takich badań wystąpiło odpowiednio 25 i 40. Badania te trudno sklasyfikować jako mierzące zmienne samoopisowe, ale trudno też je zakwalifikować do kategorii typowych badań behawioralnych. Podobnie badania „na pograniczu” to dwa badania, w których badani na bieżąco raportowali swoje przeżycia przez dłuższy okres czasu. Ogółem te cztery kategorie stanowiły 19,7 % wszystkich badań. Zmienne utajone były stosunkowo rzadkie (17 badań), oraz w jednym wypadku wystąpiła zmienna biologiczna. Zdecydowaną większość zmiennych zależnych stanowiły różne formy miar samoopisowych (76,5 %). Psychologia nie jest więc „nauką o samoopisach i ruchach palców” (jak sugerowali Baumeister i in., 2007). Jest przede wszystkim „nauką o samoopisach”.

Przeanalizowane tutaj przykłady pokazują, że badacze potrafią sięgać po dane behawioralne, szczególnie, gdy wykażą się pewną minimalną pomysłowością. Zarazem jednak w olbrzymiej większości badań, badacze idą po linii najmniejszego oporu w wyborze zmiennych zależnych.

Jak pokazałem wcześniej występowały też tendencje do tego, żeby badacze używali jednego rodzaju danych przez cały cykl badawczy. Jeśli w wyjściowym badaniu dominowały techniki korelacyjne, takie też były stosowane dalej, jeśli eksperymentalne to takie. Podobnie rzecz się ma z naturą zmiennej wyjaśnianej. Osobnym problemem jest to, że w niektórych przypadkach stosowano ciągle podobną, niewyszukaną procedurę, jak pokazałem w rozdziale 6 (przykładem może być cykl PSPB1, gdzie bardzo często badanym prezentowano ten sam zestaw bodźców o księdzu i skinheadzie, i gdzie mierzono pamiętanie informacji zgodnych lub niezgodnych ze stereotypem takiej osoby). Zarazem wprowadzano do tej procedury na tyle duże zmiany, że trudno mówić o replikacjach dokładnych, albo o możliwościach integrowania danych.

Dodatkowo pożądanym jest, żeby badacze generowali jakieś niebanalne przewidywania, które są dodatkowym testem teorii. Ten aspekt nauki nie jest najgorszy, przynajmniej w tych cyklach badawczych, w których występuje wiele badań.

Przykładowo w JPSP4 sprawdzano na kilka sposobów wyjściowy efekt (nie tylko poprzez czytanie opisów zachowań, ale też oglądania twarzy, nagranych interakcji,

albo ocenę intencji innych osób w sesji szybkich randek), do tego wygenerowano kilka niebanalnych przewidywań, które się potwierdziły (np. mężczyźni częściej deklarowali, że opatrnie zrozumieli intencje kobiet; lub kobiety deklarowały, że częściej ich życzliwe zachowania zostały odebrane jako oznaka zainteresowania seksualnego). Wskazano też kilka możliwych korelatów dyspozycyjnych efektu przeszacowywania.

Z kolei w JPSP6 również zastosowano szereg różnych sytuacji i kontekstów, w których pokazano, że ludzie mają tendencje do zawyżania szacunku ile osób zwraca na nich uwagę (np. na błędy językowe w trakcie dyskusji; udany lub nieudany wygląd danego dnia; ich nieobecność w grupie itd.).

Podobnie szereg zróżnicowanych danych pojawił się w cyklu EJPS4 (afirmacja ważnych wartości zmniejsza dysonans). Z jednej strony były eksperymenty o znanej, standardowej procedurze „napisz esej argumentując na rzecz jakiegoś niepopularnego temat i powiedz jak bardzo się zgadzasz z nim”, ale były też bardziej niebanalne, jak np. afirmacja ważnych wartości zmniejszała krytyczność oceny eseju anty-amerykańskiego czy zwiększała poczucie winy za zbrodnie dokonywane przez przedstawicieli własnego narodu. Podobnie dosyć zróżnicowany był cykl EJSP9. Pozostałe cykle oceniam jako raczej mało zróżnicowane, zdominowane przez jedną metodę (nie mówię tu o cyklu badań nad entatywnością JPSP7, gdzie było wiele różnych rodzajów badań, ale tak naprawdę było wiele różnorodnych hipotez; podobne uwagi mogę zastosować do PSPB5).

Bardzo często było tak, że gdy poziom zróżnicowania grupowego (kulturowego) był spory, tam występowało niewielkie zróżnicowanie metod badawczych (szczególnie mam tu na myśli cykle EJSP5 i, w mniejszym stopniu, EJSP1). Z kolei tam gdzie występował przyzwoity poziom zróżnicowania metod, tam badacze „nadrabiali” ten wysiłek wprowadzając w niewielkim stopniu zróżnicowanie pod względem grupy i narodowości (i byli to zwykle amerykańscy studenci, tak jak w JPSP4, JPSP6, EJSP4, EJSP9, JESP9). W trzech przypadkach (JPSP7, PSPB5 i JESP2) występuje spore zróżnicowanie zarówno metod, jak i grup badanych, ale tu z kolei występuje tak duże zróżnicowanie hipotez (szczególnie w JESP2), że można powiedzieć, że badania te są bardzo luźno powiązane ze sobą pod względem teoretycznym.

Osobną kwestią wartą omówienia jest kwestia realistyczności manipulacji zmienną niezależną, czy szerzej problem realizmu samej procedury w ogóle. Jak pisałem

wcześniej (rdz. 5.3.7) częstość stosowania manipulacji eksperymentalnych realistycznych i nierealistycznych rozłożył się prawie idealnie po połowie. W niektórych wypadkach trudno tu mówić o jakichś wielkich zasługach badaczy (np. trudno, żeby badacze kazali badanym „wyobrazić” sobie negocjacje twarzą w twarz lub przez komputer, zamiast po prostu zaaranżować takie sytuacje) lub „realizm” wymagał minimum wysiłku (tak jak w badaniu JESP5_02, gdzie manipulowano ciężkością książki, albo JPSP4_14:3 gdzie manipulacja przekonaniem o stanie matrymonialnym drugiego uczestnika badania polegała po prostu na podaniu informacji „wolny” lub „w związku”). Tym niemniej ta część analizy zaskoczyła mnie raczej pozytywnie. Badacze niejednokrotnie stosowali dosyć pomysłowe procedury eksperymentalne, dbali o przekonanie badanych, że określone fakty są prawdziwe, a czasami dodatkowo zwiększali „realizm”, gdy uczestnicy gier ekonomicznych mogli wygrać realne pieniądze.

7.6. Wnioski – ocena zdolności programów badawczych do generowania wiedzy

Zasadnicza teza tej analizy jest taka, że badacze mają tendencje do powierzchownego „skakania” od hipotezy do hipotezy, z których każda jest jakoś powiązana z wyjściową hipotezą, ale jest to najczęściej związek luźny. Przede wszystkim jednak kolejne badania w systematyczny sposób do siebie nie nawiązują.

Zanim rozwinę tę myśl, chcę pokazać przykład cyklu badań, który doskonale ilustruję te problemy. Wybrałem cykl JESP2, w którym dla przypomnienia główną hipotezą jest to, że interakcje twarzą w twarz (w porównaniu do interakcji telefonicznych lub drogą komunikacji elektronicznej) zwiększają poziom kooperacji (a więc też ich wyniki) i zadowolenia z interakcji w negocjacjach lub innych sytuacjach konfliktowych.

W wyjściowym artykule opisano dwa badania, gdzie studenci uczestniczyli albo w symulowanych negocjacjach (temat: negocjacje „strajkowe”), albo w odmianie gry w dylemat więźnia. Zmienną zależną była samooceniane zadowolenie z interakcji i jej wynik. W kolejnych badaniach w tym cyklu:

- mierzono liczbę ujawnianych informacji o sobie [JESP2_01:1] w symulowanych negocjacjach w zależności od trybu negocjacji - podobne badanie, ale inna zmienna zależna, więc w zasadzie jest to już nieco inna hipoteza;

- sprawdzano czy umożliwienie chwilowego kontaktu twarzą w twarz poprawia poziom zadowolenia i wynik negocjacji [JESP2_01:2] - znowu, można dyskutować, czy to ciągle ta sama hipoteza, co w wyjściowym badaniu;
- badano która strona bardziej ustąpiła w zależności od trybu negocjacji i jej pozycji przetargowej [JESP2_02] – luźno nawiązuje do wyjściowego badania, ale na dobrą sprawę jest to znowu nieco inna hipoteza;
- sprawdzano poziom lubienia się i osiągnięte wyniki podczas negocjacji cenowych [JESP2_03] – dosyć zbliżone badania wyjściowego, ale inny temat negocjacji;
- w kolejnych negocjacjach (przy pomocy komputera lub twarzą w twarz) sprawdzano wynik, ale też samoocenianą rzetelność i wiarygodność [JESP2_04]; okazało się, że tutaj jednak „e-negocjacje” powodowały wyższy wynik, ale niższą samoocenianą wiarygodność badanych [JESP2_04] (nie podano dokładnie jaki był przedmiot negocjacji);
- w innych negocjacjach cenowych [JESP2_05] nie znaleziono różnic w zależności od trybu negocjacji (temat negocjacji tutaj: sprzedaż wycieczki);
- w kolejnym badaniu osoby w czteroosobowych grupach grały w grę ekonomiczną, w której mogli zyskiwać dzięki współpracy, ale mogli też zyskiwać na egoistycznych decyzjach (kosztem innych); konceptualnie był to rodzaj gry w dylemat więźnia [JESP2_06:1]; zwraca uwagę, że jest to mocno różna procedura od wyjściowej, gdyż można sobie wyobrazić, że czteroosobowa grupa ma inną dynamikę niż dwuosobowa;
- w następnym badaniu procedura była konceptualnie zbliżona, ale gra była nieco bardziej wyrafinowana i symulowała korzystanie ze wspólnego dobra (tutaj: łowiska) [JESP2_06:2];
- w negocjacjach cenowych (nie podano przedmiotu) manipulowano też poziomem ich złożoności; nie podano przy tym wyniku negocjacji, tylko poziom zaufania [JESP2_07];
- w kolejnym badaniu oprócz modalności komunikacji manipulowano możliwością wyjścia z negocjacji [JESP2_08];

- w kolejnych (i ostatnich) badaniach miały miejsce negocjacje (temat: kupno nieruchomości), a sędziowie kompetentni na podstawie ich zapisu oceniali liczbę nieetycznych zachowań [JESP2_09].

Mógłbym w ten bardziej szczegółowy sposób przeanalizować każdy cykl, ale nie ma potrzeby powtarzać informacji, które przedstawiłem w rozdziale 6 i w załączniku 2. Tak czy owak, pokazuję dokładniej ten cykl, gdyż moim zdaniem oddaje on bolączki praktycznie wszystkich pozostałych cykli.

A więc z jednej strony każde badanie musi mieć jakiś element „nowości”, nie ma więc dokładnego powtórzenia któregośkolwiek badania (w innych cyklach zdarzają się sporadycznie). Dalej, poza ogólną hipotezą (modalność kontaktu wpływa na wyniki i poziom satysfakcji w sytuacjach negocjacji) trudno znaleźć w tym cyklu jakąś systematyczność, szczególnie nawiązywanie jednych badań do drugich, rozszerzanie badań wcześniejszych, wyjaśnianie niespójności między badaniami itd. Często badania na tyle się od siebie różnią, że w zasadzie trudno porównać wyniki jednych do drugich tj. istnieje tyle różnic między badaniami, że nie wiadomo, którym przypisać to, że pojawiają się ewentualne różnice w wynikach.

W tym konkretnym cyklu przebadano 1300 osób, w dodatku w aranżowanych w laboratorium negocjacjach i grach ekonomicznych. Taki cykl wymaga to więc dużo większego wysiłku niż proste rozdanie kwestionariuszy do wypełnienia. Dużo pracy zostało włożone, badacze uzyskali publikacje, czasami w przyzwoitych czasopismach, zapewne posunęli nieco swoje kariery do przodu. Nie wątpię, że na podstawie takiego cyklu badań można napisać artykuł przeglądowy lub popularnonaukowy, w którym zestawia się te dane w sposób atrakcyjny, a do każdego twierdzenia dorobi odpowiednie odwołanie w literaturze. Laik może mieć wrażenie dynamicznie rozwijającej się nauki. W moim przekonaniu to jednak ciągle tylko „zbieractwo faktów”, z którego niewiele wynika. Mimo dużej ilości danych, nie ma tutaj dużej ilości dojrzałej wiedzy, w więc takiej, o której można powiedzieć, że jest:

- pewna (co szczególnie wyraża się replikacjami o wysokiej mocy; pomijam tu fakt, że w przypadku niektórych badań są pośrednie wskaźniki psychologii fałszywie pozytywnej, takie jak małe kratki, arbitralny dobór zmiennych itd.),

- wyjaśniająca niespójności pomiędzy poszczególnymi wynikami (w niektórych przypadkach nie było różnicy między modalnościami, albo była tylko dla niektórych zmiennych),
- pozwalającej na wysuwanie jakichś niebanalnych przewidywań (np. jakie mogą być wyniki jakiegoś innego badania w tym nurcie),
- zintegrowanej tj. systematycznie opisującej we w miarę pewny sposób wyjściowy efekt i jego uwarunkowania, korelaty dyspozycyjne, moderatory sytuacyjne itd.

Uzupełnieniem tych uwag, jest to, że – jakżeby inaczej – wszystkie badania odbyły się na studentach (ewentualnie studentach MBA) i prawie wszystkie próby były amerykańskie (z wyjątkiem jednej izraelskiej). To dodatkowe potencjalne ograniczenie pewności wiedzy. W moim odczuciu zarówno samo rozumienie wyjściowego efektu jest bardzo słabe, a już w szczególności ograniczone są możliwości rozwoju teorii wyższego poziomu.

Dalej już bez przywoływania wszystkich badań przyjrzę się tym wybranym cyklom, gdzie przeprowadzono stosunkowo dużo badań nawiązujących (przynajmniej 5). W większości cykli badawczych liczba badań jest niewielka i trudno w takim wypadku mówić o systematycznym programie badawczym.

Najszerzy cykl badawczy to badania poświęcone entytatywności (JPSP7). Przyjęcie tego cyklu do analizy może być uznane za kontrowersyjne (por. 5.2.7), gdyż cykl ten wyróżnia się na tle innych, tym, że testuje się tutaj bardzo wiele bardzo różnych hipotez. Przy czym większość tych hipotez ogranicza się do jednego artykułu (z jednym lub kilkoma badaniami). Niewiele jest natomiast hipotez przetestowanych w więcej niż jednym artykule. Badań replikujących lub wysoce zbliżonych do badania wyjściowego też jest bardzo niewiele (i też nie zawsze mnie przekonują, np. w jednym badaniu, gdzie sprawdzano uwarunkowania entytatywności, za bodźce służyły animowane postacie z kreskówek). Jaki wniosek wypływa z tych danych? Wydaje się wysoce pewne, że entytatywność grupy wywiera jakiś wpływ na poznanie społeczne oraz pewne czynniki sytuacyjne zmieniają jej postrzeganie, a sporadycznie badania pokazują, że entytatywność może mieć znaczenie w przypadku zjawisk o dużym znaczeniu praktycznym np. obwinianie członków grupy odpowiedzialnością zbiorową. Jednak poza tym ogólnym wnioskiem nie wydaje się, żeby można było wyciągnąć dużo szczegółowych i wysoce pewnych wniosków (tj.

opartych na kilku niezależnie przeprowadzonych badaniach). Każda hipoteza jest testowana bardzo pobieżnie. Rozszerzenie analizy na cytowania drugiego i wyższego stopnia nie zmieniłoby zasadniczo obrazu rzeczy. W bazie PsycINFO w latach 2000-2012 roku występuje 132 artykuły ze słowem „entytatywność” (wyszukiwanie jest tu ułatwione bo to słowo ma tylko to jedno, konkretne znaczenie). Tymczasem badań cytujących artykuł jest 195. Pobieżny przegląd badań z PsycINFO ze słowem „entytatywność” wskazuje, że prawie wszystkie cytują wyjściowy artykuł JPSP7. Danych więc raczej nie ma dużo więcej. Jest to w moim przekonaniu kolejny cykl dobitnie uzasadniający moją tezę, że badania robi się „szeroko”, ale nie „głęboko”.

W przypadku cyklu badań poświęconego zniekształconej percepcji zainteresowania seksualnego u mężczyzn (JPSP4) mamy do czynienia z wygenerowaniem kilku niebanalnych przewidywań i jest to największa zaleta tego cyklu. Ale znowu, ani wyjściowe badanie, ani badania nawiązujące nie są replikowane (co sprawdziłem też analizując cytowania drugiego stopnia i analizując artykuł przeglądowy i meta-analizę por. rdz. 4.2). Tak więc pewna wiedza, jakiej udało się dorobić nie wykracza poza to, że efekt najprawdopodobniej występuje, objawia się w codziennym życiu, oraz, że może być modyfikowany przez pewne czynniki sytuacyjne i dyspozycyjne. Znowuż, do każdej szczegółowej hipotezy mamy na ogół tylko jedno badanie. W niektórych badaniach pojawiają się dane, że efekt nie występuje w pewnych okolicznościach, ale nie ma potem prób wyjaśnienia tych nieścisłości (szczególnie, czy jest to wynik przypadku czy faktycznie pewne moderatory zmieniają jego siłę). W dodatku praktycznie wszystkie badania są przeprowadzone na studentach amerykańskich.

Bardzo podobny w moim odczuciu jest cykl badań dokumentujący to, że ludzie przeceniają stopień, w jakim inni ludzie zwracają uwagę na ich zachowania i cechy, szczególnie negatywne (JPSP6). Znowu wystąpiło sporo mocno zróżnicowanych badań powiązanych oczywiście z wyjściową hipotezą, jednak te konkretne badania nie były dalej replikowane i rozwijane. I podobnie jak w przypadku wcześniejszego cyklu badań zróżnicowanie grupowe i narodowościowe było praktycznie zerowe.

Również cykle badań wokół hipotez mówiących, że:

- a) afirmacja ważnych wartości redukuje dysonans poznawczy (EJSP4),

- b) różne typy myśli o alternatywnym biegu wypadków zmieniają poziom przygotowania i nastrój (EJSP9)

wykazują niemałe i często dosyć pomysłowe zróżnicowanie metod oraz źródeł danych, jednak podobnie jak wyżej prawie wszyscy badani to studenci, większość pochodzi z USA, a poszczególne badania nie są replikowane.

Cykl badań dokumentujących fakt bardziej skromnego przedstawiania się przez Japończyków niż Kanadyjczyków (PSPB5) również oceniam wysoko pod względem zróżnicowania metod i źródeł danych. Cykl ten też pozytywnie wypada ze względu na to, że jest relatywnie zróżnicowany kulturowo (trudno jednak, żeby inaczej w przypadku takiej hipotezy). To zróżnicowanie kulturowe obniża fakt, że olbrzymia większość dorosłych prób badanych (31 na 32) to studenci. Znowu mamy niestety sytuację, że poza ogólną hipotezą niewiele te badania łączy.

Za przykład serii badań, której moim zdaniem zdecydowanie nie warto naśladować, uważam cykl badań dokumentujących związki między (deklarowanymi) wartościami, a (deklarowanym) subiektywnym dobrostanem (EJPS5). Pomijam tu nawet kwestie metodologiczne, czyli pytanie o to, na ile wyniki takich badań nie mogą być rodzajem artefaktu, jakich wiele przy badaniach samoopisowych. Przede wszystkim nie wierzę, żeby dużo wiedzy można było uzyskać z badań, które sprowadzają się do korelowania testów z testami. Jedyną „zaletą” takich badań jest łatwość ich przeprowadzania, stąd też pojawia się próby z różnych krajów i z różnych grup społecznych. Efekty są prawie zawsze słabe, co tym bardziej zwiększa prawdopodobieństwo, że subtelne artefakty mogą wpływać na wyniki lub mogą być one po prostu przypadkowe. Zzór wyników jest niejednoznaczny, co oczywiście może wskazywać na to, że nieco inne rodzaje testów i inne grupy badane różnicują w pewnym zakresie te wyniki, ale te rozbieżności nie są systematycznie wyjaśniane. Badacze wykazywali się minimalną kreatywnością, jeśli chodzi o procedurę badawczą i prawie zawsze sprowadzała się ona do rozdania pliku kwestionariuszy (co oczywiście, patrząc z punktu widzenia badacza szukającego wyników do publikacji jest wielką „zaletą”). Używając tym razem metafory budowlanej, każdy dokłada swoją cegielkę (lub nawet kilka cegieł), ale bardziej na zasadzie, że zrzuca je na plac budowy i ma nadzieję, że dom sam się z tego ułoży. Nie czuję się ekspertem od tematu wartości i subiektywnego dobrostanu, ale mam nieodparte wrażenie bardzo dużego wysiłku, który prowadzi do minimum wiedzy.

Bardziej pozytywnym przykładem są badania nad zróżnicowaniem natężenia orientacji na dominację społeczną w zależności od płci (EJSP1). Wyjściowe badanie podobnie jak we wcześniejszych przypadkach również było korelacyjne i kwestionariuszowe. Jednak zaletą jest to, że skalę samoopisową korelowano ze zmiennymi o wysoce „obiektywnej” naturze (płcią biologiczną i grupą społeczną). Po drugie wzór wyników w tym cyklu badań jest jednoznaczny: prawie zawsze płęć różnicuje wyniki i to w jednym kierunku (mężczyźni mają wyższe natężenie). Nawet wtedy, gdy badacze twierdzą, że jest to rodzaj zależności pozornej, tak czy owak występują różnice. Po trzecie, choć większość badań nawiązujących jest korelacyjna, to gdzieś tam badacze wykazują się pewną kreatywnością (np. przeprowadzano eksperymenty, w których podkreślano koncept płci, albo manipulowano poczuciem własnej pozycji społecznej). Tak czy owak, badacze mogli poszukać znacznie większej liczby danych np. związanych z różnicami międzypłciowymi w poglądach politycznych, gotowości do udziału w działalności na rzecz zmniejszania nierówności (niewykluczone, że efekt ten jest pośrednio testowany przez wiele badań, w których nie wymienia się terminu „orientacja na dominację społeczną” i których taka analiza nie obejmuje np. badań wokół innych różnic płciowych, psychologii konfliktów społecznych itd.). Ogółem więc, cykl ten wskazuje dosyć wyraźnie, że taki efekt występuje i jest to jego niewątpliwa zaleta. Niestety, nie daje dużo większej liczby informacji. Szczególnie powierzchownie potraktowano tezę o tym, że inne zmienne niwelują efekt – kilka artykułów testuje po jednej takiej tezie, ale nikt nie zabrał się do tego, by to ujednoznaczyć (na marginesie, ten cykl pokazuje, że badacze w psychologii mają wielkie możliwości swobodnego doboru tekstów do swoich przekonań – jeśli są zwolennikami tezy, że płęć różnicuje SDO, to znajdują na to kilka solidnych źródeł, ale jeśli teza im się nie podoba, to mogą przytoczyć źródła mówiące „ale efekt ten zanika, gdy kontrolować...”).

Interesująco zaczyna się badanie dotyczące współczesnych postaw rasistowskich (JESP4). W badaniu wyjściowym przeszło 700 osób pisało eseje związane z tym tematem, z których następnie wyodrębniano jakościowo główne wątki. Wydaje się to być niezłą egzemplifikacją też Rozina (2001) o konieczności większego stosowania metod opisowo-obserwacyjnych. W badaniach nawiązujących jednak już wrócono na „stare dobre tory” tj. badani w większości wypadków wypełniali kwestionariusze, w których ustosunkowywali się do różnych twierdzeń o zachowaniach osób innej rasy i

własnych postawach. W jednym wypadku miało miejsce badanie otwarte, w drugim kwestionariusze takie utworzono na podstawie otwartego badania pilotażowego (JESP4_01). W dodatku większość prób była studencka (co jest poważnym ograniczeniem trafności w badaniach nad uprzedzeniami: Henry, 2008), i wszystkie pochodziły z USA lub sporadycznie Kanady. A byłoby fascynujące poznać wyniki dla innych krajów, gdzie żyją obok siebie duże grupy osób białych i czarnoskórych (np. RPA).

Przykładem serii badań, która wyróżnia się pod względem wewnętrznego zróżnicowania jest grupa badań dotyczących związków między piciem alkoholu, a postrzeganymi niektórymi aspektami relacji romantycznej (JESP6). Miało miejsce relatywnie niewiele badań nawiązujących (5) tym niemniej warto zwrócić uwagę, że były to zarówno badania eksperymentalne, korelacyjne (w tym podłużne), oraz dodatkowo reanaliza dużej bazy danych armii amerykańskiej. Tylko dwie próby były studenckie, poza tym badano grupy takie jak ogólna populacja, osoby nadużywające alkoholu czy żołnierze armii. Pewną wadą tej serii jest z kolei to, że następuje przeskok między niektórymi hipotezami (nie jest oczywiste, że badane eksperymentalnie zależności między wypiciem alkoholu a negatywnymi interpretacjami wydarzeń w związku, są tym samym procesem co udokumentowane dalej zależności między piciem, a przemocą w związku).

Badania o związku słuchania agresywnego rapu ze zmianą percepcji społecznej osób czarnych (JESP9) są kolejnym przykładem na serię badań bardzo luźno powiązanych z wyjściową hipotezą. W każdym badaniu okazywało się, że słuchanie rapu na coś wpływa, ale zwykle na coś innego, a to na percepcję czarnych kobiet jako rozwiązyliwych, a to na poglądy względem relacji damsko-męskich, a to na przekonania o tym, czy warto przyznawać świadczeń osobom czarnym – każdorazowo inna szczegółowa hipoteza. Poza tym badanymi byli tylko Amerykanie i najczęściej studenci. Trudno byłoby tu doszukać się jakiegось systematyczności w programie badawczym.

W kilku dalszych cyklach (JPSP8, PSPB1, PSPB2, PSPB8, JESP1, JESP2, JESP3, JESP8, JESP10) liczba badań była umiarkowana (5-12). Choć są to bardzo różne cykle (pod względem wyjściowych hipotez) to mają kilka cech wspólnych. W obrębie tych cykli w poszczególnych badaniach zwykle testuje się podobne, ale nieco inne hipotezy, albo nawet przy testowaniu zbliżonej hipotezy robiono to na tyle różnymi

metodami, że trudno wyniki jednego badania odnosić do drugiego. Interpretacja takich wyników jest szczególnie szczególnie utrudniona, gdy badanie sporadycznie nie przynosi zbliżonych rezultatów – w takim wypadku trudno mówić o falsyfikacji wyjściowej hipotezy, bowiem ilość modyfikacji jest tak duża, że trudno jednoznacznie określić, które czynniki odpowiadają za zmianę wyników.

Ogółem więc z czterech celów jakie uważam za podstawowe przy badaniach na jakimś efekcie psychologicznym (dokładna replikacja, „behawioralna” weryfikacja odkrycia, zróżnicowanie grupowe, zróżnicowanie metod badawczych) opisywane w tej analizie cykle badań w minimalnym stopniu spełniają cel pierwszy i drugi, w niewielkim cel trzeci (i to raczej wybrane cykle, a nie wszystkie). Nie można jednak zaprzeczyć, że sporadycznie zdarzają się cykle które przyzwolicie realizują cel czwarty, tj. hipotezy testuje się na bardzo różne sposoby. Ktoś mógłby argumentować, że to wystarcza, bo skoro określona hipoteza jest potwierdzona różnymi metodami, to chyba świadczy to tym mocniej o jej prawdziwości. Owszem, do pewnego stopnia zgoda. W przypadku niektórych cykli badawczych nie mam wątpliwości, że danych jest tyle i są tak mocne, że w omawianych efektach coś musi być na rzeczy. Jednak wydaje się, że psychologia, nawet jako względnie mało dojrzała nauka, może pozwolić sobie na dużo więcej. Jeżeli określony efekt jest weryfikowany za pomocą mocno różnych metod lub hipotez szczegółowych utrudnia lub wręcz uniemożliwia to składanie cegiełek systemu wiedzy w jedną całość. Przykładowo, jeśli badania są różne i wyniki się różnią (lub nawet są podobne), wówczas nie wiadomo, które czynniki oddziałują na końcowy wynik. Bardziej pożądana byłaby sytuacja większego nawiązywania badań do siebie, tj. badacze używają zasadniczo tej samej procedury, w której modyfikują jeden element na raz, w takim wypadku można stwierdzić, czy element, którym się modyfikuje, ma jakieś (i jakie) znaczenie dla badanego efektu.

Mogę sobie wyobrazić, że duża liczba takich badań może nawet pozwolić na zaprojektowanie ostatecznie badania, gdzie badacza starają się w miarę precyzyjnie przewidzieć reakcje konkretnego Kowalskiego (a nie np. tylko kierunek i ewentualnie siłę różnic między grupami eksperymentalnymi). W analizowanych badaniach nie dostrzegam natomiast żadnego cyklu badawczego, w którym badacze systematycznie staraliby się zwiększać precyzję przewidywania określonego zachowania lub reakcji. Dla równowagi dodam, że nie znam takiego cyklu we

wszystkich znanych mi badaniach psychologii społecznej (nie tylko tutaj analizowanych). W przyszłości chciałbym się zmierzyć z pytaniem, czy psychologia jeszcze nie dojrzała do tego rodzaju działania, czy po prostu siłą przyzwyczajenia psychologowie nie dążą do tego.

Oczywiście fundamentalnym problemem związanym z dominacją replikacji konceptualnych nad dokładnymi jest to, że wiedza o wynikach pojedynczego badania ma niski stopień pewności. Jakaś jej część może być nieprawdziwa z racji przypadku, artefaktów lub świadomych zafałszowań (optymistycznie zakładam, że ewidentne fałszerstwa, typu stapelowskiego, należą do rzadkości). Nawet, jeśli większość badań jest przeprowadzana i przedstawiana rzetelnie, to nie ma narzędzi pozwalających na stuprocentowo trafne odróżnianie danych poprawnych od błędnych. Jediną drogą do tego są w ostatecznym rachunku replikacje dokładne. Nie potrafię sobie wyobrazić rozwoju sensownych koncepcji teoretycznych bez możliwości bazowania ich na korpusie wiedzy, która jest pewna tak bardzo jak to tylko możliwe (metoda naukowa nie gwarantuje wiedzy absolutnie prawdziwej, ale potrafi czasami zapewnić wiedzę wysoce pewną).

W rozdziale 3 postawiłem sobie kilka dodatkowych celów. Chciałem m.in. sprawdzić, czy w analizowanych cyklach badawczych nastąpiły jakieś inspiracje i duże „zmiany paradygmatu” dzięki rozwojowi teorii i/lub metod badawczych z innych dyscyplin naukowych (w obrębie psychologii i poza nią). Nie dostrzegam takich sytuacji, choć oczywiście można się zastanawiać czy takie sytuacje często występują, szczególnie w relatywnie krótkim okresie czasu (13 lat). Nie dostrzegam również takiego wyrafinowanego zwiększania wiedzy o teorii jak zarysowałem w rdz. 3.8.

Nie dostrzegłem też wielu przykładów, gdy badacze na podstawie systematycznych obserwacji, chcą wygenerować hipotezy do właściwego badania. Za takie badanie można uznać JESP4_01, gdzie eseje badanych pozwoliły na wygenerowanie puli pytań do kwestionariuszowego badania na temat treści współczesnych postaw rasistowskich. Sporadycznie badanym dawano możliwość zaprezentowania jakiejś spontanicznej reakcji w formie wypowiedzi pisemnej (np. PSPB5_22; EJSP9_03). Niektóre badania (wspomniane już PSPB1_09; PSPB10; JESP2_09; PSPB5_13) polegały na obserwacji zachowań i ocenie ich przez sędziów kompetentnych, jednak obserwacje te odbywały się pod kątem konkretnych wskaźników. Nie podawano w tych artykułach informacji, że sędziowie starali się zaobserwować coś

nieoczekiwanego, coś co wykraczało poza zadane skale szacunkowe. Można więc powiedzieć, że postulat Rozina (2001) o większym stopniu obserwacji w psychologii jest realizowany w stopniu minimalnym.

7.7. Wskazania na przyszłość

Omawiając wnioski na przyszłość można łatwo popaść w banał poprzez wyliczenie wniosków słusznych, ale tysięcy razy już powtórzonych w literaturze („przeprowadzajcie dokładne replikacje”; „testujcie hipotezy na członkach innych kultur” itd.). Bez uwzględnienia szerszego kontekstu te „słuszne” zalecenia będą zupełnie nieskuteczne. Weźmy problem replikacji dokładnych. Co do zasady jest w literaturze konsensus, że replikacje dokładne są czymś wysoce pożądanym i ważnym w rozwoju nauki. Nie wspomina się jednak, że nie ma żadnych szans przy obecnych praktykach badawczo-publikacyjnych, żeby ten postulat zacząć realizować dosłownie w całej rozciągłości. Badań i opisanych w literaturze efektów jest całe multum (ja wyodrębniłem z części literatury 40 oryginalnych efektów, ale tak naprawdę w literaturze nawiązującej znajduje się dalszych kilkaset efektów, które też można od biedy zakwalifikować jako „oryginalne”, bowiem w tej formie nikt ich wcześniej nie testował). Choćby ze względów logistycznych nie ma możliwości, żeby powtórzyć wszystkie badania z ostatnich paru lat (przykład projektu wielkiej replikacji Briana Noseka jest tu wymownym przykładem).

Dlatego uważam, że w pierwszym rzędzie powinna być fundamentalnie zmieniona praktyka publikacyjna. Przykład Stapela (i nie tylko) pokazują, że do tej pory kluczowe znaczenie dla publikowalności artykułów mają istotne wyniki („sukces”) i ich oryginalność („nowatorstwo”). Sądzę, że w tym przeglądzie udało mi się wskazywać kilka przykładów badań, które są „sukcesem” i „nowatorskie”, ale które testują banalne hipotezy (np. JPSP1, JPSP2), robią to w sztabowy sposób (np. EJSP5) lub wyniki wydają się być podkreślone metodami psychologii fałszywie pozytywnej (np. PSPB1, PSPB2, EJSP3).

Nie sądzę, żeby warto było replikować takie badania dla samej zasady replikacji. Takie badania w ogóle w pierwszym rzędzie nie powinny się były ukazywać. Dlatego jestem zdania, że zanim rozpocznie się wprowadzenie wszystkich słusznych zaleceń,

należy w pierwszym rzędzie zacząć kłaść akcent na coś innego niż odkrycia („nowatorstwo”) i istotność wyników („sukces”).

Podstawowy akcent powinien być położony na ważność pytania badawczego – czy ujawnia ono coś ważnego o ludzkim zachowaniu, czy przyczynia się do rozwoju teorii (również poprzez to, że wzmacnia jakąś teorię), czy odpowiedź na nie ma istotne implikacje praktyczne. Równie duży akcent powinien być położony na pewność wiedzy: czy badanie ma dużą moc, czy nie ma wskazówek psychologii fałszywie pozytywnej (np. 10 na 10 porównań działa tak jak przewiduje teoria, choć statystycznie jest znikome prawdopodobieństwo takiego wyniku). Przy decyzjach publikacyjnych premiowany powinien też być wysiłek badawczy: czy grupy były dosyć liczne, czy zastosowano nietrywialne, pracochłonne metody, czy sięgnięto po próby, które są trudniej dostępne niż studenci na wydziale. Być może również warto większą uwagę zwracać na wielkość efektów, szczególnie w przypadku mało wyszukanych badań (np. czy warto publikować wszystkie badania, gdzie po korelacji testów z testami osiągnęliśmy wielkości efektów rzędu $r = 0,10-0,15$?; tak jak w EJSP5), choć tu pojawia się ryzyko, że badacze zaczną z kolei „podkręcać” wielkości efektów, a nie tylko same wskaźniki istotności statystycznej. Również ważne jest, że jeżeli badacz przedstawia serię badań, to żeby przedstawiał wszystkie wyniki, jakie udało mu się zebrać, nawet jeśli część jest nieistotna i obniża „atrakcyjność” manuskryptu. Redakcje powinny akceptować takie niedoskonałości, a badacze nie powinni bać się ich przedstawiać (pod ryzykiem odrzucenia). Jeśli, dajmy na to, uzyskuje się w serii rzetelnych badań o wysokiej mocy 5 na 9 istotnych wyników, to mimo wszystko jest to niezły prognostyk, że określony efekt może występować i warto go dalej badać (a nieistotne dane mogą ujawniać coś ważnego o samym efekcie). Nie jest to tak „atrakcyjne” jak przedstawić serię tylko 5 samych udanych badań, ale nauka to nie sztuka, atrakcyjność treści ma znaczenie drugorzędne.

Stąd największej nadziei upatruję w propozycji czasopism *Cortex* i *Attention, Perception & Psychophysics* (rdz. 1.10), żeby badacze zgłaszali projekt badania przed zebraniem danych, i żeby wtedy następowała zasadnicza recenzja. Po akceptacji, badacz dostawałby glejt na publikację i nie miałby żadnego interesu w tym, żeby upiększać wyniki.

Ogólnie więc postulowałbym w pierwszym rzędzie zacieśnienie niektórych kryteriów dopuszczania do publikacji oryginalnych odkryć, przy równoczesnym rozluźnieniu

innych kryteriów, szczególnie dotyczących publikacji replikacji i wyników negatywnych. Jeśli odkrycia będą nawet 100 razy rzadsze niż dziś, ale będą bardziej pewne i rzetelne, nie sądzę, żeby nauka na tym straciła. Gdy odkrycia będą wysoce prawdopodobne, dopiero wówczas duża część wysiłku badawczego może być skupiona na replikacjach (tych badań) i innych oczywistych zaleceniach (jak replikacja międzykulturowa i międzykulturowa, bardziej systematyczne nawiązywanie wyników do siebie, weryfikacja behawioralna efektu itd.). Przy czym podstawowymi kryteriami publikacyjnymi powinny być zawsze ważność pytania badawczego i poprawność metodologiczna. Sam wynik powinien mieć znaczenie marginalne.

Jesteśmy teraz bardzo odlegli od takiego systemu. Jestem przekonany, że w analizowanych przeze mnie badaniach jest aż nadto przykładów, że główną bolączką psychologii jest nastawienie na (pozorne) „nowatorstwo” i „sukces” (razem tworzące atrakcyjność manuskryptu wyżej cenioną niż jego rzetelność). Obserwując praktyki badawcze w swoim kręgu uczelniano-towarzyskim, odnoszę wrażenie, że jesteśmy socjalizowani do myślenia, że „sukces” w postaci istotnych wyników jest ostatecznym celem. Konsekwencją jest coś, co nazywam „kulturą kombinowania”. Jeśli wynik jest nieistotny szuka się tych elementów danych, które są jednak istotne (nawet jeśli nie pojawiały się początkowo jako hipoteza), ewentualnie próbuje się zmieniać elementy procedury aż do skutku. (A może trzeba użyć podobnej, ale innej skali? A może warto zmienić kilka zdań w instrukcji? A może... itd.). I oczywiście, w momencie znalezienia wyników pozytywnych zaprzestaje się dalszych poszukiwań. Ta „kultura kombinowania” nie ma drugiej strony medalu, czegoś co można by nazwać „kulturą falsyfikacji”, to znaczy po znalezieniu wyników pozytywnych co do zasady raczej nie podejmuje się już replikacji dokładnych lub takich rodzajów badań, które samej hipotezy nie wzmocnią, a mogą ją potencjalnie sfalsyfikować.

Siła przyzwyczajęń jest wielka w każdym obszarze życia. Zasygnalizowałem już (rdz. 1.10), że pojawiają się inicjatywy zmian, niekiedy daleko idące. Nie zdziwię się, jeśli się okaże, że zmiany mające charakter bardziej techniczny staną się czymś trwałym (np. udostępnianie danych surowych, prejestracja badań, publikowanie materiałów użytych w badaniu itp.). W przypadku zmian bardziej fundamentalnych jestem większym pesymistą. W pewnym sensie praktyki badawcze i publikacyjne są złożoną grą między badaczami, redaktorami, instytucjami przyznającymi środki na badania i opinią publiczną.

Badacze chcą robić karierę, a w obecnym systemie konkurencji akademickiej jej podstawą są liczne publikacje (najlepiej często cytowane). Nasz umysł jest tak skonstruowany, że faktycznie badania „nowatorskie” z jednoznacznym „sukcesem” mogą wywoływać większy podziw i przyciągać większą uwagę niż badania w stu procentach rzetelne, a więc (jak to zwykle bywa) mniej efektowne. W tej chwili ukazuje się za mało replikacji, żeby móc to w pełni ocenić, ale może się z czasem okazać, że typowa replikacja przyciąga średnio dużo mniej cytowań niż badanie „nowatorskie”.

Redakcje czasopism mogą szczerze chcieć badań rzetelnych, zarazem jednak mogą faktycznie zwracać uwagę na te artykuły, które dadzą największą cytowalność. Jeżeli nie widać oczywistych wskaźników nierzetelności, to dlaczego nie uwierzyć autorom, że przeprowadzili faktycznie serię kilku samych udanych badań? Opinii publicznej i decydentom łatwiej będzie też „sprzedać” historie efektowne i jednoznaczne, tak jak to miało miejsce niedawno w przypadku stanowiska APA na zagrożenia związane z grami komputerowymi (czego jednak Sąd Najwyższy USA łatwo nie „kupił”). Być może – spekulowałbym – działa tu też przekonanie redaktorów i badaczy, że zasadnicza praca nad „odkryciem” kończy się w momencie jego dokonania. W nauce, w której istnieją bardzo wyrafinowane procedury i narzędzia badawcze (jak w fizyce) i ryzyko artefaktów lub przypadku jest względnie niewielkie w dużej mierze może tak to wyglądać (jednak ciągle należy mieć się na baczności przed oszustami, jak pokazał przypadek Jana Hendrika Schöna; Reich, 2009). Psychologia natomiast z pewnością nie jest tego rodzaju nauką. Ostatecznym problemem nie jest to, że większość badań jest sfabrykowana (wątpię, żeby tak było). Problemem jest to, że nie ma stuprocentowo trafnych metod odróżniania badań całkowicie rzetelnych od tych, które są w ten czy inny sposób „upiększone”.

Badaczom oczywiście też zależy jakoś na rzetelności własnych badań, ale zapewne również zależy im też na tym, żeby móc utrzymać etat, awansować w hierarchii akademickiej, albo zdobywać wyróżnienia, najlepiej przy jak najmniejszym ryzyku. Jak się wydaje zrobienie pięciu małych „nowatorskich” badań, wiąże się z dużo mniejszym ryzykiem niż jednego dużego, bo jest większa szansa na znalezienie jakichś wyników istotnych. Dodatkowo, jeżeli, tak jak do tej pory, były nikłe szanse na kontrolę poprzez replikacje (dzięki czemu Stapel mógł działać ponad 15 lat w swoim procederze), badacze mogą mieć poczucie względnej bezkarności. Nawet, jeśli nie

decydują się na jawne fałszerstwa, to drobne przekłamania są powszechnie akceptowane (John, Loewenstein, Prelec, 2012), a wybiórcze publikowanie jest wręcz wymagane przez redakcje (Levelt, 2012, s. 53).

Redukowanie ryzyka porażki dla badacza oznacza też jak najmniejszy wysiłek przy pojedynczym badaniu, bowiem badacz, który chce mieć wiele badań, (z których wybierze sobie „wisienki”), musi siłą rzeczy ograniczać się przy każdym pojedynczym badaniu (względnie jeśli robi jedno badanie, sposobem na ograniczenie ryzyka będzie multum zmiennych i wybór z nich tylko działających porównań). Stąd najczęściej badani są studenci lub inne grupy względnie łatwo dostępne (jak dzieci w szkole, czy badani w internecie). Stąd też będą dominowały raczej metody samoopisowe, a nie behawioralne. Stąd też wielkości grup będą takie, że moc badań jest umiarkowana. Dodam jednak dla równowagi, że analiza procedur badawczych, szczególnie ich realizmu, skłania mnie do wniosku, że w tym akurat obszarze badacze decydują się często na mniejszą niż minimalna złożoność i pracowitość procedury. Być może ostateczne znaczenie ma to, że pewne niepożądane cechy badań (np. wykorzystywanie głównie studentów) mogą być szeroko akceptowane, ale oczekiwane jest stosowanie niebanalnych procedur badawczych (tu jednak należy się dodatkowe zastrzeżenie, że tak czy owak rzadko korzysta się z zachowania jako zmiennej zależnej).

Przezwyciężenie tego błędnego koła może być dużo trudniejsze niż się wydaje. Pamiętajmy, że były już w historii czasopisma nastawione na replikacje i wyniki negatywne (opisywałem je w rozdziale 1.8), które jednak nie odniosły żadnego sukcesu.

Organizacjom naukowo-zawodowym takim jak APS czy APA może zależeć na tym, żeby czasopisma systematycznie publikowały badania, które będą atrakcyjne dla opinii publicznej i instytucji rządowych przyznających środki na naukę. Niestety, znowu można powiedzieć, że replikacje dokładne takimi badaniami nie są, bo „nie pokazują nic nowego”.

Jestem przekonany, że sama dobra wola grupki redaktorów i badaczy nie wystarczy. Zasadniczą rolę powinny właśnie odegrać instytucje przyznające granty i organizacje takie jak APA albo APS. W tej chwili wg mojej wiedzy większość organizacji przyznających granty stawia na „innowacyjność”. Przykładowo na naszym podwórku Narodowe Centrum Nauki zamieszcza ogłoszenia na konkursy grantowe. Przy

konkursie najwyższego stopnia Maestro ogłasza na samym wstępie, że muszą to być badania „mające na celu realizację pionierskich badań naukowych [...] ważnych dla rozwoju nauki, wykraczających poza dotychczasowy stan wiedzy, i których efektem mogą być odkrycia naukowe”. W przypadku konkursów niższego szczebla nie ma co prawda na wstępie tak wygórowanych oczekiwań, ale jednym z wymienianych kryteriów jest „nowatorski charakter projektu”.

Dlatego więc na chwilę obecną jestem pesymistą w kwestii tego, czy nastąpi fundamentalna zmiana. Nawet, jeśli grupka badaczy podejdzie poważnie do problemu i za kilka lat 5 % publikowanych badań będzie replikacjami dokładnymi (lub międzykulturowymi), ciągle olbrzymia większość badań nie będzie mogła doczekać się replikacji.

Zmiana kultury badawczej i publikacyjnej wymaga prawdopodobnie radykalnych rozwiązań. Jakies przykłady? APA i APS mogłyby ogłosić, że tylko 1/3 publikowanych artykułów w ich czasopismach będą stanowiły badania „nowatorskie”, resztę miejsca przeznaczy na replikacje dokładne, bardzo zbliżone replikacje konceptualne i replikacje międzykulturowe. Redakcje miałyby obowiązek przyjąć do druku wszystkie poprawnie przeprowadzone (i o dużej mocy) replikacje badań u nich opublikowanych w ciągu pięciu lat od daty publikacji oryginalnego artykułu. Kryterium awansu na najwyższe stanowiska mogłoby być przedstawienie replikacji swoich kluczowych badań (przeprowadzonych w kooperacji z niezależnymi badaczami). Organizacje grantowe przyznają granty na duże replikacje w tym wszechstronne międzykulturowe sprawdzenie określonego efektu itp.

Niedawno pojawiła się radykalna propozycja, żeby badacze mieli „zakaz” publikowania więcej niż jednego artykułu rocznie (Nelson, Simmons, Simonsohn, 2012). Na pierwszy rzut oka, szczególnie z punktu widzenia badacza z Polski, propozycja ta brzmi absurdalnie, u nas raczej problemem są osoby wykazujące zbyt małą aktywność publikacyjną. Tym niemniej po dłuższym zastanowieniu nabiera ona pewnego sensu w moich oczach. W „wyścigu zbrojeń” najbardziej pracowici badacze publikują tak dużo artykułów, że siłą rzeczy musi na tym tracić ich jakość. Stapel publikował średnio 10 artykułów rocznie (oczywiście część jako współautor) i nie powiedzielibyśmy, że świat nauki dobrze na tym wyszedł. Być może, gdyby faktycznie najbardziej kreatywni badacze zamiast produkować dziesiątki artykułów z pobieżnie przetestowanymi hipotezami, skupili się na tym, żeby wszechstronnie

zbadać jeden efekt i opisać go jak najszerzej w jednym artykule, psychologia mogłaby tylko zyskać.

W tej podsumowującej części chciałem - nieco spekulatywnie – pokazać jakie metacynniki mogły się przyczynić się do kryzysu, którego ukoronowaniem było oszustwo Stapela i kilka pomniejszych oszustw. Mam nadzieję, że udało mi się pokazać, że proste porady w rodzaju „róbmy więcej replikacji” mogą nie wystarczyć, jeśli nie weźmiemy pod uwagę szerszych uwarunkowań obecnego stanu rzeczy.

Tak jak zauważyłem, osobiście jestem na chwilą obecną raczej pesymistą, co do tego czy nastąpi fundamentalna zmiana. Z drugiej jednak strony u podstaw pracy badawczej leży często autentyczna pasja poznawania, a nie wierzę, że może ona być w pełni realizowana przy obecnie systemie badawczo-publikacyjnym. Badacze stanowią elitę intelektualną społeczeństwa, więc być może prawdziwa zmiana zostanie zainicjowana, gdy tylko to sobie w pełni uświadomią.

Załącznik 1

Wykaz badań z oryginalnymi odkryciami uwzględnionych w analizie i uzasadnienie nieuwzględnienie niektórych artykułów

Artykuł	Decyzja
JPSP	
Wigboldus, D. H., Semin, G. R., Spears, R. (2000). How do we communicate stereotypes? Linguistic bases and inferential consequences. <i>Journal of Personality and Social Psychology</i> , 78(1), 5-18.	Brak dostatecznej nowości
Stapel, D. A., Koomen, W. (2000). How far do we go beyond the information given? The impact of knowledge activation on interpretation and inference. <i>Journal of Personality and Social Psychology</i> , 78(1), 19-37.	Uwzględniono w analizie (JPSP1)
Boldero, J., Francis, J. (2000). The relation between self-discrepancies and emotion: the moderating roles of self-guide. <i>Journal of Personality and Social Psychology</i> , 78(1), 38-52.	Uwzględniono w analizie (JPSP2)
Aarts, H., Dijksterhuis, A. (2000). Habits as knowledge structures: automaticity in goal-directed behavior. <i>Journal of Personality and Social Psychology</i> , 78(1), 53-63.	Brak lub niejasne implikacje behawioralne
Reynolds, K. J., Turner, J. C., Haslam, S. (2000). When are we better than them and they worse than us? A closer look at social discrimination in positive and negative domains. <i>Journal of Personality and Social Psychology</i> , 78(1), 64-80.	Uwzględniono w analizie (JPSP3)
Haselton, M. G., Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. <i>Journal of Personality and Social Psychology</i> , 78(1), 81-91.	Uwzględniono w analizie (JPSP4)
van Dijk, E., Wilke, H. (2000). Decision-induced focusing in social dilemmas: give-some, keep-some, take-some, and leave-some dilemmas. <i>Journal of Personality and Social Psychology</i> , 78(1), 92-104.	Uwzględniono w analizie (JPSP5)
Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. <i>Journal of Personality and Social Psychology</i> , 78(1), 105-119.	Brak lub niejasne implikacje behawioralne
Rothbart, M. K., Ahadi, S. A., Evans, D. E. (2000). Temperament and personality: origins and outcomes. <i>Journal of Personality and Social Psychology</i> , 78(1), 122-135.	Badanie podłużne
Eisenberg, N., Fabes, R. A., Guthrie, I. K., Reiser, M. (2000). Dispositional emotionality and regulation: their role in predicting quality of social functioning. <i>Journal of Personality and Social Psychology</i> , 78(1), 136-157.	Badanie podłużne
Caspi, A. (2000). The child is father of the man: personality continuities from childhood to adulthood. <i>Journal of Personality and Social Psychology</i> , 78(1), 158-172.	Badanie podłużne
McCrae, R. R., Costa Jr, P. T., Ostendorf, F., Angleitner, A., Hřebíčková, M., Avia, M. D., ... Smith, P. B. (2000). Nature over nurture: temperament, personality, and life span development. <i>Journal of Personality and Social Psychology</i> , 78(1), 173-186.	Badanie podłużne
Fazio, R. H., Ledbetter, J. E., Towles-Schwen, T. (2000). On the costs of accessible attitudes: detecting that the attitude object has changed. <i>Journal of Personality and Social Psychology</i> , 78(2), 197-210.	Brak lub niejasne implikacje behawioralne
Gilovich, T., Medvec, V., Savitsky, K. (2000). The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. <i>Journal of Personality and Social Psychology</i> , 78(2), 211-222.	Uwzględniono w analizie (JPSP6)

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., Uhles, A. (2000). Varieties of groups and the perception of group entitativity. <i>Journal of Personality and Social Psychology</i> , 78(2), 223-246.	Uwzględniono w analizie (JPSP7)
Chiu, C. Y., Morris, M. W., Hong, Y. Y., Menon, T. (2000). Motivated cultural cognition: the impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. <i>Journal of Personality and Social Psychology</i> , 78(2), 247-259.	Uwzględniono w analizie (JPSP8)
Dolinski, D. (2000). On inferring one's beliefs from one's attempt and consequences for subsequent compliance. <i>Journal of Personality and Social Psychology</i> , 78(2), 260-272.	Uwzględniono w analizie (JPSP9)
Aron, A., Norman, C. C., Aron, E. N., McKenna, C., Heyman, R. E. (2000). Couples' shared participation in novel and arousing activities and experienced relationship quality. <i>Journal of Personality and Social Psychology</i> , 78(2), 273-284.	Uwzględniono w analizie (JPSP10)
EJSP	
Higgins, E. T. (2000). Social cognition: learning about what matters in the social world. <i>European Journal of Social Psychology</i> , 30(1), 3-39.	Artykuł przeglądowy
Sidanius, J., Levin, S., Liu, J., Pratto, F. (2000). Social dominance orientation, anti - egalitarianism and the political psychology of gender: an extension and cross - cultural replication. <i>European Journal of Social Psychology</i> , 30(1), 41-67.	Uwzględniono w analizie (EJSP1)
Platow, M. J., Mills, D., Morrison, D. (2000). The effects of social context, source fairness, and perceived self - source similarity on social influence: a self - categorisation analysis. <i>European Journal of Social Psychology</i> , 30(1), 69-81.	Artykuł przeglądowy
Reid, S. A., Ng, S. H. (2000). Conversation as a resource for influence: evidence for prototypical arguments and social identification processes. <i>European Journal of Social Psychology</i> , 30(1), 83-100.	Uwzględniono w analizie (EJSP2)
Thill, E. E., Cury, F. (2000). Learning to play golf under different goal conditions: their effects on irrelevant thoughts and on subsequent control strategies. <i>European Journal of Social Psychology</i> , 30(1), 101-122.	Uwzględniono w analizie (EJSP3)
Galinsky, A. D., Stone, J., Cooper, J. (2000). The reinstatement of dissonance and psychological discomfort following failed affirmations. <i>European Journal of Social Psychology</i> , 30(1), 123-147.	Uwzględniono w analizie (EJSP4)
Schwarz, N. (2000). Social judgment and attitudes: warmer, more social, and less conscious. <i>European Journal of Social Psychology</i> , 30(2), 149-176.	Artykuł przeglądowy
Sagiv, L., Schwartz, S. H. (2000). Value priorities and subjective well-being: direct relations and congruity effects. <i>European Journal of Social Psychology</i> , 30(2), 177-198.	Uwzględniono w analizie (EJSP5)
Todorov, A., Lalljee, M., Hirst, W. (2000). Communication context, explanation, and social judgment. <i>European Journal of Social Psychology</i> , 30(2), 199-209.	Uwzględniono w analizie (EJSP6)
Niedenthal, P. M., Halberstadt, J. B., Margolin, J., Innes - Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. <i>European Journal of Social Psychology</i> , 30(2), 211-222.	Brak lub niejasne implikacje behawioralne
Castelli, L., Zogmaister, C. (2000). The role of familiarity in implicit memory effects: the case of exemplar activation. <i>European Journal of Social Psychology</i> , 30(2), 223-234.	Brak lub niejasne implikacje behawioralne
Harris, P., Middleton, W., Joiner, R. (2000). The typical student as an in-group member: eliminating optimistic bias by reducing social distance. <i>European Journal of Social Psychology</i> , 30(2), 235-253.	Uwzględniono w analizie (EJSP7)

Giebels, E., De Dreu, C. W., Van de Vliert, E. (2000). Interdependence in negotiation: effects of exit options and social motive on distributive and integrative negotiation. <i>European Journal of Social Psychology</i> , 30(2), 255-272.	Uwzględniono w analizie (EJSP8)
Sanna, L. J., Turley-Ames, K. J. (2000). Counterfactual intensity. <i>European Journal of Social Psychology</i> , 30(2), 273-296.	Uwzględniono w analizie (EJSP9)
Fiske, S. T. (2000). Stereotyping, prejudice, and discrimination at the seam between the centuries: evolution, culture, mind, and brain. <i>European Journal of Social Psychology</i> , 30(3), 299-322.	Artykuł przeglądowy
Roccas, S., Horenczyk, G., Schwartz, S. H. (2000). Acculturation discrepancies and well - being: the moderating role of conformity. <i>European Journal of Social Psychology</i> , 30(3), 323-334.	Uwzględniono w analizie (EJSP10)
PSPB	
Räikkönen, K., Katainen, S., Keskivaara, P., Kelikangas-Järvinen, L. (2000). Temperament, mothering, and hostile attitudes: a 12-year longitudinal study. <i>Personality and Social Psychology Bulletin</i> , 26(1), 3-12.	Badanie podłużne
Wyer, N. A., Sherman, J. W., Stroessner, S. J. (2000). The roles of motivation and ability in controlling the consequences of stereotype suppression. <i>Personality and Social Psychology Bulletin</i> , 26(1), 13-25.	Uwzględniono w analizie (PSPB1)
Sherman, J. W., Frost, L. A. (2000). On the encoding of stereotype-relevant information under cognitive load. <i>Personality and Social Psychology Bulletin</i> , 26(1), 26-34.	Uwzględniono w analizie (PSPB2)
Batson, C., Bowers, M. J., Leonard, E. A., Smith, E. C. (2000). Does personal mortality exacerbate or restrain retaliation after being harmed?. <i>Personality and Social Psychology Bulletin</i> , 26(1), 35-45.	Uwzględniono w analizie (PSPB3)
Gordon, A. K., Miller, A. G. (2000). Perspective differences in the construal of lies: is deception in the eye of the beholder?. <i>Personality and Social Psychology Bulletin</i> , 26(1), 46-55.	Brak nowości (zbliżona replikacja innego badania)
Gibson, B., Sachau, D. (2000). Sandbagging as a self-presentational strategy: claiming to be less than you are. <i>Personality and Social Psychology Bulletin</i> , 26(1), 56-70.	Uwzględniono w analizie (PSPB4)
Heine, S. J., Takata, T., Lehman, D. R. (2000). Beyond self-presentation: evidence for self-criticism among Japanese. <i>Personality and Social Psychology Bulletin</i> , 26(1), 71-78.	Uwzględniono w analizie (PSPB5)
Paese, P. W., Gilin, D. A. (2000). When an adversary is caught telling the truth: reciprocal cooperation versus self-interest in distributive bargaining. <i>Personality and Social Psychology Bulletin</i> , 26(1), 79-90.	Uwzględniono w analizie (PSPB6)
Greenberg, J., Arndt, J., Simon, L., Pyszczynski, T., Solomon, S. (2000). Proximal and distal defenses in response to reminders of one's mortality: evidence of a temporal sequence. <i>Personality and Social Psychology Bulletin</i> , 26(1), 91-99.	Uwzględniono w analizie (PSPB7)
Harkins, S. G., White, P. H., Utman, C. H. (2000). The role of internal and external sources of evaluation in motivating task performance. <i>Personality and Social Psychology Bulletin</i> , 26(1), 100-117.	Uwzględniono w analizie (PSPB8)
Forrest, J. A., Feldman, R. S. (2000). Detecting deception and judge's involvement: lower task involvement leads to better lie detection. <i>Personality and Social Psychology Bulletin</i> , 26(1), 118-125.	Uwzględniono w analizie (PSPB9)
Nelson, L. J., Klutas, K. (2000). The distinctiveness effect in social interaction: creation of a self-fulfilling prophecy. <i>Personality and Social Psychology Bulletin</i> , 26(1), 126-135.	Uwzględniono w analizie (PSPB10)
JESP	

van Dijk, E., Vermunt, R. (2000). Strategy and fairness in social decision making: sometimes it pays to be powerless. <i>Journal of Experimental Social Psychology</i> , 36(1), 1-25.	Uwzględniono w analizie (JESP1)
Drolet, A. L., Morris, M. W. (2000). Rapport in conflict resolution: accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. <i>Journal of Experimental Social Psychology</i> , 36(1), 26-50.	Uwzględniono w analizie (JESP2)
Fine, G. A., Elsbach, K. D. (2000). Ethnography and experiment in social psychological theory building: tactics for integrating qualitative field data with quantitative lab data. <i>Journal of Experimental Social Psychology</i> , 36(1), 51-76.	Artykuł koncepcyjny
Otten, S., Moskowitz, G. B. (2000). Evidence for implicit evaluative in-group bias: affect-biased spontaneous trait inference in a minimal group paradigm. <i>Journal of Experimental Social Psychology</i> , 36(1), 77-89.	Brak nowości (zbliżona replikacja innego badania)
Ryan, C. S., Bogart, L. M., Vender, J. P. (2000). Effects of perceived group variability on the gathering of information about individual group members. <i>Journal of Experimental Social Psychology</i> , 36(1), 90-101.	Brak nowości (zbliżona replikacja innego badania)
Hewstone, M., Hamberger, J. (2000). Perceived variability and stereotype change. <i>Journal of Experimental Social Psychology</i> , 36(2), 103-124.	Uwzględniono w analizie (JESP3)
Monteith, M. J., Spicer, C. V. (2000). Contents and correlates of whites' and blacks' racial attitudes. <i>Journal of Experimental Social Psychology</i> , 36(2), 125-154.	Uwzględniono w analizie (JESP4)
Croizet, J., Fiske, S. T. (2000). Moderation of priming by goals: feeling entitled to judge increases judged usability of evaluative primes. <i>Journal of Experimental Social Psychology</i> , 36(2), 155-181.	Uwzględniono w analizie (JESP5)
MacDonald, G., Zanna, M. P. (2000). An experimental test of the role of alcohol in relationship conflict. <i>Journal of Experimental Social Psychology</i> , 36(2), 182-193.	Uwzględniono w analizie (JESP6)
Mussweiler, T., Neumann, R. (2000). Sources of mental contamination: comparing the effects of self-generated versus externally provided primes. <i>Journal of Experimental Social Psychology</i> , 36(2), 194-206.	Uwzględniono w analizie (JESP7)
Jost, J. T., Thompson, E. P. (2000). Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and European Americans. <i>Journal of Experimental Social Psychology</i> , 36(3), 209-232.	Uwzględniono w analizie (JESP8)
Johnson, J. D., Trawalter, S., Dovidio, J. F. (2000). Converging interracial consequences of exposure to violent rap music on stereotypical attributions of blacks. <i>Journal of Experimental Social Psychology</i> , 36(3), 233-251.	Uwzględniono w analizie (JESP9)
Idson, L., Liberman, N., Higgins, E. (2000). Distinguishing gains from nonlosses and losses from nongains: a regulatory focus perspective on hedonic intensity. <i>Journal of Experimental Social Psychology</i> , 36(3), 252-274.	Uwzględniono w analizie (JESP10)

Załącznik 2

Dane surowe

Dane surowe załączono w dołączonej do pracy płycie CD. Zawiera ona dwa pliki: wyniki.xls oraz wyniki.sav. Pierwszy plik zawiera dane w formacie tabeli Excel, drugi w formacie SPSS. W przypadku pliku Excela legenda do wyników znajduje się w zakładce „Legenda”. W przypadku pliku .sav (format SPSS) kolumna etykieta zawiera jednoznaczny opis zmiennej.

LITERATURA

- Abbott, A. (2013). Disputed results a fresh blow for social psychology. *Nature*, 497(7447), 16.
- Acker, F. (2008). New findings on unconscious versus conscious thought in decision making: additional empirical data and meta-analysis. *Judgment and Decision Making*, 3(4), 292-303.
- Ackerman, J. M., Griskevicius, V., Li, N. P. (2011). Let's get serious: communicating commitment in romantic relationships. *Journal of Personality and Social Psychology*, 100(6), 1079-1094.
- Ajzen, I. (2000). Nature and operation of attitudes. *Annual Review of Psychology*, 52, 27–58.
- Albarracin, D., Johnson, B. T., Fishbein, M., Muellerleile, P. A. (2001). Theories of reasoned action and planned behavior as models of condom use: a meta-analysis. *Psychological Bulletin*, 127(1), 142-161.
- Aldhous, P. (2011). Journal rejects studies contradicting precognition. Pozyskano z: <http://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition.html#.VJdussDM>
- Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., ..., McCoy, S. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9(5), 556-578.
- American Psychological Association (APA) (2005). Resolution on violence in video games and interactive media. Pozyskano z: <http://www.apa.org/about/policy/interactive-media.pdf>
- American Psychological Association, Ethics Committee. (2010). Report of the Ethics Committee, 2009. *American Psychologist*, 65, 483–492.
- Anderson, C. A., Dill, K. (2000). Video games and aggressive thoughts, feelings and behavior in the laboratory and in life. *Journal of Personality and Social Psychology*, 78(4), 772–790.
- Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A.,..., Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: a meta-analytic review. *Psychological Bulletin*, 136(2), 151–173.

- Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355-365.
- Anderson, J. R. (2014). *Rules of the mind*. New York: Psychology Press.
- Anderson, J. R., Matessa, M., Lebiere, C. (1997). ACT-R: a theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4), 439-462.
- Anderson, K. (2005). *Predicting the weather: Victorians and the science of meteorology*. Chicago: University of Chicago Press.
- Anderson, M. L. (2003). Embodied cognition: a field guide. *Artificial Intelligence*, 149(1), 91-130.
- Armitage, C. J., Conner, M. (2001). Efficacy of the theory of planned behaviour: a meta-analytic review. *British Journal of Social Psychology*, 40(4), 471-499.
- Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *American Psychologist*, 63(7), 602-614.
- Asch, S. E. (1952/1987). *Social psychology*. New York: Oxford University Press.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ..., Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2), 108–119.
- Association for Psychological Science (2013). *Registered Replication Reports*.
 Pozyskano z: <https://www.psychologicalscience.org/index.php/replication>
- Bakker, M. Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3), 666-678.
- Bakker, M., Cramer, A. O., Matzke, D., Kievit, R. A., Van Der Maas, H. L. J., Wagenmakers, E. J., Borsboom, D. (2013). Dwelling on the past. *European Journal of Personality*, 27(2), 120-144.
- 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.
- Bargh, J. A., Chen, M., Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2), 230-244.
- Barnett, C. (2015). *Rain: A Natural and Cultural History*. New York: Crown Publishing Group.

- Bartlett, T. (2012). Is psychology about to come undone?. *Chronicle of Higher Education*. Pozyskano z: http://chronicle.com/blogs/percolator/is-psychology-about-to-come-undone/29045?sid=at&utm_source=at&utm_medium=en
- 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(4), 396-403.
- Begley, C. G., Ellis, L. M. (2012). Improve standards for preclinical cancer research. *Nature*, 483(7391), 531–533.
- Belles, S., Kunde, W., Neumann, R. (2010). Timing of sexual maturation and women's evaluation of men. *Personality and Social Psychology Bulletin*, 36(5), 703-714.
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407-425.
- Bergsieker, H., Shelton, J., Richeson, J. (2010). To be liked versus respected: Divergent goals in interracial interactions. *Journal of Personality and Social Psychology*, 99(2), 248-264.
- Bering, J. M., Shackelford, T. K. (2004). The causal role of consciousness: a conceptual addendum to human evolutionary psychology. *Review of General Psychology*, 8(4), 227-248.
- Bermúdez, J. L. (Ed.). (2007). *Philosophy of psychology*. New York: Routledge.
- Bernardo, A. I., Zhang, L., Callueng, C. M. (2002). Thinking styles and academic achievement among Filipino students. *The Journal of Genetic Psychology: Research and Theory on Human Development*, 163(2), 149-164.
- Bartolucci, A. D., Zeichner, A. (2003). Sex, hypermasculinity, and hyperfemininity in perception of social cues in neutral interactions. *Psychological Reports*, 92, 75–83.
- Bevan, W. (1991). Contemporary psychology: a tour inside the onion. *American Psychologist*, 46(5), 475-483.
- Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times*. Pozyskano z: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?_r=0
- Blanton, H., Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, 61(1), 27-41.

- Blanton, H., Mitchell, G. (2011). Reassessing the predictive validity of the IAT II: reanalysis of Heider & Skowronski (2007). *North American Journal of Psychology*, 13(1), 99-106.
- Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., Tetlock, P. E. (2009). Strong claims and weak evidence: reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, 94(3), 567–582.
- 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.
- Botterill, G., Carruthers, P. (1999). *The philosophy of psychology*. Cambridge: Cambridge University Press.
- Boyack, K.W., Klavans, R., Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Bressan, P., Stranieri, D. (2008). The best men are (not always) already taken. Female preference for single versus attached males depends on conception risk. *Psychological Science*, 19(2), 145-151.
- Brochu, P. M., Morrison, M. A. (2007). Implicit and explicit prejudice toward overweight and average-weight men and women: testing their correspondence and relation to behavioral intentions. *Journal of Social Psychology*, 147(6), 681-706.
- Budzicz, Ł. (2008). *Replikacja eksperymentu Apa Dijksterhuisa dotyczącego nieświadomego myślenia*. Niepublikowana praca magisterska. Uniwersytet im. Adama Mickiewicza w Poznaniu.
- Budzicz, Ł. (2010). Dlaczego u ludzi pojawiła się świadomość? Hipoteza Uzasadniania Gregga Henriquesa. *Poznańskie Forum Kognitywistyczne. Teksty pokonferencyjne, Tom 4*, 42-49.
- Budzicz, Ł. (2011). Metody badań jakościowych czy metody pisanie swobodnych esejów? Refleksje w obronie „świętej krowy oświecenia” - obiektywnej nauki. *Nauka*, 4, 163-174.
- Budzicz, Ł. (2012a). Znaczenie myślenia ewolucyjnego we współczesnej psychologii. *Nauka*, 3, 91-118.
- Budzicz, Ł. (2012b). Czas reakcji w badaniu złożonych konstruktów psychologicznych: emocjonalne zadanie Stroopa i test decyzji leksykalnych. W: W. J. Paluchowski, A. Bujacz, P. Haładziński, L. Kaczmarek (red.)

- Nowoczesne metody badawcze w psychologii* (s. 11-25). Poznań: Wydawnictwo Naukowe Wydziału Nauk Społecznych.
- Buhrmester, M., Kwang, T., Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, 6(1), 3-5.
- Burgard, D. E. (2001). Journals of the century in psychology. *The Serials Librarian*, 39(3), 41-56.
- Burger, J. M. (2009). Replicating Milgram: would people still obey today?. *American Psychologist*, 64(1), 1-11.
- Buss, D. M. (1989). Sex differences in human mate preferences: evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, 12(01), 1-14.
- Calvillo, D. P., Penaloza, A. (2009). Are complex decisions better left to the unconscious? Further failed replications of the deliberation-without-attention effect. *Judgment and Decision Making*, 4(6), 509-517.
- Carter, E. C., McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated?. *Frontiers in Psychology*, 5, 1-11.
- Chen, G., Zhang, L. (2010). Mental health and thinking styles in Sternberg's theory: an explanatory study. *Psychological Reports*, 107(3), 784-794.
- Cobey, K. D., Stulp, G., Laan, F., Buunk, A. P., Pollet, T. V. (2013). Sex differences in risk taking behavior among Dutch cyclists. *Evolutionary Psychology*, 11(2), 350-364.
- Cohen, A. B., Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology*, 81(4), 697-710.
- 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., Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14(2), 165-176.

- 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.
- Correia, I., Vala, J., Aguiar, P. (2007). Victim's innocence, social categorization, and the threat to the belief in a just world. *Journal of Experimental Social Psychology*, 43(1), 31-38.
- Coyne, J. (2009). *Ewolucja jest faktem*. Warszawa: Prószyński i S-ka.
- Craig, J. R., Reese, S. C. (1973). Retention of raw data: a problem revisited. *American Psychologist*, 28(8), 723.
- Cronk, L., Leech, B. L. (2012). *Meeting at Grand Central: understanding the social and evolutionary roots of cooperation*. Princeton: Princeton University Press.
- Cross, K. P. (1977). Not can, but will college teaching be improved?. *New Directions for Higher Education*, 1977(17), 1-15.
- Crump, M. J., McDonnell, J. V., Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, 8(3), e57410.
- Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, 25(1), 7-29.
- Dabbs Jr, J. M., Hargrove, M. F., Heusel, C. (1996). Testosterone differences among college fraternities: well-behaved vs rambunctious. *Personality and Individual Differences*, 20(2), 157-161.
- Dandurand, F., Shultz, T. R., Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40(2), 428-434.
- Darwin, K. (1859/2009). *O powstaniu gatunków*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.
- Deary, I. J., Strand, S., Smith, P., Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13-21.
- DeScioli, P., Christner, J., Kurzban, R. (2011). The omission strategy. *Psychological Science*, 22(4), 442-446.
- Dijksterhuis, A., van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology*, 74(4), 865–877.

- Dijksterhuis, A. (2004). Think different: the merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, 87(5), 586-598.
- Dijksterhuis, A., Bos, M. W., Nordgren, L. F., van Baaren, R. B. (2006). On making the right choice: the deliberation-without-attention effect. *Science*, 311(5763), 1005–1007.
- Dijksterhuis, A., Bos, M. W., Van der Leij, A., Van Baaren, R. B. (2009). Predicting soccer matches after unconscious and conscious thought as a function of expertise. *Psychological Science*, 20(11), 1381-1387.
- Dijksterhuis, A., Nordgren, L. (2006). A theory of unconscious thought. *Perspectives on Psychological Science*, 1(2), 95-109.
- Douven, I., (2007). Underdetermination. W: S. Psillos, M. Curd (red.), *The Routledge Companion to the Philosophy of Science*, (s. 292-301). London: Routledge.
- Doyen, S., Klein, O., Pichon, C.-L., Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PLoS One*, 7(1), e29081.
- Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(04), 681-694.
- Dunbar, R. I., Marriott, A., Duncan, N. D. (1997). Human conversational behavior. *Human Nature*, 8(3), 231-246.
- Durante, K. M., Arsene, R., Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24, 1007–1016.
- Dutton, D. G., Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, 30(4), 510-517.
- Eagly, A. H., Wood, W. (1999). The origins of sex differences in human behavior: evolved dispositions versus social roles. *American Psychologist*, 54(6), 408-423.
- Edlund, J. E., Sagarin, B. J. (2010). Mate value and mate preferences: an investigation into decisions made with and without constraints. *Personality and Individual Differences*, 49(8), 835-839.
- Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, 30(10), 967-976.

- Enserink, M. (2012). Final report on Stapel also blames field as a whole. *Science*, 338(6112), 1270-1271.
- European Association of Social Psychology (2012). *Statement EASP on Levelt December 2012*. Pozyskano z:
http://www.easp.eu/news/Statement%20EASP%20on%20Levelt_December_%202012.pdf
- Faith, M. S., Wong, F. Y., Carpenter, K. M. (1995). Group sensitivity training: update, meta-analysis, and recommendations. *Journal of Counseling Psychology*, 42(3), 390-399.
- Fan, X., Miller, B. C., Park, K. E., Winward, B. W., Christensen, M., Grotevant, H. D., Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18(3), 223-244.
- Fanelli, D., Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLoS One*, 8(6), e66938.
- 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. (2010a). "Positive" results increase down the hierarchy of the sciences. *PLoS One*, 5(4), e10068.
- Fanelli, D. (2010b). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS One*, 5(4), e10271.
- Farber, M. L. (1952). The college student as laboratory animal. *American Psychologist*, 7(3), 102-102.
- Farris, C., Treat, T. A., Viken, R. J. (2010). Alcohol alters men's perceptual and decisional processing of women's sexual interest. *Journal of Abnormal Psychology*, 119(2), 427-432.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C, Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229-23.
- Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: a meta-analytic review. *Aggression and Violent Behavior*, 12(4), 470-482.
- Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.

- Ferguson, C. J. (2013). Violent video games and the Supreme Court: lessons for the scientific community in the wake of *Brown v. Entertainment Merchants Association*. *American Psychologist*, 68(2), 57–74.
- Ferguson, C. J. (2014). Comment: why meta-analyses rarely resolve ideological debates. *Emotion Review*, 6(3), 251-252.
- Ferguson, C. J., Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17(1), 120-128.
- Ferguson, C. J., Kilburn, J. (2009). The public health risks of media violence: a meta-analytic review. *Journal of Pediatrics*, 154(5), 759–763.
- Fiedler, K. (2011). Voodoo correlations are everywhere: not only in neuroscience. *Perspectives on Psychological Science*, 6(2), 163–171.
- Flore, P. C., Wicherts, J. M. (2014). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, 53(1), 25-44.
- Fraley, R. C., Shaver, P. R. (1998). Airport separations: a naturalistic study of adult attachment dynamics in separating couples. *Journal of Personality and Social Psychology*, 75(5), 1198-1212.
- Francis, G. (2012a). Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, 19(2), 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(25), E1587.
- Francis, G. (2012d). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & 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 & Review*, 21(5), 1180-1187.
- 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(1), 3-12.
- Furr, R. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, 23(5), 369-401.
- 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.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., Simmons, J. P. (2012). Correcting the past: failures to replicate psi. *Journal of Personality and Social Psychology*, 103(6), 933-948.
- Gallander-Wintre, M., North, C., Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: a developmental perspective. *Canadian Psychology/Psychologie Canadienne*, 42(3), 216-225.
- Galperin, A., Haselton, M. (2009). Predictors of how often and when people fall in love. *Evolutionary Psychology*, 8(1), 5-28.
- Gangestad, S. W., Garver-Apgar, C. E., Simpson, J. A., Cousins, A. J. (2007). Changes in women's mate preferences across the ovulatory cycle. *Journal of Personality and Social Psychology*, 92(1), 151-163.
- Gawronski, B., Deutsch, R., Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. W: K. C. Klauer, C. Stahl, A. Voss (red.), *Cognitive methods in social psychology* (s. 78-123). New York, NY: Guilford.
- Gawronski, B., LeBel, E. P., Peters, K. R. (2007). What do implicit measures tell us?: scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, 2(2), 181-193.
- Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, 6(8), 653-659.
- Geake, J. (2008). Neuromythologies in education. *Educational Research*, 50(2), 123-133.

- Gervais, W. M., Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, 336(6080), 493–496.
- Gilbert, D.T., Pinel, E.C., Wilson, T.D., Blumberg, S.J., Wheatley, T.P. (1998). Immune neglect: a source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3), 617–638.
- Gildersleeve, K., Haselton, M. G., Fales, M. (2014). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*, 5, 1205-1259.
- Gillath, O., Landau, M. J., Selcuk, E., Goldenberg, J. L. (2011). Effects of low survivability cues and participant sex on physiological and behavioral responses to sexual stimuli. *Journal of Experimental Social Psychology*, 47(6), 1219-1224.
- Gintis, H., Bowles, S., Boyd, R., Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3), 153-172.
- Glasman, L. R., Albarracín, D. (2006). Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. *Psychological Bulletin*, 132(5), 778-822.
- Gollwitzer, M., Denzler, M. (2009). What makes revenge sweet: seeing the offender suffer or delivering a message?. *Journal of Experimental Social Psychology*, 45(4), 840-844.
- González-Vallejo, C., Lassiter, G. D., Bellezza, F. S., Lindberg, M. J. (2008). "Save angels perhaps": a critical examination of unconscious thought theory and the deliberation-without-attention effect. *Review of General Psychology*, 12(3), 282-296.
- Grahe, J. E. (2014). Announcing open science badges and reaching for the sky. *The Journal of Social Psychology*, 154(1), 1-3.
- Greenberg, J. (1987). The college sophomore as guinea pig: setting the record straight. *Academy of Management Review*, 12(1), 157-159.
- Greenwald, A. G. (1980). The totalitarian ego: fabrication and revision of personal history. *American Psychologist*, 35(7), 603-618.
- Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, 7(2), 99-108.

- Greenwald, A. G., Poehlman, T. A., Uhlman, E. L., Banaij, M. R. (2009). Understanding and using the implicit association test III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17–41.
- Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress?. *Psychological Review*, 93(2), 216-229.
- 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.
- Griskevicius, V., Cialdini, R. B., Kenrick, D. T. (2006). Peacocks, Picasso, and parental investment: the effects of romantic motives on creativity. *Journal of Personality and Social Psychology*, 91(1), 63-76.
- Grobler, A. (2006). *Metodologia nauk*. Warszawa: Wydawnictwo Aureus.
- Gueguen, N. (2007). Women's bust size and men's courtship solicitation. *Body Image*, 4(4), 386-390.
- Guéguen N. (2008a). The effects of women's cosmetics on men's courtship behavior. *North American Journal of Psychology*, 10(1), 221-228.
- Guéguen, N. (2008b). The effect of a woman's smile on men's courtship behavior. *Social Behavior and Personality: an International Journal*, 36(9), 1233-1236.
- Guéguen, N. (2009). Man's uniform and receptivity of women to courtship request: three field experiments with a firefighter's uniform. *European Journal of Social Sciences*, 12(2), 235-240.
- Guéguen N. (2011). The effect of women's suggestive clothing on men's behavior and judgment: a field study. *Psychological Reports*, 109(2), 635-638.
- Guéguen, N., Lamy, L. (2012). Men's social status and attractiveness: women's receptivity to men's date requests. *Swiss Journal of Psychology*, 71(3), 157-160.
- Guéguen, N., Meineri, S., Fischer-Lokou, J. (2013). Men's music ability and attractiveness to women in a real-life courtship context. *Psychology of Music*, 42(4), 545-549.
- Hafer, C. (2000). Do innocent victims threaten the belief in a just world? Evidence from a modified Stroop task. *Journal of Personality and Social Psychology*, 79(2), 165-173.

- Hagger, M. S., Chatzisarantis, N. L. D., Biddle, S. J. H. (2002). A meta-analytic review of the theories of reasoned action and planned behavior in physical activity: predictive validity and the contribution of additional variables. *Journal of Sport & Exercise Psychology*, 24(1), 3-32.
- Hagger, M. S., Wood, C., Stiff, C., Chatzisarantis, N. D. (2010). Ego depletion and the strength model of self-control: a meta-analysis. *Psychological Bulletin*, 136(4), 495-525.
- Hall, L., Johansson, P., Strandberg, T. (2012). Lifting the veil of morality: choice blindness and attitude reversals on a self-transforming survey. *Plos ONE*, 7(9).
- Haselton, M. G., Miller, G. F. (2006). Women's fertility across the cycle increases the short-term attractiveness of creative intelligence compared to wealth. *Human Nature*, 17(1), 50-73.
- Haselton, M. G., Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, 78(1), 81-91.
- Haselton, M. G., Nettle, D. (2006). The paranoid optimist: an integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, 10(1), 47-66.
- Hatta, T., Ohbuchi, K. I. (2003). An experimental study of electronic negotiation: Content analysis of verbal interactions focusing on the effects of correctability and excitability. *Psychologia*, 46(2), 129-139.
- Heider, J., Skowronski, J. (2007). Improving the predictive validity of the Implicit Association Test. *North American Journal of Psychology*, 9(1), 53-76.
- Heim, J., Astring, T., Schliemann, T. (2002). Medium effects on persuasion. *Proceedings of the second Nordic conference on Human-computer interaction*, 259-262.
- Hendrie, C. A., Mannion, H. D., Godfrey, G. K. (2009). Evidence to suggest that nightclubs function as human sexual display grounds. *Behaviour*, 146(10), 1331-1348.
- Henrich, J., Heine, S. J., Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, 33(2-3), 61-83.
- Henriques, G. (2003). The tree of knowledge system and the theoretical unification of psychology. *Review of General Psychology*, 7(2), 150-182.

- Henry, P. J. (2008). College sophomores in the laboratory redux: influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, 19(2), 49-71.
- Hergenhahn, B., Henley, T. (2013). *An introduction to the history of psychology*. Belmont, CA: Cengage Cengage Learning.
- Higbee, K.L., Lott, W.J., Graves, J.P. (1976). Experimentation and college students in social-personality research. *Personality and Social Psychology Bulletin*, 2, 239-241.
- Hock, R. (2003). *40 prac badawczych, które zmieniły oblicze psychologii*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Huffman, M., Nahallage, C., Leca, J. (2008). Cultured monkeys: social learning cast in stones. *Current Directions in Psychological Science*, 17(6), 410-414.
- Hugenberg, K., Bodenhausen, G. V. (2004). Ambiguity in social categorization the role of prejudice and facial affect in race categorization. *Psychological Science*, 15(5), 342-345.
- Huizenga, H. M., Wetzels, R., van Ravenzwaaij, D., Wagenmakers, E. J. (2012). Four empirical tests of unconscious thought theory. *Organizational Behavior and Human Decision Processes*, 117(2), 332-340.
- Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640-648.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8), e124.
- Jamruška, K. (2013). Replikacja eksperymentu dotyczącego preferencyjnego przetwarzania uwagowego bodźców ożywionych. *Niepublikowana praca magisterska*. Uniwersytet im. Adama Mickiewicza w Poznaniu.
- 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.
- Johnson, C. S., Stapel, D. A. (2007). No pain, no gain: the conditions under which upward comparisons lead to better performance. *Journal of Personality and Social Psychology*, 92(6), 1051-1067.
- Joly, J. F., Stapel, D. A., Lindenberg, S. M. (2008). Silence and table manners: when environments activate norms. *Personality and Social Psychology Bulletin*, 34(8), 1047-1056.

- Jonas, E., Fischer, P. (2006). Terror management and religion: Evidence that intrinsic religiousness mitigates worldview defense following mortality salience. *Journal Of Personality And Social Psychology*, 91(3), 553-567.
- Kahneman, D., (2012). *Pułapki myślenia: o myśleniu szybkim i wolnym*. Poznań: Media Rodzina.
- Kay, A., Jost, J. (2003). Complementary justice: effects of 'poor but happy' and 'poor but honest' stereotype exemplars on system justification and implicit activation of the justice motive. *Journal of Personality and Social Psychology*, 85(5), 823-837.
- Keizer, K., Lindenberg, S., Steg, L. (2008). The spreading of disorder. *Science*, 322(5908), 1681-1685.
- Kenrick, D. T., Neuberg, S. L., Cialdini R. B. (2006). *Psychologia społeczna. Rozwiązane tajemnice*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Kepes, S., McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology?. *Industrial and Organizational Psychology*, 6(3), 252–268.
- 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 doping. O upiększaniu psychologii społecznej. *Psychologia społeczna*, 3(22), 213-217.
- Koenig, B. L., Kirkpatrick, L. E. E. A., Ketelaar, T. (2007). Misperception of sexual and romantic interests in opposite-sex friendships: four hypotheses. *Personal Relationships*, 14(3), 411–429.
- Kosslyn, S. M., Cacioppo, J. T., Davidson, R. J., Hugdahl, K., Lovallo, W. R., Spiegel, D., Rose, R. (2002). Bridging psychology and biology: the analysis of individuals in groups. *American Psychologist*, 57(5), 341-351.
- Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. *Psychological Bulletin*, 133(3), 464-481.
- Kramer, D., Goldman, M. (2003). Using a modified Stroop task to implicitly discern the cognitive organization of alcohol expectancies. *Journal of Abnormal Psychology*, 112(1), 171-175.

- Krueger, J. I., Funder, D.C. (2004). Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27(3), 313-328.
- 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.
- Kuhle, B. X. (2011). Did you have sex with him? Do you love her? An in vivo test of sex differences in jealous interrogations. *Personality and Individual Differences*, 51(8), 1044-1047.
- Kuipers, T. A. (2007). Laws, theories and research programs. General philosophy of science—focal issues. W: D. M. Gabbay, P. Thagard, J. Woods (red.), *General Philosophy of Science: Focal Issues* (s. 1-95). North Holland: Elsevier.
- Kukla, A. (2008). Observation. W: S. Psillos, M. Curd (red.), *The Routledge companion to philosophy of science* (s. 396-404). London: Routledge.
- Kurzban, R. (2010). *Why everyone (else) is a hypocrite: evolution and the modular mind*. Princeton, NJ: Princeton University Press.
- Lane, K. A., Banaji, M. R., Nosek, B. A., Greenwald, A. G. (2007). Understanding and using the Implicit Association Test: IV. What we know (so far). W: B. Wittenbrink, N. S. Schwarz (red.). *Implicit measures of attitudes: Procedures and controversies* (s. 59–102). New York: Guilford Press.
- La France, B. H., Henningsen, D. D., Oates, A., Shaw, C. M. (2009). Social–sexual interactions? Meta-analyses of sex differences in perceptions of flirtatiousness, seductiveness, and promiscuousness. *Communication Monographs*, 76(3), 263-285.
- Lassiter, G. D., Lindberg, M. J., González-Vallejo, C., Bellezza, F. S., Phillips, N. D. (2009). The deliberation-without-attention effect evidence for an artifactual interpretation. *Psychological Science*, 20(6), 671-675.
- Leary, M. R., Hoyle, R. H. (Eds.). (2009). *Handbook of individual differences in social behavior*. Nowy Jork: Guilford Press.
- LeBel, E. P., Paunonen, S. V. (2011). Sexy but often unreliable: the impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, 37(4), 570-583.

- LeBel, E. P., Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15(4), 371–379.
- LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., Smith, C. T. (2013). PsychDisclosure.org: grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, 8(4), 424–432.
- Levelt Committee, Noort Committee, Drenth Committee (2012). *Flawed science: the fraudulent research practices of social psychologist Diederik Stapel*. Pozyskano z <http://www.tilburguniversity.edu/nl/nieuws-enagenda/finalreportLevelt.pdf>
- Levesque, M. J., Nave, C. S., Lowe, C. A. (2006). Toward an understanding of gender differences in inferring sexual interest. *Psychology of Women Quarterly*, 30(2), 150-158.
- Li, N. P., Valentine, K. A., Patel, L. (2011). Mate preferences in the US and Singapore: a cross-cultural test of the mate preference priority model. *Personality And Individual Differences*, 50(2), 291-294.
- Li, N. P., Yong, J. C., Tov, W., Sng, O., Fletcher, G. O., Valentine, K. A., ... Balliet, D. (2013). Mate preferences do predict attraction and choices in the early stages of mate selection. *Journal of Personality and Social Psychology*, 105(5), 757-776.
- Lilienfeld, S. O., Wood, J. M., Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, 1(2), 27-66.
- Lincoln, D. (2010). *Kwantowa granica. LHC - Wielki Zderzacz Hadronów*. Warszawa: Prószyński i S-ka.
- Lindgren, K. P., Parkhill, M. R., George, W. H., & Hendershot, C. S. (2008). Gender differences in perceptions of sexual intent: A qualitative review and integration. *Psychology of Women Quarterly*, 32(4), 423-439.
- 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.
- Lipton, P. (2003). *Inference to the best explanation*. Florence, KY, US: Routledge.
- 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: Biological Sciences*, 278(1714), 2032-2039.
- Loftus, E. F. (2005). Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4), 361-366.
- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197), 869-878.
- Lydon, J. E., Menzies-Toman, D., Burton, K., Bell, C. (2008). If-then contingencies and the differential effects of the availability of an attractive alternative on relationship maintenance for men and women. *Journal of Personality and Social Psychology*, 95(1), 50-65.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2), 163-203.
- Macrae, C. N., Milne, A. B., Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: a peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66(1), 37-47.
- 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 & Motor Skills*, 112(2), 331-348.
- 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.
- Matczak, A. (2000). Style poznawcze. W: J. Strelau (red.), *Psychologia. Podręcznik akademicki, t. 2* (s. 761-782). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9(2), 147-163.
- Mayr, E. (2004) *What makes biology unique. Considerations on the autonomy of a scientific discipline*. Cambridge: Cambridge University Press.

- McConnell, A. R., Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology*, 37(5), 435-442.
- McGregor, I., Nash, K., Prentice, M. (2010). Reactive approach motivation (RAM) for religion. *Journal of Personality and Social Psychology*, 99(1), 148-161.
- Mehl, M. R., Gosling, S. D., Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5), 862-877.
- Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., Price, J. H. (2001). The Electronically Activated Recorder (EAR): a device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4), 517-523.
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., Pennebaker, J. W. (2007). Are women really more talkative than men?. *Science*, 317(5834), 82-82.
- Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research*, 20, 431-440.
- Meston, C. M., Buss, D. M. (2007). Why humans have sex. *Archives of Sexual Behavior*, 36(4), 477-507.
- Meston, C. M., Buss, D.M. (2010). *Dlaczego kobiety uprawiają seks? Motywacje seksualne - od przygody po zemstę*. Sopot: Smak słowa.
- Mikulincer, M. (1998). Adult attachment style and individual differences in functional versus dysfunctional experiences of anger. *Journal of Personality and Social Psychology*, 74(2), 513-524.
- Mikulincer, M., Shaver, P. R., Gillath, O., Nitzberg, R. A. (2005). Attachment, caregiving, and altruism: boosting attachment security increases compassion and helping. *Journal of Personality and Social Psychology*, 89(5), 817-839.
- Miller, S. L., Maner, J. K. (2011). Ovulation as a male mating prime: subtle signs of women's fertility influence men's mating cognition and behavior. *Journal of Personality and Social Psychology*, 100(2), 295-308.
- Mitchell, G. (2012). What is wrong with social psychology?. *Dialogue. The Official Newsletter of the Society for Personality and Social Psychology*, 26(2), 12-17.

- 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.
- Nave, C. S., Sherman, R. A., Funder, D. C. (2008). Beyond self-report in the study of hedonic and eudaimonic well-being: correlations with acquaintance reports, clinician judgments and directly observed social behavior. *Journal of Research in Personality*, 42(3), 643-659.
- Nelson, L. D., Simmons, J. P., Simonsohn, U. (2012). Let's publish fewer papers. *Psychological Inquiry*, 23(3), 291-293.
- Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on Psychological Science*, 7(6), 643-644.
- Neville, L. (2012). Do economic equality and generalized trust inhibit academic dishonesty? Evidence from state-level search-engine queries. *Psychological Science*, 23(4), 339-345.
- New, J., Cosmides, L., Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42), 16598-16603.
- Newell, B. R., Wong, K. Y., Cheung, J. C., Rakow, T. (2009). Think, blink or sleep on it? The impact of modes of thought on complex decision making. *The Quarterly Journal of Experimental Psychology*, 62(4), 707-732.
- Nguyen, H. H. D., Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93(6), 1314-1334.
- Nisbett, R.E., Wilson, T.D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.
- Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, 34(7), 913-923.
- Norenzayan, A. (2013). *Big gods: how religion transformed cooperation and conflict*. Princeton: Princeton University Press.
- Norenzayan, A., Hansen, I. G. (2006). Belief in supernatural agents in the face of death. *Personality and Social Psychology Bulletin*, 32(2), 174-187.
- Nosek, B. A., Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, 23(3), 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(6), 615-631.
- Öhman, A., Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review* 108(3), 483-522.
- Okasha, S. (2002). *Philosophy of science: a very short introduction*. Nowy Jork: Oxford University Press.
- Oksanen, A. (2014). Affect and addiction in the Celebrity Rehab reality television show. *Addiction Research & Theory*, 22(2), 137-146.
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660.
- Opthof, T., Coronel, R., Piper, H. M. (2004). Impact factors: no *totum pro parte* by skewness of citation. *Cardiovascular Research*, 61(2), 201-203.
- Orliński, W. (2013a). Nienaukowa działalność naukowców. *Gazeta Wyborcza*, 132, 18.
- Orliński, W. (2013b). Zagubieni w lekach. *Gazeta Wyborcza*, 150, 28.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171-192.
- Pashler, H., Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536.
- Pashler, H., Harris, C., Coburn, N. (2011). *Elderly-Related Words Prime Slow Walking*. Pozyskano z <http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D>
- Pashler, H., Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence?. *Perspectives on Psychological Science*, 7(6), 528–530.
- Pawlowski, B., Atwal, R., Dunbar, R. I. M. (2008). Sex differences in everyday risk-taking behavior in humans. *Evolutionary Psychology*, 6(1), 29-42.
- 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(02), 187–255.

- Petersen, M. B., Sznycer, D., Sell, A., Cosmides, L., Tooby, J. (2013). The ancestral logic of politics upper-body strength regulates men's assertion of self-interest over economic redistribution. *Psychological Science*, 24(7), 1098-1103.
- Peterson, R. A. (2001). On the use of college students in social science research: insights from a second-order meta-analysis. *Journal of Consumer Research*, 28(3), 450-461.
- Piedmont, R. L., McCrae, R. R., Riemann, R., Angleitner, A. (2000). On the invalidity of validity scales: evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, 78(3), 582-593.
- Piff, P. K., Kraus, M. W., Côté, S., Cheng, B. H., Keltner, D. (2010). Having less, giving more: the influence of social class on prosocial behavior. *Journal of Personality and Social Psychology*, 99(5), 771-784.
- Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, 306(5695), 462-466.
- Prioleau, L., Murdock, M., Brody, N. (1983). An analysis of psychotherapy versus placebo studies. *The Behavioral and Brain Sciences*, 6(2), 275-85.
- Rand, K., Ilardi, S. (2005). Toward a consilient science of psychology. *Journal of Clinical Psychology*, 61(1), 7-20.
- Reich, E. S. (2009). *Plastic fantastic: how the biggest fraud in physics shook the scientific world*. New York, NY: Palgrave-Macmillan.
- Report of the Smeesters Follow-Up Investigation Comitee (2014). Pozyskano z: http://www.eur.nl/fileadmin/ASSETS/press/2014/maart/Report_Smeesters_follow-up_investigation_committee.final.pdf
- Retraction Watch (2012). *Following investigation, Erasmus social psychology professor retracts two studies, resigns*. Pozyskano z: <http://retractionwatch.com/2012/06/25/following-investigation-erasmus-social-psychology-professor-retracts-two-studies-resigns/>
- Rezaei, A. R. (2011). Validity and reliability of the IAT: measuring gender and ethnic stereotypes. *Computers in Human Behavior*, 27(5), 1937-1941.
- Rhode, T. E., Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35(1), 83-92.
- 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.

- Riding, R. J. (1997). On the nature of cognitive style. *Educational Psychology*, 17(1-2), 29-49.
- Ring, K. (1967). Experimental social psychology: some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, 3(2), 113-123.
- Ritchie, S. J., Wiseman, R., French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive facilitation of recall' effect. *PloS One*, 7(3), e33423.
- Riva, G., Teruzzi, T., Anolli, L. (2003). The use of the internet in psychological research: comparison of online and offline questionnaires. *CyberPsychology & Behavior*, 6(1), 73-80.
- Robinson, M. D., Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6), 934-960.
- Rosenthal, R., Rosnow, R. L. (2009) *Artifacts in behavioral research*. Oxford: Oxford University Press.
- Ross, L., Amabile, T. M., Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, 35, 485– 494.
- Ross, M., Conway, M. (1986). Remembering one's own past: the construction of personal histories. W: R. M. Sorrentino, E. T. Higgins (red.), *Handbook of motivation and cognition* (s. 122-144). New York: Guilford.
- Rozin, P. (2001). Social psychology and science: some lessons from Salomon Asch. *Personality and Social Psychology Review*, 5(1), 2-14.
- Rozin, P. (2006). Domain denigration and process preference in academic psychology. *Perspectives on Psychological Science*, 1(4), 365-376.
- Rozin, P. (2007). Exploring the landscape of modern academic psychology: finding and filling the holes. *American Psychologist*, 62(8), 754-766.
- Rudman, L. A., Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations*, 10(3), 359-372.
- Rudman, L. A., Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, 57(4), 743-762.
- 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.

- Ruys, K. I., Stapel, D. A. (2008). The secret life of emotions. *Psychological Science*, 19(4), 385-391.
- Sąd Najwyższy USA (2011). *Brown vs Entertainment Merchants Association*, 131 S. Ct. 2729. Pozyskano z: <http://www.supremecourt.gov/opinions/10pdf/08-1448.pdf>
- 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(2), 472–476.
- Schachter, D. L. (2003). *Siedem grzechów pamięci*. Warszawa: Państwowy Instytut Wydawniczy.
- Schaller, M., Murray, D. R. (2008). Pathogens, personality, and culture: disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *Journal of Personality and Social Psychology*, 95(1), 212-221.
- Schimmel, J., Hayes, J., Williams, T., Jahrig, J. (2007). Is death really the worm at the core? Converging evidence that worldview threat increases death-thought accessibility. *Journal of Personality and Social Psychology*, 92(5), 789-803.
- Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, 470(7335), 437.
- Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin*, 72(3), 214-228.
- Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, 54(2), 93-105.
- Schwarz, N., Strack, F., Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis. *Public Opinion Quarterly*, 55(1), 3-23.
- Sears, D. O. (1986). College sophomores in the laboratory: influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51(3), 515-530.
- Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?. *Psychological Bulletin*, 105(2), 309–316.

- Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., ... Moore, C. (2013). Priming intelligent behavior: an elusive phenomenon. *PloS One*, 8(4), e56515.
- Sherry, J. (2001). The effects of violent video games on aggression: a meta-analysis. *Human Communication Research*, 27(3), 409–431.
- 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., Nelson, L., Simonsohn, U. (2012). *A 21 Word Solution*. Pozyskano z: <http://ssrn.com/abstract=2160588>.
- 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. (2013). P-Curve: a key to the file-drawer. *Journal of Experimental Psychology: General*. Advance online publication.
- Simonton, D. K. (2004). Psychology's status as a scientific discipline: its empirical placement within an implicit hierarchy of the sciences. *Review of General Psychology*, 8(1), 59–67.
- 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.
- Smart, R. G. (1966). Subject selection bias in psychological research. *The Canadian Psychologist*, 7(2), 115-121.
- Smeesters, D., Liu, J. E. (2011). The effect of color (red versus blue) on assimilation versus contrast in prime-to-behavior effects [Retracted article]. *Journal of Experimental Social Psychology*, 47, 653–656.
- Song, C., Qu, Z., Blumm, N., Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, 327(5968), 1018-1021.
- Spellman, B. (2012). Introduction to the special section: data, data, everywhere... Especially in my file drawer. *Perspectives on Psychological Science*, 7(1), 58-59.
- Staats, A. (1999). Unifying psychology requires new infrastructure, theory, method, and a research agenda. *Review of General Psychology*, 3(1), 3-13.
- Stapel, D. A. (2012). *Ontsporing*. Amsterdam: Prometheus.

- Stapel, D. A., Lindenberg, S. (2011). Coping with chaos: how disordered contexts promote stereotyping and discrimination. *Science*, 332(6026), 251–252.
- Sterling, T. D., Rosenbaum, W. L., Weinkam, J. J. (1995). Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and *vice versa*. *American Statistician*, 49(1), 108–112.
- Sternberg, R. J., Zhang, L. F. (2005). Styles of thinking as a basis of differentiated instruction. *Theory into Practice*, 44(3), 245-253.
- Sternberg, R. J., Zhang, L. F. (red.) (2001). *Perspectives on thinking, learning and cognitive styles*. Mahwah, NJ: Lawrence Erlbaum and Associates.
- Stoet, G., Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement?. *Review Of General Psychology*, 16(1), 93-102.
- Strack, F., Martin, L. L., Schwarz, N. (1988). Priming and communication: social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, 18(5), 429-442.
- Strelau, J., Zawadzki, B. (2005). Trauma and temperament as predictors of intensity of posttraumatic stress disorder symptoms after disaster. *European Psychologist*, 10(2), 124-135.
- Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B., Nordgren, L. F. (2011). A meta-analysis on unconscious thought effects. *Social Cognition*, 29(6), 738–762.
- 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(1), 40-50.
- Sun, S., Pan, W., Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102(4), 989-1004.
- Tafforin, C. (2013). Time effects, cultural influences, and individual differences in crew behavior during the Mars-500 experiment. *Aviation, Space, and Environmental Medicine*, 84(10), 1082-1086.

- Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 223(5), 96-102.
- Tavris, C., Aronson, E. (2008). *Błądzą wszyscy (ale nie ja)*. Sopot: Smak Słowa.
- Tressoldi, P. E. (2012). Replication unreliability in psychology: elusive phenomena or “elusive” statistical power?. *Frontiers in Psychology*, 3(4), 1-5.
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35-57.
- Trivers, R. L. (1972). Parental investment and sexual selection. W: B. Campbell (red.), *Sexual selection and the descent of man* (s. 136-179). London: Heinemann.
- Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist*, 14(1), 249-264.
- Tuschman, A. (2013). *Our political nature: the evolutionary origins of what divides us*. Amherst, NY: Prometheus Books.
- Tyszka, T. (2000). *Psychologiczne pułapki oceniania i podejmowania decyzji*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.
- Ulanowski, T. (2006). Jak podjąć najlepszą decyzję. *Gazeta Wyborcza*, 78, 10.
- Vankov, I., Bowers, J., Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*. Advance Online Publication.
- Vantomme, D., Geuens, M., De Houwer, J., De Pelsmacker, P. (2005). Implicit attitudes toward green consumer behaviour. *Psychologica Belgica*, 45(4), 217-239.
- Vogel, G. (2011). Psychologist accused of fraud on ‘astounding scale’. *Science*, 334(6056), 579.
- Von Hippel, W., Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(01), 1-16.
- Wallace, D. S., Paulson, R. M., Lord, C. G., Bond Jr, C. F. (2005). Which behaviors do attitudes predict? Meta-analyzing the effects of social pressure and perceived difficulty. *Review of General Psychology*, 9(3), 214-227.
- Wang, S. W., Repetti, R. L., Campos, B. (2011). Job stress and family social behavior: the moderating role of neuroticism. *Journal of Occupational Health Psychology*, 16(4), 441-456.
- Wang, S., Repetti, R. L. (2014). Psychological well-being and job stress predict marital support interactions: a naturalistic observational study of dual-earner

- couples in their homes. *Journal of Personality and Social Psychology*, 107(5), 864-878.
- Waroquier, L., Marchiori, D., Klein, O., Cleeremans, A. (2010). Is it better to think unconsciously or to trust your first impression? A reassessment of unconscious thought theory. *Social Psychological and Personality Science*, 1(2), 111-118.
- Wellman, H. M., Cross, D., Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, 72(3), 655-684.
- Wertheimer, M. (2012). *A brief history of psychology*. New York, NY: Taylor & Francis.
- 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.
- 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.
- Williams, J. G., Mathews, A., MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, 120(1), 3-24.
- Wilson, R. E., Gosling, S. D., Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, 7(3), 203-220.
- Wilson, T. D., Gilbert, D. T. (2003). Affective forecasting. *Advances in Experimental Social Psychology*, 35, 345-411.
- Wilson, T.D., Gilbert, D.T. (2005). Affective forecasting: knowing what to want. *Current Directions in Psychological Science*, 14(3), 131-134.
- Wolfe, J. M. (2013). Registered reports and replications in attention, perception, & psychophysics. *Attention, Perception, & Psychophysics*, 75, 781-783.
- Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, 17, 657-658.
- Wood, W., Kressel, L., Joshi, P. D., Louie, B. (2014). Meta-analysis of menstrual cycle effects on women's mate preferences. *Emotion Review*, 6(3), 229-249.
- Wróblewski, A. K. (2011). *Historia fizyki. Od czasów najdawniejszych do współczesności*. Warszawa: Wydawnictwo Naukowe PWN.
- Yong, E. (2012a). Bad copy. *Nature*, 485, 298-300.

- Yong, E. (2012b). *Uncertainty shrouds psychologist's resignation*. Pozyskano z: <http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968>
- Yong, E. (2012c). *The data detective*. Pozyskano z: <http://www.nature.com/news/the-data-detective-1.10937>
- Yzerbyt, V. Y., Rogier, A., Fiske, S. T. (1998). Group entitativity and social attribution: On translating situational constraints into stereotypes. *Personality and Social Psychology Bulletin*, 24(10), 1089-1103.
- Zhang, L. (2005). Validating the theory of mental self-government in a non-academic setting. *Personality and Individual Differences*, 38(8), 1915-1925.
- Zhang, L. (2006). Thinking styles and the big five personality traits revisited. *Personality and Individual Differences*, 40(6), 1177-1187.
- Zhang, L. F. (2000). Relationship between Thinking Styles Inventory and Study Process Questionnaire. *Personality and Individual Differences*, 29(5), 841-856.
- Zhang, L. F. (2001). Thinking styles, self-esteem, and extracurricular experiences. *International Journal of Psychology*, 36(2), 100-107.
- Zhang, L. F. (2004a). Field-dependence/independence: cognitive style or perceptual ability? Validating against thinking styles and academic achievement. *Personality and Individual Differences*, 37(6), 1295-1311.
- Zhang, L. F. (2004b). Thinking styles: University students' preferred teaching styles and their conceptions of effective teachers. *The Journal of Psychology*, 138(3), 233-252.
- Zhang, L. F. (2008). Thinking styles and identity development among Chinese university students. *The American Journal of Psychology*, 121(2), 255-271.
- Zhang, L. F., Sternberg, R. J. (2006). *The nature of intellectual styles*. Mahwah, NY, US: Psychology Press.
- Zhu, C., Zhang, L. (2011). Thinking styles and conceptions of creativity among university students. *Educational Psychology*, 31(3), 361-375.
- Ziegert, J. C., Hanges, P. J. (2005). Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, 90(3), 553-562.
- Ziegler, M., MacCann, C., Roberts, R. D. (red.). (2011). *New perspectives on faking in personality assessment*. New York: Oxford University Press.

Zwarun, L., Hall, A. (2014). What's going on? Age, distraction, and multitasking during online survey taking. *Computers In Human Behavior*, 41, 236-244.