PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 49–68
DOI: 10.31648/przegldpsychologiczny.9457

Kilka uwag o stanie badań w naukach społecznych.

Rozmowa z prof. dr. hab. Jarosławem Górniakiem[1]

Jarosław Górniak[2]
Uniwersytet Jagielloński, Instytut Socjologii
https://orcid.org/0000-0001-9210-5712

Arkadiusz Białek[3]
Uniwersytet Jagielloński, Instytut Psychologii
https://orcid.org/0000-0002-9002-4764

Piotr Wolski[4]
Uniwersytet Jagielloński, Instytut Psychologii
https://orcid.org/0000-0002-7028-6142

Piotr Wolski: Głośny projekt *Open Science Collaboration* (2015) wskazał na alarmująco niską replikowalność wyników publikowanych w trzech prestiżowych czasopismach psychologicznych. W psychologii toczy się szeroka debata nad przyczynami tego niepokojącego stanu rzeczy. Szuka się sposobów jego naprawy, podejmuje próby reformowania standardów. Dyskusja zdaje się wykraczać poza

[1] Socjolog i ekonomista, profesor nauk społecznych specjalizujący się w zakresie metod badań społecznych, statystycznej analizie danych, socjologii gospodarki i organizacji oraz analizie polityk publicznych, ze szczególnym uwzględnieniem polityki nauki i szkolnictwa wyższego, rozwoju kompetencji i rynku pracy. Prorektor ds. rozwoju Uniwersytetu Jagiellońskiego, były dziekan (2012–2020) i prodziekan (2005–2012) Wydziału Filozoficznego UJ, kierownik Zakładu Socjologii Gospodarki, Edukacji i Metod Badań Społecznych w Instytucie Socjologii UJ. Twórca oraz pierwszy kierownik Centrum Ewaluacji i Analiz Polityk Publicznych UJ. W latach 2016–2018 przewodniczący Rady Narodowego Kongresu Nauki. Jest autorem i współautorem 7 książek, 80 artykułów i rozdziałów oraz redaktorem naukowym 16 prac zbiorowych. Według Google Scholar jego prace były cytowane 1675 razy, natomiast indeks h wynosi 21. Wypromował 14 doktorów.

[2] Adres do korespondencji: jaroslaw.gorniak@uj.edu.pl.

[3] Adres do korespondencji: a.bialek@uj.edu.pl.

[4] Adres do korespondencji: piotr.wolski@uj.edu.pl.

psychologię, jej echa widać np. w medycynie. Czy dostrzegasz przejawy podobnych ruchów reformatorskich w socjologii bądź innych znanych Ci naukach społecznych?

Jarosław Górniak: Myślę, że wśród nauk społecznych psychologia jest najmocniej osadzona w badaniach eksperymentalnych. Ustala swoje wyniki empiryczne na podstawie badań, które mogą uzyskiwać status konkluzywnych z punktu widzenia weryfikacji hipotez przyczynowych. W innych naukach społecznych, w tym w socjologii, którą reprezentuję, rzecz ma się odmiennie. To znaczy… oczywiście, badania eksperymentalne nie są w socjologii całkiem nieobecne, ale trzeba powiedzieć, że są rzadkie. Częściej występują na gruncie graniczącej z psychologią mikrosocjologii, która zalicza teorię dynamiki grupowej do swojego obszaru kompetencji. Mikrosocjologia nie jest w socjologii jakimś, że tak powiem, szerokim strumieniem, raczej strumyczkiem, ale jest obecna, choć pewnie chciałoby się, żeby była uprawiana na większą skalę. Niestety, u nas w Instytucie Socjologii UJ skończyła się wraz ze śmiercią Jacka Szmatki, a następnie rozproszeniem jego zespołu. Gdyby nie to, moglibyśmy lepiej porównywać, nawet na naszym własnym podwórku, jak różnią się podejścia naszych dyscyplin. Ale i tak było wyraźnie widać, że mikrosocjologia w wykonaniu Szmatki i jego zespołu szła w kierunku badań eksperymentalnych – z racji przedmiotu badań i stylu uprawiania nauki, który był oparty jednak na dążeniu do weryfikacji hipotez przyczynowych. Te zaś wynikały często ze sformalizowanych konstrukcji teoretycznych.

Gdybyśmy jednak szukali odpowiedzi na pytanie, co – w odbiorze potocznym – jest specyficzne dla socjologii, to naszą dziedzinę raczej wiąże się z badaniami o charakterze obserwacyjnym, można powiedzieć opisowym, demoskopijnym, dla którego charakterystyczne jest uzyskiwanie odpowiedzi na pytania dotyczące dominacji pewnych zjawisk, tego co Anglosasi nazywają *prevalence*, bardziej niż na ustalaniu ścisłych relacji przyczynowych. Choć nurt przyczynowy jest obecny i w psychologii, i w socjologii, to zwróćcie uwagę na charakterystyczną różnicę między naszymi dyscyplinami. W socjologii – przynajmniej w tym, dość dynamicznie rozwijającym się paradygmacie, który dąży do budowania klarownych modeli przyczynowych – weryfikacji tych modeli dokonuje się raczej na podstawie wyników pochodzących z badań obserwacyjnych, a narzędziem do takiej weryfikacji zależności przyczynowych jest modelowanie równań strukturalnych ze zmiennymi ukrytymi. To bowiem łączy psychologię z socjologią – obecność w naszych badaniach konstruktów ukrytych. Nie chcę wchodzić w głębsze debaty dotyczące statusu ontologicznego tych cech. Wiele zależy od tego, czy jesteśmy realistami, czy antyrealistami… Realiści mówiliby tu o cechach ukrytych, które rzeczywiście istnieją i są mierzone inferencyjnie za pomocą pewnego zbioru obserwowanych wskaźników. Z kolei antyrealiści twierdziliby, że to są tylko pewnego rodzaju konstrukty teoretyczne, użyteczne w obserwacji i w wypowiadaniu sądów na tematy rzeczywistości… co zresztą w sensie technicznym i tak sprowadza się później do tego samego. Spór dotyczy tu raczej fundamentów filozoficznych.

PW: Z bardziej pragmatycznego punktu widzenia ważne jest to, że konstrukty, o których mówisz, są niewątpliwie użyteczne. Jeśli taki konstrukt, jak np. czynnik inteligencji ogólnej g w badaniach inteligencji wyjaśnia większy procent wariancji

niż którykolwiek z surowych pomiarów zdolności umysłowych składających się na iloraz inteligencji, to dobitnie pokazuje tę użyteczność, prawda?

JG: No tak, bo on zyskuje pewnego rodzaju trafność predykcyjną i jest w stanie przewidywać większe spektrum pochodnych tego, co nazywamy inteligencją niż każdy pojedynczy wskaźnik. Nawet, gdyby ta cecha była obserwowalna, to – o ile mówimy o jakimś uniwersum typowych wskaźników o umiarkowanej korelacji – główna składowa, będąc po prostu liniową kombinacją tych wskaźników, będzie wyjaśniała każdy z nich lepiej niż dowolny inny pojedynczy wskaźnik.

Walor predykcyjny powoduje, że sięgamy do tego rodzaju konstruktów. Jak mówię, możemy pozostawić na boku debatę o ich status ontologiczny. Wracając do wątku specyfiki dyscypliny, socjologia, jeżeli nawet jest zdyscyplinowana metodologicznie i dąży do weryfikacji względnie mocnej hierarchii dowodów, teorii czy twierdzeń o zależnościach przyczynowych – bo to jest dla mnie teorią w naukach społecznych: twierdzenie o zależnościach przyczynowych, czyni to bardziej opierając się na wynikach pochodzących z danych o charakterze obserwacyjnym. Posługuje się przy tym modelami strukturalnymi bądź też innymi rodzajami modelowania, pozwalającymi zasadnie wypowiadać się na temat zależności przyczynowych. Mam tu na myśli modelowanie testujące na podstawie modelu kontrfaktycznego przyczynowość oddziaływania. To może być regresja ze zmiennymi instrumentalnymi, może być metoda *difference in difference* albo np. rodzaj analizy dopasowywania obiektów, takiej jak *propensity score matching*, czy mogą to być inne rodzaje podobnych dopasowań… Oczywiście, każde z tych podejść ma swoje zalety, każde ma też jakieś rozpoznane słabości. Ale generalnie, każde z nich jest lepsze, od posługiwania się metodami naiwnymi bądź też wyciągania wniosków przyczynowych *ad hoc* z czystych analiz opisowych…

Ogólnie jednak trzeba powiedzieć, że w naukach społecznych takie bardziej zdyscyplinowane podejście stanowi niestety nurt mniejszościowy. No może w ekonomii – oczywiście tej, która szuka ekonometrycznej weryfikacji swoich teorii – spotykamy różne przypadki modelowania, także takie, które spełniają warunki niezbędne dla interpretacji przyczynowej. Oczywiście, samo zrobienie modelu regresji nie daje podstaw do tego, żeby twierdzić, że nawet pozostające w silnych zależnościach cechy są powiązane przyczynowo. Do tego potrzebny jest element analizy kontrfaktycznej, takiej jaką umożliwia eksperyment. Bo to eksperyment jest tutaj złotym standardem. On jest oparty, można powiedzieć, na zasadzie poszukiwania kontrfaktyczności poprzez zderzenie próby testowej z próbą kontrolną reprezentującą sytuację kontrfaktyczną, czyli sytuację, w której bodziec nie działa na osoby objęte badaniem w porównaniu z taką sytuacją, w której on działa.

Poszukiwanie różnicy pomiędzy oddziaływaniem i nieoddziaływaniem bodźca u osób objętych interwencją, czyli *treatment effect*, pozwala właściwie ustalić siłę wpływu przyczynowego. Taką szansę daje eksperyment. Choć trzeba tu też powiedzieć, że to nic dziwnego, iż liczba udanych replikacji jest ograniczona. Na to się składa szereg rozmaitych kwestii. To nie tylko jest problem związany z niewłaściwym stosowaniem miar statystycznych, choć oczywiście i z tym możemy mieć do czynienia. Trzeba powiedzieć, że eksperymenty często nie są

przygotowywane wystarczająco starannie z punktu widzenia ich statystycznej oprawy. Na przykład nie zapewnia się przy planowaniu warunków eksperymentu odpowiedniej mocy statystycznej.

PW: Pełna zgoda.

Czy i jak używać testów istotności? Replikowalność parametru czy mechanizmu?

JG: Zapomina się, że coś takiego jak moc też jest ważnym elementem składowym oceny sytuacji eksperymentalnej. Posługujemy się zwykle istotnością. Chcemy po prostu stwierdzić, czy następuje oddziaływanie.

PW: W psychologii chyba bardziej niż w socjologii?

JG: W psychologii, w psychologii. O socjologii też zaraz powiem, ale teraz myślę o psychologii… Tam problem replikacji to jedna rzecz, ale druga, o której się mówi, to tzw. *p-hacking*, czyli nadużywanie testu istotności do stwierdzania oddziaływania.

PW: No i związany z tym *publication bias* – fiksacja autorów, redaktorów i recenzentów na magicznym $p = 0{,}05$.

JG: Tak, rzecz w tym, że tych elementów nie planuje się w powiązaniu. Wiemy przecież, że przyjęcie określonego krytycznego poziomu istotności nie pozostaje bez wpływu na moc testu. Razem z liczebnością próby wpływa on na ogólne, statystyczne warunki eksperymentu. Powinno się to łącznie ustalać na poziomie planowania. Ale mnie szczególnie zastanawia kwestia tej gwałtownej krytyki testów statystycznych i praktyki podawania prawdopodobieństwa testowego, p. Podaje się dokładną wartość p – komputery ją bez trudu wyliczają – i sprawdza się, czy ona leży poniżej progu. Ten próg umownie przyjmuje się najczęściej na poziomie pięciu setnych i oczekuje się, że testowana zależność powinna się replikować odpowiednio często, skoro prawdopodobieństwo popełnienia błędu pierwszego rodzaju jest mniejsze niż pięć na sto powtórzeń[5].

[5] Zwróćmy uwagę, że mowa tu o warunkowym prawdopodobieństwie popełnienia błędu pierwszego rodzaju, czyli odrzucenia hipotezy zerowej wtedy, kiedy jest ona prawdziwa. Przeprowadzając statystyczny test istotności, nie wiemy, czy hipoteza zerowa jest prawdziwa, czy fałszywa, dlatego błędem byłoby uznanie, że wartość p oznacza prawdopodobieństwo podjęcia błędnej decyzji albo nieudanej replikacji w tym konkretnym przypadku. Wspominamy o tym, bo ten i podobne błędy interpretacyjne są w praktyce stosowania testów istotności bardzo częste (przyp. red.).

Co zaproponowano zamiast testów istotności? Redaktorzy *Basic and Applied Social Psychology* ogłosili na fali krytyki nadużywania wyników testów statystycznych, że ich czasopismo nie będzie publikowało artykułów zawierających wartość statystyki *p* (Trafimow i in., 2015, za: Woolston, 2015), co jest reakcją skrajną i niepotrzebną. Osobiście przychylam się do stanowiska, które oddaje tytuł innej publikacji zaangażowanej w tę debatę: *The Practical Alternative to the p Value Is the Correctly Used p Value* (Lakens, 2021).

Oczekiwanie, że w badaniach psychologicznych, w jakichkolwiek badaniach z zakresu nauk społecznych, które nie mają charakteru ściśle, powiedziałbym, neuronalnego, biochemicznego (gdzie też jest zmienność, ale mniejsza), osiągniemy dokładną replikowalność parametrów, jest nierealistyczne. W badaniach np. z zakresu psychologii społecznej, a nawet częściowo poznawczej, tam, gdzie nie dysponujemy mocnymi danymi, za którymi stałaby twarda biochemia, trudno o replikowalność parametrów. Jedyne czego możemy oczekiwać – i dobrze, jeśli nam się to uda – to replikowalność mechanizmu – struktury modelu, wzoru, zależności, *pattern*. W ekonomii zwracał na to uwagę Friedrich Hayek, a jeszcze wcześniej Ludwig Mises, który w swoim słynnym tekście *Human Action* stwierdził, iż oczekiwanie, że w tym samym miejscu za rok elastyczność cenowa popytu na ziemniaki będzie taka sama, jest oczekiwaniem głupca. Co nie podważa jednak ogólnej zasady, mówiącej, że w normalnych warunkach, gdy rośnie cena, to powinien też spadać popyt na dobro. Chyba że mamy do czynienia z pewnym szczególnym rodzajem sytuacji, opisał taką np. Robert Giffen, kiedy na skutek konkurencji dóbr – mają tu znaczenie jeszcze kwestia możliwości substytucji i inne czynniki – w odniesieniu do dóbr elementarnych, zaobserwowano w Irlandii sytuację, w której wzrost ceny ziemniaków paradoksalnie powodował zwiększenie się popytu, bo ludzie nie mogli już sobie na nic innego pozwolić, więc wybierali najtańszy sposób pożywienia w warunkach zagrożenia głodem, czyli właśnie ziemniaki, i dlatego – mimo iż ich cena rosła – wzrastał też popyt. Zdarzają się więc pewnego rodzaju odstępstwa, które jednak nie podważają ogólnej zasady, są wytłumaczalne. Nikt jednak nie będzie oczekiwał, że otrzymamy identyczne parametry…

PW: Psychologia różni się jednak od ekonomii tym, że nas zazwyczaj nie interesują predykcje dotyczące parametrów, które są kluczowe w ekonomii. W ekonomii znajomość mechanizmu jest potrzebna po to, żeby predykcja była precyzyjna. Natomiast w psychologii potrzebujemy replikacji badań, które służą do zidentyfikowania mechanizmów, np. badań zależności.

JG: Tu się nie zgodzę… Na końcu i tak chodzi o predykcję. Jeżeli teoria przyczynowa jest właściwa, w nauce zawsze chodzi o predykcję. Jeżeli przyjmujemy system hipotetyczno-dedukcyjny, konfirmowany, czyli weryfikowany empirycznie i struktura teorii jest oparta na zależnościach przyczynowych, to z natury rzeczy taki system daje na końcu możliwość predykcji. Inna sprawa, czy to mają być predykcje w kategoriach konkretnych wartości…

PW: Otóż to. Jeśli znamy np. mechanizm rozwoju dziecka, to chcemy na tej podstawie móc przewidywać, czy i kiedy pojawią się w tym rozwoju określone

trudności. Ten poziom predykcji jest w psychologii obecny, ale nic bardziej ambitnego… Nie potrzebujemy np. szacować ilorazu inteligencji dziecka w wieku lat siedemnastu na podstawie tego, jaką miało inteligencję w wieku lat dwunastu…

JG: Myślę raczej o predykcjach w kategoriach wzorców. Choć ekonomia aspiruje do przewidywania na bardziej precyzyjnym poziomie i niekiedy jej się to udaje, to jednak nie zawsze… Kiedyś napisaliśmy z moimi współpracownikami tekst, w którym przeanalizowaliśmy prognozy dotyczące rynku pracy, robione przez wszystkie polskie instytucje, które się tym zajmowały. Tekst był pisany ok. roku 2008, analizowaliśmy lata 90., prawie całe, mniej więcej do pierwszego kryzysu finansowego. No i w konkluzji musieliśmy stwierdzić, że wszyscy się w swoich prognozach mylili. Nie tylko co do poziomu, ale też kierunku. I to systematycznie… Więc ja z tą zdolnością ekonomii do predykcji nie przesadzałbym. Są tworzone różne modele, np. inflacji, które mogą dostarczać czasem trafnych prognoz – w warunkach pewnej stabilności układu instytucjonalnego, o ile nie nastąpią jakieś dodatkowe zdarzenia modyfikujące ten mechanizm, który nie jest przecież izolowany. Chodzi jednak o w przewidywanie wzorców. Rzecz w tym, żebyśmy wiedzieli, że jeżeli będziemy działali w sposób A, wobec alternatywy B, to mamy szansę w danym przypadku uzyskać lepszy efekt, że np. rozwój dziecka będzie w scenariuszu A korzystniejszy niż w scenariuszu B.

PW: Pełna zgoda.

JG: Ani w przypadku socjologii, ani psychologii nie miałbym oczekiwań idących szczególnie daleko. Wyjątek stanowią tylko zjawiska związane z procesami o charakterze podstawowym, biochemicznym, jak mówiłem hasłowo wcześniej. Choć oczywiście i w ich przypadku mamy do czynienia z pewnego rodzaju losową zmiennością. Ona jest obecna zawsze, tylko w przypadku tych procesów i zjawisk jest mniejsza i dlatego można uzyskać bardziej satysfakcjonujące, bardziej precyzyjne wyniki.

Wróćmy do powodu, dla którego o tym mówimy – jako rozwiązanie problemu ograniczonej replikowalności badań zaproponowano w skrajnym wariancie podawanie wielkości efektu. Tak jakby wielkość efektu nas przed czymś ratowała, kiedy nie wiemy, czy efekt jest statystycznie istotny, bo ktoś zabronił publikowania wyniku testu statystycznego…

Nie w tym jest problem. Chodzi o to, że publikuje się wyniki testów statystycznych zastosowanych do wyników ze źle przeprowadzonych badań… Test nie jest zły sam w sobie. Może bardzo dobrze służyć, jeżeli warunki jego prawomocnego stosowania są spełnione – jeżeli eksperyment został właściwie zaplanowany i zrealizowany. Także i wtedy, oczywiście, trzeba te eksperymenty replikować. Tymczasem czasopisma odrzucały replikacje, nie zajmowały się nimi, a trzeba replikować! Bo nawet, jeżeli tylko pięć razy na sto popełnimy błąd przy podejmowaniu decyzji statystycznej na podstawie wyniku testu, to wcale nie jest powiedziane, kiedy ten błąd nastąpi…

Prawdopodobieństwo katastrofy lotniczej jest bardzo niskie, a i tak wiele osób na wszelki wypadek zażywa relanium albo – korzystając z serwisu pokładowego – znieczula się inaczej.

PW: Mimo że prawdopodobieństwo wypadku lotniczego jest dużo mniejsze niż prawdopodobieństwo błędnej decyzji w badaniach…

JG: Replikacje są potrzebne, bo nigdy nie wiadomo, w którym momencie się to prawdopodobieństwo zrealizuje.

PW: Fishera wini się za to, że wymyślił kryterium 0,05, gdy tymczasem on zawsze podkreślał nieuchronność możliwości pomyłki i konieczność replikacji. Zaakceptował fakt, że badacze przyjmują takie kryterium, uznał, iż wartość 0,05 nie jest złym pomysłem, bo to odpowiada przedziałowi o szerokości mniej więcej dwu odchyleń standardowych, co wydaje się rozsądnym kryterium. Supermałe kryterium wymagałoby nierealistycznie wyrafinowanych badań…

JG: Dużych prób, po prostu.

PW: Więc na co dzień jest to użyteczne kryterium, ale musimy pamiętać, że ono jest relatywne i że to badacz powinien decydować o tym, jakie narzędzie dobiera do jakiego problemu, jaki poziom istotności wybiera i potem stosuje. To jest decyzja jakościowa. Jeszcze w jednej kwestii Fisher bywa źle rozumiany – często myślimy o istotności statystycznej jako istotności obserwowanych *wyników*, tymczasem on nie mówił o istotnych *wynikach*, a o ich istotnej *sprzeczności* z hipotezą zerową. Fakt zaobserwowania bardzo niskiej wartości *p* silnie kłóci się z hipotezą zerową i pozwala nam ją mocno odrzucić. Jednak mocne odrzucenie hipotezy zerowej, a uznanie, że efekt jest istotny praktycznie, czytaj „znaczący", to są różne rzeczy.

JG: To problem błędnej interpretacji, złego zrozumienia terminu… Można tu przywołać teorię semantyczną Sapira-Whorfa, która mówi, że język decyduje o tym, w jaki sposób widzimy świat i generuje nasze działania. Jak w tym słynnym przykładzie z pustymi beczkami po benzynie, gdzie różnice w znaczeniu słowa „pusty" w dwu odmiennych językach sprawiały, iż użytkownicy jednego z nich częściej ignorowali przepisy bezpieczeństwa i palili papierosy przy rzekomo „pustych" beczkach, zawierających wybuchowe opary. Tak samo tu, nie powinniśmy używać terminu *istotny* statystycznie.

PW: Tak jest.

JG: Kiedy omawiam to zagadnienie ze studentami, mówię, że prawdopodobieństwo błędnego odrzucenia prawdziwej hipotezy zerowej jest mniejsze niż 0,05 albo 0,01, albo po prostu wynosi np. 0,003… To jest niewielkie prawdopodobieństwo, w związku z tym, odrzucamy hipotezę zerową. Podejmując taką decyzję, rzadko popełnimy błąd.

Trzeba też pamiętać o przygotowaniu statystycznym badań, o którym mówiłem wcześniej. Poziom istotności i moc testu są ze sobą powiązane. Na etapie planowania badań możemy przyjąć, jaka minimalna wielkość efektu będzie dla nas znacząca z punktu widzenia wagi mechanizmu, jego merytorycznego znaczenia.

To pozwoli ustalić odpowiednie progi dla liczebności próby, mocy testu i poziomu istotności. Wszystkie te elementy są ze sobą powiązane. Są aplikacje komputerowe, przeznaczone właśnie do tego. One powinny być po prostu, najnormalniej w świecie, wykorzystywane. Są komercyjne, darmowe, np. R. Akurat na naszym uniwersytecie mamy do dyspozycji kilka różnych pakietów, np. Stata ma to bardzo dobrze zrobione, pozwala zrobić bardzo ładne wykresy pomagające podjąć właściwe decyzje odnośnie do projektowanych badań[6].

Wracając do wcześniejszej myśli, dla mnie nie jest rozwiązaniem zastąpienie w publikacjach podawania wartości *p* podawaniem wielkości efektu. Bo ta goła wartość sama w sobie niewiele mówi. Nie oczekiwałbym też dokładnej replikacji zaobserwowanego poziomu efektu. W następnym badaniu będziemy mieli inną wartość tego efektu!

Arkadiusz Białek: W psychologii rozwojowej wartość predykcyjna jest dodatkowo zmniejszona przez efekt kohorty. Nie możemy oczekiwać, że za kilka lat zaobserwujemy dokładnie to samo. Podobnie, jak rozumiem, jest w ekonomii…

JG: Uniwersalność teorii może być weryfikowana właśnie przez replikację eksperymentów w różnych segmentach populacji. Jeżeli hipoteza przewiduje, że zachodzi tutaj efekt kohortowy, no to trzeba eksperyment zreplikować w różnych kohortach.

AB: Dokładnie. Są plany badawcze, które pozwalają oddzielić efekt kohorty. Jednak badania podłużne, które są w psychologii rozwojowej najistotniejsze, bo pozwalają zidentyfikować zmianę rozwojową, nie dość, że są trudne same w sobie, są też bardzo trudne do zreplikowania.

JG: To są badania panelowe, czyli powtarzane u tych samych osób. One pozwalają zmierzyć *gross change*, czyli zmianę, która zachodzi na poziomie osób pomiędzy poszczególnymi punktami pomiaru w czasie. Na przykład związaną z fazami cyklu życia. Efekt kohortowy to jednak coś innego. Kohorta jest nośnikiem pewnych procesów socjalizacyjnych. Można je badać w niezależnych próbach. I można tu robić niezależne replikacje. Po pierwsze, trzeba zreplikować badanie na tej samej kohorcie po pewnym czasie, to może być niezależna próba, byle była pobrana z tej samej kohorty. Po drugie, powinniśmy oczywiście robić za każdym razem, w każdej fali pomiarów, badania nad różnymi kohortami. Następnie, mając te wszystkie wyniki, trzeba odseparować efekt kohortowy od efektu fazy cyklu życia.

AB: Dokładnie tak.

[6] Analizę mocy testu i określenie wielkości próby umożliwia darmowe oprogramowanie G*Power (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower) (przyp. red.).

JG: Tam jest jeszcze trzeci efekt – okresu historycznego, *period effect*, działanie pewnego splotu czynników, można powiedzieć historycznych, który może oddziaływać na wszystkie kohorty.

AB: Choćby pandemia, prawda?

JG: Tak jest. Może oddziaływać na wszystkie kohorty i najgorsze jest to, że jak próbujemy później zrobić jeden model, to te parametry są wzajemnie powiązane i niestety nie mamy dość stopni swobody, żeby taki model estymować. Ale to można obejść. Są sposoby, które pomagają zmagać się z tym problemem. Widać tu, jak nauka wytwarza narzędzia do tego, żeby radzić sobie z różnorodnymi pytaniami badawczymi. *Research design*, czyli plan badawczy, jest zawsze pochodną pytania badawczego, które stawiamy, problemu, który chcemy rozwiązać.

Mam wrażenie, że zarówno w kształceniu, jak i w samokształceniu, a następnie w pracy badacza, przywiązuje się zbyt małą wagę do właściwego planowania badań, *research design.*

AB: Całkowicie się zgadzam.

JG: To tutaj tkwi problem. Nie wyleczy się go, „banując" wartość *p*. Skazanie go na banicję nie rozwiąże problemu niewłaściwego stosowania testów ani tego, że nam się nie replikują badania – bo ktoś przyjął niewłaściwe założenia albo zbyt pochopnie stwierdził, że ma do czynienia z efektem istotnym statystycznie…, bo się po prostu ucieszył…

Miałem w życiu okres, kiedy intensywnie zajmowałem się przekazywaniem wiedzy na temat analiz statystycznych. Robiłem różne szkolenia, seminaria. Kiedyś po jednym z takich spotkań, w czasie którego mówiliśmy o testach dokładnych dostępnych w pakiecie SPSS, podszedł do mnie uczestnik, który pracował dla pewnej uczelni medycznej i mówi, że on nie lubi SPSS-a, woli Statistikę, bo tam dostaje w jednej tabelce wyniki wielu różnych testów i od razu widzi, które testy wybrać, żeby potwierdziły hipotezy, na których badaczowi zależało. Klienci są z niego zadowoleni, a jemu takie zestawienie oszczędza wiele pracy… Oczywiście, jeśli ktoś prowadzi badania w taki sposób, no to rodzi wiele problemów…

PW: Chciałbym odnieść się do postulatu, o którym mówisz, zastąpienia wartości *p* wielkością efektu. Owszem, wielu metodologów przypomina o konieczności zwracania większej uwagi na wielkość efektu, ale jeśli chodzi o zastępowanie *p* inną miarą, to najbardziej wpływowy jest tu chyba głos Geoffa Cumminga, autora podręcznika *New Statistics*, który postuluje zastępowanie wartości *p* przedziałami ufności. Matematycznie jedno i drugie jest ściśle powiązane, więc zmiana jest – można by powiedzieć – kosmetyczna, ale jednak znacząca. Jeśli badacz jest przeciętnie zaznajomiony z podstawami wnioskowania statystycznego, to uważa, że jeśli *p* jest mniejsze od 0,05, wówczas może wierzyć, że prawidłowość, którą zaobserwował w próbie, zachodzi również w populacji.

JG: No, ale jeśli on będzie miał przedział ufności, to dojdzie do takiego samego wniosku, widząc, że przedział ufności nie obejmuje 0.

PW: Otóż nie, Cumming opiera swój postulat i program dydaktyczny na wynikach badań pokazujących, że jeśli badaczom pokazuje się dwa efekty takie, że dla jednego z nich wartość *p*, przekracza magiczne kryterium istotności 0,05, a dla drugiego nie, to są bardziej skłonni uznawać te efekty za jakościowo odmienne niż wtedy, kiedy im się pokazuje dwa przedziały ufności…

JG: Dla mnie to jest fałszywy trop. Żeby było jasne, nie mam nic przeciwko przedziałom ufności. Ponieważ to nie kosztuje wiele, radziłbym badaczom – pokazujcie i to, i to. Podejście, w którym coś trzeba wyrzucać, jest błędne. Ale argument z naiwności nie jest dobrym argumentem. Przecież my tu nie mówimy o tekstach popularnych, w których trzeba starannie dobierać środki w taki sposób, żeby intuicyjny odbiór czytelnika był zgodny z zamierzeniem popularyzatora, czyli żeby popularyzator uzyskiwał efekt prawidłowego odbioru treści naukowej. Jednak kiedy mowa o komunikacji wewnątrz nauki, to wymagamy, żeby strony, które się komunikują, dysponowały odpowiednią kompetencją warsztatową. Badacze muszą być dobrze przygotowani.

Mamy, oczywiście, problem z tym, że nie wszyscy badacze są dobrze przygotowani.

PW: Wyniki badań rozumienia podstaw wnioskowania statystycznego sugerują, że większość z nich nie jest.

JG: No, ale od tego są recenzenci. Powinno się dbać – i dobre czasopisma to robią – żeby recenzenci byli bardzo dobrze przygotowani od strony statystycznej. Jeśli mają z tym problem, to mogą powołać osobnych recenzentów metodologiczno-statystycznych, którzy będą sprawdzali, czy w składanych pracach wszystko jest pod tym względem w porządku.

Trzeba to kontrolować, bo wiedza naukowa to pewien specyficzny segment wiedzy ludzkiej, społecznej, który charakteryzuje się tym, że powstaje w szczególny sposób – na podstawie metody naukowej. Nie będę się rozwodził nad pojęciem metody naukowej, bo się tu pewnie dobrze rozumiemy, ale dla tej metody kluczowa jest metodyka badania naukowego, metodologia, która mówi, zgodnie z najlepszą obecną wiedzą, w jaki sposób należy wykonać badania, żeby ich wyniki były, w sensie naukowym, wiarygodne. Ta wiedza ulega zmianom, dlatego czasem stare wyniki mogą ulegać refutacji, kiedy pojawiają się nowe, lepsze metody badania jakiegoś zjawiska.

W naukach przyrodniczych też mieliśmy do czynienia z takimi przypadkami, kiedy badacze dostawali do dyspozycji lepsze instrumenty i odkrywali anomalie, niewytłumaczalne w ramach starszych teorii, które stawały się nie do utrzymania i trzeba je było odrzucić.

Dlatego w zakresie metodologii każdy z nas musi się kształcić przez całe życie. Jeżeli chce uprawiać naukę, to musi cały czas rozwijać swój warsztat.

PW: To świetna konkluzja tej części naszej rozmowy…

Znaczenie badań eksploracyjnych i konfirmacyjnych w budowaniu teorii

PW: Chciałbym, żebyśmy teraz zmienili nieco temat i porozmawiali o budowaniu teorii. Pretekstem niech będzie tocząca się ostatnio dyskusja wokół praktyki znanej jako HARKing, czyli *Hypothesizing After the Results are Known* (Kerr, 1998) – formułowania hipotez badawczych po uzyskaniu wyników. Uważa się to za błąd. Jak więc prowadzić badania, które sprzyjają budowaniu dobrych teorii?

JG: Trzeba spytać, czy rzeczywiście to jest błąd. Tutaj znów pewnie nie pójdę za głównym nurtem, bowiem bliższe od popperowskiego jest mi stanowisko, które poznałem później – Charlesa Peirce'a. Peirce zwracał uwagę na to, co się dzieje przed momentem, w którym – jak chciał Popper – stawiamy hipotezę i testujemy ją, czyli poddajemy próbie obalenia.

Popper mówił: niezbyt mnie interesuje, skąd się bierze hipoteza. To trochę nonszalanckie stwierdzenie… A Peirce'a to właśnie interesowało. Mówił o procesie abdukcji, która wynika z prowadzenia systematycznych obserwacji, jako o naturalnym etapie procesu odkrycia naukowego. Mówił o naukowym olśnieniu, *serendipity*. Rozpoznawalny dla wszystkich przykład takiego olśnienia to Archimedes wyskakujący z wanny i krzyczący: *Eureka!*

Przykładem bliższym życiu może być to, co robił serialowy doktor House ze swoimi współpracownikami. Oni zbierali pewne parametry medyczne badanej osoby, po czym stawiali hipotezy. W ich przypadku to były teorie dotyczące tego, z jaką chorobą mają do czynienia, choroby mają bowiem określone objawy. Istnieją pewne zależności przyczynowe między obserwowanymi parametrami a wystąpieniem danego schorzenia. Oni się starali odgadnąć, z czym związana jest określona konfiguracja danych, czyli hipotetyzowali po pobraniu tych parametrów. Potem, oczywiście, następował drugi etap, już popperowski, czyli, mając hipotezę dotyczącą zależności przyczynowych, wyprowadzamy z niej przewidywanie, dotyczące tego, co powinno wystąpić, jeżeli nasza teoria jest prawdziwa. Czyli stawiamy hipotezę o zależności przyczynowej: jeżeli pacjent choruje na dane schorzenie, to powinniśmy zaobserwować taki, a nie inny specyficzny element. Zróbmy zatem dodatkowe badanie i sprawdźmy. To jest coś, co występuje w owej chorobie, a nie występuje w innych, więc pozwoli nam stwierdzić, z czym mamy do czynienia. Planujemy więc badanie, które ma albo obalić naszą hipotezę, albo pozwolić nam ją utrzymać.

Przeprowadzamy test konsekwencji przyczynowych, wynikających z teorii. Ale teoria bierze się jednak z tego, że my cały czas funkcjonujemy w pewnym obszarze wiedzy. Nie jesteśmy naukową *tabula rasa*, nie czyścimy naszych mózgów przed sformułowaniem hipotez, które poddajemy falsyfikacji. Jesteśmy zanurzeni w pewnym paradygmacie… w pewnym zbiorze teorii. Teorii, które czasami ze sobą konkurują… Gromadzimy obserwacje i mówimy: spośród znanych mi teorii ta akurat do tych danych pasuje. Jeżeli ona jest prawdziwa, to co z niej powinno wynikać? Albo mówimy: ta teoria tu nie pasuje, no to wtedy zaryzykuję stworzenie własnej teorii. To też jest w jakimś sensie hipotetyzowanie *post hoc*, ale to jest tworzenie teorii, do której inspirują mnie te dane. Tylko, co ważne, ja

na tym nie poprzestaję. Moje przedsięwzięcie naukowe nie kończy się na wyprowadzeniu koncepcji z uzyskanych danych, tylko na postawieniu i zweryfikowaniu takiej hipotezy, która mnie doprowadzi do konkluzywnego wniosku. Tę pierwszą fazę nazwałbym po prostu umownie eksploracją, a tę drugą też umownie – konfirmacją.

Jestem gorącym zwolennikiem badań eksploracyjnych. Uważam, że dzięki nim narodziło się bardzo wiele ciekawych rzeczy. Eksplorowanie jest dla nauki zresztą dość naturalne. Spójrzmy, ile dzięki niemu powstało rzeczy bardzo ciekawych, nawet jeśli często nie miały później konkluzji.

Na przykład dystynkcje Pierre'a Bourdieu, bardzo popularnego socjologa, można powiedzieć ikony socjologii późnego XX wieku. Jak powstała jego koncepcja struktury klasowej społeczeństwa w powiązaniu ze stylami życia, gustami, które on nazywał *tastes*? Bourdieu sprawdzał, czym ludzie się zajmują zawodowo – traktując to jako manifestację ich sytuacji klasowej, pozycji w strukturze społecznej – i notował, na co poświęcają swój czas wolny, analizując te wybory z punktu widzenia gustów czy smaku. Korzystając z techniki analizy korespondencji, sporządził mapę percepcyjną, zobaczył, jak to się wszystko układa i zaczął wyciągać wnioski. Zbudował koncepcję teoretyczną, ale teraz pojawia się kolejny etap. Z tej koncepcji należy wyprowadzić przewidywania: jeśli ona jest prawdziwa, to pojawienie się elementu A powinno pociągnąć za sobą A', a pojawienie się B powinno skutkować zaobserwowaniem B'. Znów mowa tu nie o replikowalnych wartościach parametrów, tylko replikowalnych wzorach zależności.

AB: Bardzo się cieszę, że w naszej rozmowie dochodzimy do wniosków, które tak bardzo współbrzmią z ideami pojawiającymi się i proponowanymi w kontekście reformowania psychologii. Holenderski psycholog i metodolog, Denny Borsboom, uważa, że proces poznania naukowego ma charakter iteracyjny. Tak jak mówisz, zaczynamy od badań deskryptywnych i identyfikacji pewnych wzorców zależności…

JG: Tak, struktury. Struktury, która rządzi korelacjami obserwowanymi między zjawiskami.

AB: Zdaniem Boorsbooma (Boosboom i in., 2021) ten model należałoby przynajmniej częściowo sformalizować i przeprowadzić symulację. Moglibyśmy tu więc mówić o dodaniu jeszcze jednego etapu. Co jednak szczególnie ciekawe, to fakt, że oni też odwołują się do Charlesa Sandersa Pierce'a i mówią wprost o abdukcji jako etapie wstępnym…

JG: Cieszę się, że nie jestem samotny. (*śmiech*)

AB: Moim zdaniem w przypadku HARKingu problemem jest to, iż hipotetyczno-dedukcyjny wzorzec uprawiania nauki jest tak powszechną normą – przyjmowaną czasem nawet nieświadomie, wymuszaną też przez strukturę artykułów naukowych – że wielu badaczy czuje bardzo silną potrzebę wejścia w ów wzorzec

konfirmacyjny i formułuje hipotezy po uzyskaniu wyników. Zamiast w duchu transparentności przyznać, że hipoteza pojawiła się w następstwie poczynionych obserwacji, przedstawiają ją jednak jako coś, z czym przystępowali do badań…

JG: To jest podnoszenie statusu dowodowego swojego osiągnięcia. Jeśli się śledzi literaturę z zakresu modelowania strukturalnego, to tam widać podobny problem. Nie będę wchodził w szczegóły debat między zwolennikami różnych podejść do weryfikacji modeli, ale powiem o jednej dyskusji, toczącej się wokół pytania o zasadność wykorzystywania tzw. indeksów modyfikacyjnych do poprawiania modeli strukturalnych. Te indeksy mówią nam, jak uwolnienie danego parametru wpłynie na zmianę całościowego dopasowania modelu. Dzięki temu możemy *post hoc* dokonywać drobnych modyfikacji poprawiających dopasowanie. Moim zdaniem samo w sobie takie postępowanie nie jest czymś złym. To pewnego rodzaju eksploracja. Po takim postępowaniu naprawczym należy jednak ów zmodyfikowany model przetestować na nowych danych z niezależnej próby. To zalecają zwolennicy tej praktyki. Mówią – nie bójcie się stosować indeksów modyfikacyjnych, bo one wam pomagają. Jak śledzenie reszt pomiędzy macierzą odtworzoną a zaobserwowaną…

PW: Trochę jak w regresji...

JG: Dokładnie – poprawiamy model.

PW: Jednak ryzykując *overfitting* – za dobre dopasowanie modelu do tych szczególnych danych.

JG: Ale to wszystko wpisuje się w fazę eksploracyjną. Właśnie po to później następuje weryfikacja – sprawdzanie, czy model da się utrzymać.

Toczy się też dyskusja, co jest kryterium przyjęcia bądź odrzucenia modelu. Ortodoksi uważają, że jako takie kryterium może służyć tylko test chi-kwadrat. Inni uważają, że lepiej jest użyć miar opisowych, takich jak jedna z najbardziej popularnych, RMSEA, *root mean square error of approximation*, czy inne. Przyjmuje się tu różne, mniej lub bardziej umowne, progi. Zwolennicy tych metod zwracają uwagę, że przyjmując chi-kwadrat jako kryterium, przy dużych próbach – a w badaniach społecznych często stosuje się duże próby – praktycznie żaden model nie byłby do utrzymania.

Nie chcę tego teraz rozstrzygać, ale co ważne, nikt – ani jeden, ani drugi obóz – nie kwestionuje, że po dokonaniu przeróbek, adiustacji modelu – potrzeba niezależnej konfirmacji.

AB: Ale mało kto to robi...

JG: No tak, mało kto to robi.

Zresztą w badaniach eksperymentalnych też mamy podobne problemy. Pokazywałem kiedyś na jednej z konferencji psychologicznych – za Kennethem Bollenem i Judeą Pearlem (2013) – jakie błędy w eksperymentach psychologicznych

możemy popełniać z powodu tego, że zarówno zmienne niezależne, jak i zmienne zależne są cechami latentnymi. Natomiast nie tylko bodziec, którym bezpośrednio manipulujemy, lecz także mierzona wielkość to są tylko wskaźniki tych latentych cech. Zakładamy, że gdy dokonujemy określonego zabiegu, to wywołujemy rzeczywistą zmianę poziomu bodźca na wejściu.

Czyli np., chcąc zbadać wpływ motywacji na poziom koncentracji uwagi, chwalimy osoby badane i zakładamy, że to zwiększa ich motywację. Potem mierzymy stopę błędów popełnianych w zadaniu wymagającym uwagi, traktując ją jako (negatywny) wskaźnik koncentracji uwagi.

Ale jaką mamy pewność, że tak naprawdę oddziaływanie nie zachodzi bezpośrednio między wskaźnikami? Że zastosowana pochwała wprost nie zmniejsza prawdopodobieństwa błędu w użytym zadaniu, np. dostarczając osobie badanej użytecznej informacji zwrotnej? Zakładamy, że manipulując naszym obserwowalnym wskaźnikiem, rzeczywiście dokonujemy zmiany w poziomie bodźca na wejściu, że zwiększamy motywację (czyli zmienną latentną), która poprawia skuteczność działania uwagi, co przejawia się zmniejszeniem liczby błędów. Ale przecież tego nie wiemy.

Proponowane rozwiązanie tego problemu w przypadku modelowania strukturalnego polegało na połączeniu eksperymentu z modelowaniem, na zastosowaniu przynajmniej dwóch lub większej liczby wskaźników.

PW: W psychologii problem trafności wewnętrznej jest tak powszechny, że myśmy się nawet trochę przyzwyczaili, że zawsze go ryzykujemy. Trzeba jednak coś robić, żeby go minimalizować.

JG: Generalnie można powiedzieć, że metodologia badań jest nauką o unikaniu błędów poznawczych. Zaczynając od najprostszych błędów podejścia naturalnego, takich jak np. nieuzasadnione uogólnianie, a skończywszy na kwestiach bardziej wyrafinowanych, takich jak właśnie omawiany problem niedostrzegania istotnej struktury relacji przyczynowych, dotyczących danej sytuacji badawczej. Te rzeczy wymagają większej uwagi badaczy.

Ale żeby jeszcze krótko skonkludować problemy budowania teorii…

Według mnie naukowcy w sposób naturalny pracują eksploracyjnie. Jeden z moich przyjaciół badaczy powiedział mi kiedyś: „wiesz, dla mnie największa przyjemność jest wtedy, kiedy się przymierzam do problemu, szukam, robię różne analizy eksploracyjne, żeby stworzyć model. Ale kiedy już ten model mam i muszę się zająć jego weryfikacją, to się już robi strasznie nudne…". Tam trzeba już wykonać konkretną, żmudną pracę…

AB: No tak, a tu jest kreacja.

JG: Tak, kreacja, owo wspomniane wcześniej *serendipity*, olśnienie. I ta przyjemność, którą się odczuwa, kiedy człowiekowi wpada do głowy coś, co może zbliżyć do rozwiązania problemu…

Jest jeszcze inna ważna kwestia. Można przeciwstawić sobie dwa podejścia: naukowe i praktyczne. Nauka z natury dąży do formułowania wyjaśnień

przyczynowych, do poznania mechanizmu. Praktyka jest za to zainteresowana głównie predykcją.

I teraz, w tej chwili można powiedzieć, że potężny rozwój technik obliczeniowych, analitycznych, dokonuje się w obszarze predykcji. Teraz wszystkie te właśnie analizy *big data*, czyli to, co nazywa się *data science*, to jest potężny zwrot do indukcji, przy czym w sposób świadomy rezygnuje się z nadawania tej indukcji statusu poznawczego, poszukiwania jakiegoś przyczynowego mechanizmu. Wystarczające jest to, żeby dokonać indukcji reguł, które pojawiają się z wystarczającym prawdopodobieństwem do uzyskania określonego zwrotu na inwestycji. Koniec końców to jest istotne. Jeżeli coś nam pozwala skuteczniej o kilka procent przewidzieć, to przy wielokrotnym powtarzaniu się określonego efektu zbieramy śmietanę z mleka. Nie byłoby w tym nic dziwnego, gdyby nie to, że takie postępowanie przynosi potężny sukces w bardzo różnych dziedzinach, bo przecież sztuczna inteligencja też się opiera, *de facto*, na indukcji reguł.

AB: Czyli zastosowanie uczenia maszynowego.

JG: Tak, sieci neuronalne, uczenie maszynowe. Musimy się nauczyć wykorzystywać to, co nam przynoszą uzyskane z ich pomocą ustalenia. Natomiast, jeżeli się otworzymy na ten pełniejszy cykl, w którym mamy do czynienia, *de facto*, z eksploracją, to musimy się nauczyć wyprowadzać z tego wnioski dotyczące potencjalnych modeli przyczynowych.

PW: No właśnie, bo tamte reguły oparte na korelacjach potrafią jedynie przewidywać, natomiast nie potrafią podać mechanizmu przyczynowego.

JG: Nie jest to kwestią tego, czy potrafią, czy nie potrafią, bo one tego w ogóle nie testują.

AB: Są stworzone w innym celu.

PW: No, ale tego się nawet nie da. Sieć neuronalna, nauczona pewnego reagowania, nie zdradza tego algorytmu i programista też nie jest w stanie go wydobyć.

JG: Tak, ale próbuje się też zautomatyzować generowanie modeli przyczynowych, co jest związane z sieciami bayesowskimi i szerzej z próbą identyfikowania w sposób bardziej zautomatyzowany tego, jak należy zapewnić trafność wewnętrzną analizy przyczynowej dzięki odpowiedniemu pozamykaniu „tylnych drzwi", przez które przepływają informacje, mogące zakłócać obraz zależności przyczynowej. Przykładowo, jest dostępne bezpłatne narzędzie DAGitty, dzięki któremu możemy z sieci potencjalnych zależności wyekstrahować te punkty, które musimy skontrolować i tak zaplanować badanie, aby w jego wyniku móc stwierdzić, czy czynnik A rzeczywiście wpływa na czynnik B przy obecności innych czynników. Dodatkowo charakterystyczną właściwością tego rodzaju modeli, modeli opartych na grafach, jest to, że są to modele nieparametryczne. Czyli

nie interesuje nas wartość parametrów, ale interesuje nas, czy w ogóle występuje mechanizm przyczynowy.

AB: Natomiast w szerszej perspektywie można stwierdzić, że współcześnie w pewnym sensie dochodzi do realizacji tego, co chciał Karl Pearson, czyli zgromadzenia niemal zupełnych danych. Fisher wprowadził wnioskowanie statystyczne, gdyż uznał, że nie jesteśmy w stanie zgromadzić zupełnych danych, więc musimy pobierać próbki i na ich podstawie wnioskować o populacji. Jeśli zatem współcześnie, w dobie *big data*, jesteśmy w stanie zgromadzić pełne dane, to – być może – wnioskowanie statystyczne jest w ogóle niepotrzebne?

JG: Nie, to są analizy populacyjne, ale to tylko zdejmuje kwestię inferencji, to znaczy wnioskowanie statystyczne z prób na populacje, ważna gałąź statystyki traci tutaj trochę fundament. Aczkolwiek wykorzystuje się techniki opracowane na gruncie wnioskowania statystycznego w badaniu stabilności modeli. Zawsze można powiedzieć, że nawet jeżeli badamy populację, to moment, w którym dokonujemy pomiaru, jest jednym, losowym momentem z uniwersum stanów, w którym się mogą znajdować wszystkie obiekty danego uniwersum. Także tu może występować pewna zmienność losowa. Możemy też chcieć wyciągać wnioski dotyczące precyzji naszych oczekiwań co do przyszłości.

Przyjrzyjmy się modnej *big data analysis* na danych, które często mają charakter danych niereaktywnych, czyli takich, które zostały wygenerowane bez świadomości badanych, że są poddani obserwacji, takie dane nie podlegają zakłóceniu określanemu jako ekologiczne, które w przypadku danych reaktywnych występuje. Gdy prowadzimy wywiady, to w taki, a nie inny sposób traktujemy badanych, więc zawsze występuje pewien efekt badania. Jeżeli jest to badanie kwestionariuszowe realizowane techniką wywiadu, to konsekwencją jest efekt ankieterski. W przypadku eksperymentu to może być efekt oddziaływania sytuacji badawczej itd. Zawsze te zakłócenia musimy brać pod uwagę. Różnicę pomiędzy badaniami reaktywnymi i niereaktywnymi można obrazowo zilustrować następującym przykładem: jeśli ważka nie wie, że jest obserwowana, mamy szansę zobaczyć, że ona czasami siedzi na jakimś patyku, a nie tylko ciągle lata. Bo gdyby ważka za każdym razem widziała, że ją obserwujemy, to dla własnego bezpieczeństwa zrywałaby się do lotu, a w efekcie na bazie takich obserwacji moglibyśmy nabrać przekonania, że ona zawsze pozostaje w locie i wyciągnęlibyśmy fałszywy wniosek o niesiadającym owadzie. Badania niereaktywne są ważne, ale jeśli nawet obejmiemy nimi całą populację, to wcale nie znaczy, że wygenerowane z nich parametry wiążące pewne cechy powtórzą się w kolejnych badaniach w sposób idealny.

Predykcja z pewnym satysfakcjonującym poziomem precyzji jest bardzo ważna operacyjnie, ale jest absolutnie niewystarczająca strategicznie. Strategia zajmuje się bowiem tym, co się będzie działo z badaną zbiorowością w przyszłości, w innym czasie, w być może w nieco zmienionych okolicznościach.

Po to, żeby uzyskać zdolność przewidywania i stwierdzić, co się będzie działo wtedy, kiedy zadziałamy innymi czynnikami, które nie występowały, kiedy były zbierane dane, trzeba zrozumieć, jak działają mechanizmy. I to jest miejsce dla

nauki i tego miejsca nauce nikt nie odbierze. Zawsze będzie zapotrzebowanie na praktyczne wykorzystanie nauki do podejmowania decyzji o charakterze strategicznym. Jeżeli by taka wiedza nie była dostarczana, to decyzje o charakterze strategicznym podejmowano by pewnie na podstawie intuicji. Człowiek ma zdolność do tworzenia teorii, które są trafne bądź nie, na podstawie obserwacji, wyindukowanych reguł. Ludzie dostrzegają, że pewne rzeczy się powtarzają, ale tylko sprytni ludzie zastanowią się, dlaczego, oraz zauważą, że czasami następują w nich pewne mniejsze lub większe zakłócenia i stworzą koncepcje, dlaczego się tak dzieje. Ludzie wytwarzają w naturalny sposób koncepcje teoretyczne, tworzą modele. Są w stanie podejmować decyzje, biorąc pod uwagę niewielką ilość danych. Mogą stworzyć modele dotyczące zależności, w ogóle nie mając danych empirycznych albo mając bardzo szczątkowe, jednostkowe obserwacje. Na przykład stereotypy, które są modelami dotyczącymi spodziewanych zależności, które budujemy na podstawie jednostkowych, szczątkowych i niesystematycznych, w sensie naukowym, obserwacji, albo są elementem przekazu kulturowego i mogą dotyczyć zupełnie innej sytuacji, zupełnie innych konfiguracji i nie nadawać się do przeniesienia na stan dzisiejszy. Ludzie dzięki temu, że tworzą modele, mogą podejmować decyzje, na bazie bardzo niewielkiej ilości danych.

Ta oszczędność, jeżeli chodzi o ilość danych, która jest potrzebna do tego, żeby skutecznie działać, jest bardzo pociągająca, jeżeli chodzi o wizję budowy systemów sztucznej inteligencji, więc prace nad tym nie będą ustawały.

AB: Chciałbym, abyśmy jeszcze wrócili trochę do początku naszej rozmowy. Do numeru tematycznego mamy zaplanowany artykuł dotyczący wnioskowania przyczynowego i prac Judei Pearla. Wspomniałeś na wstępie, że w socjologii takie myślenie jest już obecne, natomiast w psychologii dominuje przekonanie, że jedyną podstawą do formułowania wniosków przyczynowych są badania eksperymentalne. Wykorzystanie danych obserwacyjnych, nawet z zastosowaniem modelowania równań strukturalnych, jest niewystarczające.

JG: To życzę powodzenia w stosowaniu eksperymentów w astronomii.

AB: Właśnie, jednocześnie nie poprosimy ludzi, żeby zaczęli palić papierosy, czyli nie przeprowadzimy manipulacji eksperymentalnych.

PW: Albo nie usuniemy im schizofrenii, albo nie nadamy jej w eksperymencie.

AB: Tak, nie nadamy eksperymentalnie schizofrenii. W związku z tym jak przekonać psychologów do tego, że przy spełnieniu odpowiednich warunków istnieje możliwość wnioskowania o zależnościach przyczynowo-skutkowych na podstawie danych obserwacyjnych.

JG: Skłonić ich do uczenia się metodologii. (*śmiech*)

AB: Czyli wracamy do edukacji.

JG: Myślę, obserwując aktywność Koła Banacha[7], że jego członków nie trzeba specjalnie przekonywać. Myślę, że są już przekonani, że ważny jest nie tylko eksperyment. Tylko też trzeba uważać, bo w drugą stronę można bardzo łatwo odejść w kierunku bylejakości. Bardzo dużo w naukach społecznych jest byle jakich badań. Oczywiście, badania są robione dla różnych potrzeb, np. po to, żeby opisać rzeczywistość. I jeśli jest takie zapotrzebowanie, to taki wynik jest dostarczany. Tyle że wtedy on musi być, na tyle, na ile to możliwe, precyzyjny i pozbawiony różnego rodzaju błędów. I to, z czym mam problem przy tak stawianych zagadnieniach w naukach społecznych, w socjologii i pokrewnych obszarach, to jest niehołdowanie podstawowym zasadom, chociażby metody reprezentacyjnej, jeśli się prowadzi badania na próbach, oraz oczywiście zasadom prowadzenia pomiaru za pomocą sondażu społecznego. Zaniedbywanie różnego rodzaju czynników, które zakłócają jakość wyniku, poczynając od niedbalstwa w zakresie gromadzenia danych, czyli samej sytuacji pomiaru... Ale przede wszystkim bylejakość dotyczy sposobów doboru i realizacji próby albo przy szacowaniu wyników radzenie sobie z sytuacją, że mamy przypadki wylosowane, a niezrealizowane. Albo że mamy braki danych w obrębie zrealizowanych obserwacji. Te wszystkie czynniki wpływają na wyniki, zakłócają je.

PW: W psychologii jest więcej takich cech, które są bardziej biologicznie uwarunkowane, i nielosowo pobrana próba też ma je losowo rozłożone…

JG: Nie pocieszałbym się. Rozumiem, to ta jedność gatunkowa człowieka, tak? (*śmiech*) Jeśli zbadamy jedną osobę, to jakbyśmy zbadali wszystkich. Albo w ogóle możemy zbadać gołębia, (*śmiech*) też jest kręgowcem, też jest stałocieplny, tak?

AB: Ale studenta psychologii to już na pewno wystarczy. (*śmiech*)

JG: Tak studenta psychologii to już na pewno, ale gołębie też są bardzo popularne, to nawet później przeniknęło do socjologii, np. w porządku dziobania.

AB: Wracając do tych ograniczeń prowadzonych współcześnie badań, o których rozmawialiśmy niedawno z Piotrem i z profesorem Brzezińskim, który jest także autorem artykułu w tym numerze, to jest kwestia badań prowadzonych online z wykorzystaniem różnych przeznaczonych do tego platform. Te ograniczenia polegają na tym, że osoby badane często – mówiąc potocznie – przeklikują kwestionariusze, tj. bardzo szybko udzielają odpowiedzi, nie wczytując się w ich treść lub w ogóle nie czytając treści. Choć można to w pewnym zakresie kontrolować...

JG: Ale pod tym względem to narzędzie nie różni się od wypełniania ankiety papierowej, nawet daje większą możliwość kontroli. Uważam, że zagrożenia

[7] Działające w Instytucie Psychologii UJ koło dyskusyjne skupione wokół zagadnień metodologicznych i kwestii integralności badań naukowych.

związane z poszczególnymi technikami akwizycji danych są rozpoznane. W przypadku badań online dramatycznym problemem jest kontrolowanie próby oraz zagadnienie reprezentatywności wyników. Mam wrażenie, że przestano o to w ogóle dbać i po prostu się gromadzi dane. To tak jak w początkach badań ankietowych, kiedy rozsyłano ankiety, jak w słynnym, opisywanym w podręcznikach statystycznych, przykładzie: *Literary Digest* rozesłał 4,5 mln ankiet do swoich czytelników, którzy deklarowali na kogo zagłosują, i na tej podstawie stworzono prognozę wyborczą. Natomiast Instytut Gallupa zrobił badanie na próbie reprezentatywnej, która liczyła ok. 1,3 tys. osób. I okazało, że ta prognoza trafnie przewidziała, że wygra Roosevelt, a *Literary Digest* na podstawie 4,5 mln zebranych ankiet przesadził o ponad 10% na rzecz jego przeciwnika. Samo zwiększenie liczebności próby nie zmniejsza błędu systematycznego, zmniejsza jedynie błąd losowy, w sposób nieliniowy.

PW: W psychologii większość badań jest na małych próbach, więc jak jakieś jest na dużej próbie, to od razu nam się wydaje, że to jest fantastyczna rzecz, można wierzyć wynikom, bo jest przecież duża próba. No, ale im jest większa, tym może być bardziej tendencyjna, niestety. Prawda?

JG: To zależy, ale krótko mówiąc, samo zwiększanie liczebności próby nie ma wpływu na błąd systematyczny, bo o błędzie systematycznym decydują czynniki kontekstowe, które ten błąd wywołują. Jeżeli one się powielają, to zwiększanie liczebności próby niczego tu nie zmienia. Wraz ze zwiększaniem liczebności próby rośnie udział błędu systematycznego, a maleje udział błędu losowego w całkowitym błędzie badania.

PW: W psychologii kluczową rzeczą jest ta nieśmiertelna wartość p, a w dużych próbach mamy dużo efektów, w których p przekracza przyjęty próg.

JG: I właśnie dlatego wam się badania nie replikują. Niekiedy trzeba byłoby założyć od razu kilka replikacji w różnych kontekstach. Może się okazać, że w określony sposób zmieniają się parametry, co może być źródłem bardzo interesującej hipotezy na temat znaczenia pewnego czynnika kontekstowego dla ogólnego wyniku badań, dla wartości parametru. Jeżeli by się okazało, że w replikacjach utrzymuje się wzór relacji, a zmieniają się parametry i pozostają w jakiejś systematycznej zależności od pewnej cechy, która różnicuje ludzi w szerszej zbiorowości, wówczas może to prowadzić do skonstruowaniu teorii mechanizmu przyczynowego oraz jego interakcji z określonym czynnikiem kontekstowym. Nie twierdzę, że respondenci muszą być dobierani losowo do badań psychologicznych. Oni mogą być dobierani celowo. Powiedziałbym nawet, że mogą być dobierani homogenicznie, ale wtedy powinno być kilka sytuacji eksperymentalnych z homogenicznymi, ale różniącymi się między sobą grupami. Wtedy możliwe stanie się zbadanie, czy pewna cecha, ze względu na którą dokonamy tego celowego wyboru, nie będzie cechą, która powoduje pewną zmienność. Wspomniany tu profesor Jerzy Brzeziński jest niezmordowanym adwokatem trafności zewnętrznej badań…

AB: Czyli wracamy do problemu generalizowalności, która w psychologii jest niejawnie przyjmowana i pomimo że badamy właśnie studentów psychologii, to uogólniamy na ludzi w ogóle.

JG: Albo nawet uogólniamy badania gołębi... Homans w teorii wymiany zaczerpnął te gołębie z psychologii.

PW: Pewnie od Skinnera.

JG: Na dzisiaj kończymy, bo już mamy następne przedsięwzięcia...

Bibliografia

Bollen, K., Pearl, J. (2013). Eight Myths About Causality and Structural Models. W: S. L. Morgan (red.), *Handbook of Causal Analysis for Social Research* (s. 301–328). Springer.

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

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

Lakens, D. (2021). The Practical Alternative to the *p* Value Is the Correctly Used *p* Value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Woolston, C. (2015). Psychology journal bans P values. *Nature*, *519*, 9. https://doi.org/10.1038/519009f