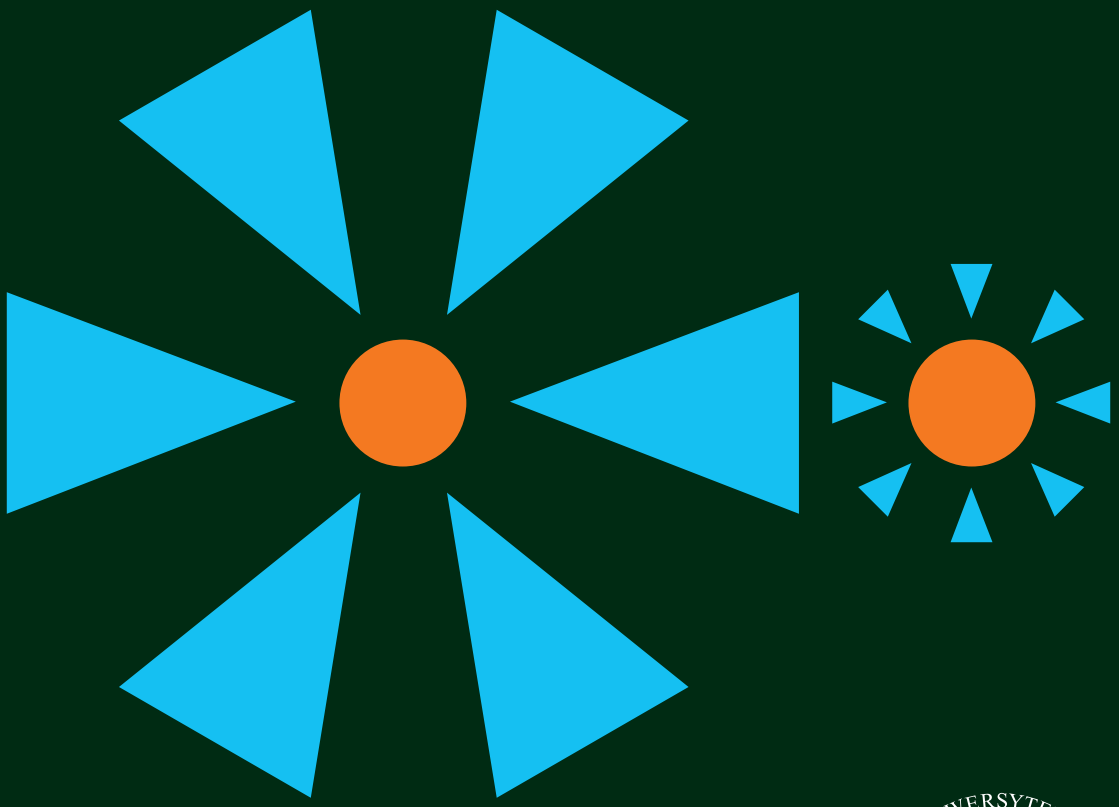
Grażyna WiW [Wieczorkowska-Wierzbińska]

Rafy metodologiczne

w naukach o zarządzaniu ludźmi



SWWZ
Sekcja Wydawnicza
Wydziału Zarządzania
Uniwersytetu Warszawskiego

Wydawnictwa Uniwersytetu Warszawskiego
WUW

Grażyna WiW
[Wieczorkowska-Wierzbińska]

Rafy metodologiczne w naukach o zarządzaniu ludźmi

SWWZ
Sekcja Wydawnicza
Wydziału Zarządzania
Uniwersytetu Warszawskiego

Warszawa 2025

WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

Grażyna WiW [Wieczorkowska-Wierzbińska], Wydział Zarządzania,
Uniwersytet Warszawski, Polska
https://orcid.org/0000-0002-3307-1679

Publikacja dofinansowana z subwencji na utrzymanie i rozwój
potencjału badawczego na Wydziale Zarządzania Uniwersytetu Warszawskiego

Recenzenci:
prof. dr hab. Wiktor Adamus
prof. dr hab. Irena Heszen-Celińska

Redakcja:
Agnieszka B. Góralczyk

Projekt okładki:
Agnieszka Miłaszewicz

ISBN pdf 978-83-235-7063-9
DOI: 10.7172/978-83-235-7063-9.swwz.29

DOM WYDAWNICZY ELIPSA

Opracowanie komputerowe:
Dom Wydawniczy ELIPSA
ul. Inflancka 15/198, 00-189 Warszawa
tel. 22 635 03 01
e-mail: elipsa@elipsa.pl, www.elipsa.pl

Książkę dedykuję współautorom moich tekstów metodologicznych:

mojemu mężowi prof. Jerzemu Wierzbińskiemu,
mojej córce Małgorzacie Siarkiewicz
oraz wypromowanym przeze mnie doktorom (w kolejności chronologicznej):
Grzegorzowi Królowi, Bartkowi Michałowiczowi, Annie Kuźminskiej,
Katarzynie Kowalczyk, Wojtkowi Karczewskiemu i Kindze Wilczyńskiej

Spis treści

Wprowadzenie

Nauki o zarządzaniu wypracowały swoją metodologię, nie różnicując jej ze względu na to, czy dotyczy ona zasobów rzeczowych czy ludzkich. Główne przesłanie tej monografii mówi, **że nie warto bezmyślnie kopiować procedur badania świata przedmiotów w badaniu ludzi.**

Rozwój zarządzania ludźmi jako dyscypliny naukowej jest znacznie spowolniony przez bezrefleksyjne naśladowanie metodologii nauk ścisłych. Biolodzy dość szybko uporali się z przygotowaniem szczepionki przeciwko COVID-19. Tysiące badań społecznych dotyczących COVID zostało sfinansowanych i opublikowanych – ale czy rozwiązały one realny problem postaw antyszczepionkowych? W tym samym czasie nie potrafimy zatrzymać narastającej polaryzacji i fali nienawiści. Produkujemy ogromną liczbę publikacji – jednak nadal brak przełomu. Nauki społeczne nie radzą sobie z polaryzacją społeczną, usieciowioną przemądrzałością i przekonaniem, że każdy – niezależnie od poziomu wiedzy – ma prawo formułować opinie na dowolny temat. Co gorsza, panoszy się coraz śmielej przeświadczenie, że to ekspert ma obowiązek przekonać oponenta, mimo że ten uczestniczy w dyskusji bez posiadania choćby minimalnego poziomu wiedzy na temat, którego dotyczą jego sądy.

Obiekt naszych badań jest dużo bardziej skomplikowany i reaktywny niż np. w naukach biologicznych czy fizycznych, na których wzorujemy swoją metodologię. Taleb[1] porównuje nauki społeczne usiłujące stosować metodologię wykorzystywaną przez fizyków do krów, które starają się latać.

Analizy statystyczne w badaniu śrubokrętów i ludzi są identyczne, i choć w obu przypadkach można wylosować próby, to jedynie śrubokręty udaje się mierzyć bez stawiania oporu. Nasze osoby badane są zdolne nie tylko do odmowy uczestniczenia w badaniu, ale – co zdarza się często – mogą starać się wpłynąć na wyniki pomiaru. W przeciwieństwie do obiektów badań fizyków czy chemików są też niestety obdarzone pamięcią.

---

[1] Taleb, 2012.

Zadawanie tego samego pytania wielokrotnie stanowi źródło problemów interpretacyjnych, ponieważ kolejne pomiary nie są od siebie niezależne. Z tego powodu przeważająca większość badań eksperymentalnych i korelacyjnych koncentruje się na pojedynczym punkcie czasowym. **Uniemożliwia to uchwycenie dynamiki zjawisk**. Nawet, jeśli wykażemy pozytywny wpływ nagrody na zaangażowanie pracownika, nie będziemy wiedzieć co zdarzy się kilka godzin później, kiedy osoby badane już opuszczą laboratorium. Jeżeli nawet warunki środowiskowe pozostaną bez zmian, **samo powtórzenie działania bodźca** może zmienić reakcję badanego.

Próbując złamać śrubokręt, możemy odnieść sukces albo porażkę. Próbując „złamać psychologicznie" pracownika, odniesiemy sukces (o wątpliwej wartości etycznej), porażkę albo wywołamy efekt paradoksalny (przeciwny do naszej intencji – próba złamania zaowocuje wzmocnieniem).

Aby zilustrować to studentom na wykładzie z medycyny psychosomatycznej – przeprowadziłam **„trening relaksacyjny"**. Prawie 300 osób siedzących na sali było proszonych o podążanie za instrukcjami podawanymi w tle relaksującej muzyki. Nie tylko obserwowałam reakcje, ale także poprosiłam ich o ocenę swojego stanu za pomocą wyboru jednej z 6 odpowiedzi. Oprócz opcji oznaczających całkowity lub częściowy sukces w wykonaniu tego zadania – zgodnie z moimi oczekiwaniami – 8% wybrało odpowiedź „**nie miałem ochoty się relaksować, więc nie próbowałem"**, a 9% wykazało efekt kontrastu, stwierdzając: „**próba relaksacji wywołała moją irytację**". Wariancja odpowiedzi umysłu jest dużo większa niż wariancja biologicznej odpowiedzi organizmu (podanie leków usypiających nie wywołałoby efektu kontrastu). To oznacza, że modele nieuwzględniające różnic indywidualnych i kontekstu sytuacyjnego nie mogą wyjaśniać wielu wariancji analizowanych zmiennych. Koncentrując się w naszych analizach statystycznych na miarach tendencji centralnej, nie dostrzegamy często efektów kontrastu.

Powtarzając – obiektami naszych badań są organizmy złożone, które w odróżnieniu od obiektów nieożywionych charakteryzują się skomplikowanymi sprzężeniami zwrotnymi.

Czynniki stresowe (przeciążenie organizmu, kryzys ekonomiczny) mogą mieć nie tylko efekty negatywne (upadek firmy, śmierć), ale także pozytywne (wzrost posttraumatyczny). Ta cecha organizmów żywych (a także organizacji) została nazwana przez Taleba **antykruchością**. **Żaden złamany śrubokręt sam się nie naprawi**, ale ludzie potrafią to zrobić. W literaturze przedmiotu[2] został opisany przykład kobiety, której z powodu ciężkich napadów padaczkowych generowanych w lewej półkuli odpowiadającej za funkcje językowe – usunięto połowę kory mózgowej. Przez pierwszy rok po operacji ledwo była w stanie cokolwiek powiedzieć. Sześć lat później odzyskała już znaczną część funkcji językowych, co pokazuje niesamowite zdolności regeneracyjne mózgu. Jej mózg przeorganizował się tak, aby umożliwić jej mówienie i czytanie.

[2] Poldrack, 2022.

Jeśli porównamy ludzki mózg z czymś bardziej złożonym niż śrubokręt – np. z komputerem – zauważymy, że **oprogramowanie komputera (software) jest zasadniczo oddzielone od sprzętu (hardware)**. Na tym samym komputerze można zainstalować różne programy, natomiast w mózgu sprzęt i oprogramowanie są nierozdzielne – „program" jest przechowywany w połączeniach między neuronami. Można **wymienić uszkodzoną część komputera** (np. kartę sieciową), natomiast w przypadku mózgu ze względu na liczne połączenia całościowe jest to niemożliwe. Gdyby nasze mózgi pracowały jak komputery: pedantycznie i wydajnie pozbawiłyby się zdolności dynamicznego kojarzenia i łączenia informacji. Dlatego tak ważne jest, aby pamiętać, że wszelkie oddziaływania psychologiczne charakteryzują się **ekwifinalnością** (**ten sam rezultat** można osiągnąć za pomocą **różnych oddziaływań**) i **ekwipotencjalnością (to samo oddziaływanie** może prowadzić do bardzo **różnych rezultatów**).

Zarządzanie przedmiotami daje zazwyczaj przewidywalne wyniki – nie przesuniemy piramidy, jeśli nie mamy wystarczających narzędzi. Tłum ludzi możemy przesunąć o wiele łatwiej, jeśli ten tłum uzna nasze żądanie za uzasadnione. W przeciwnym wypadku będziemy musieli użyć przymusu bezpośredniego, czyli dysponować odpowiednimi narzędziami do wywarcia presji.

Zarządzanie ludźmi możemy analizować na trzech poziomach wywierania wpływu na: (1) zachowania, (2) przekonania, (3) emocje. Poziom pierwszy wydaje się najłatwiejszy, bo ludzi można zastraszyć lub przekupić, aby zachowali się zgodnie z instrukcją, ale oczywiście nigdy nie możemy być pewni, że nasze oddziaływanie zakończy się sukcesem. Są tacy, którzy wolą umrzeć niż podporządkować się „terrorystom". Zmiana przekonań i emocji innych jest jeszcze trudniejsza. Co więcej, mamy takie same problemy (jeśli nie większe) z zarządzaniem sobą. Próba modyfikacji własnych zachowań, przekonań, emocji jest bardzo trudna[3].

Czy wyniki badań ludzi w XX wieku mają zastosowanie do ludzi żyjących w XXI wieku?

Pewne podstawowe potrzeby psychologiczne (np. potrzeba uznania, dotyku...) nie zmieniły się, ale sieci neuronowe tworzone z największą mocą w dwóch oknach przyspieszonej socjalizacji (we wczesnym dzieciństwie oraz przez ok. 10 lat po osiągnięciu dojrzałości płciowej) wpływają na wszystko, czego doznajemy (np. nie usłyszymy zniewagi ani komplementu wypowiedzianego w języku, którego nie znamy). W poprzedniej monografii[4] *„Rafy psychologiczne w zarządzaniu. Ludzie różnią się od śrubokrętów"*, do której odsyłam zainteresowanych, tłumaczyłam dlaczego sieci neuronowe generacji internetowych różnią się znacząco od sieci w głowach gene-

[3] Wieczorkowska, 2011; Aronson i Wieczorkowska, 2001.

[4] Wieczorkowska, 2025.

racji starszych. Świat zmienia się dużo szybciej niż wartości prototypowe wzorców pojęć, które mają bardzo wysoki potencjał aktywizacyjny i status niedyskutowalnych oczywistości. Przykładowo słowo „**masakra**" jest zrozumiałe (ma to samo znaczenie denotacyjne) zarówno dla Boomerów, jak i Millenialsów, ale ma zupełnie inne znaczenie konotacyjne w tych pokoleniach. Dla mnie „masakra" to nacechowane **silnie negatywnie** zjawisko kojarzące się z bestialstwem i brutalnością. Moje dzieci słowem *masakra* komentują trudną, nieprzyjemną sytuację. Słowo to jest klasycznym przykładem generacyjnej zmiany znaczenia konotacyjnego wyrazów[5].

Analizując dane z badań, trzeba pamiętać o zmianie trendów kulturowych[6] w cywilizacji zachodniej, które przejawiają się **w uwalnianiu** od więzi rodzinnych **i poczucia zobowiązań wobec rodziny czy ojczyzny.** Socjalizacja kulturowa w sieci powoduje słabnięcie więzi z miejscem zamieszkania i tradycjami. Zmniejsza się też poczucie solidarności pokoleniowej. Coraz więcej osób wybiera „bezdzietność", nie chcąc rezygnować z wygodnego życia bez zobowiązań. W Internecie możemy znaleźć poparcie dla dowolnych indywidualnych przekonań. Pozytywnym efektem zmian społecznych jest to, że dzięki większej mobilności społecznej status społeczny może zależeć bardziej od osobistych osiągnięć niż od pochodzenia. Zawód nie jest dziedziczony po rodzicach tak często, jak to bywało w poprzednich stuleciach.

Żyjemy w czasach **przełomu technologicznego**. Technologia wyzwoliła nas z wielu ograniczeń, wprowadzając takie rozwiązania, jak np. kuchenki mikrofalowe, pralki automatyczne, zmywarki, roboty kuchenne, drukarki 3D, Internet, smartfony z najprzeróżniejszymi aplikacjami w rodzaju nawigacji GPS, komunikatorów internetowych, pomiarów w telemedycynie... Zmiany technologiczne powodują zmiany w nas samych, których nie dostrzegamy, ponieważ procesowi zrywania z przeszłością NIE towarzyszą żadne spektakularne wydarzenia[7]. **Stopniowo zachodzą w nas zmiany,** a więc możemy mówić o „efekcie gotowanej żaby". Żaba może zginąć, gdy temperatura w płytkim akwarium będzie podgrzewana bardzo powoli, podczas gdy wrzucenie jej do gorącej wody spowodowałoby, że natychmiast by wyskoczyła. Nasze środowisko informacyjne gęstnieje systematycznie, ale powoli, więc nie wyskakujemy.

Żyjemy w świecie **gwałtownej eksplozji informacyjnej**. W erze Internetu tempo akumulacji wiedzy rośnie wykładniczo, lecz **nasze mózgi biologicznie pozostały niezmienione**[8]. Niezależnie od teoretycznych możliwości obliczeniowych mózgu, w sensie praktycznym okazują się one ograniczone. Wiedza staje się coraz bardziej rozproszona i wyspecjalizowana, dlatego zdolność do integrowania różnorodnej wiedzy jest zdecydowaną przewagą konkurencyjną dla tych, którzy potrafią to zrobić.

---

[5] https://nck.pl/projekty-kulturalne/projekty/ojczysty-dodaj-do-ulubionych/ciekawostki--jezykowe/masakra

[6] Bokszański, 2007.

[7] Marody, 2015.

[8] Goldberg, 2024.

Żyjemy w **zanieczyszczonym informacyjnie środowisku** określanym informacyjnym SMOGIEM. Aby w otwartej internetowej konkurencji poglądów wygrały najlepsze, odbiorcy muszą być zdolni oddzielić ziarno od plew. Niestety **przeciążeni sygnałowo odbiorcy** nie są w stanie tego zrobić, więc chętnie posługują się powierzchownymi wyznacznikami jakości, takimi jak atrakcyjność przekazu, kwalifikacje (przy gwałtownym wzroście liczby uczelni jakość certyfikatów spada) i zdolności retoryczne nadawców[9]. Najczęściej **wygrywa to, co łatwe i przyjemne.**

Sieć stała się naszym preferowanym miejscem spędzania czasu. Mamy dostęp do olbrzymiej liczby informacji – nieporównywalnej z tym, co kiedykolwiek było udostępniane wcześniej. Możemy mówić o **usieciowionej żarłoczności** prowadzącej do mentalnej **ociężałości** i **otyłości,** bo ciągle **chcemy więcej, nie trawiąc starannie tego, co już pochłonęliśmy**. Poszukiwanie w sieci powoduje w naszej krwi wzrost wywołujących specyficzny błogostan neurotransmiterów[10]. Nienasycenie przejawia się w nieustającym skanowaniu otoczenia w poszukiwaniu kolejnej podniety. **Świat nadmiaru** zachęca do ciągłego poszukiwania w nadziei, że **to, co nowe będzie lepsze od tego, co już znamy**. Ten nadmiar może sprawić, że zagubimy się w alternatywnych światach. Odruch szukania w wyszukiwarkach odpowiedzi na pytanie prowadzi do **poznawczego rozleniwienia**. Pisanie i czytanie multiplikowało nasze zmysłowe doświadczenie w innym wymiarze, pozwalając na **eksternalizację doznań i myśli.**

Wraz z przyzwyczajaniem naszych mózgów do przetwarzania **krótkich strzępków informacji** zanika u nas **zdolność angażowania się w procesy myślowe wymagające wyjścia poza powierzchowność**, niezbędne do pełnego zrozumienia trudnych materiałów pisemnych oraz uczestniczenia w długich rozmowach. Przeprowadzone badanie[11] 100 milionów nagłówków artykułów opublikowanych w 2017 roku wykazało, że największym powodzeniem cieszą się te, które **obiecują doznania natury emocjonalnej np.** „sprawi, że:", „pęknie ci serce", „się zakochasz", „do tego wrócisz", „się zdziwisz", „doprowadzi do łez", „wywoła dreszcz" i „serce zmięknie". Choć 76% Amerykanów[12] twierdzi, że oglądało, czytało lub słyszało wiadomości, jedynie 41% przyznaje, że wyszło poza same nagłówki. Prowadzi to do zjawiska **usieciowionej przemądrzałości** – myślenia, że wie się więcej niż jest to w rzeczywistości.

Algorytmy portali internetowych (np. te sugerujące filmy, wydarzenia czy książki, które mogą nam się spodobać) podejmują decyzje za nas w obliczu ogromnej liczby ofert, których nie jesteśmy w stanie przejrzeć samodzielnie. Uleganie ich wpływowi oddala nas od ludzi o innych poglądach.

Od listopada 2020 roku mamy do czynienia z nową jakością algorytmów. GPT (*Generative Pre-trained Transformer*) i inne duże modele językowe (LLM, czyli *Large Language Models*) zostały zaprojektowane do imitowania inteligentnego pisania.

[9] Pigliucci, 2019.
[10] Hallowell i Ratey, 2005.
[11] Cialdini, 2023.
[12] Zimbardo i Coulombe, 2015.

Okazało się jednak, że sieci neuronowe potrafią rozwiązywać zadania, których nie zostały pierwotnie nauczone. Nadal trudno przewidzieć, w jakim kierunku rozwinie się generatywna AI.

Podsumowując – trudność prowadzenia badań dotyczących zarządzania ludźmi wynika z faktu, że:

- mózgi badaczy badają inne mózgi (respondentów);
- ludzi można wylosować, ale nie można ich zmusić do udziału (choć fałszywi respondenci udają, że odpowiadają na pytania);
- ludzie nie reagują na bodźce, ale na ich własną interpretację, która zależy od wcześniejszych doświadczeń zapisanych w ich sieci neuronowej;
- zróżnicowanie reakcji ludzkich jest ogromne – większe niż reakcji biologicznych. Psychologiczne oddziaływania, np. mające na celu uspokojenie, mogą zakończyć się efektem asymilacji (uspokoimy się) lub kontrastu (poczujemy się zirytowani). Standardowe analizy tego nie wykrywają. Uspokojenie za pomocą oddziaływania farmakologicznego daje dużo mniejszą wariancję reakcji – wszyscy zasną prędzej czy później.

***

Wiek XXI to triumf badań naukowych, które pozwoliły dokonać wielkiego skoku w technice, informatyce, a także medycynie. Przeciętny Kowalski korzysta z tych dobrodziejstw, nie rozumiejąc czym jest sieć 5G, komputer kwantowy czy sztuczna inteligencja.

Ufamy wynikom badań naukowych i to zaufanie rozlewa się na wszystkie dziedziny nauki. A z powodów, które wymieniłam wcześniej – gdy obiektem badań są ludzie – wyniki są bardzo wysoko kontekstowe (tzn. że poziom generalizacji stwierdzanych zależności między zmiennymi jest bardzo ograniczony).

Gdy słyszymy o wynikach badań naukowych dotyczących ludzi, powinniśmy od razu zacząć zadawać pytania np.: Czy było to badanie eksperymentalne, czy jedynie badanie korelacyjne dostarczające często korelacji pozornych? Kto uczestniczył w badaniu? Na jaką populację mogą być generalizowane wyniki zebrane od tej próby? Czy badanie było replikowane na innych próbach, innymi metodami? Który rodzaj opisywanej w tej monografii triangulacji zastosowano? Najlepszym sposobem na ocenę jakości badania jest dokładne przyjrzenie się wykorzystywanym w nim procedurom pomiaru/manipulacji wartościami zmiennych niezależnych, doboru próby i wykonywanym analizom.

Wiedza o ludziach to podstawowe narzędzie w pracy menedżerów**. Bezkrytyczne wykorzystywanie wyników „badań naukowych**" może być poważną rafą w ich pracy.

Rafy metodologiczne, na które wszyscy napotykają przy badaniu postaw, emocji, zachowań ludzi, omówię w trzech klasach sytuacji:

- Klasa 1 to szeroko pojęte BADANIA, do udziału w których zapraszamy pracowników lub sami uczestniczymy. Nasze odpowiedzi są wtedy łączone z innymi i giną we wskaźnikach agregatowych.
- Klasa 2 to POMIAR cech np. pracownika, którego podstawą są odpowiedzi na pytania na skalach szacunkowych.
- Klasa 3 to EWALUACJA, czyli ocenianie innych ludzi lub produktów przez nich wytworzonych.

W sekcji 4. odpowiadam na pytanie, w jakim kierunku zmierzają nauki o zarządzaniu ludźmi, aby zakończyć w sekcji 5. przedstawieniem paradygmatu metodologicznego WIW dla badań prowadzonych w obszarze ZZL (HRM). Został on opracowany na potrzeby rozpraw doktorskich przygotowywanych pod moim kierunkiem i oczywiście uwzględnia fakt, że „badani ludzie różnią się od śrubokrętów".

W ostatniej sekcji zamieściłam vademecum dla badaczy opinii, w którym omówiłam (a czasem tylko zasygnalizowałam) wybrane przeze mnie pułapki metodologiczne zebrane przez 45 lat mojego doświadczenia badawczego, ilustrowane przykładami z literatury przedmiotu. To vademecum nie rości sobie pretensji do bycia wyczerpującym, ale ma pobudzać do refleksji.

****

Książka ta jest **próbą uporządkowania mojej wiedzy metodologicznej**, tak aby mogła stanowić pomoc dla innych starających się oceniać liczby zbierane w procesie zarządzania ludźmi. **Książka nie jest przewodnikiem po literaturze przedmiotu** – jestem bardziej „syntetyzatorem" niż „sprawozdawcą". Nie opisuję sporów teoretycznych. Przedstawienie wszystkich stron danego zagadnienia staje się bezpieczne, choć zniechęcające, gdy celem jest syntetyzowanie wiedzy. Jak to kiedyś ładnie ujął Aronson[13]: „**Czytelnicy pytają nas, która jest godzina**, a my pokazujemy im mapę różnych stref czasowych na całym świecie, przedstawiamy historię pomiaru czasu – od wynalezienia zegara słonecznego po supernowoczesne urządzenia pomiarowe, a na końcu szczegółowo opisujemy budowę elektronicznego zegarka." Słuchacze, którzy nie są ekspertami, zazwyczaj tracą zainteresowanie podczas takich prezentacji. W dobie lawinowego przyrostu badań i publikacji przeciążenie informacyjne dotyczy także ekspertów.

Mój proces twórczy przy pisaniu tej książki można spuentować podobnie jak to zrobił ostatnio Damasio[14]. Postanawiałam sobie zawsze to samo: pisać wyłącznie o tym, co uważam za istotne zarówno dla zarządzających, jak i zarządzanych, pomi-

[13] Aronson i Wieczorkowska, 2001.

[14] Damasio, 2022.

jając szczegóły nieistotne z punktu widzenia praktycznego zastosowania, czyli **podążając śladem dobrych rzeźbiarzy – odrzucić to, co zbędne**.

Aby kumulować wiedzę w monografii, wykorzystałam w niej fragmenty publikowanych wcześniej tekstów, poddając je restrukturyzacji, ociosaniu i uzupełnieniom. To zadanie wymagało częstego i ponownego przemyślenia wielu kwestii. Zmiany nie wynikają ze zmiany wiedzy metodologicznej (niczego opublikowanego wcześniej nie musiałam odwoływać☺), choć moje doświadczenie badawcze rośnie z roku na rok, a także pojawienie się Gen AI ułatwia naszą pracę naukową. Przede wszystkim zmieniły się możliwości przetwarzania informacji przez Czytelnika żyjącego w przeładowanym i zaśmieconym informacyjnie świecie.

Wybór tego, o czym piszę (a co pomijam), wynikał z chęci **integracji wiedzy metodologicznej tak, aby była łatwa do zrozumienia i stosowania przez konsumentów wyników badań empirycznych**. Zalew badań i publikacji sprawia, że synteza wiedzy staje się coraz trudniejszym wyzwaniem[15]. Trzeba umieć oddzielać ziarna od plew.

Nie jestem pisarzem. **Większe znaczenie przywiązuję do myśli zawartej w zdaniu niż do perfekcji językowej.** Dlatego, choć jest to niezgodne z regułami języka polskiego, **nie unikam powtarzania tego samego słowa w kolejnym zdaniu, ponieważ unikanie synonimów** znacząco zwiększa przejrzystość i jednoznaczność tekstu. **Przytoczone badania są jedynie ilustracjami toku wywodu** lub zapalnikami skłaniającymi do refleksji. Jak to bywa przy „lepieniu", pewne elementy badań lub myśli cytowanych autorów mogły zostać przeze mnie świadomie lub mimowolnie zniekształcone.

Wiem, że czytanie książki będzie wymagało od Czytelników przyswojenia nowych terminów. Starałam się usuwać z tekstu informacje zbędnie obciążające Czytelnika, takie jak nazwiska badaczy (są one umieszczone w przypisach dolnych i w bibliografii). W czasach zalewu informacyjnego zachęcam do przerw w czytaniu i **rozwiązywania znajdujących się na końcu książki krzyżówek**. Wszystkie hasła wykorzystane w krzyżówkach można znaleźć w słowniczku na końcu książki. Ułatwić lekturę mają puste linie oddzielające „kąski" informacyjne oraz wyróżnienie pojęć za pomocą wytłuszczeń i kapitalików.

[15] Nęcka, 2005; Śpiewak, 2013.

Podziękowania

Dziękuję wszystkim osobom, które uczestniczyły w naszych badaniach, wypełniając internetowy Sondaż Stylów Aktywności, oraz moim współpracownikom, którzy prowadzili ze mną badania i dyskusje naukowe. W tej grupie należy wymienić 21 wypromowanych przeze mnie doktorów (6 doktorów psychologii i 15 doktorów nauk o zarządzaniu). Są to w kolejności chronologicznej według czasu obrony: (1) Rafał Stefański, (2) Joanna Czarnota-Bojarska, (3) Grzegorz Król, (4) Dorota Król, (5) Izabella Bednarczyk, (6) Marcin Skład, (7) Marcin Jeśka, (8) Bartłomiej Michałowicz, (9) Anna Kuźmińska, (10) Katarzyna Kowalczyk, (11) Krzysztof Nowak, (12) Wojciech Karczewski, (13) Kamila Pietrzak, (14) Anna Koval, (15) Afsaneh Yousefpour, (16) Magdalena Sieradzka, (17) Daniel Pazura, (18) Marta Kabut, (19) Kinga Wilczyńska, (20) Dominik Schulze, (21) Magdalena Simkowska-Gawron.

Dziękuję też mojej najbliższej rodzinie – mężowi, córce i synowi – za nieustającą akceptację odmienności moich potrzeb i zachowań poznawczych.

Bardzo dziękuję recenzentom prof. Irenie Heszen-Celińskiej i prof. Wiktorowi Adamusowi za poświęcony czas, rozbudowane recenzje i doskonałe sugestie dotyczące rozdzielenia treści.

Kłaniam się z wdzięcznością redaktorom – paniom Agnieszce Góralczyk i Anicie Sosnowskiej za wsparcie, jakiego mi udzieliły na ostatniej prostej.

Bardzo dziękuję Wioli Kochańskiej – mojemu dobremu duszkowi – która z dużą pomocą Marioli Majchrak pomaga mi w pracy.

Grażyna WiW
Antałówka, Ustka, Muskat, Kabaty

.

Sekcja 1
Rafy metodologiczne w badaniach ludzi

W opublikowanym w 2016 roku tekście[16] opisaliśmy 4 kluczowe zagrożenia metodologiczne, które pojawiają się w badaniach ilościowych w naukach o zarządzaniu: (1) zbyt małą liczbę eksperymentów; (2) słabość modeli teoretycznych; (3) słabość pomiaru; (4) słabość analiz.

Dziewięć lat później tekst pozostaje aktualny i warto ciągle popularyzować jego treść, choć w międzyczasie pojawiły się dwa nowe zagrożenia, które dołączam do opisywanych w 2016 roku:

(1) bezmyślna popularyzacja wyników badań naukowych;
(2) problem fałszywych respondentów.

Bezmyślna popularyzacja wyników badań naukowych

Praktyka zarządzania wymaga wsparcia badań naukowych, dlatego ich liczba systematycznie rośnie. Jednak co z absorpcją wyników tych badań? Pojawia się coraz więcej stron internetowych i książek poświęconych zarządzaniu, ale ich rosnąca liczba nie zawsze przekłada się na skuteczniejsze zastosowanie w praktyce. Żyjemy w epoce nadmiaru informacji, w której motorem naszych działań stają się hasła „WIĘCEJ" i „SZYBCIEJ". Chętnie korzystamy ze skrótów myślowych i poszukujemy prostych, łatwych do wdrożenia modeli czy metod. Taka ścieżka płytkiego myślenia nie stanowi jednak solidnej podstawy do osiągnięcia długoterminowego sukcesu zawodowego.

Trwały stan rozproszonej uwagi i płytkiego myślenia[17] przejawiają się w nadmiernym uleganiu wpływom bodźców przyciągających wzrok, skłonności do kategorycznych opinii, nadmiernych uogólnień oraz skracania czasu poświęcanego na skupianie się na bodźcu. Zamiast czytać (nawet nie wspominając o przemyśleniu)

[16] Wieczorkowska i in., 2016.

[17] Carr, 2013.

skanujemy tekst (zapewne tak, jak część czytelników robi to właśnie teraz) i wyciągamy pobieżne wnioski.

Łatwy dostęp do informacji wytworzył **usieciowioną przemądrzałość** – wydaje nam się, że wszystko możemy znaleźć i zrozumieć. Niestety, na skutek spadku poziomu krytycznego myślenia wynikającego z przeciążenia poznawczego, często „kupujemy" fałszywe interpretacje wyników badań naukowych. Aby coś poprawić musimy umieć to zmierzyć, bo inaczej nie będziemy wiedzieli jaka zmiana się dokonała. Powstaje, więc pytanie, jak można zmierzyć subiektywne odczucia pracownika w organizacji? Jak pisze Gilbert[18], **jeśli dany obiekt jest niemierzalny, nie może być przedmiotem badań naukowych.**

Niemierzalny obiekt może stanowić przedmiot dociekań, ale nie zostanie zbadany w sposób naukowy, ponieważ nauka wymaga precyzyjnych pomiarów, a jeśli czegoś nie można zmierzyć – określić jego właściwości za pomocą zegara, linijki lub jakiegokolwiek narzędzia – nie może być traktowane jako potencjalny przedmiot badań naukowych.

Z tego powodu słynny eksperyment więzienny Zimbardo[19], mimo losowego przydziału uczestników do grup więźniów i strażników, nie był badaniem naukowym, ponieważ nie dokonano w nim żadnych precyzyjnych pomiarów. Choć mamy nagrania różnych scen z eksperymentu, brakuje pomiaru (rzetelnych danych ilościowych, które umożliwiłyby obiektywną analizę). Celem badania opartego na szeregu uporządkowanych procedur i wykorzystaniu informacji zebranych w procesie pomiaru jest uzyskanie wyników replikowalnych – czyli takich, które mogą być **demonstrowane wielokrotnie** – zarówno przez naukowca, który je odkrył, jak i przez inne osoby. Wyniki, których nie da się uzyskać ponownie, nie są godne zaufania.

Liczby są idealne do roznoszenia ściemy

Wyniki badania[20] zostały szeroko rozpowszechnione w formie reguły **55-38-7** oznaczającej, że PODOBNO niewerbalne sygnały ciała są podstawowym kanałem przekazywania wiadomości. Sugerują one, że 55% treści adresat odczytuje z języka ciała nadawcy, 38% ze sposobu mówienia (głos, ton, intonacja), a **jedynie 7% z treści słów**.

Kiedy studenci dowiadują się, że **reguła 55-38-7** oparta na tych badaniach jest **fałszywa,** natychmiast kwestionują wiarygodność badań przeprowadzonych na Uniwersytecie Stanforda. Trudno im zrozumieć, że badanie zostało przeprowadzone rzetelnie, jednak **sposób** jego **popularyzacji zniekształcił interpretację wyników**. Odbiorcy żyjący w smogu informacyjnym często ulegają dychotomicznemu myśle-

[18] Gilbert, 2007.
[19] Zimbardo i in., 1971; Wieczorkowska, 2018.
[20] Mehrabian, 1971.

niu: jeśli badania naukowe zostały przeprowadzone prawidłowo, to wnioski należy uznać za słuszne, w przeciwnym razie należy zakwestionować ich wiarygodność. Konsumenci wyników naukowych, chcąc je podważyć, najczęściej krytykują wielkość lub skład próby badawczej.

Przytaczane badania były przeprowadzone bardzo starannie, jednak dotyczyły wyłącznie **NIESPÓJNYCH komunikatów**, takich jak sytuacja, w której mówimy, że jesteśmy spokojni, podczas gdy ton naszego głosu wyraźnie wskazuje coś zupełnie innego. Ten warunek ograniczający zniknął w popularnych opisach i spowodował **fałszywą generalizację wyników**. Ludzie zapisują się na kursy komunikacji niewerbalnej, zapominając, że najważniejsze, aby mieć coś ciekawego do powiedzenia odbiorcy. Takich przykładów głupich generalizacji jest znacznie więcej. Internetowa „mądrość psychologiczna" prawie zawsze opiera się na wynikach badań naukowych – **grzechem pierworodnym stała się ich nadmierna generalizacja**. Trzeba jednak przyznać, że przesłanie o wadze komunikacji niewerbalnej znajduje największe zainteresowanie w grupach, które można było wskazać na podstawie niezgeneralizowanych wyników oryginalnych badań. Szkoleniami interesują się przede wszystkim politycy i sprzedawcy, którzy mniej lub bardziej świadomie zdają sobie sprawę z faktu, że słowa, które wypowiadają, mogą nie być zgodne z tym, w co wierzą i **komunikacja niewerbalna może ich zdradzić** – chcą więc nauczyć się świadomie ją kontrolować.

Każde badanie i każda teoria mają swoje warunki brzegowe stosowalności, które internetowi „mędrcy" lekceważą. **Przewidywanie porażki** podczas wykonywania **silnie zautomatyzowanego zadania**, takiego jak chodzenie po linie, jest **samospełniającym się proroctwem**, ponieważ włączamy procesy kontrolowane, które pochłaniają nasze zasoby psychoenergetyczne. Popularna, lecz fałszywa generalizacja tej zależności mówi, o zgrozo, że ZAWSZE trzeba myśleć pozytywnie. O zagrożeniach związanych z pozytywnym myśleniem szerzej pisaliśmy w pierwszej części monografii.

Powinniśmy pamiętać, że nawet dokładnie przeprowadzone badania naukowe różnią się pod względem trafności wewnętrznej i zewnętrznej. Gdy słyszymy o wynikach badań naukowych, zawsze powinniśmy przede wszystkim zapytać o trafność wewnętrzną badań, a dopiero potem zastanowić się nad jego trafnością zewnętrzną, czyli możliwością uogólnienia wyników na całą populację.

Trafność wewnętrzna badań jest ważniejsza niż trafność zewnętrzna

Podstawowym podziałem analitycznych badań ilościowych jest podział na badania eksperymentalne i korelacyjne. Różni je charakter zmiennej niezależnej. W badaniach eksperymentalnych manipulujemy wartościami zmiennej niezależnej, w korelacyjnych dokonujemy jej pomiaru bez wpływania na jej poziom.

Wyjaśnijmy to na przykładzie[21].

Wyobraźmy sobie, że chcemy zbadać czy **częstość** uzyskiwania pochwał od bezpośredniego przełożonego (X – zmienna niezależna objaśniająca) ma wpływ na **satysfakcję** pracowników (Y – zmienną zależną objaśnianą).

W badaniach eksperymentalnych manipulujemy pochwałą (wprowadzamy ją w ściśle określonych warunkach, w odmienny sposób w różnych grupach), dlatego trafność wewnętrzna badania jest bardzo wysoka.

W badaniach korelacyjnych mierzymy zarówno X [pochwałę] i Y [satysfakcję], jednak badacz NIE MA WPŁYWU na wystąpienie zmiennej niezależnej (pochwały). Oznacza to, że nie jest w stanie wykluczyć wpływu „trzecich" zmiennych (kto, kiedy, za co i w jakich warunkach pochwalił). Z tego względu immanentną cechą badań korelacyjnych jest ich niska trafność wewnętrzna.

W języku potocznym pojęcie manipulacji ma **negatywną konotację**, ponieważ kojarzy się z ukrytym sposobem wywierania na nas wpływu – wbrew naszej woli i/lub interesowi. Natomiast **w metodologii możliwość manipulowania wartościami** zmiennej **niezależnej** jest kluczowa.

Jeśli potrafimy manipulować daną zmienną, nie musimy czekać aż jej określona wartość wystąpi w sposób naturalny, możemy wprowadzić ją sami i sprawdzić, jakie są tego skutki. W badaniach **manipulujemy wartościami ZMIENNYCH, a nie ludźmi**, choć efekty tej manipulacji z definicji powinny wpływać na zachowanie osób uczestniczących w badaniach.

Reakcja na pochwałę może zależeć od wielu zmiennych, takich jak samoocena, poziom aspiracji, neurotyzm czy doświadczenia z przeszłości etc. Ludzie różnią się między sobą na wielu wymiarach, które mogą wpływać na naszą zmienną zależną (satysfakcja), niezależnie od wartości zmiennej niezależnej (pochwała). Jednak, jeśli LOSOWO podzielimy badaną próbę na 2 grupy, to średni poziom samooceny, neurotyzmu etc. będzie w obu grupach taki sam. Losowy przydział do grup eksperymentalnych to metoda tworzenia **„zbiorowych klonów"**, która chroni nas przed zakłócającym wyniki wpływem innych zmiennych. Losowy przydział do grup eksperymentalnych można wyobrazić sobie jako wielki „wyrównywacz", dzięki któremu wiemy, że wartości zmiennej niezależnej są jedyną różnicą odpowiedzialną za zróżnicowanie zmiennej zależnej (**kanon jedynej różnicy** Milla).

Eksperyment różni się od innych typów naukowych dociekań tym, że zamiast czekać na zaistnienie interesujących nas wydarzeń naturalnych, eksperymentator kreuje warunki potrzebne do obserwacji. Ma to dwie podstawowe zalety[22].

Po pierwsze: konstruowanie sytuacji eksperymentalnej pozwala uwypuklić najważniejsze elementy i pominąć te nieistotne. Na przykład zadanie można zaaranżować tak, aby badani byli chwaleni w ściśle określony sposób. W codziennym życiu wpływ chwalenia mógłby być modyfikowany np. poprzez osobowość szefa. Jeżeli

[21] Wieczorkowska i in., 2015.

[22] Aronson i Wieczorkowska, 2001.

jedna grupa będzie chwalona, a druga nie otrzyma żadnej informacji zwrotnej, nie możemy być pewni, czy różnice w poziomie satysfakcji są efektem pochwały, czy też takie same wyniki otrzymalibyśmy po zastosowaniu krytyki. Dlatego należy dokładać wszelkich starań, aby warunki eksperymentalne różniły się tylko pod jednym względem. Kanon jedynej różnicy będzie spełniony, jeżeli jedna grupa będzie np. chwalona publicznie, a druga w cztery oczy.

Po drugie: eksperymentator może kontrolować i systematycznie zmieniać warunki, aby zbadać dokładnie tę samą sytuację zawierającą (lub nie) pewne elementy (np. pochwała ustna vs. pochwała wysłana e-mailem). Gdyby badacz chciał zastosować nieeksperymentalny schemat badania, musiałby znaleźć **„naturalne" grupy chwalone w różny sposób**. Jednak szefowie stosujący te dwie metody chwalenia mogą różnić się pod wieloma innymi względami. Znalezienie dwóch grup, które byłyby do siebie podobne pod każdym względem (ten sam typ pracy, branża itp.) – z wyjątkiem jednego, interesującego badacza czynnika – jest bardzo trudne, jeżeli nie niemożliwe. Co ważniejsze, uczestnicy badań w eksperymentach są przydzielani do grup losowo. W naturalnych warunkach ludzie mogą unikać określonych typów szefów w zależności od swoich preferencji.

Nastawieni sceptycznie wobec badań eksperymentalnych twierdzą, że gromadzimy głównie wiedzę o studentach, ponieważ to oni najczęściej biorą udział w takich badaniach.

O ile badania ankietowe można przeprowadzać na tysiącach osób, eksperymenty są realizowane na małych grupach, ponieważ ich przygotowanie wymaga wielu godzin pracy i większej inwestycji czasowej ze strony uczestników. Pojawia się więc pytanie, **na ile można uogólnić wyniki uzyskane** np. w badaniu 30 studentów, innymi słowy, jaka jest **trafność zewnętrzna** takiego badania. W końcu nie interesują nas jedynie reakcje uczestników, lecz chcemy wnioskować o populacji, z której próba została wybrana.

Jeśli interesują nas np. preferencje wyborcze Polaków, to wylosowana do badań próba powinna być reprezentatywna dla dorosłych obywateli kraju.

Próby reprezentatywne są niezwykle istotne, gdy naszym celem jest **poznanie rozkładu zmiennej w populacji**. Jednak ze względu na dużą liczbę zmiennych zakłócających, często o bardzo skośnych rozkładach (np. wykształcenie), są one znacznie mniej użyteczne, gdy chcemy badać zależności między zmiennymi. Skuteczna statystyczna kontrola ważnych zmiennych (choć nieistotnych dla celu badania) jest bardzo trudna do osiągnięcia, dlatego **zamiana zmiennych towarzyszących (zakłócających**) w stałe daje dużo lepsze efekty.

Podstawowym celem badań eksperymentalnych nie jest określenie rozkładu zmiennej zależnej np. poziomu satysfakcji, lecz **stwierdzenie czy różnica w poziomie satysfakcji między badanymi grupami jest istotna statystycznie**. Do tego celu NIE jest konieczna próba reprezentatywna.

Lepszą strategię maksymalizacji trafności zewnętrznej niż badanie prób reprezentatywnych stanowi **replikowanie eksperymentów na próbach homogenicznych** – np. osobno badamy rolników, osobno – pracowników naukowych, osobno – studentów itd. Wyniki badań z udziałem 30 osób losowo podzielonych na grupy eksperymentalne mają większą wartość poznawczą niż wyniki badań korelacyjnych na próbie dogodnej z udziałem 300 tys. osób.

Randomizacja drugiego stopnia (losowy przydział do grup eksperymentalnych) zapewnia wysoką trafność wewnętrzną, jakiej nigdy nie osiągniemy w badaniach korelacyjnych. **Randomizacja pierwszego stopnia oznacza** losowanie próby z populacji.

Problemem pozostaje fakt, że ludzie niechętnie uczestniczą w badaniach eksperymentalnych, co znacząco utrudnia rozwój nauk społecznych. Nadal można też usłyszeć, że eksperymenty laboratoryjne są nierealistycznymi i „wymyślonymi" imitacjami ludzkich interakcji, nieodzwierciedlającymi w ogóle „rzeczywistego świata". Twierdzący tak, zapominają, że eksperyment może być realistyczny w dwojaki sposób[23].

O **realizmie sytuacyjnym** mówimy, gdy sytuacja w laboratorium staje się podobna do zdarzeń, które często spotykają ludzi w świecie zewnętrznym. Jednak o wiele ważniejsze jest to, czy stworzona przez badacza sytuacja angażuje badanego i skłania go do traktowania poważnie tego, co się dzieje. Tę cechę sytuacji określa się mianem **realizmu psychologicznego**. Dobrze zaprojektowane sytuacje eksperymentalne, na przykład gdy uczestnicy grają w angażującą grę, powodują, że po pewnym czasie sytuacja staje się bardzo realna. Osiągnięte wyniki, pochwały, które uzyskują, mają dla nich realną wartość. W takich warunkach uczestnicy są o wiele bardziej zaangażowani niż przy wypełnianiu ankiet. Niechęć wobec eksperymentów często wynika z faktu, że **ich prawdziwy cel musi pozostać ukryty na czas trwania badania**. Gdyby badani wiedzieli, że interesuje nas np. ich uległość wobec autorytetu i zostali poproszeni o opisanie jak zachowaliby się w danej sytuacji, dowiedzielibyśmy się jedynie, jacy chcieliby być, a nie jacy byliby w rzeczywistości.

W jednym z badań[24] przeprowadzonych w warunkach naturalnych pielęgniarki otrzymywały telefonicznie polecenie od nieznanego im lekarza, który rzekomo prowadził pacjenta przebywającego na ich oddziale. Lekarz mówił: „Przyjadę za godzinę, ale proszę już teraz podać pacjentowi 20 mg estrogenu". Na opakowaniu zawarta była informacja, że maksymalna dawka wynosi 10 mg, typowa – 5 mg. Polecenie wymagało, więc dwukrotnego przekroczenia maksymalnej dawki. Gdy poproszono grupę pielęgniarek o **przewidywanie**, jak zachowałyby się w takiej sytuacji, mniej niż 20% stwierdziło, że pewnie wykonałyby takie polecenie. W sytuacji szpitalnej aż 21 z 22 pielęgniarek, które nie wiedziały, że uczestniczą w badaniu, podało lek. Nie doceniamy wpływu, jaki sytuacja ma na nasze zachowanie, dlatego nasze wyobra-

[23] Aronson i Wieczorkowska, 2001.

[24] Gerrig i Zimbardo, 2006.

żenia na temat tego, jak postąpilibyśmy w określonych warunkach, nie mają dużej wartości badawczej.

Porównajmy dwa sposoby badania związku między oglądaniem pornografii a agresją.

Badanie korelacyjne: pornografia współwystępuje z agresją

Badacze w projekcie nazwanym przeze mnie **PORNOGRAFIA 2**[25] sprawdzali w poszczególnych stanach USA, czy istnieje związek pomiędzy liczbą sprzedawanych pism pornograficznych (zmienna X) i liczbą odnotowanych w tych stanach gwałtów (zmienna Y).

Do pomiaru liczby sprzedawanych pism pornograficznych (zmienna X) badacze wybrali osiem czasopism o wyraźnie erotycznym charakterze. Liczbę czasopism sprzedanych w każdym stanie w 1979 roku porównywano z liczbą gwałtów odnotowanych przez FBI w „Uniform Crime Reports" (zmienna Y) w tych stanach.

Należy jednak zaznaczyć, że ze względu na fakt, iż gwałt jest przestępstwem nie zawsze zgłaszanym policji, opublikowane dane są bez wątpienia zaniżone w stosunku do rzeczywistej liczby popełnionych gwałtów.

Analiza tak zebranych danych wykazała silną dodatnią korelację między X i Y. Wykazano również, że *X* nie koreluje z częstością nieseksualnych aktów przemocy. Badacze jednak nie udowodnili, że pornografia stanowi przyczynę gwałtów. **Nie sposób bowiem wykluczyć, że źródłem stwierdzonej korelacji są różnice** w kulturowych modelach „super mężczyzny" funkcjonujących w różnych stanach. Modele te mogą skłaniać mężczyzn do kupowania czasopism pornograficznych, jak i do popełniania gwałtów.

W kolejnym badaniu korelacja między liczbą gwałtów a rozpowszechnieniem pornografii w różnych stanach USA wyniosła 0,53. Okazało się, że obie te zmienne korelują z proporcją rozwiedzionych mężczyzn w danym stanie (odpowiednio: 0,67 i 0,59). Uwzględnienie tej „trzeciej zmiennej"[26] w analizie związku między „pornografią" a „gwałtami" spowodowało redukcję współczynnika korelacji z 0,53 do 0,22.

Badanie eksperymentalne: pornografia wzmaga agresję

W badaniu nazwanym przeze mnie **PORNOGRAFIA 3**[27] uczestniczyli tylko ochotnicy płci męskiej, którzy zgłosili się do badania wpływu stresu na proces uczenia się.

W czasie oczekiwania na rozpoczęcie badania byli proszeni o obejrzenie i ocenę krótkiego filmu, który badacz szykuje do innych badań.

---

[25] Wieczorkowska i Wierzbiński, 2013.

[26] Kuźmińska, 2016b.

[27] Wieczorkowska i Wierzbiński, 2013.

Potem każdy z mężczyzn brał udział w losowaniu roli, w której zostawał nauczycielem, podczas gdy aktor udający osobę badaną pełnił rolę ucznia mającego uczyć się listy słów.

Badani w roli nauczyciela sprawdzali poprawność odpowiedzi uczniów, aplikując im wstrząsy elektryczne o intensywności wybranej na skali od 1 do 8 za każdą błędną odpowiedź. W rzeczywistości nie stosowano prawdziwych wstrząsów elektrycznych.

Pierwszą zmienną niezależną była **płeć „ucznia-ofiary"** (aktora), drugą – **rodzaj** krótkiego **filmu**, który mężczyźni oglądali, zanim zaczęli karać swoich uczniów: (1) film o mężczyźnie, który włamuje się do pewnego domu i gwałci kobietę, grożąc jej pistoletem (tj. pornografię nacechowaną przemocą); (2) film pokazujący różne etapy stosunku seksualnego kobiety z mężczyzną, ale bez aktów przemocy czy agresji; (3) film neutralny nie zawierający treści seksualnych ani agresywnych.

Następnie badani mężczyźni słyszeli odpowiedzi aktora (w rzeczywistości były to nagrania głosowe odtwarzane z taśmy magnetofonowej), z których część była błędna. Zmienną zależną stanowiła tu intensywność wstrząsów, które badany mężczyzna zaaplikował „ofierze".

Wyniki tego badania są jednoznaczne. Mężczyźni nie mieli wpływu na to, jaki film oglądali – dzięki losowemu przydziałowi do grup możemy założyć, że grupy mężczyzn tworzyły „zbiorowe klony" nie różniące się średnim poziomem agresji. Oczekiwalibyśmy, że średnia intensywność wstrząsów aplikowanych „ofierze" będzie zbliżona we wszystkich grupach. Stwierdzone różnice można wyjaśnić jedynie aktywizacją myśli i zachowań przez obejrzany wcześniej film oraz płcią ucznia. Choć w dwóch grupach, które oglądały filmy niezawierające przemocy, nie zaobserwowano różnic w karaniu kobiet i mężczyzn, to pod wpływem filmu z przemocą kobiety występujące w roli ucznia były karane znacznie surowiej.

Podsumowując – choć badania korelacyjne mogą dostarczać istotnych informacji, prawdziwy eksperyment (czyli taki, który umożliwia poznanie relacji przyczynowej) jest nieoceniony.

Gdyby porównać przyrost naszej wiedzy, okazałoby się, że budują go przede wszystkim wyniki badań eksperymentalnych, a nie korelacyjnych. Mimo to fakt ten nie zyskał należytego uznania w naukach o zarządzaniu, choć rozwój ekonomii behawioralnej oraz kolejne **Nagrody Nobla** w dziedzinie ekonomii już dawno powinny zmienić nastawienie badaczy. W 2002 roku przyznano Nagrody Nobla **V. L. Smithowi** za ustanowienie eksperymentów laboratoryjnych narzędziem empirycznej analizy oraz **D. Kahnemanowi** za zastosowanie narzędzi psychologicznych w badaniach ekonomicznych ze szczególnym uwzględnieniem teorii perspektywy. Także noblista z 2013 roku **R. J. Shiller** (jak twierdzi – pod wpływem żony psychologa) zajmuje się ekonomią behawioralną, a w 2017 roku za swój wkład w jej rozwój **R. Thaler** również otrzymał Nagrodę Nobla. Nawet Nagroda Nobla w dziedzinie fizyki w 2024 roku została przyznana m.in. **G. E. Hintonowi**, który zaczynał karierę naukową od zdobycia dyplomu w 1970 roku w zakresie psychologii eksperymentalnej.

Słabość modeli teoretycznych w naukach o zarządzaniu

Słabością nauk społecznych jest brak systematyzacji pojęć. Nawet publikując w fachowych czasopismach, często unikamy używania „żargonu" naukowego, co przez lata było wzmacniane obowiązkiem publikacji prac habilitacyjnych w formie książek dostępnych dla szerszego grona czytelników. Prawdą jest, że posługujemy się pojęciami naturalnymi, które trudno zdefiniować w sposób klasyczny. Na szczęście prace przeglądowe prezentujące różnorodne definicje zniknęły z pola naszych zainteresowań, choć wciąż można spotkać przeglądy zawierające kilkadziesiąt definicji danego pojęcia np. zaufania[28]. **Pomiar zaufania jest bardzo trudny**, ponieważ – jak wykazano w wielu badaniach (np. w powstałej pod moim kierunkiem rozprawie doktorskiej)[29] – taki prosty czynnik kontekstowy, jak **liczenie pieniędzy** (w porównaniu z sumowaniem liczb) **obniża poziom zaufania** wobec nowo poznanej osoby. Na szczęście **efekt aktywizacji pieniędzy podlega habituacji**, gdyż nie zaobserwowano go u osób, które liczą pieniądze na co dzień. Jeśli porównamy pojęcia naukowe na poziomie operacjonalizacji szybko wykryjemy, że naprodukowaliśmy ich za dużo. Oto przykład:

1. Potrzeba poznania (*Need for Cognition*)
2. Potrzeba domknięcia (*Need for Closure*)
3. Nietolerancja niejednoznaczności (*Intolerance of Ambiguity*)
4. Refleksyjność poznawcza (*Cognitive Reflection*)
5. Potrzeba struktury i porządku (*Need for Structure and Order*)
6. Tolerancja niepewności (*Uncertainty Tolerance*)
7. Dogmatyzm (*Dogmatism*)
8. Złożoność poznawcza (*Integrative Complexity*)

Wszystkie te cechy korelują z konserwatyzmem[30], co nie dziwi, gdy zobaczymy, że w ich operacjonalizacjach używane są częściowo te same pozycje (itemy).

Dlatego porównanie wniosków z badań musi odbywać się na poziomie operacjonalizacji, których niestety nie da się wystandaryzować. Rozkłady odpowiedzi uzyskiwane za pomocą standardowych technik pomiaru zależą od charakterystyki próby. Przykładem może być pozycja z kwestionariusza przedziałowości **„Jestem nonszalancki w traktowaniu szczegółów"**, która w grupie pracowników Wydziału Farmacji zmieniła diametralnie swoją korelację ze skalą. Nonszalancja w traktowaniu szczegółów dla farmaceuty ma zupełnie inne znaczenie niż dla studenta. Także w badaniach eksperymentalnych operacjonalizacja manipulowania poczuciem zagrożenia wymaga innego podejścia w grupie studentów niż w grupie pracowników firmy.

Tworząc modele teoretyczne, ciągle zapominamy, że **nauki społeczne są wysokokontekstowe**. Wiedza o zależnościach między zmiennymi uzyskiwanymi w modelu

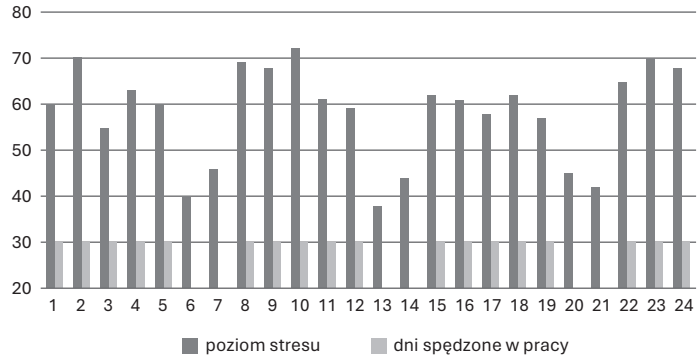
[28] Grudzewski i in., 2007.
[29] Kuźmińska, 2016a; 2019.
[30] Jost i in., 2003.

ceteris paribus jest trudno aplikowalna, gdy mamy do czynienia z problemem jednostkowym (np. „trudny" pracownik, konfliktowy szef, pojedyncza firma[31]). Dlatego powinniśmy zdecydowanie ograniczyć poziom ogólności naszych teorii. Zazdrościmy fizykom i chcielibyśmy opracować równie proste reguły, tak jak trzy zasady dynamiki Newtona, ignorując przy tym fakt, że nasz umysł rozwija się w odpowiedzi na wyzwania pojawiające się w otoczeniu. Jest on więc raczej zbiorem modułów i wyuczonych reakcji na różne klasy(!) bodźców. Poważnym problemem pozostaje również rozdział między badaniami naukowymi, które stosują głównie podejście nomotetyczne a praktyką, która wymaga wiedzy idiograficznej.

Pomimo znacznych przyrostów naszej wiedzy, jak dotąd nie przełożyły się one na zwiększenie skuteczności działań praktycznych. Wiemy znacznie więcej, ale nadal potrafimy niewiele więcej. Ponownie oczekujemy, że rozwój technologii pomoże rozwiązać ten problem. Technologia mogłaby umożliwić wielokrotny pomiar tej samej osoby **(np. ilościowe, porównawcze studia przypadków**[32]) bez wywoływania jej irytacji.

Na rys. 1 pokazano zastosowanie ilościowego (choć nie eksperymentalnego, ponieważ nie wprowadzono interwencji) studium przypadku dotyczącego poziomu stresu. Tego typu dane pozwalają łatwo zauważyć, że poziom stresu u tego pracownika istotnie maleje podczas weekendów.

Rysunek 1. Ilustracja dynamiki poziomu STRESU w ciągu kolejnych 24 dni



Źródło: Johnston i Johnston, 2023.

Ilościowe studia przypadków mogą wykorzystywać dane **EMA (*Ecological Momentary Assessment*)**[33] zbierane za pomocą telefonów. Osoba diagnozowana odpowiada na parę pytań wysyłanych przez urządzenia elektroniczne kilkukrotnie w ciągu dnia przez np. kilka tygodni. Odpowiedzi mogą być automatycznie łączone

[31] Wieczorkowska i Gonzalez, 2023.
[32] Siarkiewicz, 2007.
[33] Obarska i in., 2021.

z rejestrowanymi przez biosensory danymi fizjologicznymi (np. zmienność rytmu serca, długość poszczególnych faz snu etc.). Dzięki powtarzalności pomiarów tej samej zmiennej w warunkach naturalnych i w czasie rzeczywistym EMA eliminuje zniekształcenia retrospekcyjne. Metoda ta pozwala również opracować **spersonalizowaną interwencję** (EMI = *Ecological Momentary Intervention*), która może wychwycić w czasie rzeczywistym moment wystąpienia niebezpiecznego stanu np. głodu emocjonalnego lub zachowania, jak przykładowo kompulsywne skrolowanie mediów społecznościowych.

Do tego problemu wrócimy przy omawianiu paradygmatu WiW.

W sporze jednoteoretycznych z eklektykami uważamy, że najważniejsze jest otrzymanie interesujących poznawczo wyników. Dlatego nie widzimy przeszkód w łączeniu różnych szkół teoretycznych. Czterdzieści lat temu mówienie o procesach nieświadomych było całkowicie odrzucane przez naukę, która została wówczas zdominowana przez podejście poznawcze. Dziś jednak nikt nie ma wątpliwości, że przetwarzanie informacji w przeważającej części odbywa się poza świadomością.

Za dużo mamy w naukach społecznych **twórczości** (ciągłego wprowadzania nowych pojęć), **za mało – rzemiosła**. Nie jest tajemnicą, że najłatwiej osiągnąć „sukces" przez wprowadzenie nowego pojęcia i opublikowanie narzędzia pomiarowego. Jeśli mamy szczęście, wiele osób zacznie badać korelacje nowego narzędzia z już istniejącymi aż fala zainteresowania opadnie – co jest łatwe do przewidzenia.

Na wykładach z metodologii powtarzam do znudzenia: replikować, replikować i jeszcze raz replikować, bo prawie żadna udana replikacja w naukach społecznych nie jest banalna. Studenci najpierw powinni nauczyć się uzyskiwać wyniki przewidywalne na podstawie wcześniejszych badań, a dopiero potem próbować wychodzić poza ich granice i wprowadzać własne innowacje.

W ostatnich latach obserwujemy **kryzys replikowalności**, który uwidacznia z jak wielozmiennową materią[34] mamy do czynienia. Przykładowo wyniki badań, które wielokrotnie potwierdzały w latach 90. XX wieku **Teorię Opanowywania Trwogi** (*Terror Management Theory*, TMT), nie replikują się w trzeciej dekadzie XXI wieku[35]. W badaniach[36] przeprowadzonych w XX wieku, po przypomnieniu o śmierci ludzie bardziej wspierali „swoich" (np. rodaków, ludzi o podobnych poglądach), surowiej karali „obcych" (np. odmiennych kulturowo, światopoglądowo, religijnie), deklarowali wyższą religijność i uskrajniali swoje poglądy (liberałowie stawali się bardziej liberalni, konserwatyści bardziej konserwatywni). Czy brak replikacji oznacza, że nasze sieci neuronowe są bardziej oswojone ze śmiercią niż sieci neuronowe naszych przodków? Na pewno nastąpiło istotne zmniejszenie roli socjalizacji do wspólnoty (co pokazaliśmy także w naszych badaniach[37]). Takie replikacyjne porażki powinny być impulsem do **poszukiwania warunków brzegowych** dla wcześniej opi-

[34] Por. np. Van Bavel i in., 2016.
[35] Klein i in., 2022.
[36] Greenberg i in., 1997.
[37] Yosephpour, 2021; Wilczyńska, 2023.

sanych zależności. Informacje na temat: „Kiedy to nie działa" – mogą wzbogacać naszą wiedzę.

Trzeba podkreślić, że nie powinniśmy oczekiwać replikacji wyników badań z użyciem **tych samych metod operacjonalizacji zmiennych,** lecz skupić się na **replikacji potwierdzania hipotez**. Tak jak dowodziłam w mojej poprzedniej monografii „Rafy psychologiczne w zarządzaniu. Ludzie różnią się od śrubokrętów"[38], mózgowe sieci neuronowe interpretują wszystko, co się dzieje, także w czasie badania. Współczesne mózgowe sieci neuronowe respondentów mogą inaczej interpretować pytania kwestionariusza czy instrukcje, ponieważ były kształtowane na podstawie innych sygnałów.

Słabość analiz niepotrafiących wykryć wieloczynnikowych efektów

O ile badania eksperymentalne opierają się na dobrze ugruntowanych i oczywistych standardach analizy, o tyle w badaniach korelacyjnych publikowanych jest wiele wyników o wątpliwej ważności.

Wynika to przede wszystkim z niskiej trafności wewnętrznej będącej immanentną cechą badań korelacyjnych, co przejawia się w splątaniu (*confounding*) wpływów różnych zmiennych. Prawdą jest jednak, że **wielu zagadnień nie da się zbadać eksperymentalnie**, dlatego powinniśmy intensywnie pracować nad nowymi modelami analizy badań korelacyjnych. Jest to tym bardziej istotne zadanie, że istnieją ogromne zbiory ogólnodostępnych danych z badań prób reprezentatywnych, zgromadzone przy znacznym nakładzie kosztów, które są w niewielkim stopniu wykorzystywane poznawczo. Bardzo popularne stały się porównania międzynarodowe zakładające, że żyjący w tym samym społeczeństwie upodabniają się do pewnego stopnia do siebie, ponieważ są socjalizowani w tym samym środowisku. W naszym przeładowanym i zaśmieconym informacyjnie świecie **rankingi krajów** na różnych wymiarach – np. zaufania – wzbudzają duże zainteresowanie, bo są bardzo proste i obrazowe. Niepokoi, gdy te same kraje, np. Dania, znajdują się w czołówce szczęśliwości, zaufania i zażywania środków psychotropowych na głowę mieszkańca. Ten ostatni ranking jest najtrudniejszy do przygotowania, ponieważ metodologia zliczania środków psychotropowych jest mniej jednorodna niż pytania o zaufanie uogólnione.

Analizy wielozmiennowe są koniecznością. Problem tkwi jednak w ich wrażliwości na trafność przyjętego modelu. Jakże często powtarzam moim studentom: współczynniki regresji zależą od „towarzystwa" innych uwzględnionych w równaniu predyktorów. Tylko osoby nierozumiejące istoty tych analiz mogą stosować krokową analizę regresji, pozostawiając decyzje merytoryczne algorytmom statystycznym,

[38] Wieczorkowska, 2025.

i pisać z pełnym przekonaniem, że analiza czynnikowa czy analiza skupień dowiodła istnienia określonej liczby czynników, czy skupień. Zbyt często usunięcie jednej zmiennej zmienia strukturę macierzy korelacji.

Na moich studentach wielkie wrażenie robi analiza przykładu[39], w którym na tych samych danych stwierdza się różne zależności między X a Y zależnie od tego, które zmienne dodatkowe zostały uwzględnione lub kontrolowane w analizach wielozmiennowych.

Oczywiście, tam gdzie istnieje jasna teoria, lista zmiennych jest wyznaczona jednoznacznie. W analizach danych sondażowych na próbach reprezentatywnych, gdzie kontrolowanie zmiennych socjodemograficznych jest koniecznością, często zapomina się o uwzględnieniu w modelu efektów interakcyjnych.

Przykładowo, sprawdzając socjodemograficzne predyktory poziomu **wykształcenia** (operacjonalizowanego przez lata nauki), możemy stwierdzić brak istotnego związku z płcią, ponieważ istotnym predyktorem jest **interakcja wieku i płci**. Przewaga edukacyjna starszych mężczyzn zamienia się w przewagę młodszych kobiet, dlatego efekt główny płci nie jest istotny. Podobną zależność stwierdzono w 14 z analizowanych 33 krajów[40]. Wprowadzenie do modelu regresyjnego efektów interakcyjnych było jednak kluczowe – czego niestety badacze często nie robią. Nasze wnioski staną się zdegenerowane, jeżeli nie uwzględnimy efektów interakcyjnych.

Drastyczny przykład: ryzyko zachorowania na raka krtani[41] jest prawie:

- 37-krotnie wyższe u palących więcej niż paczkę papierosów dziennie niż u osób niepalących;
- 9-krotnie wyższe u pijących więcej niż 7 drinków tygodniowo niż u abstynentów;
- 330-krotnie wyższe u palących i pijących równocześnie.

W innych badaniach[42] wykazano, że łączne oddziaływanie nadmiernej otyłości i nadmiernej konsumpcji alkoholu na wątrobę jest wielokrotnie większe niż wynikałoby to ze zwykłego sumowania ryzyka stwarzanego przez każdy z tych czynników osobno.

Wykrywanie efektów interakcyjnych jest niesłychanie trudne, bo ich liczba rośnie wzrastająco wraz z liczbą predyktorów (np. dla 3 predyktorów trzeba przetestować 4 interakcje, dla 4 predyktorów jest ich już 10...).

Pięknie to zilustrował Lem w swojej jedynej powieści pozornie detektywistycznej „Katar", opisującej serię niewyjaśnionych śmierci amerykańskich turystów we Włoszech. Mężczyźni w wieku 40–50 lat, podróżujący samotnie, ginęli, wyskakując przez okno pokoi hotelowych. Przyczyną była niezwykle złożona interakcja wielu czynni-

[39] Wieczorkowska i Wierzbiński, 2013.

[40] Wierzbiński, 2009.

[41] Wyniki badań prof. J. Lissowskiej z Zakładu Epidemiologii i Prewencji Nowotworów Złośliwych w Instytucie w Warszawie: http://wyborcza.pl/TylkoZdrowie/7,137474,25201718,dlaczego-coraz-czesciej-chorujemy-na-raka.html

[42] Liu i in., 2010.

ków. Mordercą okazała się choroba psychiczna, której rozwój wymagał jednoczesnego spełnienia 7 warunków:

1) wiek 40–50 lat – osoby młodsze ani starsze nie walczyły z łysieniem;
2) stosowanie hormonalnej maści przeciw łysieniu, wcieranej w skórę głowy;
3) chorowanie na **katar sienny** i branie stosownych leków;
4) przebywanie w nasłonecznionym miejscu;
5) spożywanie migdałów;
6) zażywanie kąpieli siarkowych;
7) bycie turystą zagranicznym – Włosi we wstępnej fazie zatrucia trafiali pod opiekę psychiatryczną;
8) podróżowanie w samotności – zwiastunami zatrucia były zmiany usposobienia, które mógłby zauważyć towarzysz podróży i nakłonić chorego do powrotu do domu.

Lem podaje następujące wyjaśnienie**:**

Mężczyźni, walcząc z łysieniem wcierali specjalny **preparat hormonalny** w skórę głowy, który podczas nasłonecznienia (np. w Neapolu latem) zmieniał budowę chemiczną pod wpływem ritaliny zawartej w lekach przeciwalergicznych, więc preparat przekształcał się w silny depresor, który w dużych ilościach wywoływał zatrucie chemiczne prowadzące do obłędu i skłaniające do wyskakiwania z okien. **Katalizatorem** zwiększającym działanie depresora były **związki cyjanku z siarką**. Ślady cyjanku znajdowały się w **gorzkich migdałach**, które zostały również zanieczyszczone siarką – na skutek użycia w kilku neapolitańskich palarniach migdałów środka do dezynsekcji. Resztki preparatu przenikały do emulsji, w której migdały były moczone przed procesem palenia. Konieczne do zajścia reakcji wolne **jony siarkowe** pochodziły z leczniczych kąpieli siarkowych.

Ten przykład pokazuje, jak ważne jest uwzględnienie efektów interakcyjnych i jak skomplikowana może być ich analiza. Przytoczyliśmy ten opis, ponieważ w naszych analizach statystycznych taka zależność: konieczność jednoczesnego wystąpienia wszystkich 8 czynników, pozostałaby niewykryta. W XXI wieku dysponujemy wielkimi zbiorami danych, ale nadal nie potrafimy wykrywać w nich takich interakcyjnych zależności.

Problem fałszywych respondentów

Ankiety internetowe ze względu na swoją powszechność wypierają inne sposoby prowadzenia badań w naukach społecznych i zdobyły dominującą pozycję[43]. Łatwość pozyskania danych za pomocą ankiet nie idzie w parze ze starannością ich analiz, która wymaga odsiewania odpowiedzi FAŁSZYWYCH respondentów.

[43] Kabut, 2021.

Tacy respondenci są określani jako fałszywi, ponieważ, choć czasem świadomie, częściej bezrefleksyjnie udają prawdziwych respondentów, od których oczekujemy rzetelności i troski o udzielanie maksymalnie prawdziwych odpowiedzi. Fałszywy respondent (określany także jako niedbały, nieuważny lub niezmotywowany) może unikać wysiłku mentalnego[44] np.:

- wybierając bez czytania pierwszą lub najbardziej widoczną opcję z kafeterii,
- udzielając odpowiedzi zbyt szybko,
- odpowiadając losowo,
- bezwiednie poddając się tendencji do zgadzania się z treścią pytania.

W ankietach papierowych takich respondentów można było stosunkowo łatwo wykryć na etapie wprowadzania odpowiedzi do komputera. Jednak w przypadku plików elektronicznych wiersz zawierający „fałszywe" dane jest znacznie trudniejszy do zidentyfikowania.

Najprostszymi do wykrycia problemami stały się: nadużywanie odpowiedzi bezteściowej, gdy taka opcja jest dostępna czy pominięcie całego zestawu pytań (jeżeli jest to technicznie możliwe).

Zachowania FAŁSZYWYCH respondentów są w literaturze określane różnorodnymi terminami[45], takimi jak: losowość, lenistwo, niedbałość, wybieranie pierwszej satysfakcjonującej odpowiedzi, nieuważność, brak różnicowania.

Opracowana pod moim kierunkiem procedura wykrywania fałszywych respondentów wykorzystuje **4 czerwone flagi** będące **sygnałami ostrzegawczymi (ang. *Warning Signals*):**

SO1: zbyt krótki CZAS odpowiedzi

Ten wskaźnik dotyczy ankiet wypełnianych za pomocą komputera, który może mierzyć czasy. **Całkowity** czas odpowiedzi (**OAT**) to czas, który upłynął od pierwszego załadowania pierwszej strony ankiety do wyświetlenia strony końcowej. **Częściowy** czas odpowiedzi (**PAT**) to czas spędzony na udzielaniu odpowiedzi na poszczególne bloki ankiety. Słowa na minutę (**WPM**) to wskaźnik szybkości czytania obliczany przez podzielenie liczby słów na pojedynczej stronie ankiety przez czas, w którym ta część była oglądana (w minutach dla WPM i w sekundach dla WPS). U osoby umiejącej czytać długość słowa nie wpływa na czas jego przeczytania. **Szybkość czytania słowa zależy od jego długości wyłącznie na początku nauki czytania.**

[44] Krosnick, 1991; Galesic i in., 2008; Couper i in., 2004; Conrad i in., 2017; Michałowicz, 2016; Krosnick i Alwin, 1989; Kabut, 2021.

[45] Levi i in., 2021; Credé, 2010; Huang i in., 2012; Huang i DeSimone, 2021; Meade i Craig, 2012; Bowling i in., 2020; Krosnick, 1991; McKibben i Silvia, 2017; Beck i in., 2019; Steedle i in., 2019; Kabut, 2021.

SO2: zbyt duża liczba błędów nieuwagi [*Attention Check Questions-ACQ*]

Drugim najczęściej stosowanym sygnałem jest SO2 (pytania sprawdzające uwagę)[46]. Niektórzy badacze twierdzą, że jedno pytanie sprawdzające uwagę może być wystarczające, jednak ze względu na dynamikę uwagi respondenta ich liczba powinna wzrastać wraz z długością ankiety.

W przeprowadzonych w moim zespole badaniach testowaliśmy dwa typy ACQ (*Attention Check Questions*):

(1) pytania z instrukcją (tj. „Proszę wybrać „tak jak osoba A" w tym pytaniu");
(2) pytania arytmetyczne (np. „Wybierz z czterech podanych odpowiedzi poprawny wynik tego działania **23+5=?**").

Pierwszy rodzaj pytań ACQ polegający na narzuceniu odpowiedzi bez wyjaśnienia powodu może u niektórych respondentów wywołać reaktancję. Dlatego rekomendujemy stosowanie pytań arytmetycznych, uzasadniając respondentom ich obecność jako sposób na walkę z monotonią ankiety.

SO3: zbyt duża liczba odpowiedzi beztreściowych lub zbyt niska wariancja

Odpowiedzi beztreściowe (*don't know* – NIE WIEM – trudno powiedzieć) – nie dostarczają żadnych informacji o opinii, sposobie myślenia czy faktach na temat pytania. Pierwszym krokiem w analizie jest sprawdzenie liczby takich odpowiedzi. Analizuje się je nie tylko jako cechę pytania, ale również jako cechę respondenta. Pytania, które charakteryzują się dużą liczbą odpowiedzi beztreściowych, mogą zostać usunięte z dalszych analiz.

Należy także zliczyć liczbę odpowiedzi beztreściowych, jakich udzielił dany respondent w całym badaniu lub w konkretnym bloku ankiety. Jeśli takich odpowiedzi jest bardzo dużo (np. więcej niż 50%), może to świadczyć o braku zaangażowania respondenta lub braku opinii lub wiedzy (np. w przypadku pytań o atrakcyjność różnych modułów gazety, której respondent nie czytał). Takiego respondenta trzeba wykluczyć z dalszych analiz.

[46] Kuźmińska i Pazura, 2018; Kuźmińska i in., 2019; Maniaci i Rogge, 2014; Liu i Wroński, 2018a, Berinsky i in., 2014.

SO4: zbyt niski poziom zaangażowania respondenta

Niskie zaangażowanie respondentów może przejawiać się na **poziomie behawioralnym** w:

(1) **niespójnościach logicznych** (np. respondenci deklarują, że są bezrobotni w jednym pytaniu, ale w dalszej części ankiety odpowiadają: „Czuję się w pracy zawodowej wykorzystywany" zamiast „Nie dotyczy");
(2) **dziwnych odpowiedziach na pytania otwarte** (np. „Rząd nas okrada" jako odpowiedź na pytanie o sposoby wykorzystywania czasu wolnego);

Zakresy wykluczeń

Można rozważać dwa zakresy wykluczenia fałszywych respondentów: **globalne i lokalne**.

Wykluczenie lokalne zakłada, że respondent może odpowiadać niestarannie na określony blok pytań, ale na inne części ankiety udzielać odpowiedzi z należytą starannością. Natomiast przy **wykluczeniu globalnym** nieuważni respondenci w jednym bloku pytań są usuwani z całego badania.

Decyzja, jaki typ wykluczenia zastosować, zależy od długości i uciążliwości ankiety oraz liczby respondentów. W każdym przypadku należy oszacować wpływ wykluczenia, przeprowadzając analizy zarówno na całości danych, jak i po ich wykluczeniu.

Skala problemu

W badaniach opisanych w literaturze przedmiotu[47] odsetek FAŁSZYWYCH respondentów wahał się od **4 do 97,8%**. W przeprowadzonych pod moim kierunkiem analizach opisanych w rozprawie doktorskiej[48] wykazano, że odsetek respondentów oznakowanych jako „FAŁSZYWI" zależał od badania i rodzaju Sygnału Ostrzegawczego.

Analizowano rozkład częstości występowania czterech sygnałów ostrzegawczych SO dla wykluczenia w 7 zbiorach danych (N=2399 uczestników, którzy w przeważającej większości łączą studia na WZ UW z pracą zawodową i 287 pracowników z co najmniej trzyletnim stażem pracy). Wszystkie te badania zostały przeprowadzone za pomocą internetowego narzędzia SSA i udział w nich był ograniczony do osób przez nas zaproszonych (w przypadku dwóch badań stanowiły je osoby wylosowane z komercyjnego panelu badawczego, którego uczestnicy są nagradzani za udział w badaniach). Można więc powiedzieć, że wszyscy **uczestnicy badań** byli

[47] Johnson, 2005; Kurtz i Parish, 2001; Meade i Craig, 2012; Curran i in., 2010; Baer i in., 1997; Kabut, 2021.

[48] Kabut, 2021.

nastawieni **kooperatywnie**. W badaniach dostępnych w sieci bez żadnych ograniczeń procent fałszywych respondentów, np. klikających z czystej ciekawości, jest o wiele większy.

W przeprowadzonych analizach ustalono, że:

1. SO1 [zbyt krótki CZAS odpowiedzi] wykluczał z dalszych analiz średnio 6,9% (od 3,2 do 10,1% z medianą równą 7,1%) respondentów.
2. SO2 [zbyt duża liczba błędów nieuwagi – *Attention Check Questions-ACQ*] wykluczał z dalszych analiz średnio 3,7% (od 0,2 do 7,7% z mediana równą 3,1%) respondentów.
3. SO3 [zbyt duża liczba odpowiedzi beztreściowych lub zbyt niska wariancja odpowiedzi treściowych] wykluczał z dalszych analiz średnio 1,7% (od 0,6 do 5,3% z medianą równą 1%) respondentów.
4. SO4 [zbyt niski poziom zaangażowania respondenta na poziomie behawioralnym, przejawiający się w niespójnościach logicznych i dziwnych odpowiedziach na pytania otwarte, a także w deklaracji braku uważności] wykluczał z dalszych analiz średnio 17,1% (od 2,2 do 25,4% z medianą równą 18,2) respondentów.

Przeprowadzona analiza SO (sygnałów ostrzegawczych) nie pozwoliła jednoznacznie stwierdzić, który sygnał ostrzegawczy jest najważniejszy. Najczęściej stosowanym przez badaczy filtrem w ankietach internetowych jest SO1 (czas ogólny lub czas na poszczególnych stronach). Używanie tylko pierwszego sygnału: zbyt krótki czas odpowiedzi może nie pozwolić na wykrycie respondentów „przeklikujących" bez czytania, którzy robią sobie przerwy. Aby móc obliczyć SO2 i SO4 konieczne jest zaplanowanie odpowiednich pytań przed zebraniem danych.

Dwa sygnały – SO1 (czas) i SO3 (liczba odpowiedzi beztreściowych) warto analizować lokalnie, tj. osobno dla bloków pytań. Jeśli zmierzona wartość jest wyższa/niższa od progu ustalonego dla danego bloku, wszystkie odpowiedzi w tym bloku można zamienić na brakujące wartości. Metoda wykluczenia lokalnego stała się standardową procedurą stosowaną w naszych badaniach.

Sekcja 2
Rafy metodologiczne w diagnozowaniu ludzi

Drugą klasą sytuacji, w której ludzie „przekształcani są w liczby", jest diagnoza. W odróżnieniu od badań wykonywanych w paradygmacie *ceteris paribus*, gdzie wyniki dotyczące pojedynczej osoby są nieinterpretowalne, diagnoza skupia się na pomiarze cech, predyspozycji i kompetencji konkretnego pracownika.

Na pytanie, dla kogo diagnoza jest przeprowadzana, narzucająca się odpowiedź brzmi „dla pracodawców". Globalizacja i wysoki poziom bezrobocia uczyniła proces selekcji bardzo kosztownym, zarówno dla firm (ze względu na olbrzymią liczbę zgłoszeń), jak i dla kandydatów (konieczność równoczesnej aplikacji do wielu miejsc). Z przeglądu literatury[49] wynika, że pracodawcy (92% menedżerów HR) skarżą się na nadmiar e-aplikacji. Istnieje bogata literatura na temat operacjonalizacji kompetencji ważnych na różnych stanowiskach pracy, które są wykorzystywane w procesie rekrutacyjnym[50]. Narzędzia diagnostyczne są jednak potrzebne nie tylko na etapie rekrutacji czy awansowania pracowników, ale także wtedy, gdy chcemy usprawnić funkcjonowanie firmy.

O ile przygotowywanie narzędzi selekcyjnych stanowi wyzwanie mogące budzić problemy etyczne (trzeba być w pełni przekonanym o **trafności narzędzi**, żeby nie mieć wątpliwości, że nie odsiewamy cennych kandydatów do pracy), o tyle umożliwianie autodiagnozy pracownikom jest palącym wyzwaniem.

Klasyk[51] pisał: „**Zdumiewające, jak niewiele osób potrafi określić swoją metodę i styl pracy**. Większość nawet nie wie, że każdy z nas pracuje inaczej. Zapewne z tego powodu wiele osób powiela cudze metody pracy, a w konsekwencji osiąga mierne wyniki. [...] sposób naszego działania jest niepowtarzalny, wynika on bowiem z naszej osobowości. Niezależnie od tego, czy osobowość kształtują nasze geny czy wychowanie, proces formowania osobowości kończy się długo przed rozpoczęciem kariery zawodowej. Tak, jak nasze uzdolnienia i niedostatki, metoda i styl pracy są w jakiś spo-

[49] Woźniak, 2013.

[50] Np. w Assesment Center, por. wyczerpujący przegląd w: Woźniak, 2013.

[51] Drucker, 2006b.

sób zdefiniowane. Można je wprawdzie zmodyfikować, ale nie można ich całkowicie zmienić, w każdym razie nie jest to łatwe. Dobre wyniki osiągamy nie tylko wówczas, gdy wykonujemy pracę, do której predestynują nas wrodzone zdolności, lecz także wówczas, gdy metoda i styl pracy pozwalają wykonywać ją jak najlepiej".

Podstawowa teza brzmi, że wykonując pracę niedopasowaną do naszych predyspozycji pracownicy mogą ponosić koszty psychofizjologiczne, których nie muszą sobie uświadamiać. Represorzy mogą nie czuć zmęczenia, choć ich ciało i umysł mogą być zmęczone. Dlatego najbardziej zainteresowaną stroną w diagnozie własnych ograniczeń są sami pracownicy. Pracodawca też ponosi koszty braku dopasowania, które przejawiają się w zwiększonej absencji czy fluktuacji pracowników.

Narzędzia pomiarowe omawiane w tej sekcji są również wykorzystywane w badaniach zbiorowych. Do opisu tych narzędzi pomiarowych używane są synonimiczne pojęcia: **test, kwestionariusz, skala, inwentarz, ankieta, sondaż**. Cechą wspólną jest zadanie stawiane przed respondentem/badanym. Ma się ustosunkować do zbioru zdań – pytających lub oznajmujących – najczęściej za pomocą precyzyjnie określonej skali odpowiedzi/szacunkowej.

- **SKALA** powinna obejmować zestaw pytań mierzących tę samą zmienną latentną, np. Skala Aprobaty Społecznej.
- **KWESTIONARIUSZ** może składać się z kilku skal.
- Użycie terminu **INWENTARZ** może sugerować, że zawiera on wiele skal nastawionych na szerszą diagnozę.

„Test" i „ankieta" to pojęcia ogólne.

- **TEST** może zawierać pytania z prawidłowymi odpowiedziami np. test kompetencji arytmetycznych.
- **ANKIETA** to zbiór dowolnych pytań zamkniętych (z narzuconymi skalami odpowiedzi) i otwartych. Termin „sondaż" sugeruje badanie skierowane do masowego odbiorcy.

Wszystkie te terminy występują w literaturze przedmiotu często w sposób synonimiczny.

W tej sekcji przyjrzymy się problemom pojawiającym się przy „kwestionariuszowej" diagnozie.

Dobrze znane powiedzenie Gordona Allporta: **„Jeżeli chcemy wiedzieć, jak ludzie się czują – dlaczego ich o to nie zapytamy?"** zapoczątkowało rozwój samoopisowych technik pomiaru najróżniejszych cech. Tworząc narzędzia bazujące na samoopisie zapominamy jednak, że w licznych badaniach[52] pokazano, że często nie jesteśmy świadomi wpływów, jakim podlegamy, co jednak nie powstrzymuje nas od

[52] Por. przykłady w Aronson i Wieczorkowska, 2001.

przekonania, że potrafimy trafnie wskazać przyczyny naszych myśli, uczuć i zachowań. **Samoopisy odwołują się do wiedzy o Ja**. Dla niektórych osób odpowiedzi na pytania kwestionariuszy są wynikiem przemyśleń, inni nigdy wcześniej na ten temat nie myśleli, więc ich odpowiedzi powstają w czasie rzeczywistym. Jeden z moich znajomych ma wpisany w dowodzie kolor oczu: piwny. Kiedy wyraziłam zdziwienie (ma oczy szaro-niebieskie), odpowiedział: „Kazali wpisać, spojrzałem do lustra i pomyślałem, że są piwne". Różnimy się zarówno zdolnością, jak i skłonnością (chęcią wykorzystania tej zdolności) do autorefleksji (uczynienia siebie obiektem poznania), a te różnice wpływają na trafność pomiaru. Możemy przeżyć całe życie bez wglądu w mechanizmy leżące u podstaw naszych zachowań. Przyglądanie się sobie z punktu widzenia niezaangażowanego obserwatora dla wielu z nas wydaje się być niestety zadaniem sztucznym i zupełnie niepotrzebnym. W efekcie **nie jesteśmy świadomi mechanizmów obronnych, które zniekształcają nasz obraz Ja. Innych bardzo często postrzegamy dużo trafniej niż siebie**[53]. Pożytek z analizy samoopisów zależy od wielu czynników. Odpowiedzi mogą być relatywnie bardziej wiarygodne, jeśli pytania odwołują się do **postaw jawnych**[54], a diagnozowana osoba nie jest zainteresowana robieniem dobrego wrażenia, lecz zmotywowana do uzyskania wglądu w swoje predyspozycje i ma skłonność do samoobserwacji oraz autorefleksji.

W naszych badaniach 61 studentów odpowiadało na pytania dotyczące ich stylu działania, a miesiąc później na te same pytania z instrukcją określenia **jacy chcieliby być**. W ten sposób można było zauważyć, że niektóre ze skal SSA są bardziej podatne na aprobatę społeczną. Zanotowano ponad dwukrotny spadek wariancji wskaźnika dla skali odporności na przeciążenie i pedantyzm, co oznacza istnienie społecznie podzielanego ideału.

Chcemy być bardziej pedantyczni i odporni na przeciążenie niż jesteśmy.

Natomiast dla skali sekwencyjności zanotowano dwukrotny wzrost wariancji, co oznacza, że **część osób za pożądaną uważa symultaniczną realizację zadań, a część, że trzeba koncentrować się kolejno na pojedynczych zadaniach.**

Kolejnym przykładem są kwestionariusze mierzące inteligencję emocjonalną za pomocą samoopisu[55]. Techniki samoopisowe dotyczące zdolności emocjonalnych można porównać z zadawaniem pytań w teście inteligencji „szkolnej", takich jak: „Czy myślisz, że jesteś inteligentny?".

Mamy dużo większe zaufanie do pomiaru inteligencji bazującego na rozwiązywaniu zadań. Problem stanowi, że o ile łatwo jest zbadać zdolności matematyczne (bo znamy prawidłowe odpowiedzi), to **zdolności emocjonalne** przejawiają się przede

---

[53] Wieczorkowska, 2011.

[54] Na nasze odpowiedzi wpływają także prowadzone przez Holistyka pozawerbalne zapisy doświadczenia (połączenia tworzone w wyniku rejestracji częstości współwystępowania bodźców w otoczeniu). Do tych zapisów nasza świadomość (Analityk) ma dużo gorszy dostęp niż do zapisów informacji werbalnych (w tym naszych przemyśleń). Te drugie są podłożem do tworzenia się postaw jawnych, te pierwsze postaw utajonych. (por. Wieczorkowska, 2025).

[55] Salovey i in., 2004.

wszystkim w **wykorzystywaniu informacji kontekstowej** – w tym przede wszystkim komunikatów pozawerbalnych, których brakuje nawet w narzędziu audiowizualnym (np. odczucie pola energetycznego naszego rozmówcy), a co dopiero w kwestionariuszu opierającym się na komunikatach werbalnych. Bardziej subtelną metodą diagnozowania mogłaby być analiza niewerbalnych wskaźników, takich jak wyrazy twarzy, ruchy ciała czy zmiany w sposobie mówienia. Takie obserwacje mogą być cenne, ponieważ w porównaniu z samoopisami:

- wiele zachowań niewerbalnych słabiej poddaje się świadomej kontroli i cenzurze;
- zachowania niewerbalne mogą być mierzone bez zwracania uwagi diagnozowanej osoby na fakt dokonywania pomiaru.

Można sądzić, że narzędzia psychometryczne oparte **na samoopisie** okazały się być **ślepą uliczką** w nauce**,** czego dowodem jest **brak jakiegokolwiek postępu w psychometrii** w ciągu ostatnich kilkudziesięciu lat, ale choć **samoopis** to bardzo niedoskonałe narzędzie pomiaru, **jest jednak lepszy niż brak narzędzi**[56]. Powtórzymy: niedoskonałości pomiaru to problem, który dyskwalifikuje nasze wyniki tylko wówczas, kiedy ich nie dostrzegamy. Jeżeli zdajemy sobie sprawę z nieuniknionych zniekształceń, którym podlegają nasze samoopisy, możemy starać się o dokonanie odpowiednich korekt.

Osoby poddające się diagnozie – dokonując często kreacji swojego wizerunku – prezentują się w taki sposób, aby chronić samoocenę i spełnić oczekiwania pracodawcy. Jeden z moich znajomych aplikujących na stanowisko dyrektora chwalił się tym, jak dobre uzyskał wyniki w zakresie odporności na stres (konkurs wygrał, więc uzyskał swobodny dostęp do materiałów rekrutacyjnych). Obydwoje wiedzieliśmy, że ze stresem radzi sobie kiepsko, ale jako inteligentny człowiek wiedział, jak odpowiadać na pytania kwestionariusza.

Odpowiedź na pojedyncze pytanie wiąże się z dużym błędem pomiaru – dlatego w fizyce, chemii czy biologii przeprowadza się wiele powtórzeń pomiaru. Niestety, jest to zabieg możliwy jedynie w badaniach bytów nieożywionych – a dokładniej, w badaniu obiektów pozbawionych pamięci i wolnej woli. Ludzie pamiętają, że przed chwilą byli o coś pytani, a powtarzanie pytań wywołuje ich irytację. Od ankietera zadającego pytania oczekujemy przestrzegania zasad komunikacji kooperatywnej[57]. Dlatego zamiast zadawania tego samego pytania wielokrotnie, pytamy o różne powiązane kwestie. Nawet przy wypełnianiu dokumentów czy podań w urzędach, często pojawia się pytanie o wiek lub datę urodzenia, a zaraz potem o numer PESEL, który przecież zawiera już te dane. Przy tworzeniu **wskaźnika syntetycznego**, np. poprzez sumowanie lub uśrednianie odpowiedzi na kilka–kilkanaście pytań dotyczących

[56] Tak jak użycie kiepskiej maszyny do szycia jest lepsze niż szycie garnituru ręcznie, Gilbert, 2007.

[57] Grice, 1977.

pewnego zjawiska lub cechy, najczęściej dokonuje się dodatkowej weryfikacji jakości takiego wskaźnika, obliczając **alfę Cronbacha**. Jest to najpopularniejsza miara rzetelności (jednorodności) wskaźnika syntetycznego, pozwalająca na określenie spójności pozycji wchodzących w jego skład.

Przykłady kwestionariuszowych narzędzi diagnostycznych

Kwestionariusz NEO_FFI (*NEO-Five Factor Inventory*)

Najbardziej popularny kwestionariusz oparty jest na Pięcioczynnikowym Modelu Osobowości[58], który powstał na podstawie badań leksykalnych. Skrócona wersja NEO_FFI (*NEO-Five Factor Inventory*) składa się z 60 pytań pozwalających opisać respondenta na pięciu wymiarach: neurotyczności, ekstrawersji, otwartości na doświadczenie, ugodowości i sumienności. Alfa Cronbacha dla tych wymiarów, obliczona na grupie ponad 1000 studentów[59], wyniosła odpowiednio: 0,84; 0,70; 0,69; 0,81; 0,71. Żadna z tych 12-itemowych skal nie jest jednoczynnikowa. Ekstrawersja i otwartość na doświadczenie wykazują aż cztery czynniki. Nie dziwi to, gdy przyjrzymy się treści pytań. Przykładowo, aby uzyskać maksymalny wynik w skali otwartości na doświadczenie, należy zgodzić się z pięcioma pytaniami i nie zgodzić się z pozostałymi.

Odmienne procesy poznawcze uruchamiane przy zgodzie i zaprzeczaniu[60] bardzo często prowadzą do tego, że itemy pozytywne (wymagające zgody) i negatywne (wymagające zaprzeczenia) w analizie czynnikowej są separowane na osobne czynniki. Największym problemem pozostaje jednak **niejednorodność konstruktów teoretycznych**. W przypadku otwartości na doświadczenie aż ¼ pytań dotyczy zainteresowania sztuką/poezją. W skali sumienności zawarta jest zarówno potrzeba osiągnięć, odpowiedzialność, jak i pedantyczność. Taka niejednorodność utrudnia wyobrażenie sobie osoby, która uzyskała wysoki lub niski wynik na danej skali. Na ten problem zwracał uwagę już Allport[61], twierdząc, że uzyskane w analizie **czynniki** stanowią **wymiary osobowości uśrednionej**, która jest **całkowitą abstrakcją nieprzydatną** dla psychologa pragnącego badać osobowości konkretnych osób.

Warto podkreślić, że mierząc w ten sposób własności osoby, **tracimy związek z poziomem obserwacji zachowań, analizując jedynie statystyczne abstrakty.** Fetyszyzacja alfy Cronbacha jako miary wartości psychometrycznej wskaźnika sprawiła, że zapominamy, iż bardzo łatwym sposobem uzyskania wysoce jednorodnej skali o wysokim współczynniku rzetelności alfa, jest zadanie kilku pytań o to samo,

[58] Costa i McCrae, 1992; Siuta, 2006.
[59] Wieczorkowska i in., 2014; Turska, 2016.
[60] Por. np. badania nad asymetrią Wänke, Schwarz i Noelle-Neumann, 1995.
[61] Allport, 1939 za: Szarota, 2008.

sformułowanych w różny sposób. Nie gwarantuje to jednak w żaden sposób trafności takiego wskaźnika.

Czasami powinniśmy zoperacjonalizować konstrukt teoretyczny za pomocą pytań o szereg niezależnych przejawów interesującej nas zmiennej. Szukając wskaźnika ryzyka nadwagi możemy pytać o częstotliwość jedzenia słodyczy, picia piwa, jedzenia w nocy. Wszystkie te aktywności mogą prowadzić do nadwagi, występując samodzielnie. **Elementy wskaźnika syntetycznego nie muszą być skorelowane,** abyśmy mogli je zliczać lub uśredniać, jeśli pytania dotyczyły częstości. Podobnie zbudowany został syntetyczny wskaźnik aktywności ekonomicznej[62]. Pytał on respondentów o 8 możliwych przejawów zmiennej latentnej, m.in. o to, czy inwestują w usługi, handel, akcje, podnoszą swoje kwalifikacje, wykazują aktywność pozazawodową, mają plany na przyszłość itd. Jeżeli respondent odpowiedział twierdząco na co najmniej dwa pytania z ośmiu, był przypisywany do kategorii aktywnych ekonomicznie „lisów". Pozostali respondenci uznani zostali za bierne ekonomicznie „jeże". Poszczególne pytania lub pozycje takiej szczególnej skali są przejawami tej samej zmiennej latentnej, ale mogą występować zupełnie niezależnie od siebie. Na tych przykładach widać, że dla części wskaźników syntetycznych ich elementy nie muszą być ze sobą wysoko skorelowane, a w efekcie alfa Cronbacha dla takiego wskaźnika syntetycznego może być niska.

Nie należy więc – choć to się powszechnie czyni – utożsamiać wielkości alfy Cronbacha z **jakością wskaźnika syntetycznego**. Nie sposób odpowiedzieć, jaka musi być wartość alfy Cronbacha, ponieważ nawet bardzo wysoka alfa nie gwarantuje ani jednowymiarowości konstruktu, ani wysokiej trafności. Niestety, przy ocenie jakości wskaźników syntetycznych złożonych z wielu pytań zbyt często ograniczamy się do walidacji rzetelności, nie zastanawiając się nad trafnością pomiaru.

Problemy diagnozy za pomocą bardzo popularnych kwestionariuszy o wątpliwej trafności diagnostycznej przedstawimy na przykładzie czterech popularnych kwestionariuszy[63].

Test osobowości Myers–Briggs

Test osobowości Myers-Briggs (MBTI, *Myers-Briggs Type Indicator*) jest jednym z najczęściej stosowanych narzędzi do badania osobowości w zastosowaniach biznesowych (według szacunków około 2,5 miliona osób rocznie poddaje się temu testowi)[64]. Najpopularniejsza wersja MBTI składa się z 93 pozycji, w których uczestnicy dokonują wyboru między dwoma samoopisami. Wyniki są następnie klasyfikowane na czterech dychotomicznych wymiarach[65]: (W1) Ekstrawersja (E) vs. Intro-

[62] Czapiński, 1996.
[63] Kosslyn i Miller, 2019.
[64] Myers, 1995; Kosslyn i Miller, 2019.
[65] Quenk, 2009; Kosslyn i Miller, 2019.

wersja (I), (W2) Poznanie (S) vs. Intuicja (N), (W3) Myślenie (T) vs. Odczuwanie (F), (W4) Osądzanie (J) vs. Obserwacja (P).

Wynik oznacza zaklasyfikowanie do jednego z 16 typów (2^4=16typów), np. respondent zakwalifikowany jako ENFP wykazuje wyższe nasilenie: ekstrawersji, polegania na intuicji, odczuwania i obserwacji. Autorki testu[66] przygotowały narzędzie podczas II wojny światowej, aby pomóc kobietom odkryć, jakie role zawodowe będą dla nich najbardziej komfortowe i odpowiednie. Podstawą narzędzia była teoria opisana w książce szwajcarskiego psychiatry (i byłego współpracownika Freuda) Carla G. Junga „Typy psychologiczne", wydanej po angielsku w 1923 roku.

Mimo że MBTI (ze względu na wykorzystanie wyborów odpowiedzi) jest wolny od zniekształceń związanych z ***response bias*** (omówiony w poprzedniej sekcji **styl odpowiedzi = sposób wykorzystania skali**), zarzuca mu się, że[67]:

- nie opiera się na badaniach naukowych – w dużej mierze wyrósł on ze sformułowanej na podstawie intuicji i niewielu obserwacji klinicznych teorii Junga;
- założenia teoretyczne („intuicja" różni się od „odczuwania") są sprzeczne z najnowszą wiedzą psychologiczną pokazującą rolę emocji w przeczuciach;
- analiza czynnikowa nie odtwarza założonej czteroczynnikowej struktury, na której opiera się test;
- rozkłady analizowanych cech nie są dychotomiczne – najwięcej osób uzyskuje wyniki skupione wokół środka wymiarów;
- test ma niską rzetelność: powtórny pomiar daje zbyt często inne wyniki.

Mimo krytyki, MBTI jest wypełniany przez miliony ludzi każdego roku i nadal wykorzystywany w wielu firmach do oceny obecnych i potencjalnych pracowników. Jeszcze bardziej kontrowersyjny i jednocześnie bardziej popularny jest test 4-kolorowy.

Pseudonaukowy test 4-kolorowy

Sukces książki Thomasa Eriksona „Otoczeni przez idiotów" stał się jednym z największych skandali pseudonaukowych w najnowszej historii[68]. Jedyną podstawą teoretyczną, na którą powołuje się autor, jest koncepcja różnic indywidualnych w strukturze sieci „energii psychonicznej" sformułowana w 1928 roku przez Williama Moultona Marstona, a spopularyzowana prawie 30 lat później przez Waltera Clarka w postaci 4-kolorowego testu DISC.

---

[66] Katharine Cook Briggs i jej córka Isabel Briggs Myers.

[67] Pittenger, 2005; Lloyd, 2008; Damasio, 1999/2020; Sipps i in., 1985; Bess i Harvey, 2002; Pulver i Kelly, 2008; Kosslyn i Miller, 2019; Feldman-Barrett, 2022.

[68] https://soccermatics.medium.com/how-swedes-were-fooled-by-one-of-the-biggest-scientific-bluffs-of-our-time-de47c82601ad

Mimo ponad 60 lat stosowania podziału na 4 kolorowe typy:

- czerwony (dominujący, „zakręcony", skoncentrowany na rozwiązaniu),
- niebieski (analityczny, ostrożny, skrupulatny),
- zielony (cierpliwy, troskliwy, miły),
- żółty (ekstrawertyczny, kreatywny, werbalny),

nie istnieją żadne badania naukowe potwierdzające jego trafność. Choć Erickson twierdzi, że „80% osób ma dwa kolory!", nie przedstawia na to żadnych dowodów. Brak postępu w rozwoju wiedzy w ciągu tylu lat (podobnie jak w badaniach leków homeopatycznych) dyskwalifikuje ten test jako narzędzie.

Pomimo że kwalifikacje zawodowe Eriksona nie są w żaden sposób potwierdzone, popularność jego książek jest zatrważająca. Powstały już kolejne tomy opisujące osoby otoczone przez kłamców, psychopatów, wampiry, narcyzów... Książki sprzedają się w milionach egzemplarzy, ponieważ doskonale wykorzystują naszą **skłonność do szukania prostych wyjaśnień własnych problemów przez przypisanie ich cechom innych ludzi**, którymi jesteśmy otoczeni. Zdecydowanie odradzam lekturę tych książek. Lepiej już sięgnąć po horoskopy. ☺

SSA – Sondaż Stylów Aktywności[69]

Podstawowym celem stworzenia w 1994 roku Inwentarza Stylów Aktywności (ISA) było dostarczenie narzędzi pomiarowych dla różnych zmiennych teoretycznych opisujących różne aspekty sposobu organizacji działania. Jego internetowa wersja nosi nazwę Sondażu Stylów Aktywności – SSA. W odróżnieniu od Pięcioczynnikowego Modelu Osobowości (BIG FIVE)[70], który został opracowany na podstawie badań leksykalnych, Inwentarz Stylów Aktywności[71] stworzony w 1994 roku powstał w oparciu o obserwacje[72] różnych sposobów organizacji realizacji zadań.

Głównym wymaganiem psychometrycznym ISA/SSA jest spełnianie przez tworzone skale założeń modelu pomiarowego (tak jak to jest rozumiane w modelowaniu strukturalnym), a więc przede wszystkim tego, aby były one **jednoczynnikowe**. ISA składa się z pozycji zawierających przeciwstawne opisy zachowania dwóch osób: A i B oraz pytania: „Czy Twoje zachowanie/odczucia w tej sytuacji byłyby podobne bardziej do A, raczej do A, do B czy raczej do B?". Respondent może też wybrać opcję „trudno powiedzieć", która zawsze znajduje się **poza skalą odpowiedzi.**

Ten sposób formułowania pytań ma niezaprzeczalne zalety: respondent nie musi mieć doświadczeń związanych z konkretną sytuacją, o którą pytamy, a ponadto informacja, że ktoś tzn. A lub B zachował się w określony sposób, niejako legitymizuje to zachowanie, przez co osłabia się wpływ zmiennej aprobaty społecznej. Ważne jest,

---

[69] Wieczorkowska, 1998.

[70] Costa i McCrae,1992; Siuta, 2006.

[71] Wieczorkowska, 1998.

[72] Podobne podejście zastosowali Jurek i Olech (2017), wyodrębniając wymiary ich kwestionariusza na podstawie opinii praktyków.

że pytania dotyczą różnych reakcji na tę samą sytuację, dzięki czemu bardzo łatwo można sobie wyobrazić sposób zachowania osoby, która uzyskała wysokie/niskie wyniki na skali. Trudnością w konstruowaniu pytań zawierających binarne wybory jest fakt, że nie wszystkie interesujące aspekty da się przedstawić w formie prostej alternatywy.

SSA składa się z **kilkunastu bloków pytań.** Każdy blok zawiera 5–6 pytań będących wskaźnikami konkretnej cechy. Taki zestaw pytań nazywany jest **skalą/wymiarem**. Przy konstrukcji zestawu pytań dąży się do tego, aby liczba pytań diagnostycznych wymagających wskazania osoby A była równa liczbie pytań wymagających wskazania osoby B, co eliminuje wpływ tendencji do potakiwania[73]. Edycje SSA używane w badaniach w kolejnych latach są modyfikowane w zależności od celu badania i badanej próby. W ostatnich latach do skal opisujących styl aktywności/pracy zostały dodane m.in. skale do pomiaru: trzech potrzeb (afiliacji, dominacji, osiągnięć) i temperamentu (reaktywności, ekstrawersji, bilansu emocjonalnego w pracy i w czasie wolnym).

Odpowiedzi na pytania SSA są poddawane **procedurze wykrywania fałszywych respondentów**[74]. Pierwszym krokiem jest sprawdzenie liczby odpowiedzi beztreściowych TRUDNO POWIEDZIEĆ [TP], która jest analizowana nie tylko jako cecha pytania, ale także jako cecha respondenta. W pojedynczym pytaniu odpowiedź TP, jeśli jest dodatkowo związana z dłuższym czasem odpowiadania, może być wskaźnikiem elastyczności respondenta, ponieważ ze względu na niewyspecyfikowany wystarczająco w pytaniu kontekst respondent może uważać, że raz zachowuje się jak osoba A, a w innej sytuacji jak osoba B. W takich przypadkach odpowiedzi TP są przekodowywane na środek skali odpowiedzi. Najpierw jednak trzeba zliczyć liczbę odpowiedzi TP, jakich udzielił respondent – jeśli jest ich bardzo dużo (np. więcej niż 50%) to jest to wskaźnikiem lenistwa poznawczego lub lekceważenia badania i takiego respondenta trzeba usunąć z dalszych analiz.

W badaniach mojego zespołu do pomiaru przedziałowości stylu pracy wykorzystuje się najczęściej 4 skale SSA – **nazwy wymiarów opisują lewy kraniec wymiaru, czyli punktowość.**

Skala: Metodyczność

Wysokie wyniki otrzymuje osoba, która myśli o tym, co jest do zrobienia. Dzieli zadanie na części, planuje je w czasie i zaczyna realizować zadanie, gdy obmyśliła dokładnie sposób jego wykonania. Uważa, że podejmowanie decyzji powinno być metodycznym (ustrukturyzowanym i sekwencyjnym) procesem.

---

[73] Opisana po raz pierwszy przez Cronbacha tendencja do potakiwania oznacza skłonność respondentów do odpowiadania "prawda", „zgadzam się", „tak" niezależnie od treści pytania. Jest to wynik automatycznej tendencji umysłu do wyszukiwania informacji potwierdzających.

[74] Por. Kabut, 2021; Wieczorkowska i Wierzbiński, 2013.

Niskie wyniki otrzymuje osoba, która rozpoczyna zadania, nie wiedząc, jak je wykona, myśląc, że jakoś to będzie, nie analizuje ile jest do zrobienia i ile czasu jej to zajmie. Uważa, że w podejmowaniu decyzji ważne jest, aby nie stosować powtarzalnych schematów, tylko zostawić sobie pełną swobodę.

Skala: Sekwencyjność

Wysokie wyniki otrzymuje osoba, która denerwuje się, gdy musi równolegle myśleć o kilku różnych sprawach. Osoba nisko symultaniczna lubi koncentrować się tylko na jednym zadaniu jednocześnie. Gdy różne zadania konkurują ze sobą co do ważności, osoba nisko symultaniczna stara się najpierw skończyć to, co zaczęła.

Niskie wyniki otrzymuje osoba, która stara się mieć kilka rzeczy rozpoczętych równocześnie, aby „przerzucać się" z jednej na drugą. Gdy różne zadania konkurują ze sobą co do ważności, osoba wysoko symultaniczna stara się w jakiś sposób realizować je równolegle. Często przerywa ważną dla niej pracę, gdy pojawia się coś interesującego, choć nie związanego z tym, co robi.

Skala: Precyzja

Wysokie wyniki otrzymuje osoba, która dba o detale, lubi zadania, przy których trzeba zwracać uwagę na szczegóły. Jej wiedza jest bardzo dokładna, jeśli coś wie, to ze szczegółami. Niskie wyniki otrzymuje osoba, która pomija szczegóły, szukając ogólnego obrazu problemu. Bardziej zależy jej na ogólnym wyniku niż na szczegółach zadania, które ma wykonać. Jej wiedza jest mało precyzyjna, sporo wie, ale niezbyt dokładnie.

Skala: Rutynizacja

Wysokie wyniki otrzymuje osoba, która lubi wykonywać zadania według jasno określonej procedury. Lubi pracę, która wymaga ścisłego stosowania otrzymanych wytycznych co do sposobu jej realizacji. Męczy ją chaos i nadmiar informacji. Niskie wyniki otrzymuje osoba, która lubi mieć swobodę co do wyboru sposobu wykonania zadań. Lubi pracę, która pozwala realizować zadania za każdym razem inaczej. Męczy ją monotonia.

Warto dodać, że w badaniach wykazano dodatnie korelacje tych wymiarów z samokontrolą (brak prokrastynacji i kończenie rozpoczętych zadań), pedantyzmem i dobrą estymacją czasu. Szczególnie ten ostatni wymiar jest istotny w warunkach pracy.

Wskaźniki dla poszczególnych skal są jednoczynnikowe i bliskie poziomowi obserwacji – łatwo można sobie wyobrazić zachowanie np. osoby nisko metodycznej. Wymiary metodyczności, precyzji, sekwencyjności i rutynizacji korelują ze sobą, ale nie na tyle wysoko, aby można było jeden z nich usunąć. W indywidualnych diagnozach pracowników są np. wysoko metodyczne i mało precyzyjne osoby, choć jest ich zdecydowanie mniej niż osób precyzyjnych i metodycznych. Na potrzeby diagnozy psychologicznej pracownik otrzymuje wyniki na wymiarach cząstkowych, ponieważ one pokazują obszary konieczne/warte modyfikacji.

Na potrzeby zbiorczych analiz statystycznych skorelowane teoretycznie i empirycznie wymiary są agregowane we **wskaźniki drugiego rzędu**. Czynniki drugiego stopnia nie przekładają się tak łatwo na poziom obserwacji jak czynniki pierwszego stopnia, ale pozwalają na testowanie hipotez. Do analiz porównawczych często stosowany jest medianowy podział wskaźnika wyodrębniający grupę punktowców i przedziałowców.

Niedoskonałości pomiaru to problem, który dyskwalifikuje nasze wyniki tylko wówczas, kiedy ich nie dostrzegamy. Jeżeli zdajemy sobie sprawę z nieuniknionych zniekształceń, którym podlegają nasze samoopisy, możemy starać się o dokonanie odpowiednich korekt.

Tymczasem **biznes korzysta z niezweryfikowanych naukowo narzędzi diagnostycznych**, ponieważ są pięknie opakowane, z utajnionym algorytmem generowania informacji zwrotnej.

Naukowe narzędzia do diagnozy są ciągle poddawane testom i udoskonalane.

Przyszłością jest diagnoza wykorzystująca technologię[75].

Przykłady wykorzystania technologii w diagnozie

Monitoring w czasie pracy

Chociaż w badaniu eksperymentalnym możemy wykazać, że ekspozycja na ocenę zwiększa poziom stresu pracowniczego, mimo że klony utworzone w wyniku losowania wyeliminowały wpływ różnic indywidualnych, nadal nie wiemy czy lepszym rozwiązaniem byłoby, gdyby nowy pracownik (nazwijmy go Kowalski) pracował w izolacji, czy pod nadzorem kontrolera, który śledziłby jego działania.

Nowoczesne technologie, takie jak śledzenie aktywności na komputerze pracownika, sensory noszone w postaci zegarków czy elektrod, pozwalają na prowadzenie badań nawet na jednej osobie (ilościowe ESPERYMENTALNE studium przypadku). Każdego ranka możemy rzucić monetą (lub uruchomić odpowiednią aplikację): orzeł – Kowalski pracuje bez nadzoru, reszka – Kowalski jest nadzorowany. Codziennie monitorujemy poziom stresu, wydajność i liczbę popełnianych błędów. Po kilku tygodniach podsumowujemy wyniki, obliczając średnią osobno dla dni z nadzorem i bez.

Zebrane do tej pory dane mogą ujawnić, że wyniki Kowalskiego mogą być dla niego zaskakujące, ponieważ nasz wgląd w to, co nas stresuje, jest często bardzo ograniczony.

W badaniach grupowych korelacja między pomiarami fizjologicznymi a ocenami subiektywnymi jest bliska zeru. Dzięki takim pomiarom może się okazać, że nadzór będzie korzystny dla Zielińskiego, ale nie dla Kowalskiego. Ten wgląd jest możliwy tylko dzięki nowym technologiom.

---

[75] Wieczorkowska i Wilczyńska, 2019.

Korzystanie z czujników snu, monitorów tętna itp. w tak wygodnej formie jak zegarki lub opaski na nadgarstki, o których użytkownik łatwo zapomina, a także stosowanie coraz mniejszych kamer do śledzenia wzroku, sprawia, że badana osoba może szybko przyzwyczaić się do urządzenia i w rzeczywistości zapomnieć o tym, że jest testowana. To oczywiście istotne dla badania, ponieważ podczas eksperymentów zawsze chcemy, aby badany obiekt nie miał stałego wrażenia, że jest monitorowany lub kontrolowany, ponieważ takie wrażenie wprowadza dodatkową zmienną do badania i może wpływać zarówno na zachowanie, jak i wyniki pomiarów.

Zastosowanie nowoczesnych urządzeń pomiarowych to olbrzymia szansa na monitorowanie zmęczenia i przeciążenia poznawczego. Stres w miejscu pracy jest główną przyczyną złego stanu zdrowia i chorób. W Polsce w 2024 roku nastąpił wzrost (o 13,8% w porównaniu do 2023 roku) liczby zwolnień lekarskich[76] z powodu stresopochodnych zaburzeń psychicznych. Problemy, takie jak trudności z zasypianiem, koncentracją czy objawy fizyczne (np. bóle głowy, napięcie mięśni) są niestety powszechne. Najbardziej „narzekają" i cierpią młodsi i narażeni na silną presję zawodową.

Wierzymy, że dzięki technologii pracownicy mogą stać się bardziej świadomi momentów swojej największej wydajności i odpowiednio wykorzystywać ten czas. Ja poczułam się o wiele lepiej odkąd śpię „w zegarku", który mierzy pieczołowicie różne fazy snu. Rano sprawdzam czy przespałam „dobrym" snem co najmniej 90 minut – jeśli tak, to wiem, że ewentualne złe samopoczucie nie wynika z niewyspania.

Badanie różnic indywidualnych w zakresie predyspozycji do pracy pod presją, przywiązywania wagi do szczegółów czy czasu może stanowić ważny aspekt w procesie rekrutacji, zwłaszcza na stanowiska, które wiążą się z określonymi wymaganiami i charakterem pracy.

Nowe technologie mogą pomóc w zmniejszeniu liczby wypadków w miejscu pracy. Samochody zawodowych kierowców wyposażone w systemy monitorujące prędkość ich reakcji (zmiana biegów, hamowanie, użycie pedałów) i ruchy źrenic są coraz bardziej powszechne. Algorytm oceniający poziom zmęczenia kierowcy na podstawie pomiarów w czasie ciągłym może sugerować czas przerwy.

Coraz więcej firm chce monitorować pracę swoich pracowników biurowych za pomocą oprogramowania, które rejestruje to, co dzieje się na monitorze, robi zdjęcia, a także przygotowuje podsumowanie czasu spędzonego na korzystaniu z różnych aplikacji, odwiedzanych stron internetowych oraz klasyfikuje czas jako produktywny lub nieproduktywny. Przykładem może być badanie[77], w którym porównywano automatycznie monitorowaną **prywatną aktywność komputerową (*cyberloafingu*) w godzinach pracy** 379 pracowników w ciągu 4 miesięcy z ich deklaracjami. Korzystając z takiego oprogramowania można nawet wprowadzać reguły dotyczące

[76] https://businessinsider.com.pl/praca/zwolnienia-lekarskie-z-tytulu-zaburzen-psychicznych-alarmujace-dane-zus/mgcz84m

[77] Hensel i Kacprzak, 2020, 2021.

powiadomień, które będą wysyłane do bezpośredniego przełożonego, gdy zostaną uruchomione określone (zakazane) strony lub aplikacje – takie jak Facebook, Allegro, Skype – które z pewnością nie są używane do wypełniania zobowiązań zawodowych. Tego typu monitorowanie może prowadzić do bezprecedensowej utraty kontroli pracowników nad danymi ich dotyczącymi.

Taki monitoring może jednak budzić wiele zastrzeżeń natury etycznej. Gromadzone dane są wrażliwe i powinny być odpowiednio chronione przed dostępem osób trzecich. Istnieje ryzyko nadużyć. Przełożeni, mając dostęp do tak szczegółowych informacji o swoich pracownikach mogą dojść do wniosku, że np. Kowalski robi sobie za dużo przerw, więc należy mu dołożyć obowiązków. W wielu przypadkach takie działania mogą być nieskuteczne, a nawet szkodliwe, szczególnie, gdy menedżer ma złe intencje i brakuje mu wystarczającej wiedzy o różnicach indywidualnych (np. niektórzy pracują szybko i wydajnie, ale potrzebują długich przerw).

Nie ulega jednak wątpliwości, że narzędzia monitorowania pozwalają na uzyskiwanie ważnych danych naukowych. Przykładem może być eksperyment[78] porównujący wpływ w warunkach naturalnych dwóch interwencji: przypomnienia o możliwej karze oraz samej kary, którym objęto 230 pracowników przez dziewięć miesięcy.

Karanie osób naruszających politykę organizacyjną zawierającą zakaz prywatnej aktywności komputerowej w godzinach pracy (*cyberloafingu*), wpłynęło zarówno na ukaranych, jak i nieukaranych pracowników (spadek odpowiednio o 41 i 24%, utrzymujący się do końca prowadzonej obserwacji).

Warto jednak podkreślić, że narzędzia umożliwiające zaawansowane monitorowanie pracowników powinny przede wszystkim służyć osobom monitorowanym. Pracownicy mogliby omawiać swoje odczyty z coachami, aby zwiększyć samoświadomość na temat predyspozycji do wykonywania określonych zadań oraz momentów swojej największej wydajności lub największego obciążenia poznawczego. Takie dane z pewnością nie powinny być wykorzystywane w sposób opresyjny wobec pracowników. Neuroplastyczność naszych mózgów daje możliwość rozwoju pod warunkiem, że wiemy nad czym powinniśmy pracować i jesteśmy gotowi się rozwijać.

Tak, jak przełom w medycynie nastąpił nie dzięki nowym teoriom, ale dzięki postępowi technologicznemu, tak samo będzie w przypadku nauk społecznych. Pozostaje tylko cierpliwie czekać, intensywnie pracując z obecnie dostępnymi narzędziami. Dobry przyrząd diagnostyczny może pomóc nam uzyskać większą samoświadomość naszych predyspozycji, a tym samym lepiej zrozumieć koszty, które ponosimy podczas wykonywania naszej pracy.

[78] Hensel i Kacprzak, 2020, 2021.

Neuroobrazowanie, elektromiografia

Można określać odporność pracownika na presję czasową, pytając go o jego odczucia, można też mierzyć różne parametry fizjologiczne, takie jak zmiana tętna w sytuacji stresowej. Trzeba jednak pamiętać, że pomiary fizjologiczne, takie jak pomiar tętna i reakcji galwanicznej skóry (GSR – *galvanic skin response*), **są wskaźnikami ogólnego poziomu pobudzenia bez różnicowania jego rodzaju.** Nowe możliwości, które daje neuroobrazowanie mózgu, nie pozwalają jeszcze na masowe zastosowanie.

Dobrze wiedzieć, że to metoda mająca wiele ograniczeń, bo nawet specjaliści[79]:

1) podkreślają[80], że **biologia nie działa jak fizyka**, ani – tym bardziej – jak **inżynieria** (systemów biologicznych nie da się łatwo zredukować do odrębnych elementów składowych, które po połączeniu tworzą z powrotem całość) i twierdzą, że wielu naukowców zajmujących się mózgiem nie do końca zdaje sobie z tego sprawę, ponieważ ich opisy mózgu pełne są etykiet wskazujących, że dany region mózgu pełni funkcję X lub Y, tak jakby poszczególne części funkcjonowały quasi-autonomicznie. Analiza wyników neuroobrazowania mózgu przypomina obserwację wskazań szybkościomierza w jadącym samochodzie. Dowiadujemy się, z jaką prędkością porusza się samochód, ale nadal nie wiemy jak działa. Korelacje między aktywnością danego regionu mózgu a wykonywanym przez mózg zadaniem mogą być pozorne;
2) przypominają[81], że kiedy komórki mózgu lub neurony „odpalają", to często taka stymulacja prowadzi do sytuacji, w której sąsiedni neuron też odpala. Jednak podczas, gdy część neuronów uruchamia w ten sposób aktywację całego obszaru mózgu, inne działają hamująco i ją tłumią. Efekt jest taki, że na obrazie możemy **zobaczyć wzmożoną aktywację** jakiegoś obszaru, **kiedy w istocie aktywne są głównie połączenia hamujące, których funkcją jest tłumienie**, a nie stymulowanie aktywności sąsiadów;
3) twierdzą[82], że **nie istnieje proste odwzorowanie jeden-do-jednego** pomiędzy stanami psychologicznymi i aktywnością specyficznych obszarów mózgu. Czasem możliwe jest odkodowanie zawartości umysłu osoby przy użyciu fMRI, ale wymaga to wyrafinowanych analiz statystycznych i ostrożnej interpretacji;
4) podkreślają[83], że „plamy" na obrazie mogą być **też czarne**, mimo że w danym obszarze mózg wykazuje **podwyższoną aktywację – ale w mniejszej skali** niż woksel (to około trzy milimetry sześcienne), który mimo to może pełnić kluczowe funkcje;
5) wykazali[84], że gdy badani w skanerze potwierdzają fałszywą datę urodzin, można za pomocą fMRI wykryć to z ogromną dokładnością, o ile „oszuści" nie będą

[79] Wieczorkowska, 2025.
[80] Pessoa, 2024.
[81] Satel i Lilienfeld, 2017.
[82] Poldrack, 2022.
[83] Satel i Lilienfeld, 2017.
[84] Ganis i in., 2011.

stosować sztuczek. Wystarczy bowiem, że odpowiadając na pytanie wyobrażają sobie jednocześnie np. poruszanie palcem, aby dokładność wykrycia oszustwa spadła do 33%. Wstrzymanie oddechu lub potrząsanie głową powoduje wielkie zmiany w sygnale fMRI, znacznie większe niż jakiekolwiek zmiany spowodowane zadaniem, co oznacza, że zebrane dane bardzo łatwo mogłyby się stać zupełnie bezwartościowe.

Jeden z badaczy[85], który wykonał na sobie w ciągu 18 miesięcy 104 sesje skanowania fMRI, na pytanie czego się dowiedział, odpowiedział, że **żenująco mało**. Dla mnie ważnym wnioskiem z tych badań jest to, że różnice między skanami jego mózgu na przestrzeni 1,5 roku były o wiele mniejsze niż różnice między skanami mózgów różnych ludzi. Przykładem ostrzegającym przed bezrefleksyjną akceptacją wyników neuroobrazowania może być zgłoszenie na konferencję referatu, w którym przedstawiono wyniki skanowania mózgu łososia. Zadaniem umieszczonego w skanerze fMRI łososia było określenie, jakiej emocji doświadczała osoba przedstawiona na zdjęciu. Analiza zebranych w standardowy sposób danych wykazała **aktywację w mózgu łososia w reakcji na zadanie,** mimo że łosoś był martwy, zanim umieszczono go w skanerze. Autorzy referatu chcieli w ten sposób zwrócić uwagę na konieczność zmiany sposobu analizy. Szczegóły pomijam, bo nie są one tutaj istotne.

Nie możemy też zapominać o habituacji (*practice-suppression*) oznaczającej, że zapotrzebowanie na utlenowaną krew jest niższe u osób rutynowo wykonujących jakieś zadanie niż u tych, które mają z nim do czynienia po raz pierwszy.

Także **porównywanie reakcji niewerbalnych** różnych osób nie ma większego sensu, ponieważ różnice międzyosobnicze są w tym zakresie bardzo duże. Ekspresja niewerbalna niektórych osób jest bardzo ograniczona, u innych bardzo żywiołowa. Metaanaliza[86] wykonana na dużej liczbie dostępnych badań wykazała, że próbując rozpoznać cudzą emocję na podstawie wyrazu twarzy, po prostu zgadujemy – z wynikami na poziomie tylko nieco wyższym od losowego. Trochę lepsza, choć ciągle niska trafność, dotyczy oceny intensywności emocji, a nie jej charakteru. **Reakcja mięśni twarzy** towarzysząca stanom afektywnym może nie być możliwa do wykrycia gołym okiem, ale jest wykrywalna z użyciem techniki elektromiografii[87]. Każdy aktywny mięsień generuje słaby prąd elektryczny, który po zarejestrowaniu i wzmocnieniu pozwala na analizę związków reakcji elektromiograficznej z różnymi bodźcami wyzwalającymi. Reakcje emocjonalne są widoczne na twarzy, gdy obserwujemy mięsień jarzmowy (reagujący na sygnały pozytywne) oraz mięsień marszczący brwi (reagujący na sygnały negatywne). Niestety badanie **EMG (elektromiografia)** informuje nas tylko o **walencji**, nie dając możliwości odróżnienia różnych stanów afektywnych o tym samym znaku – pozytywnym lub negatywnym.

[85] Poldrack, 2022.
[86] Durán i Fernández, 2021; Nęcka, 2023.
[87] Larsen i in., 2003.

Interpretując sygnały niewerbalne, porównania możemy prowadzić tylko w oparciu o obserwacje poczynione „wewnątrz osoby" – analogicznie jak w badaniach na wariografie nie porównuje się zapisów zmian fizjologicznych z „normami", lecz porównuje się różnice „wewnątrz osoby", gdy odpowiada PRAWDA lub FAŁSZ na pytania o znanej i nieznanej prawidłowej odpowiedzi. Znając charakterystyczny dla danej osoby zapis reakcji fizjologicznych towarzyszący tym odpowiedziom, możemy próbować przewidywać, kiedy osoba kłamie.

Monitorowanie obciążenia allocentrycznego

Jednym z ważniejszych wskaźników fizjologicznych wykorzystywanych w diagnozie jest **zmiana poziomu pobudzenia**. Analizuje się aktywność dwóch działających antagonistycznie systemów autonomicznego układu nerwowego: (1) **współczulnego** związanego ze wzrostem pobudzenia, (2) **przywspółczulnego** związanego ze spadkiem pobudzenia. Oba te układy współdziałają w celu regulowania poziomu pobudzenia organizmu, tak aby był on adekwatny do wymagań sytuacyjnych. Długotrwałe, zbyt wysokie pobudzenie może prowadzić do zwiększonego obciążenia **allostatycznego**, czyli **skumulowanych skutków przewlekłego stresu**.

W badaniach analizuje się następujące wskaźniki fizjologiczne:

1) **zmiany średnicy źrenicy oka** – źrenice zwężają się, gdy pobudzenie wzrasta i rozszerzają się, gdy pobudzenie spada;
2) **galwaniczną reakcję skóry (GSR)** – odzwierciedlającą zmiany w potliwości skóry dłoni, które odpowiadają zwiększonej aktywności układu współczulnego;
3) **chwilową częstość tętna (*heart rate*, HR)** – liczbę uderzeń serca na minutę w danym momencie;
4) **zmienność rytmu serca (*heart rate variability*, HRV)** – wahania w czasie pomiędzy kolejnymi uderzeniami serca.

W sytuacji relaksu rytm serca charakteryzuje się stosunkowo dużą zmiennością, co świadczy o dynamicznym reagowaniu organizmu na zmieniające się wymagania wewnętrzne i środowiskowe (np. tętno przyspiesza podczas wdechu względem wydechu ze względu na zwiększone zapotrzebowanie na tlen). Pod wpływem stresorów **zmienność rytmu serca maleje**. Obciążenie allostatyczne podczas wykonywania zadań może się również wiązać ze zmianami: (1) poziomu hormonów w układzie endokrynologicznym (wzrost stężenia kortyzolu i katecholamin); (2) zmianami w układzie sercowo-naczyniowym (podwyższone ciśnienie tętnicze, zmiany w chwilowej częstotliwości pracy serca).

W naszym zespole, który założył Laboratorium Interaktywnej Diagnozy (LID) Wydziału Zarządzania szukamy lepszych niż samoopisy form diagnozy, wykorzystując m.in. **okulograf** rejestrujący **ruchy gałek ocznych, zmiany wielkości źrenicy** oraz opaskę zapisującą **zmiany rytmu serca, temperaturę nadgarstka, reakcję skórno-galwaniczną**.

Wskaźniki okulograficzne

W literaturze można znaleźć wyniki porównujące parametry okulograficzne w wyodrębnionych na podstawie pomiaru kwestionariuszowego grupach pracowników. Dokonując takich porównań wykazano[88], że pracownicy:

- o skłonnościach depresyjnych dłużej utrzymują wzrok na bodźcach negatywnych oraz neutralnych niż pozytywnych;
- o obniżonym nastroju częściej wracają do obrazu o zabarwieniu negatywnym;
- z zaburzeniami nastroju mają większe trudności przy utrzymaniu uwagi na zadaniu centralnym, co przejawia się w poszerzonym zakresie uwagi przy zadaniach wykorzystujących bodźce zarówno emocjogenne, jak i neutralne.

Neurotyzm pracownika przejawia się w krótszym czasie trwania oraz z szybszym spadkiem czasu trwania sakkad wraz ze zwiększaniem się trudności zadania. To wskazuje na wyższe obciążenie poznawcze oraz szybszy wzrost obciążenia poznawczego przy trudniejszych zadaniach u neurotycznych pracowników.

Stopień neurotyzmu jest związany z istotnie niższą intensywnością sakkadyczno-fiksacyjną oraz z jej spadkiem przy wzroście trudności zadania, co wskazuje, że przy trudniejszych neurotyczni pracownicy „pobierali mniej danych" z otoczenia. Co ciekawe ekstrawertywnych pracowników charakteryzowała niższa intensywność sakkadyczno-fiksacyjna niż osoby introwertywne.

W badaniach wykazano, że osoby o tendencjach depresyjnych:

- dłużej utrzymują wzrok na zdjęciach, które zawierają bodźce oceniane jako negatywne niż osoby bez takich zaburzeń[89];
- mają rozszerzony zakres uwagi także w przypadku prezentowania materiałów neutralnych[90].

Pracownicy z zaburzeniami nastroju charakteryzują się m.in**. poszerzonym zakresem uwagi** oraz częściej kierują wzrok na obrazy lub wyrazy o negatywnym zabarwieniu emocjonalnym. Przekłada się to także na wskaźniki okulograficzne, które wskazują na istotnie **większą liczbę fiksacji** lub rewizyt na bodźce negatywne w porównaniu z osobami z grupy kontrolnej. Czas sakkad u pracowników o wyższym poziomie **neurotyzmu** ulegał stopniowemu **skróceniu**, osiągając najniższą wartość **w warunkach dużego obciążenia poznawczego**. Średnia czasu fiksacji, czyli czasu w jakim diagnozowani mogli aktywnie przetwarzać bodźce wzrokowe, była najwyższa u neurotycznych pracowników w zadaniu, które w największym stopniu wymagało aktywności poznawczej.

---

88 Kołodziej i Brzezicka, 2015.

89 Caseras i in., 2007.

90 Kołodziej i Brzezicka, 2015.

Badania w Laboratorium Interaktywnej Diagnozy

Załóżmy, że rekrutujemy asystentów, którzy muszą być zdolni do wyłapywania najdrobniejszych literówek czy błędów, a także sprawnego przełączania się między różnymi zadaniami. Kandydaci do pracy świadomi tych oczekiwań mogą zniekształcać w pożądanym kierunku odpowiedzi, np. w skalach precyzji, symultaniczności, metodyczności i reaktywności SSA. Warto więc przetestować sposób wykonywania zadań przez pracownika, np. za pomocą przygotowanego przez nasz zespół[91] narzędzia diagnostycznego FOKUS składającego się z 5 zadań wykonywanych na komputerze:

- TE – Test elastyczności poznawczej
- TW – Test wyodrębniania figury z tła
- TZ – Test zestawiania
- TS – Test koncentracji na szczegółach
- TC – Test staranności czytania.

W ciągu 100 minut spędzonych na wielokrotnym rozwiązywaniu 5 zadań (konfiguracje są losowane) można określić m.in:

1) **szybkość uczenia** (poprawianie wyników w kolejnych iteracjach);
2) **szybkość męczenia się** (dynamika zmian wielkości źrenicy, temperatury nadgarstka, częstotliwość mrugania, długość sakkad i fiksacji[92], wskaźniki psychofizjologiczne na bazie zmienności rytmu serca);
3) **wpływ** na wykonanie i koszty **presji czasowej** (czasem wykonania zadań można manipulować) i presji społecznej (nasze badania pokazują, że stojąca za plecami osoba bardzo podnosi wskaźniki pobudzenia emocjonalnego).

Teoretycznie kandydaci do pracy nie powinni być zainteresowani fałszywą autoprezentacją, ponieważ to oni przede wszystkim będą ponosić koszty wykonywania pracy niezgodnej z ich predyspozycjami temperamentalnymi i poznawczymi. Czasami jednak chęć zdobycia pracy jest silniejsza. Prowadzone przez nas badania[93] pokazują, że pracownicy często nie są świadomi ponoszenia zwiększonych kosztów psychofizjologicznych.

Stwierdzono brak korelacji między subiektywną miarą stopnia zestresowania operacjonalizowaną za pomocą kwadratu emocji RAG[94] (*Russell's Affect Grid*) z miarą obiektywną operacjonalizowaną jako dynamika zmienności pulsu. Podczas wykonywania zadań w warunkach presji czasowej nie zmieniały się miary subiektywne (wskaźniki na kwadracie emocji), ale **zmiany wskaźników fizjologicznych** wskazywały na przeżywany **stres.**

---

[91] Wieczorkowska i in., 2018.

[92] Im wyższy stres, tym częstsze mruganie i tym częstsze wybieganie wzrokiem poza ekran.

[93] Nowak, 2019; 2021.

[94] Russel i in., 1989.

Diagnoza może uchronić pracownika przed podejmowaniem niedopasowanej do jego predyspozycji pracy. Osoby, które wolno się uczą i szybko męczą monotonią powinny być skierowane do innej pracy, ponieważ można przewidywać, że albo zaczną chorować, albo będą usiłowały zmienić pracę. W obu przypadkach koszty przyuczenia do wykonywania zadań będą stracone. Narzędzie diagnostyczne FOKUS pozwala nie tylko na ocenę osiąganych wyników i ponoszonych kosztów, ale także na poznanie sposobu wykonania zadania. Przykładowo w teście Stroopa, w którym trzeba wskazać kolor czcionki, abstrahując od treści słowa np. widząc słowo „czerwony" należy nacisnąć klawisz „**czarny**", a nie „**czerwony**", niektórzy z diagnozowanych wykorzystują **strategię patrzenia tylko na ostatnią literę wyrazu** – w ten sposób nie czytają wyświetlanego słowa.

W dotychczasowych badaniach nad obciążeniem poznawczym stosowano następujące parametry okulograficzne[95]:

- wielkość źrenicy oka,
- częstotliwość i czas trwania mrugnięcia oka,
- liczba, czas trwania fiksacji i położenie fiksacji.

Najczęściej wykorzystywanym parametrem jest **wielkość źrenicy**, która w stałych warunkach oświetleniowych zwiększa swoją średnicę nawet do 20%[96] wraz ze wzrostem trudności zadań wykonywanych przez pracownika.

Przykładowo, w naszych badaniach[97]:

- wielkość źrenicy **osoby X** malała podczas wykonywania zadań, poczynając od trzeciego. Wskazuje to na rosnące zmęczenie po ukończeniu drugiego zadania. Na początku każdego zadania wielkość źrenicy wyraźnie rosła, co może świadczyć o wyższym obciążeniu poznawczym – na przykład na skutek większego zaangażowania lub potrzeby zrozumienia sposobu wykonania zadania;
- wielkość źrenicy **osoby Y** była względnie stała w trakcie realizacji zadań, co sugeruje, że poziom zmęczenia pozostawał stabilny. Niewielkie rozszerzenie źrenic przy pierwszym i piątym zadaniu może wskazywać, że to właśnie te zadania wiązały się z największym obciążeniem poznawczym.

Dodatkowo można obliczać Indeks Aktywności Poznawczej (**ICA** – *Index of Cognitive Activity*) bazujący na obliczaniu niewielkich, krótkotrwałych rozszerzeń źrenicy, który okazał się być dobrym wskaźnikiem obciążenia poznawczego w zadaniach językowych oraz związanych z zapamiętywaniem ciągów liczb.

Przejawem zmęczenia może być:

- wzrastająca proporcja czasów fiksacji wzroku poza obiektami potrzebnymi do wykonania zadania;

[95] Tsai i in., 2007; Łazowska i in., 2014.
[96] Holmqvist, 2011 za: Łazowska i in., 2014.
[97] Wieczorkowska i in., 2018.

- rozszerzenie się źrenicy w drugiej połowie zadania w porównaniu do pierwszej połowy zadania;
- zmniejszenie się częstotliwości mrugnięcia powiekami.

Przejawem stresu może być:
- rozszerzenie źrenicy w czasie diagnozy w stosunku do poziomu bazowego;
- zwiększenie stosunku sakkad do fiksacji;
- uciekanie wzrokiem poza granice ekranu.

W badaniach testowania związków wskaźników okulograficznych z wynikami w SSA założyliśmy, że: przejawem **wysokiej reaktywności** może być gwałtowna zmiana wielkości źrenicy przy zmianie zadania [w porównaniu z odchyleniem standardowym wielkości źrenicy podczas całego badania].

Przejawem **metodyczności** może być:
- mała wariancja czasu fiksacji i sakkad podczas wykonywania testów;
- mała wariancja odległości pomiędzy kolejnymi fiksacjami;
- rzadkie fiksacje w miejscach, gdzie już wcześniej były fiksacje;
- podobne wzory przeszukiwania pola niezależnie od zadania;
- występowanie charakterystycznych wzorców podczas czytania instrukcji (np. trzykrotne skanowanie tekstu przed wciśnięciem przycisku: „dalej", skanowanie tekstu na dwie linijki naprzód przed jego dokładniejszym przeczytaniem itp.);
- planowanie strategii przed wykonaniem zadania: zadanie jest wykonywane dopiero po uprzednim wodzeniu wzrokiem po ekranie.

Przejawem zamiłowania do precyzji pracownika może być:
- kilkukrotna fiksacja w tym samym miejscu;
- stosunkowo częste fiksacje na wzorcu.

Trzeba pamiętać, że diagnoza powinna być zaprogramowana po dokładnej analizie zadań, jakie będzie wykonywał pracownik. Pozwoli to na określenie kluczowych umiejętności poznawczych wpływających na efektywność realizacji tych zadań. Warto sprawdzić, czy pracownik wcześniej wykonywał podobne zadania. Nawet, jeśli w badaniach okulograficznych widzimy, że pracownicy nie czytali dokładnie instrukcji, można to interpretować jako brak staranności, brak motywacji albo wcześniejszą znajomość treści instrukcji.

Zapisy okulograficzne i inne pomiary psychofizjologiczne pozwalające na przewidywanie predyspozycji pracownika do pracy wymagającej długotrwałej koncentracji uwagi, **koniecznie** należy konfrontować z pomiarem kwestionariuszowym oraz subiektywnymi odczuciami diagnozowanych pracowników.

Podstawowym parametrem diagnozującym pracownika jest szybkość uczenia się: skracanie czasu wykonania testów przy jednoczesnym zmniejszaniu liczby

popełnianych błędów oraz obniżaniu psychofizjologicznych kosztów mierzonych za pomocą okulografu (przede wszystkim wielkości źrenicy). W naszych badaniach *metodyczność* w rozwiązywaniu testów zmieniała się w zależności od rodzaju zadania, więc nie można oczekiwać prostej zależności między kwestionariuszowym pomiarem metodyczności a wykonaniem testu. Pojawił się też problem, że diagnozowani pracownicy nie widzieli poszukiwanego obiektu (np. „Gdzie jest Waldo?"), mimo że fiksowali na nim wzrok.

Inaczej było w przypadku analizy zmienności rytmu serca jako wskaźnika stresu. Okazało się[98], że na podstawie pomiaru kwestionariuszowego przeprowadzonego dwa tygodnie przed badaniem można było przewidzieć wpływ cech pracownika (mierzonych za pomocą SSA) na wzrost pulsu podczas wykonywania zadań w warunkach stresowych.

Podsumowując, **miary obiektywne** – takie jak wysokość zarobków, wskaźniki okulograficzne, zmienność rytmu serca – pozwalają dotrzeć do czynników nieuświadamianych przez pracownika, ale są silnie kontekstowe. Wysokość zarobków może być słabym wskaźnikiem satysfakcji finansowej, jeśli pracownik ma na przykład bardzo liczną rodzinę na utrzymaniu. Fiksacja wzroku na obiekcie nie oznacza, że obiekt ten jest świadomie dostrzegany.

Miary subiektywne zawierają holistyczną ocenę, ale są podatne na – opisane np. w sekcji 6 – zniekształcenia poznawcze. Po spotkaniu z zamożnymi koleżankami nasza satysfakcja finansowa może być znacznie niższa niż po wizycie w schronisku dla bezdomnych.

Warto więc w analizach łączyć miary obiektywne (niezależne od woli obiektu pomiaru) z subiektywnymi. Mimo wątpliwości co do trafności i rzetelności testu postaw utajonych IAT, liczne badania z użyciem tego narzędzia wywarły silny, pozytywny wpływ na rozwój metodologii mierzenia postaw.

Potrzebujemy także nowych metod badania postaw utajonych, bo dotychczasowe[99] badające zapisane w sieci neuronowej ewaluatywne skojarzenia z obiektem, wykazały jak ważne jest to, czego sobie nie uświadamiamy.

[98] Nowak, 2019; 2021.

[99] Np. https://implicit.harvard.edu/implicit/takeatest.html

Sekcja 3
Rafy metodologiczne w ewaluacji ludzi i efektów ich działań

W wielu dziedzinach naszego życia wartość obiektu **może być oceniona wyłącznie przez innych ludzi** (jurorzy w programie „Mam Talent", krytycy kulinarni, profesorowie oceniający prace studentów, sędziowie na zawodach łyżwiarstwa figurowego czy komisje decydujące o przyznaniu grantów badawczych). Ocena jakości produktu materialnego (np. samochodu, układów scalonych) czy usług (np. biura podróży, banku) jest o wiele łatwiejsza niż ocena produktu niematerialnego, takiego jak np. projekt badawczy, abstrakt wystąpienia konferencyjnego czy dorobek naukowy. Do oceny większości przedmiotów materialnych (takich jak samochody, farby...) istnieją precyzyjnie określone normy (np. parametry techniczne), w ocenach produktów niematerialnych (np. takich jak dzieło sztuki) proces obiektywizacji oceny jest bardzo dużym wyzwaniem.

Ewaluatorem lub sędzią nazywamy osobę oceniającą obiekty podlegające ewaluacji.

Obiektami ewaluacji mogą być kandydaci do pracy, teksty naukowe, granty badawcze, popełnione przestępstwa, zajęcia akademickie lub świadczone usługi. Ewaluacje mogą przyjmować formę werbalną, wtedy nazywamy je **opiniami** lub liczbową, gdy są **ocenami na skalach numerycznych** dostarczonych przez zlecającego ewaluację.

Głównym celem tworzenia narzędzi ewaluacyjnych jest **zwiększenie** obiektywizmu ocen poprzez skwantyfikowanie w ujęciu ilościowym proponowanych kryteriów[100]. Próbuje się też mierzyć to, co niemierzalne.

W naukach ścisłych przedmiotem pomiaru są byty realne (rzeczywiste wielkości np. długość położonych linii światłowodu). Istnieje **ustalony wzorzec** np. metra, który pozwala obliczyć, jak dobrze nasz pomiar odpowiada wzorcowi. W naukach

[100] Król i Kowalczyk, 2014.

społecznych natomiast przedmiotem pomiaru są byty wirtualne (konstrukty teoretyczne np. zaangażowanie pracowników, cechy czy stany afektywne). Istnieje zasadnicza różnica między pomiarem wielkości fizycznych a pomiarem konstruktów teoretycznych. Modele pomiarowe wypracowane w naukach fizycznych nie w pełni przystają do potrzeb pomiaru w naukach społecznych. **Nie dysponujemy wzorcem postawy czy inteligencji, z którym moglibyśmy porównać wyniki pomiaru**. Ekonomista, badając ilość pieniędzy dostępnych na rynku, będzie mierzył obiektywną wielkość zarobków.

Pojawia się pytanie: na ile my – ludzie, potrafimy różnicować to, co oceniamy? Zarówno w życiu codziennym, jak i w nauce, **pomiar** polega na **użyciu instrumentu** w celu przypisania przedmiotowi lub zdarzeniu **wartości na określonej skali**[101]. Ważymy się przy użyciu wagi łazienkowej, sprawdzamy temperaturę za pomocą termometru. Nawet, gdy nie mamy instrumentów pomiarowych, staramy się ocenić różnice w natężeniu różnych cech. Najczęściej zadajemy pytanie o różnice, mimo że od dawna wiadomo, że nie odczujemy różnicy w ciężarze między sztabką złota ważącą 1000 g a 1015 g, ale odczujemy różnicę między sztabką ważącą 100 g a 115 g. W obu przypadkach różnica wynosi 15 g.

Skąd ta zmiana naszej zdolności detekcji? Wykazano[102], że **ledwie dostrzegalna różnica** między dwoma obiektami na danym wymiarze oceny (wielkości, wagi, jasności, głośności) **jest stałą proporcją cechy**. Dla ciężaru ta proporcja wynosi 0,02 – więc abyśmy poczuli różnicę w porównaniu z kilogramową sztabką złota, musiałaby być ona co najmniej o 20 g cięższa. Prawo: „**Ledwie dostrzegalna różnica** między dwoma różnymi obiektami na danym wymiarze oceny (np. wielkości, wagi, jasności, głośności itp.) jest **proporcjonalna do natężenia**" – można też przedstawić w postaci logarytmicznej (w tej wersji to prawo Fechnera). Kolejne przyrosty intensywności bodźca muszą być coraz większe, aby można było odczuć zmianę. Jeżeli zna się PROPORCJĘ (wartość ułamka Webera dla ciężaru 0,02, dla głośności dźwięku 0,048) to można utworzyć skalę pozwalającą na przewidywanie stopnia, w jakim natężenie danej cechy musi się zwiększać, abyśmy mogli odczuć różnicę.

Podobnie oceniamy pieniądze – **ledwie dostrzegalna różnica w zarobkach** zależy od wyjściowej wysokości tych zarobków. Czy w związku z tym nasze oceny na skalach szacunkowych używanych powszechnie w badaniach ankietowych można uznać za skale ilościowe? To pytanie o stałą jednostkę pomiaru. Gdy analizujemy złotówki, nikt nie kwestionuje, że różnica między 10 a 11 jest równa różnicy między 10 000 a 10 001. Jednak, jak wspomniano wcześniej, w naszej percepcji są to różne wielkości.

Rygoryści metodologiczni twierdzą, że skale szacunkowe, np. w pytaniu o poczucie szczęścia, nie są skalami ilościowymi, ponieważ nie mają stałej jednostki pomiaru. Na przykład odległość między odpowiedziami [4 = szczęśliwy] a [3 = raczej szczęśliwy] oraz [2 = raczej nieszczęśliwy] nie jest taka sama, choć wynosi 1 punkt

[101] Kahneman i in., 2022.

[102] Mackiewicz, 2011.

na skali. Ci sami badacze zapominają jednak o wymaganiach stałości jednostki pomiaru, licząc średnią z ocen szkolnych. Nie ma dowodów na to, że różnice między oceną dobrą, bardzo dobrą i dostateczną są takie same [uśrednianie ocen od różnych wykładowców...].

Niestety wartość predyktywna opiera się na łatwiejszych do zmierzenia wskaźnikach obiektywnych (np. dochody w zł) niż wskaźnikach subiektywnych (np. satysfakcja z dochodów). Dlatego cały czas w badaniach musimy polegać na tym, co powiedzą nam respondenci.

Prosząc ewaluatorów o oceny, oczekujemy, że rozbieżności w ich ocenach będą mieściły się w pewnych granicach. Zakładamy, że mają te same informacje i te same cele, więc ich oceny nie powinny się znacznie różnić. W każdym systemie, w którym przyjmujemy, że ewaluatorzy np. recenzenci, są wymienni i często przydzielani losowo, duże różnice zdań na temat tej samej sprawy podważają oczekiwania co do sprawiedliwości i przewidywalności[103].

Kiedy jeden klient reklamujący uszkodzony laptop otrzymuje pełny zwrot pieniędzy, a inny – jedynie przeprosiny albo gdy jeden pracownik, który od pięciu lat pracuje w firmie i prosi o awans, otrzymuje go, a inny w identycznej sytuacji dostaje uprzejmą odmowę, odczuwamy, że system ewaluacji zawodzi.

Pułapka rozmowy kwalifikacyjnej

Rozmowa rekrutacyjna[104], czyli spotkanie przedstawiciela organizacji z kandydatem, podczas którego osoby te prowadzą dialog dotyczący kandydata i proponowanej przez organizację pracy, jest najczęściej stosowaną techniką w procesie rekrutacji[105]. Celami weryfikacyjnymi rozmowy kwalifikacyjnej są ocena kompetencji, zaangażowania oraz dopasowania do kultury kandydata[106] czy kultury organizacji?

Wiele badań pokazuje niską trafność selekcyjną wyników rozmów kwalifikacyjnych. Klasyczne badanie wykazało, że gdy 12 doświadczonych ewaluatorów oceniało przydatność kandydatów na handlowców na podstawie rozmowy kwalifikacyjnej, okazało się, że „wśród 57 kandydatów ta sama osoba była jednocześnie oceniana jako najlepszy i najgorszy kandydat”[107].

W badaniach nad predyktorami osiągnięć studentów na studiach doktoranckich[108] lepszymi wskaźnikami niż rozmowa kwalifikacyjna okazały się średnia ocen oraz wyniki egzaminu GRE. Oba te wskaźniki są częściowo niezależne od siebie, więc ich połączenie zwiększa trafność prognostyczną. Wykorzystanie listów polecających dodatkowo podnosi trafność formułowanych przewidywań.

---

103 Kahneman i in., 2022.
104 Woźniak, 2013.
105 Armstrong, 2011; Woźniak, 2013.
106 Armstrong, 2011; Woźniak, 2013.
107 Schulz, 2002.
108 Nisbett, 2016.

Warto podkreślić, że badanie trafności predykcyjnej jest zawsze ograniczone przez selekcję kandydatów (nie wiemy, jak poradziliby sobie ci, których nie przyjęto). Tymczasem korelacja między przewidywaniami opartymi na półgodzinnej rozmowie kwalifikacyjnej a ocenami osiągnięć studentów studiów magisterskich i doktoranckich jest znacznie niższa.

Podobnie niską trafność prognostyczną rozmów kwalifikacyjnych zaobserwowano w badaniach dotyczących oficerów wojskowych, biznesmenów, studentów medycyny, wolontariuszy Korpusu Pokoju i innych grup poddanych takiej analizie. Charakterystyczną cechą ewaluatorów prowadzących rozmowy kwalifikacyjne jest przecenianie swoich umiejętności oceny kandydatów. Niestety, często obniżają oni trafność własnych przewidywań, przeceniając wartość rozmowy kwalifikacyjnej w porównaniu do innych, bardziej istotnych źródeł informacji.

Ludzie znacząco przeceniają wartość rozmowy kwalifikacyjnej, uważając, że pozwala ona trafniej przewidywać osiągnięcia akademickie w college'u niż średnia ocen ze szkoły średniej. Ponadto sądzą, że jest lepszym prognostykiem jakości pracy w Korpusie Pokoju niż listy polecające, które opierają się na wielogodzinnych obserwacjach kandydata.

Przegląd 41 badań[109] dotyczących trafności ewaluacji rewidentów, patologów, psychologów, dyrektorów i przedstawicieli innych zawodów sugeruje brak rzetelności wystawianych ocen (brak spójności ocen ewaluatorów pojawia się nawet w sytuacji, kiedy ten sam przypadek jest oceniany ponownie kilka minut później).

Trzeba jednak pamiętać, że oprócz funkcji preselekcyjnych (do których rozmowa kwalifikacyjna słabo się nadaje), pełni ona również funkcje: motywacyjną i informacyjną. Powinna zachęcić do pracy „właściwych" kandydatów i zniechęcić pozostałych. Rola realistycznego przedstawienia wymagań stanowiska jako skutecznego narzędzia do zmniejszenia odejść pracowników po rozpoczęciu pracy została szeroko opisana w literaturze[110]. Ważnym aspektem dla kandydatów jest również sam styl prowadzenia wywiadu – kandydat buduje sobie opinię o firmie także na podstawie sposobu, w jaki był traktowany podczas rozmowy kwalifikacyjnej. Dążenie do traktowania kandydata z szacunkiem i profesjonalną rzetelnością jest niezbędne dla uzyskania pozytywnej opinii o firmie. Przeprowadzając rozmowę kwalifikacyjną, powinniśmy być świadomi zawodności naszych odczuć – jak łatwo ulegamy efektowi zakotwiczenia, traktując ją bardziej jako nawiązanie kontaktu niż obiektywną ocenę.

Przyszły laureat Nagrody Nobla Daniel Kahneman podczas służby wojskowej opracował metodę oceny przydatności poborowych do wojska[111]. Dla każdej z sześciu cech, takich jak odpowiedzialność, towarzyskość czy poczucie dumy, uznanych za istotne dla walki na froncie, przygotował zestaw konkretnych pytań dotyczących

[109] Shanteau, 1988.
[110] Breaugh, 2008.
[111] Kahneman, 2012.

życia w cywilu, m.in. częstotliwości zmiany pracy, uprawiania sportu, punktualności oraz metodyczności w nauce. Celem była możliwie obiektywna ocena poborowego.

Ewaluatorzy, dokonując szczegółowych ocen, byli proszeni o zamknięcie oczu i wyobrażenie sobie ocenianego rekruta w roli żołnierza, a następnie o określenie na pięciostopniowej skali jego przydatności. Kahneman pisze, że zarówno suma sześciu ocen cząstkowych, jak i intuicyjna ocena z zamkniętymi oczami formułowana po uporządkowanym gromadzeniu obiektywnych informacji i zdyscyplinowanej ocenie poszczególnych cech, pozwalały przewidzieć późniejsze wyniki poborowych.

Liczbom, niezależnie od ich pochodzenia, przypisuje się niemal magiczną moc, ponieważ bardzo łatwo je porównywać. Gdy musimy dokonać wyboru spośród dwóch kandydatów na dane stanowisko pracy, możemy długo dyskutować, który z nich jest lepszy. Jeśli jednak przypiszemy kandydatom liczby, które można poddać różnym działaniom arytmetycznym, proces decyzyjny staje się znacznie prostszy.

Nawet Danielowi Kahnemanowi nie przeszkadza fakt, że tak uzyskane liczby nie pochodzą ze skali pomiarowej, która pozwalałaby na liczenie średniej (co wymagałoby wykazania, że jednostka pomiaru jest stała). Jego zdaniem proste algorytmy liczbowe są lepsze niż sądy ekspertów, nawet jeśli liczby wprowadzone do algorytmu pochodzą właśnie od tych niedoskonałych ekspertów.

Dalej przyjrzymy się problemom, które pojawiają się, gdy eksperci generują ewaluacje liczbowe.

W ewaluacji ODRĘBNEJ – dla każdego obiektu wybierany jest niezależnie zespół ewaluatorów. W ewaluacji SERYJNEJ – wszystkie konkurujące ze sobą obiekty są oceniane przez ten sam zespół ewaluatorów. W wariancie SZTYWNYM nie można modyfikować wcześniejszych ocen (np. natychmiastowe ogłaszanie wyników egzaminu ustnego lub ocen sędziów oceniających występy sportowców w zawodach łyżwiarstwa figurowego). W wariancie ELASTYCZNYM pozwala się ewaluatorowi zmieniać kolejność oceny i poprawiać wcześniejsze oceny (przykładem są oceny trzydziestu wypracowań przez nauczyciela).

Zacznijmy od bardzo ciekawego przykładu badania sędziów ferujących wyroki skazujące.

Wyroki sędziowskie[112]

W trwającym 90 minut badaniu poproszono 208 aktywnych zawodowo sędziów federalnych z całych Stanów Zjednoczonych o ustalenie wyroków dla oskarżonych uznanych za winnych w 16 hipotetycznych procesach sądowych. Scenariusze dotyczące napadów rabunkowych lub oszustw różniły się pod niektórymi względami, np. czy oskarżony był głównym sprawcą, czy tylko pomocnikiem; czy była to recydywa; jakiej użyto broni...

[112] Kahneman i in., 2022.

Wyniki tego rodzaju badań przedstawia się w tabeli z 16 kolumnami (dla każdego procesu sądowego jedna kolumna) oraz 208 wierszami (dla każdego sędziego jeden wiersz).

Na podstawie analizy takiej tabeli można policzyć:

- średnią ze wszystkich orzeczonych wyroków kary więzienia we wszystkich 16 procesach sądowych. W przeprowadzonym badaniu wyniosła ona **7 lat;**
- średnie dla każdej kolumny przedstawiające „**sprawiedliwy wyrok”** będący uśrednieniem dla każdego wykroczenia wyroków wydanych przez 208 sędziów. Średnie kary w kolumnach dla 16 wykroczeń zmieniały się od 1 do 15,3 lat;
- średnie dla każdego z 208 wierszy obrazują tendencyjność każdego z 208 sędziów. Ich różnice pokazują zróżnicowanie ewaluatorów (*level noise*).

W idealnym świecie, w którym nie występuje tendencyjność sędziów, wszystkie wyroki w danej kolumnie powinny być identyczne, ponieważ wszyscy sędziowie orzekaliby taki sam wyrok. Niestety, wszyscy wiedzą, że jedni sędziowie są bardzo surowi, inni bardziej pobłażliwi.

Poszukiwanie korelatów tendencyjności sędziów (potwierdzanych przez ich reputacje jako bardziej surowych lub bardziej pobłażliwych) przyniosło dość oczywiste ustalenia:

- długość wyroku zależy od wyznawanych przez sędziego wartości: sędziowie, którzy wierzą w resocjalizacyjną rolę więzienia zasądzają krótsze wyroki niż ci, którzy kładą większy nacisk na odstraszającą rolę surowych kar lub konieczność izolowania przestępców od społeczeństwa;
- sędziowie pracujący na południu Stanów Zjednoczonych orzekali znacznie dłuższe kary więzienia niż ich koledzy z innych części kraju.

Podsumowując: tendencyjność sędziego przejawiająca się jako średni poziom orzekanych kar, przypomina cechy osobowości – koreluje z czynnikami genetycznymi, doświadczeniami życiowymi i nie zależy od konkretnej sprawy sądowej, w której sędzia orzeka.

Zróżnicowanie obiektowe (p*attern noise*) przeanalizujmy na przykładzie sędziego X, który w procesie C zasądził karę więzienia 4 lat.

- Analiza wszystkich danych pokazuje, że średni **(„sprawiedliwy”) wyrok** wydany przez 208 sędziów dla wykroczenia C wynosi **3,7 roku**. Średnia tendencyjność **sędziego X** to **5 lat** (czyli o dwa lata mniej niż średnia ogólna wynosząca 7 lat, więc sędzia X jest bardziej pobłażliwy niż reszta).
- Gdyby na wyrok sędziego X wpływała wyłącznie jego pobłażliwość, powinien orzec w sprawie C karę **1,7 roku** (3,7 – 2). Jednak X orzekł **4 lata**, co oznacza, że w tej sprawie sędzia X wydał wyrok surowszy niż zwykle.

Wniosek? Poszczególni sędziowie nie są jednakowo surowi we wszystkich sprawach: w niektórych orzekają surowiej niż wskazywałaby średnia ich wyroków, a w innych łagodniej. Takie odchylenia niezależne od ogólnego poziomu surowości

danego sędziego, nazywamy **zróżnicowaniem obiektowym** (*pattern noise* określanym też błędem/szumem prawidłowościowym). Może to oznaczać na przykład, że zwykle surowy sędzia stosunkowo łagodniej traktuje wykroczenia popełniane przez osoby wykształcone. Inny, na ogół łagodny sędzia, może takie osoby karać bardziej surowo. Jeszcze inny może być wyjątkowo surowy, gdy ofiarą przestępstwa jest osoba starsza. Statystycznym określeniem dla tego rodzaju zróżnicowania jest **„interakcja między sędzią a wykroczeniem".**

Przytoczyliśmy ten przykład[113], aby pokazać jak ważna jest analiza cech ewaluatora. W następnej części wrócimy do problemów związanych z ewaluacją mniej drastycznych kwestii niż skazywanie na więzienie, takich jak ocena projektów badawczych, artykułów naukowych czy zajęć akademickich. Zadziwiające jest to, że praca ewaluatorów w nauce tak rzadko bywa poddawana ocenie.

Styl ewaluatora można określić jedynie w sytuacji, gdy ocenia on wiele obiektów (tak, jak w omawianym wcześniej przykładzie, gdy każdy sędzia ferował wyroki w tych samych 16 procesach). Styl ten może zostać zdefiniowany za pomocą miary tendencji centralnej (np. średniej arytmetycznej) oraz miary zróżnicowania (np. odchylenie standardowe lub rozstęp) rozkładu wystawionych ocen.

Najczęściej występujący schemat ewaluacji odrębnej, gdy np. recenzent określa tylko jeden artykuł, nie pozwala na wnioskowanie o stylu ewaluatora.

Miara **tendencji centralnej rozkładu ocen ewaluatora** (jego **tendencyjność**) może być interpretowana jako pozycja w wymiarze surowości – łagodności/pobłażliwości.

Rozbieżności spowodowane tendencyjnością ewaluatorów można znaleźć w dowolnym zadaniu wymagającym formułowania ewaluacji[114]. Przykładami mogą być oceny okresowe pracowników (bo jedni kierownicy oceniają łagodniej od innych), prognozy udziału rynkowego (ponieważ niektórzy progności są większymi optymistami) czy skierowania na operację pleców (bo jedni ortopedzi preferują bardziej agresywne leczenie niż inni).

Systematyczne różnice indywidualne w tendencyjności ewaluatorów wykazano w wielu badaniach[115], m.in. w ocenach kandydatów na lekarzy, policjantów, pielęgniarki, pracowników społecznych[116], mechaników samolotowych[117]. W interpretacji tych wyników wskazuje się na różnice w sposobie przetwarzania informacji, cechy osobowości oraz uwarunkowania sytuacyjne[118].

Podsumowując, można stwierdzić, że styl oceniania ewaluatora przejawia się w stopniu:

- **pobłażliwości** operacjonalizowanej przez miarę tendencji centralnej np. średnią – niektórzy ewaluatorzy oceniają obiekty średnio wyżej niż inni;

[113] Kahneman i in., 2022.
[114] Kahneman i in., 2022.
[115] Dewberry i in., 2013.
[116] Kane i in., 1995.
[117] Borman i Hallam, 1991.
[118] Dewberry i in., 2013; Judge i Ferris, 1993.

- **różnicowania ocen** – niskim, gdy wszystkie opiniowane artykuły są oceniane podobnie: równie dobrze, równie źle lub równie umiarkowanie[119], w tym także skrajności ocen – ewaluator unika ocen skrajnych lub wyłącznie ich używa[120].

Bardzo wiele zależy od skali liczbowej dostarczonej ewaluatorom – co omówiliśmy wcześniej. Tutaj przedstawimy wyniki badań przeprowadzonych w moim zespole.

Tendencyjność ewaluatorów w schemacie ewaluacji odrębnej

Wśród 33 ewaluatorów[121] oceniających w schemacie ewaluacji odrębnej średnio 61 abstraktów na międzynarodową konferencję (nagrodą było stypendium pokrywające koszty pobytu i podróży) były zarówno osoby, których średnia z kilkudziesięciu ocen na skali dziesięciostopniowej wyniosła 8,40, jak i takie, dla których średnia wyniosła 3,44. Sukces aplikacyjny zdecydowanie zależy zatem od tego, który z ewaluatorów będzie oceniał abstrakt.

W kolejnym badaniu[122] analizowano oceny ewaluatorów 126 projektów badawczych konkurujących o granty w wysokości do 500 tys. zł w programie LIDER. W tym programie większość ewaluatorów (67%) recenzowała tylko jeden projekt (schemat ewaluacji odrębnej).

Procedura składała się z dwóch etapów:

- dla każdego projektu komisja oddzielnie wybierała zespół trzech ewaluatorów (recenzentów), którzy mieli ocenić projekt na skali stustopniowej,
- po zapoznaniu się z recenzjami komisja konkursowa zapraszała na rozmowę liderów projektów, które uzyskały średnią co najmniej 65 punktów.

Do oceny 126 projektów zaangażowano 252 ewaluatorów, z których większość (67%) recenzowała tylko jeden projekt. Analiza ocen przyznanych różnym projektom przez tych 252 recenzentów pokazała, że 13 recenzentów wystawiło projektom oceny poniżej 20 punktów (na skali stustopniowej), a 54 recenzentów wystawiło oceny powyżej 80 punktów. W takim przypadku trudno stwierdzić czy recenzent, który ocenił projekt na 2,5 (jeden przypadek), był nadmiernie surowy, czy też trafił na bardzo słaby projekt i pobłażliwych współrecenzentów.

W sytuacji, gdy ewaluator ocenia tylko jeden projekt, nie można określić czy jego ocena wynika z surowości przyjętych standardów, czy raczej z niskiej wartości ocenianego projektu. Aż w przypadku 48 projektów (co stanowi 38%) **różnica między**

[119] Król i Kowalczyk, 2014.
[120] Clarke, 2000.
[121] Kowalczyk, 2019.
[122] Wieczorkowska i Kowalczyk, 2021; Kowalczyk i Wieczorkowska, 2022.

oceną maksymalną a minimalną wyniosła ponad 30 punktów na stustopniowej skali. Niska ocena przyznana przez jednego z ewaluatorów mogła wyeliminować projekt.

Przykładowo lider projektu ocenionego przez ewaluatorów na **90, 80 i 10** punktów (średnia < 65) nie zostanie zaproszony na spotkanie, podczas gdy oceny **66, 65 i 65** punktów (średnia > 65) dają już szansę na zaprezentowanie projektu przed komisją.

Dla celów dydaktycznych policzono ranking projektów oparty na dwóch ocenach (wykluczono ocenę najbardziej surowego ewaluatora). Wybrano 14 najwyżej ocenionych projektów (średnia > 66), które zyskałyby znacząco, gdyby ocena najbardziej surowego ewaluatora została pominięta. W przypadku projektu #7 „wylosowanie" surowego ewaluatora wyeliminowało potencjalnie innowacyjny projekt z możliwości ubiegania się o finansowanie na drugim etapie konkursu.

Warto podkreślić, że decyzja końcowa zależała od komisji eksperckiej, przed którą prezentował się lider projektu. Niestety lider projektu #7 nie otrzymał takiej szansy, ponieważ przyznanie 10 punktów przez jednego z trzech ewaluatorów spowodowało, że uzyskana średnia < 65 uniemożliwiła mu znalezienie się na liście zaproszonych.

Nawet jeśli projekt przekroczył próg 65 punktów, otrzymując oceny 90, 80 i 25, jego niska średnia mogła stanowić dla członków komisji konkursowej podstawę do dalszej oceny.

Tendencyjność ewaluatorów projektów badawczych w schemacie ewaluacji seryjnej

Przedmiotem kolejnej analizy[123] były oceny ewaluatorów oceniających serię projektów badawczych dla młodych naukowców w ramach programu START przyznającego roczne stypendia naukowe w wysokości ok. 24 tys. zł. Liczba ocenianych projektów zależała od dziedziny i wahała się od 16 projektów matematycznych do 59 projektów chemicznych. W każdej z sześciu dziedzin wszystkie projekty oceniało trzech ewaluatorów.

Dzięki temu, że każdy ewaluator oceniał serię co najmniej 16 propozycji badawczych, możliwe było określenie jego stylu oceny (średnia, mediana, zmienność). Przeprowadzone analizy pokazały, że ewaluatorzy różnią się średnimi ocenami (niektórzy byli bardziej pobłażliwi niż inni), jak i zmiennością ocen. Przykładowo oceny dwóch ewaluatorów (#1 i #2) oceniających ten sam zestaw 22 projektów w fizyce przy użyciu tej samej skali 0–5 znacząco się różniły.

[123] Wieczorkowska i Kowalczyk, 2021; Kowalczyk i Wieczorkowska, 2022.

Ewaluator #1 użył do oceny 22 projektów siedmiu wartości ze skali: 2; 2,5; 3; 3,5; 4; 4,5 oraz 5.

Ewaluator #2 ocenił te same 22 projekty, wykorzystując tylko trzy wartości ze skali: 4; 4,5; 5 i przyznając najwyższą ocenę dziewięciu projektom, najniższą zaś ocenę 4 – trzem projektom.

Ewaluator #1 był w swoich ocenach o wiele surowszy od drugiego – aż dziesięć projektów ocenił poniżej 4. Nie ma wątpliwości, że wolelibyśmy być oceniani przez eksperta #2 niż eksperta #1. Jednakże w przypadku ewaluacji seryjnej wpływ surowego ewaluatora obniża oceny wszystkich projektów konkurujących w tym samym konkursie.

Sposób na eliminację tendencyjności ewaluatorów

Badania[124] pokazują, że potrafimy ustawić źródło światła tak, aby świeciło „dwa razy jaśniej" niż inne i zgadzamy się, że wyrok jednego miesiąca pozbawienia wolności wydaje się czymś znacznie gorszym niż 1/10 wyroku wynoszącego dziesięć miesięcy. Podwyżki wyrażone w procentach mają dla nas dużo większą wartość informacyjną niż wyrażone w złotówkach. Oznacza to, że w naszych ewaluacjach **sprawnie** posługujemy się **skalami ilorazowymi/stosunkowymi.**

Gdy w laboratorium pokazywano światło o określonej jasności, prosząc o uznanie tego natężenia za 10 (albo 50, albo 200 w innych warunkach eksperymentalnych), oceny jasności kolejnych świateł były proporcjonalne do pierwszej liczby (10, 50 albo 200). Gdy arbitralnie narzucona kotwica wynosiła 200, wszystkie ewaluacje były dwudziestokrotnie wyższe niż wtedy, gdy narzucona kotwica początkowa wynosiła 10.

W wielu badaniach wykazano pułapkę **zakotwiczenia**: cena wywoławcza wpływała na cenę końcową, a wynik z ruletki (wysoki lub niski) wpływał na ocenę dowolnej wielkości. Pierwotna kotwica oddziaływała również na ceny, jakie uczestnicy byli gotowi zapłacić za inne artykuły. Na przykład, jeśli uczestnik zgodził się zapłacić wysoką kwotę za bezprzewodową mysz, był również skłonny zapłacić odpowiednio więcej za bezprzewodową klawiaturę.

Jesteśmy znacznie bardziej **wrażliwi** na **relatywną wartość** porównywalnych artykułów niż na ich wartość **bezwzględną** (zjawisko „koherentnej arbitralności"). Gdy po raz pierwszy używamy danej skali pomiarowej i nie mamy żadnych wskazówek kalibracyjnych, dokonujemy **arbitralnego wyboru kotwicy**.

Rozważmy przykład skali ilorazowej, na której mierzymy wysokość kary w dolarach[125] dla 10 przypadków. Taka skala ma określony dolny kraniec (zero dolarów), ale nie ma górnej granicy. Kwoty dolarowe są często zakotwiczone na arbitralnej liczbie, którą poszczególni ewaluatorzy wybierają przy swoim pierwszym orzeczeniu.

[124] Stevens, 1958.

[125] Kahneman i in., 2022.

Kotwica arbitralnie wygenerowana przez danego ewaluatora może wywierać duży wpływ na wartości bezwzględne jego kolejnych ewaluacji. Jeśli jednak zastąpimy każdą z 10 kwot dolarowych odszkodowania jej rangą (najniższa kwota otrzymuje rangę 1, a najwyższa rangę 10), wyeliminujemy różnice między ewaluatorami. Po przekształceniu zasądzanych kwot dolarowych w rangi wykazano większą zgodność między ewaluatorami.

Niestety w wielu przypadkach, np. wśród przysięgłych czy recenzentów, ewaluacja nie odbywa się w schemacie seryjnym. Prawo amerykańskie wyraźnie też zakazuje kalibracji poprzez zakaz informowania przysięgłych o wysokości odszkodowań karnych przyznawanych w innych sprawach.

Efekt pozycji w sztywnym wariancie ewaluacji seryjnej

Ewaluacja seryjna, która pozwala uniknąć negatywnego wpływu doboru recenzentów, jest zdecydowanie lepsza od ewaluacji odrębnej, jednak jej sztywny wariant wiąże się z trudnym do uniknięcia efektem pozycji w serii.

Przedstawienie **psychologicznego modelu procesu oceny**[126] na przykładzie ewaluacji **artykułów** naukowych i grantów pozwoli zrozumieć, skąd się ten efekt bierze.

Dokonanie ewaluacji, czyli udzielenie odpowiedzi na pytanie, czy dany artykuł zasługuje na publikację, a projekt grantowy na dofinansowanie, wymaga **porównania ocenianego obiektu** z jakimś standardem/**wzorcem**.

Do podstawowych zdolności naszego umysłu należy kategoryzowanie doświadczeń, odkrywanie zależności i łączenie informacji krystalizujących się w postaci wzorców.

Jak pisaliśmy w pierwszej części książki, **mające budowę prototypową** wzorce tworzą się pod wpływem doświadczenia zapisywanego bezwiednie przez Holistyka, mogą też być wynikiem przemyśleń Analityka (w tym także edukacji).

Wzorzec składa się z **prototypu** i zakresu dopuszczalnych transformacji (na ile jesteśmy w stanie zaakceptować artykuły różniące się od prototypu).

Prototypem dla wzorca „dobry artykuł naukowy" używanym do porównań kolejnych obiektów może być[127]:

- wyabstrahowana ze szczegółów reprezentacja będąca uśrednieniem cech napotkanych wcześniej obiektów (bazujemy na wszystkich poprzednio przeczytanych artykułach z danej dziedziny) – prototyp **stanu typowego**;

[126] Wieczorkowska i Wierzbiński, 2013; Król i Kowalczyk, 2014; Michałowicz, 2016; Kowalczyk, 2019.

[127] Oberauer i in., 2012; Wieczorkowska i Wierzbiński, 2013.

- teoretyczna wizja idealnego, nieistniejącego w rzeczywistości egzemplarza (mamy wyobrażenie tego, jak powinien wyglądać idealny artykuł) – prototyp **stanu idealnego**;
- jeden ze spotkanych wcześniej egzemplarzy danej kategorii (opieramy się na artykule, który z jakiegoś powodu utkwił nam w pamięci) – prototyp **egzemplarzowy**.

Metaforycznie rzecz ujmując, porównując podobieństwo dostarczonej dokumentacji z prototypem dokonujemy mentalnej transformacji jednej reprezentacji w drugą. Końcowa decyzja może przyjąć dwie wartości – artykuł spełnia wymagania lub nie. Jednak to recenzenci są w stanie określić „odległość" artykułu od prototypu wzorca „dobry artykuł naukowy". Jeśli „odległość" **przekracza zakres dopuszczalnych transformacji prototypu**, czyli w zbyt dużym stopniu od niego odbiega, recenzent odmawia swojego poparcia dla publikacji.

Wzorzec „dobry artykuł" może mieć kilka prototypów – można jednakowo wysoko cenić systematyczny przegląd bardzo wielu artykułów, jak i propozycję awangardowego modelu kwestionującego ogólnie przyjęty paradygmat. To, który prototyp (systematyczny przegląd vs. awangardowe ujęcie tematu) zostanie zaktywizowany w procesie ewaluacji, zależy od kilku czynników, m.in. od **nawykowych wyborów recenzenta** (niektórzy mają tendencję do koncentrowania się tylko na jednym prototypie idealnego tekstu naukowego i nie tolerują od niego odstępstw), **zmiennych sytuacyjnych** (np. ostatnio oceniany projekt grantowy) oraz dostępnych zasobów psychoenergetycznych wyznaczających pojemność SUA [Sieć Uwagi Analityka], które maleją, gdy recenzent jest przemęczony[128]. Wystarczy, aby ktoś przypomniał recenzentowi podobny przypadek oceniany rok wcześniej i wtedy ten projekt (precyzyjniej zapis w sieci neuronowej z nim związany) staje się standardem/**prototypem porównawczym.**

W **ewaluacji odrębnej** nie możemy kontrolować, jaki standard porównawczy (*reference pattern*) został zaktywizowany w umyśle eksperta. W ewaluacji seryjnej można założyć, że obiekty oceniane wcześniej wpływają na modyfikacje wzorca będącego podstawą oceny. Efekt pozycji ocenianego obiektu w serii (*serial position effect*) został potwierdzony w różnych badaniach[129]. W badaniu eksperymentalnym[130] te same abstrakty, wcześniej ocenione przez niezależnych ekspertów jako słabe lub dobre, były przedstawiane w różnej kolejności: połowie ewaluatorów prezentowano je na początku, drugiej połowie na końcu. Zgodnie z hipotezą ewaluatorzy różnicowali oceny obiektów przedstawianych na końcu znacznie silniej niż tych na początku. Wartymi podkreślenia są wyniki badań niemieckiego zespołu[131]

[128] Oberauer i in., 2012.

[129] Antipov i Pokryshevskaya, 2017; Bruine de Bruin, 2006; Fasold i in., 2012; Unkelbach i in., 2012.

[130] Kowalczyk, 2019.

[131] Unkelbach i Memmert, 2008.

ukazujące, że zniekształceniu wynikającemu z pozycji w serii podlegają wszyscy ewaluatorzy – zarówno nowicjusze, jak i eksperci.

W badaniu sprawdzającym[132] świadomość efektu pozycji wśród 646 absolwentów szkół wyższych pytano ich, w jakiej kolejności chcieliby, aby ewaluator oceniał ich pracę.

Ponad **30%** z grupy **646** respondentów z wyższym wykształceniem chciało być ocenianych **na początku** w porównaniu do **32%**, którzy preferowali bycie ocenianym **na końcu** serii. Tylko trzy osoby z 646 uzasadniły swoje preferencje problemem kalibracji ewaluatorów.

Przykładowe cytaty z wypowiedzi respondentów:

- „Lepiej jest być ocenianym na początku, gdy nie ma punktu odniesienia; łatwiej jest uzyskać wyższą ocenę, choć często nie najlepszą, ponieważ oceniający pozostawia pewien margines na górę i na dole."
- „Osoby na początku serii są oceniane dokładniej, ponieważ – z jednej strony – oceniający nie jest jeszcze zmęczony, a z drugiej strony nie mają jeszcze skali porównawczej, aby móc ocenić ją bardziej pozytywnie."

Osoby, które preferowały bycie ocenianym pod koniec serii, zwróciły uwagę na możliwość zmęczenia ekspertów:

- „Znam ludzi, którzy wolą być oceniani jako ostatni, ponieważ uważają, że oceniający jest już zmęczony i chce mieć wszystko z głowy tak szybko jak to możliwe. Dlatego na koniec wszystkiego oceniają lepiej."

Żaden z ekspertów w wypowiedziach na pytanie otwarte nie wskazał na proces kalibracji – aż połowa z nich wybrała opcję „kolejność w serii nie ma znaczenia".

Efekt HALO

Przedstawiona w dalszej części tekstu analiza stopnia skorelowania odpowiedzi na pytania ankiet ewaluacyjnych zajęć akademickich wykazała, że powszechnie występuje efekt HALO/aureoli. Jeśli przykładowo wykładowcę uważa się za przyjaznego studentom, to jest on również postrzegany jako prezentujący materiał w sposób jasny oraz czytelny i odwrotnie, jeśli jest postrzegany jako nieprzyjazny studentom, to odmawia mu się jednocześnie umiejętności dydaktycznych.

Analogiczne analizy jak dla ankiet ewaluacyjnych zostały przeprowadzone[133] na ocenach 673 abstraktów zgłoszonych na międzynarodową konferencję. Pochodzące od 33 ewaluatorów oceny na poszczególnych wymiarach wykazały wysokie skorelowanie (jeden czynnik wyjaśniał od 37 do 91%). Bardzo wysoki procent wyjaśnio-

[132] Kowalczyk, 2019.

[133] Król i Kowalczyk, 2014.

nej wariancji przez jeden czynnik oznacza, że ewaluator oceniał abstrakty prawie tak samo na wszystkich wymiarach cząstkowych. Skoro odpowiedzi są tak wysoko skorelowane, to zadawanie ewaluatorom wielu pytań nie wnosi wiele do oceny, jednocześnie będąc źródłem przeciążenia poznawczego stanowiącego zagrożenie dla trafności ocen.

Sposób eliminacji efektu HALO w ewaluacji seryjnej

Ocena grantów czy publikacji naukowych jest zadaniem wymagającym dużych zasobów poznawczych. Tworzenie zbyt wielu kryteriów zamiast ułatwić pracę oceniającym, może ją wręcz utrudnić.

Od ewaluatorów wymaga się nie tylko podjęcia decyzji o zaakceptowaniu lub odrzuceniu projektu/publikacji, lecz także oceny obiektów na wielu wymiarach kryteriów cząstkowych. U podstaw tej praktyki leży milcząco przyjmowane założenie, że oceny cząstkowe zobiektywizują cały proces, a obiekty będą opisywane w postaci wielowymiarowych profili. W praktyce, nawet jeśli wymiarom przypisuje się zróżnicowane wagi, decyzje najczęściej opierają się na wartości średniej. Konieczność dokonywania wielu ocen cząstkowych stanowi jednak duże obciążenie poznawcze dla oceniającego.

Na siłę oddziaływania efektu HALO może wpływać sposób oceniania serii obiektów na kilku wymiarach cząstkowych, który można przeprowadzić w następujący sposób:

1) **obiektowy** – najpierw oceniamy obiekt 1 na wszystkich wymiarach, potem obiekt 2 itd.;
2) **wymiarowy** – najpierw oceniamy wszystkie obiekty na wymiarze 1, potem wszystkie obiekty na wymiarze 2 itd.[134].

Hipoteza zakładająca, że zastosowanie wymiarowego zamiast obiektowego sposobu oceny spowoduje redukcję efektu HALO, została potwierdzona w badaniach[135], w których losowo podzieleni na dwie grupy ochotnicy oceniali na pięciu wymiarach (**zrozumiałość projektu, opis, ciekawość, czytanie, ogólna rekomendacja**) dwa projekty badawcze na podstawie otrzymanych streszczeń.

Grupa E1 oceniała projekty **wymiarowo** (najpierw wszystkie obiekty na pierwszym wymiarze, potem wszystkie obiekty na wymiarze drugim itd.), a grupa E2 – **obiektowo** (najpierw oceniano obiekt pierwszy na wszystkich wymiarach, potem obiekt drugi itd.).

Taki wymiarowy sposób oceny prac studenckich można łatwo wprowadzić na platformie e-learningowej. Jeżeli chcielibyśmy utrzymać dotychczasowy system

[134] Tyszka, 2010.

[135] Kowalczyk, 2019.

wielowymiarowej oceny zajęć akademickich (co moim zdaniem jest marnotrawstwem energii ewaluatorów, bo przekłada się na żenująco niski odsetek studentów oceniających zajęcia), należałoby zadbać, aby student oceniał wszystkie zajęcia na pierwszym wymiarze, a następnie przechodził do kolejnego pytania. Odpowiadanie na ciąg pytań dotyczących tych samych zajęć powoduje, że oceny są bardzo wysoko skorelowane.

Podsumowując: wyniki wieloletnich badań prowadzonych w Katedrze Psychologii i Socjologii Uniwersytetu Warszawskiego pozwoliły potwierdzić cztery zagrożenia ewaluacji liczbowych.

A. Stała **tendencyjność ewaluatorów,** która może być traktowana jako cecha osobowości, sprawia, że w typowych ewaluacjach odrębnych, gdzie dla każdego obiektu są niezależnie wybierani jego sędziowie/recenzenci, bardzo dużo zależy właśnie od tego wyboru. Szanse na publikację artykułu wysłanego do czasopisma zależą od tego czy trafi on do recenzentów surowych (charakteryzujących się bardzo wysokim wskaźnikiem odrzuceń), czy też bardziej pobłażliwych (bardziej skłonnych do akceptacji).
B. Wyniki badań wskazują na **wyższość ewaluacji seryjnej**, w której wszystkie oceniane obiekty (projekty badawcze, artykuły naukowe) są oceniane przez ten sam zespół ewaluatorów.
C. W **sztywnym wariancie ewaluacji seryjnej**, w którym nie można modyfikować wcześniejszych ocen (przykładem są natychmiast ogłaszane wyniki egzaminu ustnego czy oceny sędziów oceniających występy sportowców w zawodach łyżwiarstwa figurowego), powszechnie występuje efekt pozycji w serii. Powoduje on, że ewaluatorzy wystawiają mniej zróżnicowane oceny na początku serii niż na jej końcu. W efekcie gorsze jakościowo obiekty są oceniane wyżej na początku serii niż gdyby były oceniane na końcu, natomiast lepsze obiekty są oceniane wyżej na końcu serii niż na jej początku. Dlatego, tak jak w ewaluacji odrębnej, tak i w sztywnym wariancie ewaluacji seryjnej konieczne jest przeprowadzenie szkolenia kalibracyjnego dla ewaluatorów, które mogłoby zwiększyć świadomość występowania efektu pozycji w serii i dać szansę na jego korektę.
D. W ewaluacji seryjnej zestawów odpowiedzi na pytania obiektowy sposób oceniania skutkuje wysokim skorelowaniem odpowiedzi. Zmiana sposobu oceniania na wymiarowy pozwala na minimalizację efektu HALO.

1. Opinie są podatne na **różnorodne błędy** i **zniekształcenia.** Ulegają wpływowi sposobu sformułowania pytania, wcześniejszych pytań oraz torowania wywołanego przez przypadkowe bodźce obecne w czasie zadawania pytań. Są formułowane na bieżąco i podlegają fluktuacjom.

2. Wszystkie ewaluacje opierają się na **ukrytym porównaniu z jakąś kotwicą/ grupą odniesienia**. Przykładowo, odpowiedź na pytanie o pracowitość zależy od tego, z kim lub czym porównujemy się w danym momencie.
3. W ewaluacjach **liczbowych**, w przypadku ewaluacji odrębnej nie wystarczy przekazać ewaluatorom skali – należy również zapewnić im **identyczną kalibrację** tej skali (np. poprzez zastosowanie punktu kotwiczącego, który będzie służył jako punkt odniesienia na skali).
4. W ewaluacji seryjnej zamiast uśredniania ocen pochodzących od różnych ewaluatorów warto rozważyć **wprowadzenie rankingów**.

Wnioski dla prowadzenia ewaluacji zajęć akademickich

Choć od publikacji tekstu, w którym opisywaliśmy „Mit usługi edukacyjnej i kluczowej roli studentów w wytyczaniu kierunku zmian" minęło 10 lat, nic się w tej materii nie zmieniło. Moje poglądy też nie uległy zmianie.

Kiedy efekty naszego zawierzenia specjalistom (trenerom na siłowni, chirurgom, gdy jesteśmy chorzy) są szybko widoczne (poprawiamy wyniki sportowe, zdrowiejemy), jesteśmy gotowi poddać się ich zaleceniom (i cierpieć z powodu zakwasów czy bólu pooperacyjnego). Wyniki „kształcenia mięśni mentalnych" (co jest celem edukacji uniwersyteckiej) nie są natychmiast zauważalne, więc w umysłach studentów rzadko powstaje połączenie: **„duży trud mentalny = duże korzyści".** W konsekwencji nie dziwi fakt, że studenci preferują zajęcia opisywane jako „atrakcyjne i łatwe", a nie „trudne i wymagające", choć to te drugie dają im większą szansę na rozwój umysłowy. Młodzi ludzie często nie zdają sobie sprawy, że odniesienie sukcesu na globalnym rynku wymaga przede wszystkim **sprawności myślenia**, **a tę kształci się, wychodząc poza strefę komfortu.**

Demokratyzacja zarządzania uczelniami spowodowała, że udział studentów we wszystkich ciałach kolegialnych (w tym np. w kolegium elektorskim wybierającym rektora) jest bardzo znaczący – na wielu uczelniach to ich głos decyduje o wyborze rektora, ponieważ w odróżnieniu od akademików potrafią doprowadzić do jednomyślnego głosowania swojej grupy elektorów. Czy studenci rzeczywiście powinni mieć tak znaczny wpływ na to, kto zarządza uczelnią? Jaką odpowiedzialność za losy uczelni może ponosić student, który dopiero rozpoczął edukację lub niebawem ją ukończy?

Uważam, że dotychczasową rolę studentów w ciałach kolegialnych powinni przejąć ABSOLWENCI, ponieważ oni mogą spojrzeć na edukację z koniecznego dystansu. Edukacja to proces przebiegający w relacji między nauczycielem a studentami.

Próba standaryzacji zachowań nauczyciela (podobnie jak standaryzacja zachowań kasjera) uniemożliwia elastyczność tej relacji, co obniża efektywność procesu edukacji. Niestety, coraz częściej pojawiają się studenci, którzy przyjmują rolę klien-

tów oczekujących obsługi. W ramce poniżej znajduje się fragment petycji niezadowolonych słuchaczy studiów podyplomowych prowadzonych przez jedną z renomowanych uczelni:

> „Usługa edukacyjna jest umową cywilnoprawną o wykonanie usług, które w tym przypadku wykonywane są przez Usługodawcę nierzetelnie. Nieuwzględnienie naszych wniosków może skutkować dochodzeniem roszczeń z tytułu nierzetelnego wykonania umowy przez..."

Zarzuty dotyczyły – w ocenie studentów – zbyt małej praktycznej przydatności treści zajęć.

Uczyłam tych studentów i mój przedmiot został oceniony wysoko, jednak nie mogę niestety odwzajemnić tej oceny oceną pracy studentów na podobnym poziomie. Prace zaliczeniowe, które czytałam, pokazały bardzo niski poziom wysiłku wkładanego w pracę poza zajęciami. Zapisywanie się na kolejne studia podyplomowe staje się coraz częściej przejawem chęci kolekcjonowania dyplomów oraz sposobem na ciekawe spędzenie weekendów. Niestety, coraz częściej pojawiają się studenci, którzy zachowują się, jakby byli klientami zakładu fryzjerskiego („Płacę, więc wymagam"), zapominając, że edukacja **wymaga współpracy obu stron** i że **nie można nauczyć kogoś**, kto **nie jest gotów do wysiłku.** Traktowanie tego procesu jako **usługi** jest **nieporozumieniem.**

Odpowiedzialność za wybór treści i metod nauczania spoczywa na prowadzącym, ale nie oznacza to, że nie warto pytać studentów o ich opinie. Trzeba jednak pamiętać, że **studenci nie mają wystarczającej wiedzy,** aby ocenić **wagę zajęć** dla przyszłych etapów edukacji**.** Również pytanie o ocenę kompetencji wykładowcy, występujące w niektórych ankietach, wydaje się kuriozalne. Warto pamiętać, że ankieta jest formą dialogu między pytającym (wykładowcą, przełożonym itp.) a odpowiadającym (studentem, pracownikiem).

Odpowiadający musi rozumieć nie tylko **O CO**, ale także **PO CO** jest pytany. Powinien, więc wiedzieć, jaki jest cel zbierania danych.

Możliwe są co najmniej dwa powody przeprowadzania takiej ankiety:

- wychwycenie zajęć/wykładowców, którzy wymagają wsparcia lub nadzoru oraz tych, którzy powinni zostać wyróżnieni (ważne przy awansach);
- uzyskanie informacji pozwalających udoskonalić dane zajęcia.

Aby osiągnąć pierwszy cel, wystarczy zadać studentom jedno pytanie: „Czy wykładowca – ich zdaniem – zasługuje na nagrodę dydaktyczną?" (np. na siedmiostopniowej skali odpowiedzi od „zdecydowanie nie" do „zdecydowanie tak") z prośbą lub możliwością podania uzasadnienia. Warto jednak zacząć od pytania o stopień zaangażowania studenta w pracę na zajęciach z opcją rozbudowania odpowiedzi. Takie

pytanie uświadamia ewaluatorom, że jakość zajęć zależy również od ich aktywności. Przy tak zadanym pytaniu studenci często piszą, że „sytuacja rodzinna lub zawodowa uniemożliwiła im pełne zaangażowanie".

Aby osiągnąć drugi cel, należy zadać JEDNO pytanie otwarte z prośbą o podzielenie się z wykładowcą swoimi odczuciami, np.: „Czy jest coś, co Twoim zdaniem należałoby zmienić w następnej edycji kursu? Co najbardziej zapadło Ci w pamięć? Czy inni studenci oceniają zajęcia podobnie jak Ty?". Zadając pytania warto uświadomić ewaluatorom, że pierwszym czytelnikiem ich opinii jest wykładowca, więc powinni przekazywać swoje uwagi w sposób konstruktywny – tak jak wykładowcy oceniają ich pracę, a nie obraźliwy.

Nasze doświadczenie pokazuje, że odpowiedzi na pytania otwarte dostarczają znacznie więcej informacji niż oceny liczbowe. Przesunięcie średniej na skali odpowiedzi na pytanie: „Czy zajęcia były ciekawe?" niesie niewiele informacji, ponieważ nie wiemy czym zostało ono spowodowane.

Analizując oceny liczbowe uzyskiwane przez osoby prowadzące ćwiczenia do mojego wykładu, mogę stwierdzić brak różnic – mimo że różnią się one stylem nauczania.

Zwolennicy skal liczbowych twierdzą, że oceniający zawsze mają do dyspozycji rubrykę „inne uwagi". Jednak z badań psychologicznych wiemy, że osoby wypełniające ankietę stosują reguły konwersacji, które nakazują unikać przekazywania tej samej informacji dwukrotnie.

Sformułowanie oceny liczbowej (przekazanie informacji w formie zaprojektowanej przez autora ankiety) sprawia, że rubryka „inne uwagi" pozostaje najczęściej pusta. Zaletą skal liczbowych jest to, że liczby ranią prowadzących mniej niż niektóre sformułowania używane przez studentów, którzy – korzystając z anonimowości – często dają upust emocjom bez zahamowań, nie zwracając uwagi na odbiorcę komunikatu. Czytanie roszczeniowych opinii słuchaczy może być dla wykładowcy bardzo demotywujące. Problem doboru pytań i sposobu ich sformułowania został wyczerpująco omówiony w innych pracach[136], więc tu zostanie pominięty.

W tym miejscu postaramy się odpowiedzieć na pytanie, czy wyniki ankiet ewaluacyjnych mogą być traktowane jako wskaźnik jakości pracy nauczyciela. W niektórych szkołach wyższych osoby, które uzyskują najlepsze oceny od studentów w danym semestrze, otrzymują nagrody finansowe.

Należy jednak brać pod uwagę stopień reprezentatywności próby studentów.

---

136 Wieczorkowska i Wierzbiński, 2013; Michałowicz, 2016.

137 W rozprawie doktorskiej B. Michałowicza (2016) zostały opisane wykonane przez niego analizy ankiet dotyczących 1294 zajęć (642 ćwiczeń, 224 seminariów, 152 warsztatów, 237 wykładów trzonowych i 38 specjalizacyjnych) przeprowadzonych w ciągu 4 semestrów (lata 2009/2010 i 2010/2011). Zajęcia prowadzone były przez 225 wykładowców. W analizowanym zbiorze jest 31 113 ankiet ewaluacyjnych (na ponad 69 801 zapisanych na zajęcia).

Przeprowadzona analiza danych[137] z ponad 31 tys. ankiet rozdawanych w formie papierowej na ostatnich zajęciach w jednej ze szkół wyższych w ciągu czterech semestrów wykazała, że stopień realizacji próby na zajęciach ze sprawdzaną listą obecności jest nieporównywalnie wyższy (ponad 70%) niż na wykładach (około 30%). W ankietach internetowych stopa zwrotu jest dramatycznie niska.

Stopa zwrotu internetowych ankiet ewaluacyjnych 2023/2024

Analiza ankiet ewaluacyjnych z 25 zajęć (prowadzonych przez tę samą osobę w naszej katedrze w semestrze letnim, w roku akademickim 2023/2024), w których uczestniczyło od 31 do 164 osób (mediana = 61, średnia = 76,2) wykazała, że stopa zwrotu przekroczyła 50% tylko w przypadku zajęć jednej z prowadzących. Dla 5 zajęć mieściła się w przedziale od 20 do 30%, dla 4 w przedziale od 11 do 18%, dla pozostałych zajęć w przedziale od 1 do 9%. Pominięto zajęcia, w których uczestniczyło mniej niż 30 studentów, ponieważ w tych mniej licznych grupach stopa zwrotu jest jeszcze mniejsza.

W analizowanym rozkładzie są przypadki wypełnionych 3 ankiet na 116, 6 ankiet na 135, czy 9 ankiet na 131 studentów. Oznacza to, że ankiety ewaluacyjne narzucone przez ekonomistów, którzy wszystko chcą wskaźnikować, powinny zostać zlikwidowane, ponieważ wypaczają postawy studentów wchodzących w role usługobiorców, podczas gdy są równoprawnymi agentami wpływu w procesie budowania, do czego wrócimy w sekcji 4. Powinien być oczywiście otwarty **anonimowy kanał przepływu informacji od studentów do władz**, w którym mogliby bez problemu zgłaszać – jeśliby chcieli – swoje obserwacje, postulaty, podziękowania.

Jeżeli jednak chcemy utrzymać dotychczasowy system, **liczba pytań** w ankietach ewaluacyjnych bywa stanowczo zbyt duża (najczęściej co najmniej 5). Jeśli weźmiemy pod uwagę, że student w ciągu semestru może uczestniczyć w kilkunastu zajęciach, oznacza to, że aby być sumiennym ewaluatorem musiałby odpowiedzieć na kilkadziesiąt nudnych pytań. Badania pokazują[138], że tylko **12% studentów** ocenia **wszystkie zajęcia**, w których uczestniczyli w danym semestrze, **43% nie ocenia ŻADNYCH zajęć**. Analiza czynnikowa głównych składowych odpowiedzi na pytania zawarte w ponad 94 tys. (N = 94130) studenckich ankiet ewaluacyjnych wykazała istnienie **jednego czynnika**, który wyjaśniał 66,3% wariancji zmiennych wejściowych. Oznacza to, że możemy mówić o efekcie HALO. Pytaniem najwyżej korelującym z tym jednoczynnikowym rozwiązaniem było pytanie o ogólną ocenę wykładowcy. Osobne analizy dla 50 zajęć wykładowych, z których każde zostało ocenione przez co najmniej 100 osób, pokazały, że w 48 przypadkach wyodrębniono jeden

[138] Michałowicz, 2016.

czynnik wyjaśniający od 54,6 do 80,7% zmienności pytań ankiety, co potwierdza bardzo wysokie skorelowanie ocen cząstkowych.

Podsumowując: odpowiedzi na pytania w ankietach ewaluacyjnych są wysoko skorelowane – można więc mówić o efekcie HALO/aureoli – co oznacza, że można zmniejszyć liczbę pytań.

Ocena nauczyciela na podstawie wyników ankiet ewaluacyjnych, oprócz pomijania informacji o stopniu reprezentatywności próby, często nie uwzględnia faktu, że proces dydaktyczny jest wynikiem interakcji wkładu prowadzącego i wkładu grupy.

Ilustracją tej tezy są wyniki analiz, w których porównywano oceny ewaluacyjne tych samych zajęć prowadzonych przez tego samego wykładowcę w różnych grupach studentów. Okazało się, że TEN SAM wykładowca uczący TEGO SAMEGO przedmiotu uzyskuje INNE oceny od różnych grup studentów. To pokazuje, że proces dydaktyczny nie zależy tylko od umiejętności prowadzącego, ale również od **charakterystyki uczącej się grupy**. Jeżeli chcielibyśmy nagradzać najlepszych wykładowców, musimy pamiętać, że integracja ocen z poszczególnych kursów stanowi istotny problem. Dlatego porównania wykładowców należy prowadzić w schemacie, w którym uwzględnia się **zagnieżdżenie czynnika „grupa studencka" w czynniku „wykładowca"**, a nie – jak to się powszechnie praktykuje – przez uśrednianie średnich z różnych grup[139].

Porównywanie wyłącznie średnich też nie jest dobrym pomysłem, nawet gdy porównuje się wyniki pochodzące od jednej grupy studentów. Rozkłady ocen tych samych zajęć bywają silnie skośne. Nawet w przypadku rekordowego kursu, który otrzymał maksymalne oceny od 60% studentów, znaleźli się tacy, którzy dali wykładowcy cztery najniższe oceny. Trzeba o tym pamiętać, gdy czytamy wypowiedzi pojedynczych studentów w Internecie. Rozkłady różnią się też często wariancją (przy tej samej liczebności próby zakres zmienności ocen potrafi się różnić aż dwukrotnie). Do tematu pytań w ankietach ewaluacyjnych powrócimy w sekcji 6. Tym, którzy mimo przedstawionych przez nas argumentów, chcą nadal oceniać pracę wykładowców wyłącznie na podstawie wyników ankiet ewaluacyjnych, dedykujemy cytat z Konfucjusza:

> *Co sądzić o człowieku, którego kochają wszyscy w jego wiosce?... Mistrz powiedział:* ***Nic to jeszcze nie znaczy. Dobrze by o nim świadczyło, gdyby go kochali wszyscy dobrzy, a wszyscy źli nienawidzili****.*

[139] Michałowicz, 2016.

Sekcja 4
Rafy metodologiczne w ewaluacji rozwoju nauk o zarządzaniu ludźmi

Przez wszystkie lata pracy zastanawiam się również nad kierunkiem rozwoju naszej dyscypliny. Zajęci spełnianiem uznanych w środowisku standardów ignorujemy zmiany, które zachodzą w naszym otoczeniu informacyjnym. Uważam, że już niedługo czeka nas **zmiana jakościowa** w naszej dziedzinie naukowej, która dokona się (już dokonuje?) nie tyle pod wpływem nowych teorii, ile **rozwoju technologii**, tak jak miało to miejsce np. w medycynie.

Wbrew powszechnym przekonaniom, że nauka jest jedna, uważam, że **kryteria oceny rozwoju nauk ścisłych** różnią się dramatycznie od kryteriów w naukach społecznych, a tym bardziej w naukach humanistycznych. Dlatego od razu zastrzegam, że wszystko, co tu piszę, **dotyczy wyłącznie nauk społecznych**, ponieważ tylko w tej dziedzinie mam odpowiednią wiedzę.

Zmieniony świat edukacyjny

Żyjemy w czasach przełomu technologicznego. Czy odczuwamy wyjątkowość tej zmiany? Niekoniecznie, ponieważ zmiany zachodzą stopniowo – można więc mówić o efekcie „gotowanej żaby". Żaba może zginąć, jeśli temperatura w akwarium jest podnoszona bardzo powoli, podczas gdy wrzucenie jej do gorącej wody spowodowałoby, że natychmiast by wyskoczyła.

Nasze środowisko informacyjne gęstnieje z dnia na dzień w niewielkim tempie, dlatego nie „wyskakujemy". Te małe zmiany kumulują się. To, o czym piszemy, dzieje się zawsze i wszędzie, ale zdarza się wystarczająco często, by warto było uczynić to przedmiotem refleksji. **Nauczyciele socjalizowani przed erą internetową** często zapominają, że ich rola się zmieniła. **Próbują uczyć w sposób, w jaki sami byli uczeni**. Mamy już na studiach kolejne pokolenie socjalizowane w przekonaniu, że

nie potrzebuje pośredników w dostępie do wiedzy. Rola nauczyciela ewoluuje z dostawcy wiedzy w role menedżerów, trenerów, coachów. Podstawowe funkcje **menedżera procesu uczenia**: (1) zaplanować czego i w jakim czasie chcemy się nauczyć; (2) zorganizować i koordynować; (3) przewodzić (motywować, wydawać polecenia); (4) kontrolować przyrosty wiedzy i umiejętności. W tej ostatniej funkcji świetnie może zastąpić nauczyciela „komputer" – ale wcześniej nauczyciel musi go zaprogramować – ustalić zasady konstrukcji informacji zwrotnej.

W tej sekcji pojawiają się różne statystyki, które mają zilustrować to zjawisko, jednak nie pretendują do pełności – nie jest to naszym celem.

W 2016 roku, myśląc o przyszłości edukacji akademickiej, pisaliśmy, że:

KIEDYŚ uniwersytet był definiowany jako wspólnota ludzi poszukujących prawdy. Było mało zajęć fakultatywnych i przeważały stałe grupy studenckie na wszystkich zajęciach, co sprzyjało nawiązywaniu trwałych relacji.

DZIŚ uniwersytet jest odczuwany jako wspólnota ludzi poszukujących sukcesu biznesowego, który osiąga się przez konkurencję między jednostkami, uczelniami. Dominuje wyidealizowany model wolnego rynku w wyborze zajęć przez studenta. Można zaobserwować uwiąd sfery normatywnej zastępowanej przez regulacje prawne. Dominuje kultura audytu, studenci wybierają zajęcia jak w supermarkecie, intrygujący tytuł, tak jak opakowanie, przyciąga uwagę. Żmudne, mało popularne zajęcia znikają z siatki zajęć.

Dokąd zmierzamy: nauki, dyscypliny i subdyscypliny

Kiedyś uniwersytety stanowiły niewielkie, spójne społeczności składające się głównie z naukowców, którzy oddawali się nauce – jak to określił prof. Kula – z „uszkami i nogami włącznie". Przed XIX w. uniwersytety w Europie nie miały podziału na odrębne dyscypliny. Student był zwykle szkolony przez mistrza, który uczył go języków klasycznych, chemii, astronomii, filozofii, geometrii oraz teologii, a dodatkowo miał być dla niego wzorem standardów moralnych.

Wraz z rosnącą ilością wiedzy, którą należało przekazać, uniwersytety zaczęły zatrudniać wykładowców posiadających wystarczająco głęboką wiedzę w JEDNYM obszarze tematycznym. Tak powstały uniwersytety badawcze zorganizowane wokół odrębnych dyscyplin, które stanowią trzon struktury uniwersyteckiej aż do dzisiaj[140].

Większość wykładowców uniwersyteckich przejawia lojalność przede wszystkim wobec wydziału, który ich zatrudnia, a nie całego uniwersytetu czy też całej społeczności akademickiej[141].

Podział na dyscypliny i subdyscypliny ma sens w naukach matematycznych i technicznych – ekspert w geometrii może nie znać rachunku różniczkowego. Jednak

[140] Kreber, 2009.

[141] Locke, 2008; Poole, 2009; Hensel, 2017.

w przypadku nauk społecznych taka subdyscyplinarność jest szkodliwa, ponieważ obiektu naszych badań nie da się podzielić na odrębne części. Nie można na przykład, badając zachowania ekonomiczne, ignorować procesów poznawczych, emocjonalnych czy społecznych.

Dziwi więc powstawanie nowych dyscyplin wiedzy, takich jak nauki o rodzinie. Śmieszny wydaje się rygorystycznie przestrzegany w Polsce (np. przy określaniu minimów kadrowych czy efektów kształcenia) rozdział nauk humanistycznych i społecznych. O sztuczności tego podziału świadczy fakt, że w 2011 roku jednym rozporządzeniem ministra przeniesiono psychologię z obszaru nauk humanistycznych do społecznych.

Dlatego do dziś nie wiem, czy jako osoba, która otrzymała tytuł profesora belwederskiego w 2000 roku, jestem profesorem nauk humanistycznych, czy – zgodnie z obowiązującymi obecnie regułami i moim własnym poczuciem – nauk społecznych.

Segmentacja uczelni wyższych na wydziały reprezentujące różne dyscypliny jest tak głęboka, że często opisuje się jednostki organizacyjne tego samego uniwersytetu jako **konkurujące plemiona walczące o posiadanie jak największego terytorium** i przywilejów kosztem swoich sąsiadów.

Niektórzy[142] jeszcze bardziej krytykują wąską specjalizację, twierdząc, że większość współczesnych intelektualistów odnosi sukcesy w środowisku akademickim właśnie dlatego, że specjalizuje się w niezwykle wąskich dziedzinach. Jeśli ktoś musi spędzać większość swojego produktywnego czasu na czytaniu i pracy nad bardzo szczegółowymi problemami (np. uczony przez całe życie na analizie dzieł jednego tylko filozofa lub badacz inwestujący setki tysięcy dolarów i lata pracy doktorantów, aby dowiedzieć się, czy nieznany gatunek grzybów nie występuje na Antarktydzie), trudno oczekiwać, że będzie intelektualistą zdolnym do kompetentnego wypowiadania się o wielkich problemach współczesnego, coraz bardziej złożonego społeczeństwa.

W efekcie intelektualizm wydaje się być w odwrocie – paradoksalnie w czasach, gdy tak wiele mówi się o „gospodarce opartej na wiedzy". Tymczasem to programy telewizyjne z udziałem celebrytów często decydują, która książka stanie się z dnia na dzień kolejnym bestsellerem.

Podsumowując: podział na subdyscypliny ma sens w naukach matematycznych czy technicznych, ale w przypadku nauk społecznych taka subdyscyplinarność jest szkodliwa, ponieważ obiektów naszych badań nie da się sztucznie podzielić.

142 Pigliucci, 2019.

Dokąd zmierzamy – zmienione wymagania wobec kadry uniwersyteckiej

Przytoczmy parę liczb[143]:

- **Liczba studentów** uczących się na uczelniach wyższych wzrastała na wszystkich kontynentach w ciągu ostatnich 3 dekad. Od 1990 do 2018 roku zanotowano wzrost z 70 do 200 mln. Zmiana liczby studentów w Polsce z 350 tysięcy w 1990 roku do 1,2 mln w 2022 roku ze szczytową wartością wynoszącą 1,8 mln w 2010 roku.
  Współczynnik skolaryzacji brutto w Polsce (relacja # studentów do # ludności w wieku 19–24 lata) wynosił w 1990 roku 12,9, w 1998 33,5, osiągając szczyt w 2010 roku wartością 53,7 i lekko spadając do 52,3 w roku 2022[144].
- W Wielkiej Brytanii w ciągu 60 lat (1962–2022) **liczba uczelni wyższych** wzrosła z 25 do 285[145], a liczba studentów ze 125 tys. do 2,94 mln[146].
- Pod koniec XIX wieku zaledwie 1% młodzieży w USA zdobywało wyższe wykształcenie. Do roku 1925 liczba ta utrzymywała się poniżej 5% populacji. W roku 2022 ok. 63% Amerykanów w wieku co najmniej 25 lat posiadało jakąś formę wykształcenia wyższego (więcej niż high school).

Procent Polaków legitymujących się wyższym wykształceniem w 1988 roku wynosił 6,5, w 2011 – 17,1, a w 2021 roku już 24,5. Za wzrost w 2005 roku odpowiada także uznanie 3-letnich studiów (licencjackich) za wyższe wykształcenie.

Wraz ze wzrostem liczby studentów liczba nauczycieli akademickich na świecie wzrosła **2,5-krotnie w ciągu zaledwie 23 lat** (z ok. 6,5 mln w 1999 roku do 16 mln w 2022 roku).

Największy wzrost zanotowano w Chinach, gdzie liczba wykładowców w tym czasie podwoiła się, a obecnie jest ich więcej niż w USA, gdzie zatrudnionych jest ponad milion nauczycieli akademickich. Przewiduje się, że liczba studentów i nauczycieli akademickich będzie rosła jeszcze przez co najmniej dwadzieścia lat.

Większa liczba nauczycieli akademickich nie tłumaczy dramatycznie wysokiego wzrostu liczby badań i publikacji naukowych. Wynika to z faktu, że współcze-

---

143 Billig, 2013; Trow, 2007; Altbach i in., 2010; 2011; Altbach i Reisberg, 2018.

144 https://stat.gov.pl/obszary-tematyczne/roczniki-statystyczne/roczniki-statystyczne/maly-rocznik-statystyczny-2000-r-,1,1.html?contrast=black-white; https://stat.gov.pl/obszary-tematyczne/roczniki-statystyczne/roczniki-statystyczne/maly-rocznik-statystyczny-2000-r-,1,1.html?contrast=black-white; H. Dmochowska (red.), *Polska 1989–2014*, opracowanie zbiorcze, GUS, 2014, https://stat.gov.pl/obszary-tematyczne/inne-opracowania/inne-opracowania-zbiorcze/polska-19892014,13,1.html; https://stat.gov.pl/obszary-tematyczne/roczniki-statystyczne/roczniki-statystyczne/maly-rocznik-statystyczny-polski-2024,1,26.html, dostęp: 18.11.2024.

145 https://www.universitiesuk.ac.uk/latest/insights-and-analysis/higher-education-numbers, dostęp: 11.11.2024.

146 https://commonslibrary.parliament.uk/research-briefings/cbp-7857/, dostęp: 11.11.2024.

sne uniwersytety stały się wspólnotami ludzi dążących nie do poszukiwania prawdy, jak to było kiedyś, lecz do sukcesu biznesowego osiąganego poprzez konkurencję – zarówno między jednostkami, jak i między uczelniami, które zaczynają działać jak firmy w warunkach wolnorynkowych.

Zatrudnieni na zachodnich uczelniach menedżerowie, zarabiający znacznie więcej niż profesorowie, dążą do maksymalizacji wydajności pracowników, często zwalniając tych, którzy nie spełniają przyjętych wskaźników efektywności. Mówi się o „**kapitalizmie akademickim**”[147]. Badania wykazały, że nauczyciele akademiccy w Stanach Zjednoczonych w przeciwieństwie do innych specjalistów w sektorze publicznym pracują każdego dnia dłużej niż robili to trzydzieści lat temu[148]. Na podstawie własnych doświadczeń przypuszczam, choć nie znamy wyników żadnych systematycznych badań, że w Polsce sytuacja wygląda podobnie. Naukowców prowadzących intelektualne dysputy w kawiarniach można dziś obejrzeć już tylko na starych filmach.

Naukowcy pracują intensywniej, ponieważ muszą więcej uczyć (wzrost liczby studentów następuje dużo szybciej niż wzrost kadry). Coraz częściej próbuje się minimalizować koszty przez zwiększanie liczebności grup wykładowych i ćwiczeniowych. Jednak w tym przypadku zapomina się o psychologicznych ograniczeniach. Podczas wielkiego koncertu jeden artysta może oddziaływać na tłumy, a bycie częścią takiego tłumu może być dla odbiorców bardzo angażujące. W edukacji rządzą jednak inne zasady – nie chodzi w niej o zapewnianie poczucia bycia w grupie, ale o **zwiększanie motywacji do wysiłku poznawczego**. Wykład nie polega jedynie na przekazywaniu informacji – często lepszym źródłem są podręczniki czy dostępne wykłady w sieci. Zadaniem prowadzącego jest zainteresowanie słuchaczy przedmiotem, wzbudzenie i utrzymanie motywacji do nauki, a także **skłonienie do wysiłku intelektualnego**. Jest to szczególnie ważne w edukacji, która nie jest obwarowana przymusem zaliczenia (np. kursy dla osób pracujących, pragnących z własnej inicjatywy zdobywać nowe kwalifikacje). W takich przypadkach prowadzący musi zdobyć zaufanie słuchaczy, aby skutecznie angażować ich w proces nauki.

Pamiętajmy, że badania mózgu[149] wykazały, iż spojrzenie innej osoby może uaktywniać nasz ośrodek nagrody, podobnie jak kokaina, czekolada czy ulubiona melodia w przypadku muzyki. Dlatego wykładowca powinien mieć możliwość nawiązania kontaktu wzrokowego z całą salą.

Pojawia się jednak pytanie, z iloma osobami w licznej grupie może nawiązać taki kontakt? Szukając odpowiedzi można odwołać się do wyników badań, które porównywały rozmiar mózgu zwierząt (relatywnie do masy ciała) i wielkość grup, które tworzą.

---

147 Slaughter i Rhoades, 2004, 2010.

148 Schuster i Finkelstein, 2006.

149 Spitzer, 2007.

Okazało się, że na podstawie wielkości mózgu da się przewidywać maksymalną wielkość stada dla danego gatunku – dla człowieka ta liczba została określona na 150 osób. Z naszego doświadczenia wynika jednak, że preferujemy jeszcze mniejsze grupy. Wykład dla 300 słuchaczy prowadzi do poczucia alienacji – studenci przestają być postrzegani jako jednostki, a stają się tłumem. Nie sposób już zapamiętać ich twarzy, a brak kontaktu wzrokowego z wykładowcą może zniechęcać do uczestniczenia w zajęciach. Dążenie do minimalizowania kosztów edukacji powoduje właśnie poczucie alienacji – zarówno wśród wykładowców, jak i studentów. Prowadzący wykłady semestralne dla bardzo licznych grup czują się bardziej zmęczeni, mimo że teoretycznie nie powinno mieć znaczenia, czy mówi się do 150 czy 300 osób. Jednak brak możliwości nawiązania więzi z tak dużą grupą sprawia, że kontakt wykładowcy z uczestnikami jest ograniczony. Warto pamiętać, że z grupą studentów nie spotykamy się jednorazowo, ale przez cały semestr – dlatego nawiązanie więzi staje się kluczowym elementem procesu edukacji. O znaczeniu bezpośredniego kontaktu są przekonani nawet biznesmeni, którzy, mimo dostępu do doskonałego sprzętu wideokonferencyjnego, wolą prowadzić negocjacje „w realu".

Wraz z urynkowieniem edukacji chęć urozmaicenia programów studiów doprowadziła do wprowadzenia modelu „składankowego", w którym każdy wąski specjalista prowadzi krótkie zajęcia w swojej dziedzinie. Jednak my zdecydowanie podkreślamy wagę budowania relacji między wykładowcą a studentami, co wymaga czasu. Dlatego liczba godzin spędzonych razem nie powinna być mniejsza niż 30, a jeszcze lepsze efekty przynoszą zajęcia trwające dłużej.

Kultura audytu

Zarządzający uczelniami dążą również do zwiększenia wydajności pracy naukowej. Zysk uczelni zależy głównie od liczby studentów płacących czesne, a ci, zanurzeni w morzu informacji, szukają prostych kryteriów wyboru szkoły wyższej. Kiedyś nikt nie czuł potrzeby **ilościowego** porównywania Uniwersytetu Warszawskiego z Uniwersytetem Jagiellońskim. Na jednym z nich lepsza była informatyka, na drugim medycyna... Aż wymyślono algorytmy, które agregują różne wskaźniki w jedną liczbę. Teraz emocjonujemy się, zastanawiając się, czy w tym roku UW wyprzedzi UJ o 2 punkty, czy może będzie odwrotnie.

Im wyżej instytucja znajduje się w rankingu, tym większa szansa na przyciągnięcie studentów zamożnych rodziców, którzy mogą opłacać wysokie czesne oraz na zdobycie kontraktów badawczych od zewnętrznych organizacji. Zarządzający uczelniami nie tylko chwalą się, że ich placówki przesunęły się w rankingu, ale również uwzględniają zmiany miejsca w rankingu w swoich planach strategicznych.

Skąd popularność rankingów? Jest tak wiele uniwersytetów, że nieliczbowa reputacja przestaje mieć znaczenie. Tworzenie rankingów to ewidentna próba

redukcji nadmiaru informacji do jednego wymiaru. Naszym zdaniem – całkowicie bezsensowna.

Konsekwencje „punktozy" są znaczące. Naukowiec proszony o przygotowanie tekstu do publikacji, jako pierwsze zadaje pytanie: „Za ile punktów?". Bo punkty nie zależą od tego, jak dobry tekst napisze, ale od tego, gdzie zostanie on opublikowany.

Prawo opisujące **korupcjogenną naturę ilościowych celów** ma trzech niezależnych od siebie autorów[150]: Goodharta, Strathern i Campbella. Mówi ono, że **miara, która staje się celem, przestaje być dobrą miarą**. Im częściej jakikolwiek ILOŚCIOWY wskaźnik społeczny jest wykorzystywany do podejmowania decyzji, tym bardziej **staje się podatny na presję korupcyjną**, co może prowadzić do zniekształcenia i korumpowania procesów społecznych, które miał monitorować.

Na przykład standaryzowane testy osiągnięć mogą być wartościowymi wskaźnikami ogólnych wyników szkolnych w warunkach normalnego nauczania ukierunkowanego na rozwój ogólnych kompetencji. Dzieje się tak jednak tylko do momentu, gdy nauczyciele nie zaczną być motywowani do uzyskiwania lepszych wyników za wszelką cenę i nie zaczną „uczyć pod testy" zamiast rozwijać u uczniów zasady myślenia krytycznego. Wśród akademików nie jest lepiej. Wykryto[151] oszustwa nawet w wykorzystywaniu portalu internetowego SSRN (*Social Science Research Network*), który umożliwiał akademikom zamieszczanie tekstów przed ich publikacją i sprawdzenie, ile razy dany artykuł został pobrany. Niektóre uniwersytety wykorzystywały liczbę pobrań SSRN jako miarę jakości badań naukowych. Analiza podejrzanych pobrań (np. szybkie pobrania w stałych odstępach czasowych) wykazała, że pojawiają się one w szczególnych momentach etapu kariery naukowca (np. przed promocją *tenure*).

Podobnie jest z „punktozą", której konsekwencje trafnie spuentował Billig[152]:

- dramatyczny **spadek poziomu intelektualnego publikacji** wynika nie z braku talentu do twórczego myślenia i pisania, lecz z nacisków ilościowych;
- nacisk wywierany na naukowców, aby publikowali w konkurencyjnym otoczeniu, w którym konieczna jest promocja wyników ich pracy powoduje, że uczą się oni **promować słowa**. Nagradzani są nie ci, którzy **piszą wtedy, gdy mają coś do powiedzenia** i wkładają dużo wysiłku w przygotowanie klarownego tekstu, ale ci, którzy **piszą i publikują masowo**.

W efekcie mamy zalew publikacji ocenianych nie na podstawie ich jakości, ale miejsca wydania. Jak pokazaliśmy to wcześniej proces recenzji działający nieźle w XX wieku, gdy nie było presji publikacyjnej, wymaga ponownego przemyślenia[153].

W opublikowanej w 2017 roku monografii **„Return to Meaning: a Social Science with Something to Say"**[154] autorzy wskazują, że jesteśmy świadkami nie tylko

[150] Bergstrom i West, 2021.
[151] Edelman i Larkin, 2009.
[152] Billig, 2013.
[153] Kulczycki, 2023.
[154] Alvesson i in., 2017.

spadku jakości badań, ale także mnożenia się bezsensownych prac, które nie mają żadnej wartości dla społeczeństwa i niewielką wartość dla samych autorów – poza zapewnieniem im zatrudnienia i możliwości awansu.

Eksplozja liczby publikacji prowadzi do powstania „zanieczyszczonego środowiska informacyjnego", które utrudnia rozwój dyscypliny. Coraz częściej używane jest określenie **„naukowy bełkot"** (scientific *bullshit*), opisujące teksty i badania – zwłaszcza w naukach społecznych – które powierzchownie spełniają naukowe standardy XX wieku, ale w rzeczywistości są nastawione na wywarcie oszukańczego wrażenia istotności i głębi. Takie prace charakteryzują się przesadnym skomplikowaniem oraz nadmiarem zbędnych informacji, które odwracają uwagę od istoty problemu.

Brytyjscy akademicy[155] twierdzą, że nigdy wcześniej ich praca nie była tak przesiąknięta biurokracją, a uniwersytety nigdy wcześniej nie poświęcały **tylu zasobów na promowanie swojej „marki"**. Wzrost „naukowego bełkotu" w szkolnictwie wyższym odzwierciedla **wiarę we wszechmoc wizerunku** oraz w plastyczność, a nawet nieistotność prawdy. Jako przykład rosnącej wagi problemu „bełkotu" w naukach społecznych można przytoczyć listę ostatnich publikacji poświęconych temu zagadnieniu:

- Ball, J. (2017). *Post-truth: how bullshit conquered the world*. Biteback Publishing.
- Davis, E. (2017). *Post-Truth: Why We Have Reached Peak Bullshit and What We Can Do About It*. Little Brown.
- Frankfurt, H. G. (2005). *On bullshit*. Cambridge University Press.
- Spicer, A. (2013). Shooting the shit: the role of bullshit in organisations. *M@n@gement, 16* (5), 653–666. https://doi.org/10.3917/mana.165.0653
- Spicer, A. (2017). *Business bullshit*. Routledge.
- Sims, M. (2020). *Bullshit towers: Neoliberalism and managerialism in universities in Australia*. Peter Lang.
- Levy, N. (2023). *Philosophy, Bullshit, and Peer Review: Elements in Epistemology*. Cambridge University Press.
- Bergstrom, C. T., West, J. D. (2021). *Calling Bullshit. The Art of Skepticism in a Data-Driven World*. Penguin Books Ltd.

Zalew publikacji powoduje, że starsze prace, choć często głębiej przemyślane, bo powstające w wolniejszym tempie, przestają być cytowane. Redaktorzy czasopism wymagają powoływania się przede wszystkim na najnowsze prace, co prowadzi do absurdów. Na przykład autorstwo teorii dylematów społecznych przypisuje się czasem badaczowi, który urodził się kilkadziesiąt lat po jej powstaniu.

Nagradzane jest publikowanie, które wymaga dostosowania się do nowych reguł gry zamiast czytania, analizowania i pogłębiania wcześniejszej wiedzy.

---

155 http://www.yiannisgabriel.com/2018/06/bullshitology-growing-discipline-or.html?fbclid=IwAR1zdgW3GZmtB_5AHvclYpaG7aqyAaUEY-6cYsGubmqFC_VilrpkIHXzcYCs

Rezultatem stał się powszechny **cynizm wśród naukowców** wobec wartości publikacji, czasami także ich własnych. Publikowanie zaczyna być postrzegane jako gra trafień i chybień, pozbawiona wewnętrznego znaczenia i wartości, a także bez troski o praktyczne zastosowania.

Naukowcy prowadzą badania nie po to, aby rozwiązać istotny problem czy przekazać coś społecznie znaczącego, lecz aby zdobyć punkty publikacyjne. **Wzrost „naukowego bełkotu"** w naukach społecznych stanowi poważny problem, który podważa sens uprawiania takiej nauki. Problem ten nie jest wyłącznie „akademicki", ponieważ prowadzi do znacznego marnotrawstwa zasobów, zawyżania opłat studenckich oraz zwiększania kosztów ponoszonych przez podatników.

W kulturze audytu miarą wartości naukowca stała się liczba jego publikacji w czasopismach z wysokim Impact Factor (IF), na podstawie którego ministerstwo przypisało punkty parametryzacyjne. Nic więc dziwnego, że w obecnym systemie naukowiec musi systematycznie publikować, nawet jeśli nie ma nic ciekawego do powiedzenia. **Prawdziwa nauka wymaga akceptacji porażek – mogą mijać długie lata pracy, zanim uzyskamy przełomowy wynik.** Naukowcy to inteligentni ludzie, więc potrafią się przystosować. Dysponując wynikami badań korelacyjnych, zawsze można zmieniać zmienne zależne i w ten sposób przygotować z jednego badania kilka publikacji, zgodnie z regułą najmniejszej publikowalnej jednostki (*least publishable unit*)[156]. Redaktorzy czasopism określają tę strategię ***slice and dice***.

Od 2017 roku na szczęście została ograniczona liczba publikacji do 4 w ciągu 4 lat.

W nauce, tak jak w ekonomii, rośnie zróżnicowanie dochodów naukowców – ci najwięksi stają się jeszcze więksi. Ten sam tekst podpisany znanym nazwiskiem ma dużo większe szanse na publikację niż tekst nieznanego autora. Teoretycznie proces *blind review* powinien być całkowicie „ślepy", ale w praktyce nie jest to takie proste, ponieważ dość łatwo można zgadnąć nazwisko autora po liście cytowanych pozycji. Dlatego niektórzy Polacy unikają cytowania prac polskich, aby nie zdradzać kraju pochodzenia autora. Podobnie granty otrzymują ci, którzy w przeszłości już kierowali grantami.

Mimo, że mamy uczyć studentów innowacyjności, sami jesteśmy bardzo konserwatywni. Nadal obowiązują standardy publikacyjne sprzed ery Internetu. Kiedyś, gdy dostęp do informacji był bardzo ograniczony, prace przeglądowe korzystające z wielu źródeł były cenne. Ceniono wysiłek zbierania trudno dostępnych informacji. Jednak wciąż pojawiają się prace przeglądowe, które niewiele wnoszą, a jedynie spełniają kryteria z czasów przed Internetem. Efekt inercji przejawia się także w braku istotnej zmiany standardów publikacyjnych.

W naukach społecznych odwołania bibliograficzne do myśli innych autorów mają na celu podkreślenie ich pierwszeństwa (*to give credit*) we wprowadzeniu danego terminu do obiegu naukowego. Dzięki temu ten kto pierwszy posłużył się danym terminem np. „smog informacyjny", zazwyczaj ma wysokie indeksy cytowań. Choć

[156] Scott-Lichter, 2011.

chętnie używamy metafory smogu i bagna informacyjnego, nie pamiętamy kto pierwszy użył tych terminów. Moglibyśmy to znaleźć, jednak zostawiamy tę pracę historykom nauki.

Podsumowując: próba oceny wydajności naukowej za pomocą liczby publikacji czy indeksu cytowań prowadzi do patologii, ponieważ nie uwzględnia faktu, że sukces uczelni jest efektem pracy zespołów, w których jedni piszą więcej, inni spędzają więcej czasu na organizacji badań, zdobywaniu funduszy czy dydaktyce. Zdecydowanie uważamy, że oceniane powinny być zespoły, a nie pojedyncze osoby. Należy także pamiętać o inteligencji pracowników, których chcemy oceniać – oni znajdą sposoby, aby wskaźniki rosły w górę.

Wpływ dominującej kultury neoliberalnej na edukację

Jeżeli przyjrzymy się założeniom kultury neoliberalnej[157] dominującej w XXI wieku, to przestajemy się dziwić opisanym wyżej zjawiskom.

Celem edukacji w kulturze neoliberalnej ukierunkowanej na kształtowanie tożsamości producenta-konsumenta nie jest rozwijanie umiejętności krytycznego postrzegania rzeczywistości ani rozpoznawania funkcjonujących mechanizmów, lecz raczej **trening do bezrefleksyjnego angażowania się w zastane praktyki społeczne**. Tego rodzaju szkolenia częściej tłumią wyobraźnię niż rozwijają refleksyjność i wynikającą z niej kreatywność.

Edukacja ma **przygotowywać do wymagań rynku pracy**, a **kwalifikacje są postrzegane jako towar**. Wartość jednostki ocenia się przez pryzmat sukcesu zawodowego, który jest nierozerwalnie związany z sukcesem finansowym. Zalecenie uczenia się przez całe życie (*Life Long Learning*, LLL) oznacza, że osoby niezdolne do przystosowania się do zmieniających się potrzeb rynku pracy zostaną zastąpione przez bardziej zaradnych.

Od nauczycieli akademickich oczekuje się, aby weszli **w rolę skutecznych marketingowców** tworzących i promujących odpłatne oferty kursów. Budując swoją markę, nawet szkoły wyższe naśladują „producentów szamponów", wykorzystując mechanizmy emocjonalne, ponieważ zakładają, że kandydaci wybierają uczelnię pod wpływem emocji, a nie rozumu.

Do polityki oświatowej zaadaptowano wolnorynkowe założenie: **od każdego według tego, jak wybiera, każdemu według tego, jak go wybierają.**

Edukacja szkolna coraz częściej jest usługą komercyjną, w której eksponuje się „efektywność" nauczania, „wydajność" nauczycieli i całych szkół, „skuteczność" prowadzonych praktyk wychowawczych. Lansowana i podsycana jest rywalizacja między uczniami, pogoń w karierze edukacyjnej za najwyższymi wynikami w testach.

[157] Michałowska, 2013.

Edukacja zinstytucjonalizowana staje się częścią rynku. Szkoły to małe przedsiębiorstwa usługowe. Nauczyciele to usługodawcy, od których wymaga się fachowego świadczenia edukacyjnych usług i jak w handlu można składać im reklamacje. Tymczasem sedno edukacji – rozwój poznawczy i psychofizyczny, stymulowanie ciekawości, rozwijanie kreatywności i pasji – zostaje zastąpione walką o najlepsze wyniki w testach i zdobycie dyplomu renomowanych szkół średnich, potem wyższych, a następnie wysokich i dobrze płatnych stanowisk pracy.

Zgodnie z zasadą ekonomiczną: im większa specjalizacja produkcji, tym mniejsze koszty, a kształcenie ogólne jest wypierane przez kształcenie specjalistyczne.

Neoliberalizm wymusza finansowanie oparte na wymiernych wskaźnikach. Ponieważ **cech, takich jak refleksyjność czy mądrość nie da się łatwo zmierzyć**, ocenie poddaje się to, co mniej istotne, ale łatwe do kwantyfikacji (np. punkty publikacyjne czy procentowy udział studentów zagranicznych – co często faworyzuje małe szkoły). Przykłady takich absurdów można mnożyć. Jeden z amerykańskich profesorów[158] w następujący sposób krytycznie ocenia system wyceny profesorów na wiodącej uczelni MIT:

> *Każdy profesor musiał zdobyć sto dwanaście punktów, które wyliczało się w oparciu o czas spędzony w sali wykładowej, liczbę kształconych studentów, liczbę godzin przepracowanych przez studentów, ogólny wymiar godzin zajęciowych i tak dalej. Jak się okazało, dość łatwo było ograć system i skupić się na własnych badaniach. Byłem w tym tak dobry, że któregoś roku udało mi się przetrwać, prowadząc jedynie dwanaście trzygodzinnych sesji wykładowych dla dużej grupy osób.*

Potrzeba reformy?

Czy to, co napisaliśmy wcześniej, ma być zachętą do reformy uniwersytetów? Niestety nie – za sobą mamy już zbyt wiele reform. Wciąż wierzymy w piękne programy zmian, słuchając kandydatów na najwyższe urzędy, lekceważąc nie tylko koszty finansowe, ale również koszty psychologiczne.

Niestety, coraz więcej pomysłów decydentów wprowadza się bez wcześniejszego przeprowadzania badań pilotażowych. Przedwcześnie zmarły filozof brytyjski[159] przestrzegał przed fałszywymi założeniami, które staraliśmy się wyłuskać z jego bardzo interesującego wywodu.

Przykłady niebezpiecznych iluzji wielbicieli reform to:

- Możemy kolektywnie zbliżyć się do naszych celów, przyjmując wspólny plan i działając na rzecz jego realizacji pod kierunkiem władzy centralnej. Ludzi można zorganizować na wzór wojska z hierarchicznym systemem rozkazodaw-

[158] Ariely, 2021.

[159] Scruton, 2012.

stwa i obowiązków, który zapewnia skuteczną koordynację działań ogromnej liczby osób realizujących plan zaprojektowany przez nielicznych.
- Świat NIE jest postrzegany jako złożony z ludzi wraz z ich moralnymi słabościami i egoistycznymi dążeniami, lecz coraz częściej traktowany jak zbiór wykresów i wskaźników symbolizujących odległość od wyimaginowanego celu (np. efektywność procesu kształcenia mierzona czasem trwania studiów czy miejsce w rankingach).

Realiści wiedzą jednak, że:
1. System edukacji nie jest budynkiem, który można zburzyć i odbudować od podstaw według nowego, lepszego planu. Zwyczaj, który przetrwał próbę czasu, opiera się na zakamuflowanych doświadczeniach minionych pokoleń i stanowi barierę dla iluzji, że można wszystko stworzyć od nowa, w pełni zgodnie z jakimś stuprocentowo racjonalnym planem.
2. Racjonalny plan, który narzuca kolektywny cel tam, gdzie nie da się go sensownie dostosować do zmieniających się pragnień i potrzeb jednostek, zawsze prowadzi do nieprzewidywalnych i niekorzystnych konsekwencji. Może to doprowadzić do zapaści, jeśli w systemie zmian brakuje mechanizmów umożliwiających korektę planu.
3. Efektywne rozwiązania problemów zbiorowych nie są narzucane, lecz odkrywane i negocjowane w długiej perspektywie czasowej.
4. Gdy ktoś proponuje poważne zmiany, obiecując ogromne korzyści, to na nim spoczywa obowiązek udowodnienia, że prawdopodobieństwo ich osiągnięcia jest wysokie.

Niestety, coraz częściej podporządkowujemy się planom reformatorów, którzy sądzą, że mogą narzucić nam kolektywne cele, a następnie wymyślić sposoby ich realizacji. Nikogo nie zaskoczy, gdy wspomnę, że Scruton, oprócz standardowych przykładów wprowadzenia „jedynie słusznego ustroju politycznego" (młodszym Czytelnikom przypominamy, że tak określano socjalizm), wskazuje również na udział Unii Europejskiej.

Polska zdolność adaptacji, nawet do absurdalnych rozporządzeń – wyćwiczona w poprzednim systemie politycznym – spowodowała, że bardzo sprawnie wdrożyliśmy na uczelniach proces boloński (program podpisany w 1999 roku przez ministrów z 29 krajów europejskich). Cel był szczytny: **ujednolicenie i porównywalność europejskiego systemu edukacji**. Tylko jak wygląda jego implementacja w Polsce?

Student, który ma licencjat z japonistyki, zapisuje się i płaci za magisterskie studia zarządzania. Oznacza to, że studiuje metody statystyczne razem z osobami, które mają licencjat z zarządzania. Absurd. Po prostu przy implementacji zapomniano o kursach wyrównawczych dla osób zmieniających profil kształcenia. Wiedza i umiejętności dwóch magistrów zarządzania – jednego z licencjatem z zarzą-

dzania, a drugiego z licencjatem z innej dziedziny – są nieporównywalne, a dyplomy są takie same. Ten problem występuje na wszystkich studiach dwupoziomowych. Ujednolicone efekty kształcenia to świetny pomysł, ale przy braku standaryzacji pomiaru tych efektów przekładają się one na zabawę słowami.

Nawet najlepszy program bez planu jego implementacji to przepis na absurdalne wyniki, bowiem – jak dowodzone jest dalej – podstawa **pokojowej koegzystencji** nie opiera się na realizacji wspólnych wizji, lecz **akceptacji samoograniczeń**. Sprawdzające się od tysiącleci dziesięć przykazań w dużej części składa się z zakazów, a nie z wizji powszechnej szczęśliwości. Dlatego zamiast kolejnej reformatorskiej wizji, ważniejsze byłoby uzyskanie konsensusu w kwestii samoograniczeń w szkolnictwie (np. nie promujemy nikogo, kto nie spełnia warunków).

Podsumowując:

1. Współczesne uniwersytety są wspólnotami ludzi poszukujących nie prawdy, jak to było kiedyś, lecz sukcesu biznesowego, który osiąga się poprzez konkurencję – między jednostkami, uczelniami, ponieważ funkcjonują jak firmy w warunkach wolnorynkowych.
2. W czasach masowego kształcenia akademickiego, masowej produkcji badań i publikacji, nagradzani są nie ci, którzy piszą, gdy mają coś istotnego do powiedzenia, ale ci, którzy piszą i publikują bez przerwy.
3. Naukowcy są bardziej lojalni wobec własnego wydziału niż całej uczelni. Jednostki organizacyjne tego samego uniwersytetu zachowują się jak plemiona walczące o jak największe terytorium i przywileje kosztem swoich sąsiadów.
4. Podział na subdyscypliny ma sens w naukach matematycznych czy technicznych, ale w przypadku nauk społecznych taka subdyscyplinarność jest szkodliwa, ponieważ obiektu naszych badań nie da się podzielić na niezależne fragmenty.
5. Model siłowni intelektualnej zakłada, że indywidualizacja procesu edukacji powinna dotyczyć bardziej poziomu wymagań, a nie przedmiotów. Każdy powinien uczyć się matematyki, choć od jednych należy wymagać minimum, a zdolniejszych obciążać większymi wymaganiami.
6. **Absurdem** jest **twierdzenie**, że **uniwersytety świadczą usługi edukacyjne**, skoro jakość tej „usługi" zależy w równym stopniu od wysiłku wykładowców („usługodawców") i studentów („usługobiorców").
7. Studenci preferują zajęcia nastawione na zdobycie potrzebnych na rynku pracy umiejętności praktycznych, opisywane przez nich często jako „atrakcyjne i łatwe", zamiast zajęć „trudnych i znojnych", nie zdając sobie sprawy, że do odniesienia sukcesu w konkurencji na globalnym rynku potrzebna jest przede wszystkim sprawność myślenia, którą kształci się, wychodząc poza strefę komfortu.
8. Edukacja jest aktywnością społeczną z niezbywalną rolą nauczyciela, który przejmuje kontrolę nad przebiegiem procesu uczenia się i stanowi punkt odniesienia dla postępów grupy.

9. Rola nauczyciela akademickiego zmieniła się z **dostarczyciela wiedzy** na **menedżera procesu uczenia się**, który powinien – podobnie jak trener – planować naukę, biorąc pod uwagę możliwości studenta, motywować go i kontrolować efekty.
10. Uczelnia przyszłości to przede wszystkim miejsce zdobywania doświadczeń w kontrolowanych warunkach.
11. W dobie nadmiaru informacji nagradzaną umiejętnością powinna być integracja i synteza istniejących idei.

Cieszą nas ostatnie głosy Rady Doskonałości Naukowej[160], które warto rozpowszechniać:

- artykuły w czasopismach naukowych albo monografie same w sobie NIE są osiągnięciami naukowymi. Osiągnięciami takimi mogą być uzyskane i opisane w artykułach lub książkach **wyniki badań** i wyciągnięte na ich podstawie **wnioski**;
- o wartości naukowej osiągnięcia nie decydują:
  - wskaźniki bibliometryczne ani żadne inne parametry naukometryczne;
  - miejsce jego opublikowania (czasopismo, wydawnictwo), ale treść pracy;
- o wartości naukowej osiągnięcia decyduje treść publikacji opisującej uzyskane wyniki i wyciągnięte wnioski;
- nie należy oczekiwać od recenzentów analizy bibliometrycznej prac naukowych, ponieważ tego typu analiza może być dokonana przez dowolnego pracownika administracyjnego zatrudnionego w instytucji naukowej.

Aby zakończyć RAFY METODOLOGICZNE w sposób konstruktywny, przedstawiam paradygmat metodologiczny opracowany na potrzeby rozpraw doktorskich przygotowywanych pod moim kierunkiem. Uwzględnia on fakt, że „**ludzie, których badamy, różnią się od śrubokrętów**”.

Ze względu na to, że ten tekst był zamieszczany jako **samodzielny dokument** w rozprawach doktorskich – **niektóre zdania muszą być powtórzone z wcześniejszych rozważań**.

[160] Gorynia, 2025.

Sekcja 5
Paradygmat metodologiczny WiW dla badań prowadzonych w obszarze ZZL (HRM)

Wyniki badań w obszarze ZZL nie prowadzą do konstrukcji niezmiennych praw, a jedynie pozostają ograniczonymi społecznie, kulturowo i historycznie uogólnieniami[161]. Sformułowanie programu badawczego wymaga nie tylko ustalenia obszaru badań, lecz także sprecyzowania samego celu tych badań[162]. To, jakie instrumentarium badawcze zastosujemy, zależy od przyjętego celu badań oraz możliwości jego realizacji. Badamy to, co jest obserwowalne, mierzalne oraz podatne na eksperymenty. Nauka opiera się na dowodach empirycznych.

Ustalenia terminologiczne

Wszystkie dane uzyskane poprzez zadawanie pytań pracownikom nazywamy danymi ankietowymi. Osoby biorące udział w badaniach ankietowych, eksperymentach czy wywiadach są nazywane respondentami, ponieważ ich reakcje (odpowiedzi) stanowią przedmiot analizy. Dane pochodzące z pomiaru osób mogą być wyrażone liczbami, wówczas mówimy o badaniach/analizach ilościowych, lub słowami, które są najczęściej składnikiem badań/analiz jakościowych.

Dane ilościowe to zbiory liczb, które poddawane są analizie statystycznej. Dane jakościowe to zbiory słów, które stanowią próbę opisu różnych wizji badanego zjawiska (rzeczywistość jest w oku patrzącego), a następnie są poddawane analizie interpretacyjnej przez badacza. Analiza ta może obejmować elementy obiektywizujące, takie jak klasyfikacja wypowiedzi przez niezależnych sędziów czy zliczanie częstości używania różnych sformułowań.

---

[161] Sułkowski, 2011.

[162] Niemczyk, 2011.

Badania ilościowe różnią się od badań jakościowych stopniem proceduralizacji metod analizy. Celem badań ilościowych jest zazwyczaj obiektywne testowanie hipotez zakładających relacje między zmiennymi, podczas gdy celem badań jakościowych jest głównie rozpoznanie indywidualnych sposobów postrzegania rzeczywistości.

Pluralizm/eklektyzm metodologiczny + pragmatyzm w wyborze problemu

Paradygmat WiW odrzuca zarówno anarchizm (akceptujący dowolne metody i techniki, nawet zaczerpnięte z jednostkowych doświadczeń), jak i fundamentalizm metodologiczny, który nie dopuszcza mieszania różnych metod badawczych. Jest zgodny z postulatem, że metody badawcze w ZZL powinny być stosowane refleksyjnie, ponieważ mają charakter heurystyczny, co uniemożliwia ich algorytmizację. Dlatego zaleca pluralizm, a nawet eklektyzm metodologiczny, który akceptuje stosowanie metod zaczerpniętych z różnych dyscyplin i podejść teoretycznych w celu rozstrzygnięcia problemu badawczego[163].

Na etapie wyboru problemu badawczego zaleca się stosowanie podejścia pragmatycznego zakładającego, że jeśli analizowany problem badawczy nie ma ważnych konsekwencji praktycznych, to nie warto się nim zajmować, zostawiając tego typu rozważania naukom podstawowym.

Specyfika obiektu badań

Metodycy zapominają, że badanie przedmiotów nieożywionych rządzi się innymi prawami niż badanie ludzi. Co gorsza, mamy do czynienia z prowadzeniem badań „ludzi przez ludzi". Specyfika badań ZZL polega na tym, że obiektem pomiaru są ludzie, którzy tworzą znaczenia, czyli ich reakcje na bodźce powstają za pośrednictwem ich oczekiwań i interpretacji wyznaczonych w dużym stopniu przez zapis wcześniejszych doświadczeń. Dlatego w odróżnieniu od nauk ścisłych w ZZL każda replikacja badania jest sukcesem, ponieważ zawsze zmienia się grupa badanych pracowników, ich doświadczenie i kontekst kulturowy. Obiektem analiz w badaniach ZZL są **fakty psychiczne**, czyli najczęściej odpowiedzi ludzi (słowne lub skategoryzowane na skalach liczbowych) na zadane pytania. Trzeba pamiętać, że tego typu dane ilościowe są prawie zawsze zniekształcone, co wykazano w wielu badaniach[164].

Model procesu odpowiadania na pytania[165] pokazuje, dlaczego odpowiedzi respondentów są tak zróżnicowane. Odpowiadanie na pytanie o ocenę, na przykład

[163] Sułkowski, 2011.
[164] Wieczorkowska i Wierzbiński, 2013.
[165] Slaughter i Rhoades, 2004, 2010.

satysfakcji z pracy, wymaga aktywizacji różnych informacji zawartych w pamięci długotrwałej – zarówno w jej części semantycznej (np. co to znaczy być zadowolonym), jak i epizodycznej (np. przypomnienia sobie różnych stanów emocjonalnych). Przywoływane informacje zgodnie z koncepcją świadomości nazwaną modelem szkiców wielokrotnych, podlegają ciągłemu redagowaniu. W żadnym momencie tego procesu nie można powiedzieć, że edytowanie jest ukończone i świadomie doświadczany jego ostateczny wynik. W danej chwili przypominamy sobie najgorsze epizody, za godzinę możemy przywołać informacje, które diametralnie zmieniają nasz osąd. Będąc w dobrym nastroju szukamy pozytywnych aspektów pracy w firmie, a w złym „szukamy dziury w całym". Respondenci, wypełniając ankietę, niezwykle rzadko mają gotowe oceny satysfakcji „w głowach". Założenie, że na stałe archiwizujemy różne opinie, jest mało przekonujące. Alternatywne założenie mówi, że konstruujemy je na bieżąco, kiedy są potrzebne. Specyficzne cele, standardy, oceny i postawy generują dalsze informacje. Mamy zakodowane w naszych umysłach różne ogólne opinie, cele, standardy i postawy, które pozwalają generować następne. Są one niezbędne do powstania emocji, gdyż bez nich nie sposób nadać jakiegokolwiek znaczenia napotykanym zdarzeniom. Większość reprezentacji poznawczych (np. poglądy na temat roli pracy w życiu), o które pytamy, nie jest reprezentowana w umyśle przed zainicjowaniem oceny. Takie reprezentacje można określić jako wirtualne (bo nie istnieją przed zadaniem pytania). Nasze podejście różni się istotnie od tradycyjnego podejścia teorii pomiaru, w którym zakłada się, że badany ma już ustaloną „prawdziwą" odpowiedź – taką, którą udzieliłby sobie sam, więc podstawowym problemem jest minimalizacja błędu pomiaru spowodowanego przez formę pytania lub kontekst społeczny.

Każda ocena wymaga umiejętności koncentracji uwagi, która umożliwia selekcję informacji, pomijanie lub przynajmniej zablokowanie tych, które mają mniejsze znaczenie. W trakcie przekształcania myśli w wypowiedź pojawia się łańcuch asocjacji. Każde słowo, zwłaszcza wieloznaczne, uruchamia sekwencje skojarzeń, które mogą biec w różnych, nawet bardzo rozbieżnych kierunkach. W pamięci trwałej zakodowanych jest wiele schematów poznawczych, „gotowych" do interpretacji danego słowa. Umysł zazwyczaj przesiewa skojarzenia i wybiera tylko te, które mają związek z myślą, jaką chcemy wyrazić. Im dokładniejszy jest ten odsiew informacji, tym skuteczniej może przebiegać kolejne stadium przetwarzania związane ze świadomą uwagą. Jedynie niewielka część tego procesu może zostać uświadomiona – co nie znaczy, że nie możemy przejąć kontroli i skierować naszą uwagę na różne aspekty zagadnienia. Świadomość modyfikuje działanie tego filtra. Jeśli chcemy, wywołujemy odpowiednie informacje z pamięci długotrwałej, które będą następnie filtrować napływające informacje.

Podsumowując – musimy zdawać sobie sprawę, że respondenci bardzo często nie mają gotowej odpowiedzi i tworzą ją dopiero wtedy, gdy padają pytania. Często nie odtwarzają swoich opinii, ale je konstruują. To, jaką opinię sformułują, zależy od

tego, którą z czterech strategii formułowania sądu zastosują: 1) odtwarzanie gotowych ocen, 2) przetwarzanie zmotywowane, 3) przetwarzanie heurystyczne (uproszczone), 4) przetwarzanie analityczne (szczegółowe).

O wyborze strategii przetwarzania informacji decydują możliwości poznawcze respondenta (np. poziom refleksyjności), stan organizmu (np. przeciążenie, nastrój) oraz cele, które wyznaczają stopień zaangażowania. Wybór zależy również od cech przedmiotu oceny (stopień znajomości i złożoności) i cech sytuacji (np. presja czasowa, społeczna aprobata, kosztowność błędów).

W badaniach ankietowych respondenci z powodu ograniczeń czasowych oraz braku konsekwencji związanych z udzieleniem nietrafnego sądu, niezwykle rzadko stosują strategię analityczną. Dlatego powinniśmy pamiętać o następujących kwestiach:

1. Zachowanie realizmu psychologicznego[166] badania – pytania powinny wzbudzać zainteresowanie. Respondent chce zrozumieć nie tylko O CO jest pytany, ale także PO CO? Bardzo ważne jest, aby zadbać o odpowiedni poziom motywacji, oferując tam, gdzie to możliwe, spersonalizowaną informację zwrotną.
2. Pytani nie mają w głowach gotowych odpowiedzi i powinni mieć prawo do udzielenia beztreściowej odpowiedzi – „nie wiem", „nie dotyczy" lub pominięcia odpowiedzi. Zmuszanie ich do udzielenia odpowiedzi może prowadzić do irytacji i udzielania przypadkowych odpowiedzi na kolejne pytania.
3. Pytany, jeśli może unika wysiłku – chętnie korzysta z opcji środkowych, dlatego należy ich unikać, oferując opcję „trudno powiedzieć" poza skalą odpowiedzi. W badaniach pokazano, że brak opcji środkowej nie powoduje istotnego wzrostu liczby odpowiedzi beztreściowych.

Konkludując: odpowiedzi respondentów, które poddajemy dalszym analizom, mają różną wartość poznawczą. Na nic zdadzą się wyrafinowane metody analizy danych, jeśli same dane będą zniekształcone w sposób losowy.

Pojęcia naukowe i definicje operacyjne

W nauce posługujemy się równolegle językiem obserwacji i językiem teorii. W języku teorii używamy pojęć naukowych (konstrukty teoretyczne, zmienne latentne), takich jak styl przywództwa, potrzeba dominacji, dobrostan emocjonalny pracownika itp., które muszą zostać przetłumaczone na język obserwacji.

W paradygmacie WiW uznaje się, że badane konstrukty teoretyczne są pojęciami naturalnymi, których nie można zdefiniować w sposób klasyczny, za pomocą warunków koniecznych i wystarczających. Rozwiązaniem tego problemu jest operacjonizm[167], który zakłada, że pojęcia naukowe nie wychwytują istoty rzeczy, lecz

[166] Aronson i Wieczorkowska, 2001.

[167] Tatarkiewicz, 1951/2014.

jedynie określają działania badacza oraz jego operacje psychofizyczne potrzebne do zdefiniowania badanej rzeczy.

Do budowania wskaźników wykorzystujemy różne narzędzia pomiarowe. Przykładem mogą być zestawy pytań opracowane do pomiaru cech pracownika. Takie zestawy pytań nazywane są skalami (np. Skala Lęku) lub testami psychologicznymi, które można traktować jako odmianę *narzędzi kalibrowanych*[168].

Do analizy badań ilościowych stosuje się podejście pozytywistyczne[169], które zakłada, że przedmiotem badań są fakty przedstawiane w języku wartości zmiennych. W dociekaniach naukowych ZZL, w których obiektem badań są ludzie (pojedynczo lub zbiorowo), opisano setki zmiennych i ich operacjonalizacji. Można odnieść wrażenie, że wprowadzenie kolejnego pojęcia naukowego do opisu osoby jest nadmiernie akceptowane. Dlatego badacz musi starannie wybrać zmienne, które będą przedmiotem jego dociekań, opisując model teoretyczny badanego zjawiska oraz model pomiaru konstrukcji teoretycznych.

Zadanie badacza nie ogranicza się jedynie do rejestrowania faktów i praw rządzących tymi faktami, ale polega na takim ich porządkowaniu w modelach teoretycznych, aby na ich podstawie móc przewidywać kolejne fakty.

Modele teoretyczne

W ZZL poznanie dokonuje się głównie poprzez testowanie modeli, a nie obserwacje[170]. Pierwszym krokiem jest więc wybór modeli na podstawie przeglądu literatury zmiennych teoretycznych (pojęć naukowych), które posłużą do modelowania interesującego badacza zjawiska.

Model teoretyczny powinien:

- cechować się prostotą – z faktu, że rzeczywistość jest skomplikowana, nie wynika, że model powinien być skomplikowany[171];
- nie być sprzeczny z dostępnymi faktami naukowymi, chyba że jego celem jest przedstawienie alternatywnej interpretacji tych faktów;
- być logiczny i wewnętrznie spójny[172];
- dawać możliwość przewidywania;
- być weryfikowalny empirycznie.

Model teoretyczny, który został potwierdzony w wielu badaniach, może być nazwany teorią. Każdy model w ZZL składa się z części apriorycznej – założenia, że wybrane zmienne są ważne i istotne, oraz zbioru hipotetycznych związków między

[168] Brzeziński, 2019.
[169] Tatarkiewicz, 1951/2014.
[170] McKelvey, 2017; Czakon, 2011.
[171] Jak mawiał prof. Robert Zajonc – wybitny psycholog pieczołowicie dbający o wartość metodologiczną badań.
[172] Burnewicz, 2021.

zmiennymi, które są poddawane precyzyjnym testom empirycznym. Oprócz modelu teoretycznego, należy wyspecyfikować model pomiaru – czyli sposób operacjonalizacji wszystkich zmiennych. Hipotezy to zdania o zależnościach między zmiennymi wyspecyfikowanymi w modelu teoretycznym, które są możliwe do falsyfikacji.

Pięć rodzajów triangulacji

Paradygmat WiW zaleca 5 rodzajów triangulacji: (1) metod, (2) danych, (3) operacjonalizacji, (4) sposobów analizy, (5) badacza.

Triangulacja metod: Nawet w ankietach internetowych możemy łączyć metody korelacyjne, eksperymentalne i jakościowe. Numeryczne odpowiedzi na pytania zamknięte analizujemy metodami ilościowymi, słowne odpowiedzi na pytania otwarte analizujemy metodami jakościowymi.

Triangulacja danych: Dostępność reprezentatywnych prób losowych w naukach społecznych jest bardzo ograniczona, ponieważ, choć ludzi można wylosować, nie można ich zmusić do udziału w badaniach. W związku z tym, w większości przypadków badania prowadzone są na próbach dogodnych składających się z osób, które zgodziły się wziąć udział w badaniu. Trafność zewnętrzną zwiększamy poprzez replikowanie badań na różnych próbach dogodnych. Oznacza to, że powinniśmy testować te same hipotezy na różnych zestawach danych.

Triangulacja operacjonalizacji: W ZZL nie ma standardowych operacjonalizacji zmiennych. Operacjonalizacja zmiennych powinna być starannie dobrana, uwzględniając specyfikę próby. Na przykład pozycja „Łatwiej podejmuję decyzje pod presją czasu" jest dobrym wskaźnikiem niskiej reaktywności w grupie młodych pracowników, ale nie wśród menedżerów. Nawet, jeśli używamy wystandaryzowanych, gotowych narzędzi pomiarowych, należy sprawdzić ich własności psychometryczne na badanej próbie.

Triangulacja sposobów analiz: Choć w analizach ilościowych przyjmuje się założenie o neutralności aksjologicznej nauki i nieingerencji badacza, to jednak nawet w sproceduralizowanych, zobiektywizowanych analizach statystycznych badacz musi podjąć decyzję dotyczącą sposobu „czyszczenia" zbioru danych, budowania wskaźników, wyboru założeń dotyczących poziomu pomiaru oraz wyboru testów statystycznych. Decyzja, czy wynik w kwestionariuszu traktować jako zmienną ciągłą czy porządkową (np. po podziale medianowym), może doprowadzić do różnych konkluzji. Dlatego paradygmat WiW zaleca triangulację metod ilościowych w analizie zbioru danych.

Przy analizie danych jakościowych – słów, zaleca się triangulację badacza. Dane powinny być kodowane przez co najmniej dwie osoby niezależnie od siebie.

Trafność zewnętrzna i wewnętrzna badań

Trafność zewnętrzną zwiększamy poprzez stosowanie różnych typów triangulacji – w szczególności poprzez testowanie tych samych hipotez na różnych zbiorach danych.

Tam, gdzie to możliwe, powinniśmy zadbać o TRAFNOŚĆ WEWNĘTRZNĄ badania. Nawet w badaniach ankietowych możemy manipulować zmiennymi niezależnymi, czyli prowadzić badania eksperymentalne, przydzielając ochotników losowo do różnych warunków eksperymentalnych.

Tam, gdzie to możliwe, zarówno w ankietach, jak i w wywiadach, wprowadzamy OPISY obiektów, których ocenę chcemy poznać. Przykładowo, prosząc pracowników o opinię na temat ich szefa nie jesteśmy w stanie określić, na ile wynika ona z percepcji pracownika, a na ile z obiektywnych cech szefa. Prosząc o ocenę wzorcowego opisu, np. dominującego lub partnerskiego szefa, możemy badać różnice indywidualne w ocenie różnych cech, które były podstawą przy konstrukcji tych opisów.

Badanie o wysokiej trafności wewnętrznej gwarantuje, że zmierzone zmiany w zmiennej wyjaśnianej nie są wynikiem zmiennych zakłócających, które zostały pominięte w modelu teoretycznym. Jedynym rodzajem badań zapewniającym wysoką trafność wewnętrzną są dobrze przeprowadzone badania eksperymentalne. Badania korelacyjne nigdy nie są wolne od ryzyka wykrywania korelacji pozornych. Paradygmat metodologiczny WiW promuje eksperymentalne badania porównawcze z zastosowaniem zasady jednej różnicy Milla, akceptując jednak, że badania eksperymentalne są często niemożliwe do przeprowadzenia z powodu braku możliwości manipulowania wartościami zmiennych lub z powodu niemożności badania zjawisk rozciągniętych w czasie.

Jakość danych

Przed przystąpieniem do analizy zbiory danych powinny być starannie oczyszczone z „fałszywych”[173] respondentów, którzy np. udzielali losowych odpowiedzi. Standardowe narzędzia pomiarowe stosowane w badaniach powinny być sprawdzone pod względem właściwości psychometrycznych/przystosowane w badanej grupie respondentów.

[173] Wieczorkowska i Wierzbiński, 2013; Kabut, 2021.

Typologie i ilościowe eksperymentalne studia przypadków

Zastosowanie wiedzy, np. o pozytywnych zależnościach typu „bardziej demokratyczne przywództwo – lepsza produktywność", **zawsze wiąże się z niepewnością.** Na przykład musimy rozważyć czy firma, którą chcemy ratować, nie jest przypadkiem wyjątkiem od naszej pięknej „liniowej zależności". **Zamiast naśladować model fizyki, powinniśmy czerpać inspirację z medycyny.** W medycynie choroby są często klasyfikowane na podstawie konkretnych objawów, kryteriów diagnostycznych lub mechanizmów biologicznych. Takie podejście typologiczne umożliwia identyfikację i klasyfikację różnych schorzeń, a także ułatwia diagnozowanie i leczenie.

Analogicznie, w biologicznej taksonomii organizmy są klasyfikowane do różnych gatunków, rodzajów i wyższych kategorii taksonomicznych na podstawie wspólnych cech. Taka klasyfikacja typologiczna pomaga zrozumieć różnorodność i relacje między organizmami żywymi.

Wyniki wielozmiennowej analizy w paradygmacie *ceteris paribus* są trudne do zastosowania w praktyce, dlatego promowaną przez paradygmat metodologiczny WiW formą badań pracowników są ILOŚCIOWE **eksperymentalne studia przypadku**, w których manipuluje się wartościami zmiennych w wybranych punktach czasowych i przeprowadza ilościowe pomiary przez długi czas.

W tym kierunku podąża medycyna[174], w której coraz częściej możliwa jest personalizacja leczenia chorób dzięki precyzyjniejszej diagnostyce na poziomie indywidualnym. Nie jest tajemnicą, że w przypadku wielu chorób istnieją znaczne różnice w reakcjach poszczególnych osób na konkretne metody leczenia.

[174] Poldrack, 2022.

Sekcja 6
Elementy vademecum badacza opinii

Choć głęboko wierzę, że postęp w naukach o zarządzaniu ludźmi dokona się przede wszystkim dzięki wprowadzeniu technologii do pomiarów, to nadal podstawową metodą badawczą pozostaje zbieranie odpowiedzi respondentów na zadawane pytania.

W tej sekcji zebrałam wyniki mojego 45-letniego doświadczenia badawczego – ilustrowane wynikami badań i przykładami ankiet, które widziałam w ostatniej dekadzie (nie podaję ich autorów, ponieważ moja krytyka nie ma charakteru *ad personam*). Łatwiej jest uczyć się na cudzych błędach niż na własnych.

Choć powszechnie panuje przekonanie, że „każdy może ułożyć pytania do ankiety", jest to założenie fałszywe. W sieci krąży ogromna liczba źle skonstruowanych ankiet. Szkolne błędy na poziomie konstrukcji marnują wysiłek respondentów i – zgodnie z zasadą *garbage in, garbage out* – prowadzą do fałszywych wniosków wyciąganych na podstawie źle zebranych danych.

Niestety, umiejętność zadawania trafnych pytań to wciąż bardziej sztuka niż rzemiosło.

O tym czy dane pytanie jest dobre, czy słabe, możemy się przekonać dopiero po analizie odpowiedzi – dlatego nie sposób przecenić roli badań pilotażowych.

Warto na przykład zaprosić ochotników i poprosić, by odpowiadając na pytania „myśleli głośno". W ten sposób można zauważyć, w których miejscach czują się zagubieni, znudzeni (obserwując np. ich niewerbalne sygnały), a które pytania wzbudzają opór lub niechęć.

Ekspert może nas uchronić przed podstawowymi błędami konstrukcyjnymi, ale dopiero wyniki badań pilotażowych pozwalają realnie ocenić jakość narzędzia badawczego i jego funkcjonalność.

W tej sekcji omówimy pułapki metodologiczne, które mogą stać się rafami, jeśli na nie wpadniemy. To na nich właśnie rozbiją się nasze wysiłki badawcze.

Omówimy kolejno:
P1. Pułapki wyboru respondentów
P2. Pułapki na etapie układania pytań
P3. Pułapki na etapie wyboru skali odpowiedzi/kafeterii
P4. Pułapki odpowiedzi środkowych i beztreściowych
P5. Pułapki kolejności pytań
P6. Pułapki konstrukcji kafeterii odpowiedzi
P7. Pułapki analiz i interpretacji wyników

P1. Pułapki wyboru respondentów

W książce używałam skrótu „badanie ludzi", choć trzeba pamiętać, że badamy prawie wyłącznie ochotników – czyli osoby, które zgodziły się na udział w badaniu, gdy zostały wylosowane, poproszone lub same się zgłosiły. Nawet w dużych badaniach prowadzonych w zakładach pracy – np. w KGHM – dyrekcja przekazywała 6 zł na fundusz socjalny za każdą wypełnioną ankietę dotyczącą zaangażowania pracowników. Był to silny motywator, dzięki któremu stopień realizacji próby w naszych badaniach wyniósł 71,3% (z 18 607 pracowników). Oczywiście nie była to próba losowa, lecz dogodna – uczestniczyli w niej tylko ci, którzy chcieli, ale duża liczba komentarzy słownych pokazywała zaangażowanie respondentów.

Najniższy procent uczestniczących w badaniach pracowników w poszczególnych jednostkach wyniósł 53,1%, a najwyższy 90,4%. Duża liczba zmotywowanych finansowo respondentów pokazuje, jak kluczowa jest kontrola jakości. Trzeba szczególnie uważnie śledzić sygnały ostrzegawcze, o których pisałam wcześniej – i myśleć o kontroli jakości już na etapie projektowania ankiety.

Podstawą formułowania wniosków naukowych jest porównanie. Sam opis stylu zarządzania w firmie, która odniosła sukces, niewiele nam da – jeśli nie zestawimy go z inną firmą działającą w podobnych warunkach, lecz zarządzaną inaczej. Mam nadzieję, że przedstawione dalej przykłady przekonają Czytelnika, że pytania w rodzaju: „Ilu studentów jest proaktywnych, a ilu skupia się głównie na przyjemnościach?" – nie mają większego sensu, bo kontekst może znacząco wpływać na przesuwanie się rozkładu odpowiedzi. Zamiast tego – nawet w ankietach – możemy prowadzić badania eksperymentalne, które dzięki losowemu podziałowi na grupy pokażą, jakie czynniki aktywizują gotowość do działania, a jakie ją obniżają.

Zdarza się, choć raczej rzadko niż często, że chcemy poznać rozkłady zmiennych w populacji. Wtedy potrzebujemy do badania próby **reprezentatywnej** dla dorosłych Polaków, np. mających prawa wyborcze. Musi być wylosowana za pomocą operatu, który daje każdemu Polakowi równe szanse bycia wylosowanym. Takim operatem może być PESEL dodatkowo zawierający informacje o wieku i płci. Stosuje się też operaty adresowe gospodarstw domowych – wtedy respondent jest losowany

spośród osób tworzących wspólne gospodarstwo. To powoduje zmniejszone szanse bycia wylosowanym w gospodarstwach wieloosobowych i pewność bycia wylosowanym w gospodarstwach jednoosobowych.

Próba jest reprezentatywna, jeśli socjodemograficzna struktura przebadanej próby odpowiada socjodemograficznej strukturze danej populacji.

Odkąd naukowcy pokazali, w jaki sposób dobrać próbę reprezentatywną z populacji[175] możliwy stał się trafny opis populacji na podstawie informacji uzyskanych od około tysiąca starannie wylosowanych osób (bez względu na to czy losujemy z wielu miliardów Chińczyków, czy z o wiele mniej licznej grupy Maltańczyków). Niezależnie od tego czy opisujemy populację kraju, czy populację znacznie mniejszą (np. polskich pracowników naukowych), procedury są identyczne.

Próba dobierana jest na podstawie operatu losowania lub listy członków badanej populacji.

Możliwość uogólnienia wyników próby na całą populację zależy od reprezentatywności próby. Jeśli procedury doboru próby prowadzą do nadreprezentowania lub niedoreprezentowania pewnych warstw populacji, mówimy o tendencyjnym doborze próby.

Losując próbę reprezentatywną ze spisu wszystkich polskich pracowników naukowych[176], procent kobiet zależy od tego, jaka dziedzina naukowa jest przedmiotem naszego zainteresowania. W wylosowanej próbie kobiet powinno być: ok. **51%** w naukach humanistycznych i społecznych, ale tylko ok. 29% w naukach inżynieryjno-technicznych.

Jeśli czytamy w opisie, że badanie zostało przeprowadzone na ogólnopolskiej próbie 1080 osób w wieku 18+, dobranej według reprezentacji w populacji Polaków pod względem płci, wieku i wielkości miejscowości zamieszkania, z panelu badawczego (w którym często zarejestrowanych jest kilkaset tysięcy respondentów – więc jest z czego losować), to jest to próba reprezentatywna TYLKO **dla respondentów zarejestrowanych w tym panelu**.

P#1.1. Dramatycznie malejący stopień realizacji wylosowanych prób

Niestety, losowanie – choć kosztowne i skomplikowane – nie jest najważniejszym problemem. Respondenta można wylosować, ale nie można zmusić go do udziału w badaniu. Rynek badań w XXI wieku został przesycony – ankiety wszelkiego typu są przeprowadzane masowo. Komercjalizacja badań ankietowych popsuła ich społeczną reputację. Badania społeczne zamiast opisywać społeczeństwo stały się narzędziem sprzedaży i wpływu społecznego. Sondaże zaczęły rządzić politykami. Nic dziwnego, że systematycznie maleje stopień realizacji próby – zapraszani respondenci nie chcą odpowiadać na pytania. Czasy, kiedy badania przedwyborcze

[175] Shaughnessy i in., 2007.

[176] https://radon.nauka.gov.pl/analizy/nauka-w-Polsce-2023.

idealnie przewidywały wyniki wyborów, odchodzą do przeszłości właśnie ze względu na problem z uzyskaniem reprezentatywnej próby wyborców.

W latach 90. XX wieku stopień realizacji próby (*response rate*, stopa zwrotu) wynosił ponad 80%. W trzeciej dekadzie XXI wieku jest to już tylko około 25%. Różnice w skłonności do udziału w badaniu zależą od płci, wieku, sytuacji zawodowej... Można przeczytać[177], że najtrudniejsze jest pozyskanie do badań mężczyzn w wieku 30–44 lata. Wiedząc to z wyprzedzeniem możemy wylosować więcej mężczyzn w tym wieku (tzn. „zagęścić próbę"), zakładając, że tylko 1 na 10 wylosowanych weźmie udział w badaniu. Wtedy próba wylosowana nie jest reprezentatywna dla populacji, z której została wylosowana, ale reprezentatywna jest próba **zrealizowana**. Jest to lepsza metoda niż poststratyfikacyjne ważenie danych[178].

Powtarzając – próby reprezentatywne są niezbędne do określania rozkładu zmiennej w populacji, której znajomość jest kluczowa dla przewidywania wyników wyborów w społeczeństwach demokratycznych, ale ze względu na dużą liczbę zmiennych zakłócających, często o bardzo skośnych rozkładach (na przykład wykształcenie), nie są bardzo przydatne, gdy chcemy testować zależności między zmiennymi.

Dużo lepszą niż badanie prób reprezentatywnych strategią maksymalizacji trafności zewnętrznej jest replikowanie badań na próbach homogenicznych – osobno badamy rolników, osobno pracowników naukowych, osobno studentów itd.

P2. Pułapki na etapie układania pytań

Konstruując ankietę powinniśmy pamiętać, że wszystko, co napiszemy, powinno motywować respondenta do współpracy. Instrukcje powinny być precyzyjne, krótkie i zawierać ciekawe przykłady, aby były motywujące. Jeśli tylko jest to możliwe, należy stosować odpowiednie znaki graficzne wspomagające czytelność ankiety: zastosowanie pogrubień, umieszczanie tekstu w ramkach, użycie strzałek – cały czas pamiętając, że naszym celem jest ułatwienie czytania i zapamiętanie instrukcji. Nie możemy jednak nadmiernie przeciążać mózgów.

Pytanie, które respondent uzna za niegrzeczne, niezrozumiałe lub nieprzemyślane, może wpłynąć nie tylko na tę odpowiedź, ale na wszystkie wypowiedzi. Zdarza się tak, że są włączane **pytania informacyjnie nieistotne**, lecz istotne dla wytworzenia dobrej atmosfery wywiadu. W wielu ankietach na początku umieszcza się tak zwane **pytania kontaktowe** (inaczej określane jako przełamujące pierwsze lody lub „na rozgrzewkę"). W ankiecie można także umieścić **pytania buforowe** mające neutralizować/wygaszać wpływ poprzednich pytań i **pytania sprawdzające poziom koncentracji respondenta.**

[177] https://www.wum.edu.pl/Badania-epidemiologiczne-tak-to-sie-robi-na-WUM

[178] https://radon.nauka.gov.pl/analizy/nauka-w-Polsce-2023.

Powinniśmy cały czas pamiętać, że respondenci czynią nam uprzejmość, poświęcając swój czas i wysiłek – należy wczuć się w ich sytuację i dawać im odczuć, że szanujemy ich zaangażowanie. Jeśli jakiś blok pytań jest żmudny, warto to uczciwie zaznaczyć, podkreślając jednocześnie jego wagę. Im dłuższa ankieta, tym więcej respondentów odmawia jej wypełnienia. Ankiety trwające 5 lub mniej minut wypełnia 79% proszonych, gdy czas jest większy niż 15 minut, procent ten spada do 53[179]. Zamiast pytać o wszystko, co wydaje się nam interesujące, powinniśmy wyobrazić sobie co zrobimy z odpowiedziami, które uzyskamy, innymi słowy **„napisać sprawozdanie, zanim uzyskamy dane"**[180]. Niestety bardzo często pytamy o informacje, których nie jesteśmy potem w stanie wykorzystać. Powtarzając, przede wszystkim jednak powinniśmy **przeprowadzić badania pilotażowe** i sprawdzić czy informacje, które uzyskujemy z odpowiedzi respondentów są tymi, o które nami chodzi. **Trzeba pilotować każde pytanie i każdą sekwencję pytań.** Żaden ekspert nie jest w stanie ocenić pytania bez znajomości rozkładu odpowiedzi. Może on wskazać szkolne błędy, takie jak niepotrzebne wymuszanie odpowiedzi poprzez brak opcji beztreściowej, złe umieszczenie na skali tej opcji, ale jego opinia nie zastąpi pilotażu.

Układający pytanie powinien nie tylko unikać niejasności i wieloznaczności, ale także brać pod uwagę kim jest odpowiadający. Powinien on skonstruować pytanie tak, **aby odpowiadający był w stanie zrozumieć jego intencje**. Nie wystarczy, aby odpowiadający rozumiał użyte słowa (jak w pytaniu: „**Co dzisiaj robiłeś?"**). Musi być dla niego także jasne, jakiej odpowiedzi (**na jakim poziomie ogólności**) oczekuje się od niego. Inaczej zadaje się pytania o emocje studentom psychologii, a inaczej uczniom szkoły podstawowej. Często odpowiadamy na **inferowaną intencję** pytania, a nie na jego treść. Osoba śpiesząca się z wypełnieniem ankiety, zadając pytanie: „Która godzina?", może zaakceptować odpowiedź: „Jeszcze mamy czas", choć z formalnego punktu widzenia nie jest to odpowiedź na zadane pytanie.

Jakkolwiek istnieją ścisłe reguły doboru prób w sondażach, dobór słownictwa i projektowanie pytań w dalszym ciągu pozostają w dużej mierze sztuką.

Dbajmy o to, aby **pytania były interesujące**, bo to pozwala utrzymać motywację respondenta do wypełnienia ankiety. Ankieta powinna być uporządkowana i mieć logiczną strukturę. Na początku umieszczamy pytania wprowadzające (łatwe), aby nie zniechęcić ani nie przemęczyć respondenta.

Wprowadzając nowy blok pytań dodajemy krótkie wprowadzenie, pamiętając, by zaznaczyć **po co** zadajemy te pytania. Pytania powinny być ułożone tematycznie i przechodzić od ogólnych do szczegółowych.

Pytania metryczkowe zawsze umieszczamy **na końcu** ankiety. Należy w nich uwzględnić opcję: **„nie chcę odpowiadać na to pytanie"** i koniecznie wyjaśnić respondentom, że dane socjodemograficzne[181] zbierane są tylko dlatego, że są

[179] Churchill, 2002.
[180] Oppenheim, 2004.
[181] Wieczorkowska i Grzelak, 1999.

potrzebne do naukowych analiz. Zamiast o wiek pytamy o **rok urodzenia**, ponieważ odpowiedź na pytanie o wiek może budzić wątpliwości (np. „ile mam lat, jeśli za dwa miesiące kończę 40?") czy emocje. Rok urodzenia nie jest tak nasycony ewaluacją, więc podawany automatycznie. Niestety, przy prowadzeniu badań w zakładach pracy informacje, takie jak płeć, rok urodzenia, wykształcenie czy stanowisko mogą umożliwiać jednoznaczną identyfikację respondenta – czego chcemy uniknąć. Pracownicy często, choć zazwyczaj **niesłusznie**, nie ufają zapewnieniom o anonimowości i obawiają się, że przełożeni zobaczą ich odpowiedzi. W takich sytuacjach lepiej zrezygnować z pytania o rok urodzenia i zastosować **szerokie kategorie wiekowe**. Oczywiście **zawsze** należy umożliwić wybór odpowiedzi beztreściowych, takich jak: „nie chcę odpowiadać". Dla badacza **lepszy jest brak odpowiedzi** niż odpowiedź fałszywa.

Na końcu ankiety **zawsze pytamy o samoocenę rzetelności jej wypełnienia**. We wszystkich prowadzonych w moim zespole badaniach dajemy respondentom możliwość zadeklarowania poziomu swojego zaangażowania, pytając np. **„Czy odpowiadałeś uważnie, czy też np. zmęczenie spowodowało, że Twoje odpowiedzi mogłyby się zmienić, gdybyś odpowiadał na pytania jutro?**". W wielu ankietach internetowych respondenci nie mają możliwości poinformowania badacza, że odpowiadali losowo, co oznacza, że ich odpowiedzi powinny zostać usunięte ze zbioru danych.

P#2.1. Pytania otwarte czy zamknięte?

Jedna z najstarszych debat w badaniach ankietowych dotyczy dylematu czy używać **pytań otwartych, czy zamkniętych**. W pytaniach zamkniętych respondent ma wybrać odpowiedź z przedstawionego mu zbioru opcji (kafeteria), a zadając pytanie otwarte zapisujemy w całości to, co mówi pytany.

Odpowiedzi na pytania **zamknięte** są kodowane za pomocą **liczb** i poddawane różnym typom **analiz statystycznych**. Z kolei odpowiedzi na pytania **otwarte** są kodowane za pomocą **słów** i analizowane **jakościowo** (np. analiza treści, analiza językowa). Choć i w tym przypadku możliwe jest np. **zliczanie użytych słów, określonych fraz** lub innych wskaźników ilościowych.

W telefonicznym sondażu[182] przeprowadzonym w USA w 1977 roku, amerykańskich respondentów podzielono losowo na dwie grupy i zapytano o najważniejsze problemy, z jakimi boryka się kraj.

W **grupie E1**, której przedstawiono **zamkniętą kafeterię odpowiedzi**, pytanie brzmiało: „Rozwiązanie którego z następujących problemów jest aktualnie dla naszego kraju najważniejsze?". Z kafeterią zawierającą: „bezrobocie, przestępczość, inflacja, poziom polityków, upadek moralności i religii, inne...". Spośród **592 respondentów** tej grupy, **tylko 2 osoby** wybrały opcję „inne" i wskazały na **kryzys ener-**

[182] Schuman i Presser, 1996.

getyczny... **W grupie E2**, której zadano **otwarte pytanie**: „Jaki jest najważniejszy problem, z którym aktualnie boryka się nasz kraj?" aż **22%** badanych spontanicznie wskazało na **kryzys energetyczny** – co miało silne uzasadnienie w kontekście społeczno-klimatycznym (była to wyjątkowo ciężka zima).

Ten przykład jasno pokazuje, że **forma pytania** (zamknięte vs. otwarte) **może radykalnie wpłynąć na uzyskane wyniki** – zwłaszcza w sytuacjach, gdy respondenci nie znajdą w podanej kafeterii odpowiedzi odzwierciedlającej ich rzeczywiste przekonania.

Decyzja czy stosować **pytania otwarte**, czy **zamknięte,** zależy od tego, co chcemy osiągnąć.

Pytania zamknięte są łatwiejsze do późniejszych analiz i są przydatne, gdy chcemy określić relatywną ważność dużego zbioru opcji. Należy pamiętać, że **podawanie kategorii w pytaniach zamkniętych zawsze prowadzi do zwiększenia częstości występowania tej kategorii** w porównaniu z frekwencją wystąpienia tej kategorii przy pytaniu otwartym.

Pytania otwarte są bardzo pożądane, gdy chcemy oszacować dostępność (*salience*) **danej odpowiedzi.** Jest wtedy ważne, co zostało wymienione i w jakiej kolejności przyszło to do głowy respondentowi, ale nie będą mówić nam o rzeczach, które są dla nich oczywiste lub według ich mniemania nie mają znaczenia.

Bardzo trudno jest porównywać rozkłady uzyskane za pomocą pytania zamkniętego i otwartego. W pytaniach otwartych możemy otrzymywać odpowiedzi, które są wewnętrznie sprzeczne, trudne do zrozumienia i trudne do skategoryzowania. Respondenci mogą odpowiadać, wykorzystując różne punkty widzenia, z różnym stopniem pewności itd. W pytaniach zamkniętych ważne jest, aby znać wszystkie możliwe opcje. Zastanówmy się, co się stanie, gdy w pytaniu zamkniętym pominiemy jakąś kategorię. Przy kategoriach zamkniętych, nawet gdy zostanie miejsce na wpisanie: „inne", szansa, że ominięta kategoria pojawi się, jest bardzo niewielka – ponieważ respondenci mają tendencję do wybierania odpowiedzi spośród dostarczonych przez badacza.

Kategoryzacja odpowiedzi otwartych wymaga przeczytania co najmniej stu wypowiedzi, zanim możliwe będzie sensowne ustalenie kategorii przez sędziów kompetentnych. Co gorsza – w ten sposób do wyników wprowadzamy dodatkowe źródło zróżnicowania, jakim są **sieci neuronowe osoby dokonującej kategoryzacji** (czyli jej indywidualna historia, doświadczenia, preferencje i schematy poznawcze).

Nawet zatrudnienie do tego zadania sztucznej inteligencji generatywnej (gen AI) **nie gwarantuje replikowalnych rezultatów**. Nawet bardzo rozbudowane prompty nie zapewniają spójnej i powtarzalnej kategoryzacji. Trudno się temu dziwić – w końcu sama nazwa *generatywna* AI wskazuje na jej **podstawową siłę i słabość zarazem**: tworzy nowe odpowiedzi, a niekoniecznie powtarzalne klasyfikacje.

Pamiętajmy, że jeśli przedmiotem naszej analizy są **słowa wypowiedziane przez respondentów**, to mamy do czynienia z **badaniami jakościowymi**. Obowiąz-

kiem badacza jakościowego jest **szczegółowe ujawnienie** nie tylko procedur i toku analizy, ale **także własnych poglądów, postaw i wartości, które mogą wpływać na interpretację danych.**

W tekstach prezentujących wyniki badań jakościowych **należy zamieszczać materiał dowodowy** – czyli **autentyczne cytaty z wypowiedzi respondentów**. Każdy cytat powinien być opatrzony informacjami o jego autorze – nie tylko lakonicznym „menedżer", ale także danymi, takimi jak: **wiek, doświadczenie zawodowe, branża** czy **typ firmy**, ponieważ to one pozwalają lepiej zrozumieć sens analizowanej wypowiedzi[183].

Szczególnym przykładem dylematu dotyczącego pytań otwartych czy zamkniętych są **ankiety ewaluacyjne**, które mają dostarczać informacji zwrotnej dla prowadzących. W tego typu badaniach **zdecydowanie lepiej sprawdzają się pytania otwarte**. Średnia ocen na skalach liczbowych mówi nam niewiele – natomiast, analizując odpowiedzi na pytania otwarte możemy dowiedzieć się znacznie więcej, np. że zbyt często lub zbyt rzadko przerywaliśmy studentom, że pojawiało się zbyt wiele dygresji, że tempo zajęć było zbyt szybkie lub zbyt wolne itd.

Warto, formułując pytania otwarte, **aktywizować respondentów za pomocą przykładów**, ponieważ studenci mogą nie dostrzegać pewnych kwestii lub uznawać je za zbyt oczywiste, by o nich wspominać. Przykładowe pytanie mogłoby brzmieć:

„Pamiętając o różnicach indywidualnych w potrzebach i oczekiwaniach poznawczych studentów – co chcieliby Państwo przekazać wykładowcy?" (np. Czy chcieliby Państwo zwiększenia, zmniejszenia czy pozostawienia obecnego stopnia trudności zaliczenia zajęć? Proszę uzasadnić. Czego na zajęciach mogłoby być więcej, a czego mniej?).

Zgodnie z zasadami konwersacji nie możemy oczekiwać, że studenci wpiszą takie uwagi w rubryce „inne". Dlaczego? Ponieważ już wcześniej wyrazili swoje oceny w formie skal numerycznych. Dodanie komentarza byłoby **redundantne** – czyli dostarczaniem powtórzonej informacji, co w języku potocznym łamie zasady ekonomii wypowiedzi i naturalnej komunikacji.

Niestety, powszechnie w ankietach ewaluacyjnych używa się skal liczbowych i wylicza średnie – które są trudne do interpretacji. Bo co to właściwie znaczy, że jedne zajęcia zostały ocenione na 4,13, a drugie na 4,01? Zamiast motywować do wysiłku, takie liczby często demotywują oceniane osoby. Aby ankiety ewaluacyjne dobrze spełniały swoją funkcję, powinny być transparentne – tak, by wykładowca mógł omówić, czego się z nich dowiedział i wprowadzić ewentualne zmiany. Studenci powinni również widzieć, jak inni oceniają te same zajęcia. Ankiety wypełniane na zakończenie zajęć służą dziś głównie biurokratycznym, neoliberalnym wymaganiom – nic więc dziwnego, że liczba studentów je wypełniających jest bardzo niska, co z kolei nie pozwala na formułowanie rzetelnych wniosków.

---

[183] Kociatkiewicz i Kostera, 2014.

P#2.2. Za głośno?

W 145 ankietach, w których prosiłam studentów o ocenę mojego wykładu z psychologii społecznej po pierwszej połowie semestru dwie osoby napisały, że ustawienia mikrofonu spowodowały, że było zbyt głośno. Inni nic nie napisali na temat głośności. Zmiana ustawień mikrofonu nie jest dla mnie żadnym problemem, więc byłam gotowa uwzględnić uwagi studentów. Postanowiłam jednak sprawdzić powszechność tej opinii. Gdy korzystając z platformy internetowej zadałam pytanie zamknięte tej samej grupie studentów z 3 opcjami do wyboru: (1) za głośno, (2) w sam raz, (3) za cicho, okazało się, że tylko 5 studentów uznało, że było za głośno, 7 uznało, że za cicho, reszta wybrała odpowiedź „w sam raz".

To przestroga, aby pojawiających się w pytaniach otwartych uwag studentów nie traktować jako reprezentatywnych dla całej grupy – choć zgodnie z efektem fałszywej powszechności mówiącym, że przeceniamy liczbę osób, które podzielają nasze opinie – zgłaszającym uwagi osobom wydaje się, że wszyscy tak sądzą. Ten przykład sugeruje, że pytania otwarte należy zadawać w pierwszym kroku budowania ankiety.

P3. Pułapki na etapie wyboru skali odpowiedzi/kafeterii

Najbardziej popularną skalą odpowiedzi jest zaproponowana przez Likerta kafeteria z 5 opcjami:

całkowicie się nie zgadzam, nie zgadzam się, nie mam zdania, zgadzam się, całkowicie się zgadzam.

Poszczególnym wyborom przypisuje się kolejne wartości liczbowe np. 1, 2, 3, 4, 5 lub –2, –1, 0, +1, +2. Rygoryści metodologiczni uważają, że ta skala dostarcza tak naprawdę danych porządkowych, do których można stosować tylko niektóre rodzaje procedur statystycznych (ale nie można liczyć średniej arytmetycznej czy wariancji). Prawdą jest także, że Ci sami badacze nie mają oporów, aby oceny szkolne traktować jako dane interwałowe (co wymaga założenia, że różnica np. między 5 i 4 jest taka sama jak między 2 i 3). Standardem uznawanym przez najlepsze czasopisma naukowe jest traktowanie skal typu Likerta jako ilościowych[184]. **Skala Likerta stanowi zrównoważoną skalę odpowiedzi**, ponieważ ma **opcję środkową** i jednakową liczbę odpowiedzi pozytywnych i negatywnych.

[184] Porównaj stanowisko zaprezentowane w sekcji „Metodologiczna hipokryzja" w: Wieczorkowska i Wierzbiński, 2013.

Nazwa **kafeteria** wskazuje na analogię wyboru potrawy z menu. Opcje w kafeterii mogą być:

- **nieuporządkowane**, np. w pytaniu o wykonywany zawód lub
- **uporządkowane** jak w pytaniu o wykształcenie, np.: podstawowe, średnie, wyższe.

Choć wydaje nam się, że kafeteria odpowiedzi zawiera wszystkie możliwe opcje, bezpieczniej jest pozostawić respondentowi możliwość dopisania własnej odpowiedzi, umieszczając na liście punkt „inna", „inne" itp. Pytanie zamknięte z ukrytym na liście podpytaniem otwartym nazywa się **pytaniem półotwartym** (lub półzamkniętym).

P#3.1. Zbyt długa lista opcji w kafeterii

Przykładem błędu metodologicznego może być zbyt długa lista opcji w kafeterii**.** W ankiecie skierowanej do doktorantów pojawiło się następujące pytanie: w**ybierz maksymalnie 3 aspekty studiów doktoranckich, z których jesteś najbardziej zadowolony:** (1) możliwość rozwoju, (2) elastyczność godzin pracy, (3) wyjazdy zagraniczne, (4) możliwość budowania dorobku naukowego, (5) wymiar czasu pracy, (6) współpraca z promotorami, (7) wykonywane badania naukowe, (8) współpraca z grupą badawczą, (9) przygotowanie do samodzielnej pracy naukowej, (10) pogodzenie studiów doktoranckich z życiem prywatnym, (11) status doktoranta UW, (12) względy finansowe, (13) inny – jaki?

Użycie zbyt wielu opcji w kafeterii odpowiedzi jest metodologicznym błędem. Wybór z długiej listy utrudnia późniejsze analizy, ponieważ odpowiedzi nie są równoważne: waga jedynego aspektu wskazanego przez respondenta powinna być większa niż wtedy, gdy ktoś wybrał trzy. W zbiorczych zestawieniach głos osoby, która wskazała trzy opcje, wpływa na wyniki silniej niż głos osoby, która wskazała tylko jedną.

Dlatego **nie rekomendujemy** tej – dość często spotykanej – formy pytania. Lepszym rozwiązaniem jest prośba o ocenę wszystkich (lub wybranych przez badacza) aspektów, np. na skali Likerta (ale z opcją beztreściową poza skalą, a nie w środku, jak wyjaśniamy to w dalszej części sekcji). Warto zauważyć, że wymienione aspekty **nie są niezależne** – ustawienie dla każdego z nich skali ilościowej pozwoliłoby odkryć stopień ich wzajemnego powiązania (korelacji) i umożliwiłoby budowę wskaźników złożonych.

W kolejnym przykładzie lista jest jeszcze dłuższa. Plusem jest jednak to, że autorzy pytania informują respondentów, iż powinni uważnie zapoznać się z opcjami przed wyborem.

A teraz prosimy wybrać z poniższej listy maksymalnie 5 konstruktów, które Państwa zdaniem przewidują uprzedzenia bądź szerzej rozumiany stosunek wobec uchodźców, niezależnie od tego, czy ta relacja jest pozytywna czy negatywna. *Poniższa lista jest długa – proszę ją przejrzeć, a następnie wybrać maksymalnie 5 najlepszych predyktorów* (potem następuje lista kilkudziesięciu terminów m.in.: pra-

wicowy autorytaryzm, grupowy prawicowy autorytaryzm, orientacja na dominację społeczną, przekonania polityczne itd.).

Również w tym przypadku stosowniejsze byłoby użycie skali oceny dla każdego z konstruktów osobno – dzięki temu możliwe byłoby uzyskanie pełniejszego obrazu postaw respondentów oraz prowadzenie bardziej precyzyjnych analiz statystycznych.

P#3.2. Samodzielne czytanie vs. słuchanie zadawanych pytań

W badaniach pokazano, że nie bez znaczenia jest fakt, czy kafeterię odpowiedzi czytamy my czy też czyta ją dla nas ankieter. Jeżeli **odpowiedzi są prezentowane na karcie lub w kwestionariuszu**, myślimy więcej o pierwszych niż ostatnich (zanim dojdziemy do ostatnich nasze zasoby poznawcze są wyczerpane). W tym przypadku możemy oczekiwać uprzywilejowania opcji pierwszych na liście.

Gdy **odpowiedzi są czytane przez ankietera** – nie mamy możliwości zastanawiania się dopóki ankieter nie zaprezentuje ostatniej. Dlatego przy słuchaniu odpowiedzi wymieniane jako ostatnie mają większą szansę bycia wybranymi. Jeśli lista jest czytana przez ankietera, mamy więcej czasu na **myślenie o ostatnich** niż pierwszych opcjach. Jeżeli lista możliwych odpowiedzi jest długa – możemy mieć kłopot z zapamiętaniem pierwszych. Przekonałam się o tym przed wieloma laty, gdy pytałam mojego wtedy 2-letniego synka, czy woli wypić mleko czy kakao. Zauważyłam wtedy, że wybiera zawsze drugą opcję. Dla małych dzieci utrzymanie w pamięci roboczej wielu opcji jest nieosiągalne. Podobny efekt wystąpił u mojej 93-letniej Mamy. Pod koniec życia preferowała zawsze drugą z wymienianych opcji wyboru.

W badaniach[185] potwierdzono, że efekt kolejności opcji w kafeterii zależy od tego, **czy pytanie zadaje ankieter (np. telefonicznie), czy też respondent czyta je samodzielnie** (np. w ankiecie internetowej).

Ochotnicy zostali losowo podzieleni na dwie grupy i zapytani o to, kto ich zdaniem powinien być **odpowiedzialny za poziom opieki społecznej**. Grupy różniły się kolejnością opcji odpowiedzi przedstawionych na kartach wręczanych przez ankietera.

Zaobserwowano wyraźny **efekt pierwszeństwa**:

- odpowiedź **„państwo”** uzyskała więcej wskazań, gdy była pierwsza na liście (64% vs. 50%),
- odpowiedź **„organizacje charytatywne”** uzyskała więcej wskazań, gdy to ona była pierwsza na liście (50% vs. 36%).

W kolejnym eksperymencie, również z udziałem losowo podzielonych grup ochotników, ankieter pytał o **preferowaną formę rządu**. Grupy różniły się **kolejnością odpowiedzi, które wypowiadał ankieter.**

[185] Schwarz i in., 1991c.

Tym razem zaobserwowano **efekt świeżości** – czyli silniejszy wpływ ostatnio usłyszanej opcji:

- odpowiedź **„autorytarna”** uzyskała więcej wskazań, gdy była wymieniona jako druga, a nie pierwsza (26% vs. 15%),
- odpowiedź **„demokratyczna”** również uzyskała więcej wskazań, gdy była wymieniona jako druga (85% vs. 74%).

Zjawisko to pokazuje, że **forma** i **kolejność** prezentacji informacji – zwłaszcza w zależności od kanału komunikacji **(wzrokowego lub słuchowego)** – istotnie wpływa na udzielane odpowiedzi. W ankietach internetowych przeważa efekt pierwszeństwa, natomiast w ankietach ustnych (np. telefonicznych) częściej obserwuje się efekt świeżości.

P4. Pułapki odpowiedzi środkowych i beztreściowych (TP)

Podawane w mediach wyniki sondaży najczęściej nie zawierają informacji o stopniu realizacji próby – czyli o tym, ile osób, które zapytano, odmówiło udziału w ankiecie, a także ile osób nie udzieliło odpowiedzi na konkretne pytanie. Zarówno braki odpowiedzi, jak i odpowiedzi beztreściowe[186] („**trudno powiedzieć**”/„NIE WIEM”/„nie chcę odpowiadać na to pytanie”), oznaczane dalej skrótem **TP**, są trudne do interpretacji, ale **zawsze** powinny być raportowane w opisie rozkładów.

W dominujących w XXI wieku ankietach internetowych wyróżniamy trzy typy pytań:

- **bez filtra** – respondent może pominąć pytanie, jeśli nie jest ono oznaczone jako obowiązkowe;
- **z quasi-filtrem** – odpowiedź „trudno powiedzieć”/„NIE WIEM”/„nie chcę odpowiadać na to pytanie” znajduje się na skali odpowiedzi;
- **z pełnym filtrem** – najpierw pytamy o wiedzę lub poglądy, a pytanie właściwe zadaje się tylko tym respondentom, którzy odpowiedzieli twierdząco na pytanie filtrujące np. „Czy masz zdanie w tej sprawie?”, „Czy zastanawiałeś się nad tym?” itp.

Przykład zastosowanie pełnego filtra wiedzy pochodzi z badań przeprowadzonych na panelu Ariadna, gdzie zanim zadano respondentom pytanie o ocenę pomysłu wprowadzenia „Europy dwóch prędkości” w ramach Unii Europejskiej, zapytano ich, czy znają choćby ze słyszenia ten termin. Aż 53,7% respondentów Ariadny odpowiedziało **NIE**.

[186] Sułek, 2002.

P#4.1. Konflikt palestyńsko-izraelski

Losowo podzieleni na dwie grupy[187] ochotnicy mieli odpowiedzieć **„TAK" lub „NIE"** na pytanie, czy zgadzają się z opinią, że **Arabowie dążą do zawarcia pokoju z Izraelem**. Badanie przeprowadzono pod koniec XX wieku.

W grupie **E1** odpowiedzi rozłożyły się następująco: TAK – 17%; NIE – 60%; **NIE WIEM – 23%.**

W grupie **E2** pytanie właściwe zostało poprzedzone pytaniem filtrującym: **„Czy masz opinię na ten temat?"**. W odpowiedzi brak opinii zadeklarowało **aż 45%** respondentów – prawie dwukrotnie więcej niż w grupie E1. Spośród pozostałych: TAK odpowiedziało 10%; NIE – 45%.

Dostępność odpowiedzi beztreściowej („nie mam zdania"**) znacząco wpłynęła na rozkład odpowiedzi**. Wprowadzenie pytania filtrującego pozwoliło lepiej oddzielić osoby, które faktycznie mają wyrobione zdanie, od tych, które jedynie losowo lub pod presją sytuacji wybierają jedną z dostępnych opcji.

Interpretacja odpowiedzi beztreściowych nastręcza wielu problemów. Odpowiedzi te mogą wynikać zarówno z braku wiedzy i/lub zdecydowania, znudzenia wywiadem, jak i z chęci ucieczki respondenta przed udzieleniem odpowiedzi na niewygodne dla niego pytanie. Mogą być także wskaźnikiem, że skala odpowiedzi nie uwzględnia właściwej opcji.

Badania wykazały, że liczba TP koreluje negatywnie z miarą samooceny wysiłku wkładanego w odpowiadanie przez respondenta – im więcej TP, tym mniejszy deklarowany wysiłek.

Niebezpieczna jest również sytuacja odwrotna – gdy respondent udziela **treściowej odpowiedzi losowo** lub zgodnie z domniemywanym oczekiwaniem badacza albo z uznawaną normą społeczną. Dzieje się tak np. gdy:

- autorzy ankiety internetowej **nie umieścili TP** w kafeterii odpowiedzi,
- respondent **wstydzi się przyznać** do niewiedzy lub braku opinii.

Podsumowując: należy umożliwiać respondentom wybór odpowiedzi typu TP. Niestety, autorzy ankiet internetowych masowo **wymuszają udzielenie odpowiedzi treściowej** w sytuacjach, gdy TP nie jest dostępna. W odróżnieniu od ankiety papierowej, w formularzu internetowym nie można przejść dalej, jeśli nie udzieli się odpowiedzi. Takie wymuszanie stanowi ewidentny błąd metodologiczny.

[187] Plous, 1993.

P#4.2. Wpływ wieku i kompetencji poznawczych na liczbę TP

Analiza[188] liczby odpowiedzi beztreściowych (TP) w zbiorach Polskiego Generalnego Sondażu Społecznego, w których większość stanowiły pytania bez filtra (odpowiedź „NIE WIEM" nie była odczytywana przez ankietera, ale była zapisywana, jeśli respondent wypowiedział ją sam), wykazała, że:

- wiek jest dobrym predyktorem TP po 60. roku życia;
- im wyższe kompetencje poznawcze respondenta – operacjonalizowane wspólnie przez poziom wykształcenia oraz stopień złożoności wykonywanej aktualnie lub w przeszłości pracy – tym bardziej opóźniony (średnio o 10 lat) jest wpływ wieku na wzrost liczby odpowiedzi TP;
- kobiety częściej niż mężczyźni wybierają odpowiedzi beztreściowe wyłącznie w pytaniach związanych z polityką.

P#4.3. Opcja środkowa i beztreściowa

Wybierając skalę odpowiedzi, stajemy przed dylematem: czy powinna ona zawierać **punkt środkowy** (nazywany dalej *opcją środkową*) np. „ani zgadzam się, ani nie zgadzam się" oraz czy na skali powinna pojawić się **opcja odpowiedzi beztreściowej** np. „trudno powiedzieć" (oznaczana dalej jako *TP*).

Warto zauważyć, że zarówno **odpowiedzi środkowe**, jak i **beztreściowe** są **trudne do interpretacji**. Z tego względu niektórzy badacze sugerują, aby w ogóle ich nie uwzględniać – zmuszając respondentów do wyboru jednej z treściowych opcji, co zwiększa jednoznaczność danych.

Ochotników w naszych badaniach zaufania do instytucji[189] podzielono losowo na dwie grupy różniące się dostępnością opcji środkowej („ani mam, ani nie mam zaufania") i opcji beztreściowej („trudno powiedzieć"). W grupie E1 – skala zawierała tylko opcję **beztreściową**, w **E2 nie było opcji beztreściowej, ale była dostępna opcja środkowa.**

Odsetek respondentów wybierających opcję środkową (29,5%, 31,8%, 34,9%) był **wielokrotnie wyższy w E2** niż odsetek osób wybierających odpowiedź „trudno powiedzieć" w E1 (10%, 5%, 5%).

Może to sugerować, że wybór opcji środkowej często pełni funkcję **ucieczki przed zajęciem stanowiska**, a nie rzeczywistego wyrażenia neutralności czy równowagi poglądów.

[188] Jerzyński, 2008.

[189] Wieczorkowska i in., 2009.

P#4.4. Pozycja TP na skali odpowiedzi

Ochotników w naszych badaniach[190] podzielono losowo na dwie grupy różniące się **pozycją odpowiedzi „trudno powiedzieć" (TP)** na skali odpowiedzi: w grupie E1 **w środku skali**, tak jak w oryginalnej skali Likerta, w grupie E2 **poza skalą** (na końcu listy odpowiedzi).

Odsetek odpowiedzi „trudno powiedzieć" w grupie E1 (TP **w środku**), wahał się – w przypadku 5 analizowanych pytań – od **10,5% do 14%**. W grupie E2 (TP **poza skalą**) odsetek ten zmieniał się jedynie od **3,2% do 5,9%**. W obu grupach respondenci oceniali łatwość udzielania odpowiedzi na pytania (na skali od 1 – „bardzo łatwo" do 4 – „bardzo trudno"). Liczba odpowiedzi TP korelowała **z subiektywnie odczuwaną trudnością** odpowiedzi tylko w E2, w które**j TP** znajdowało się na **końcu skali.** Oznacza to, że:

- gdy TP było w środku – traktowano je jako neutralną, środkową odpowiedź,
- gdy TP było na końcu – wybierały je osoby, które **rzeczywiście nie wiedziały, co odpowiedzieć**.

P#4.5. Wpływ punktu środkowego skali

Aby porównać konsekwencje umieszczenia na skali punktu środkowego dla liczby odpowiedzi beztreściowych, ochotników w naszych badaniach[191] podzielono losowo na dwie grupy różniące się dostępnością opcji środkowej (grupa E1) lub nie (grupa E2) na skali odpowiedzi. W obu grupach opcja TP była umieszczona na końcu skali.

Respondenci odpowiadali na ile zgadzają się z 5 stwierdzeniami składającymi się na wskaźnik patriotyzmu:

1) wolę być obywatelem Polski niż jakiegokolwiek innego kraju na świecie;
2) są pewne sprawy w dzisiejszej Polsce, których jako Polacy możemy się wstydzić;
3) ogólnie rzecz biorąc Polska jest lepszym krajem niż większość innych krajów;
4) trzeba popierać swój kraj, nawet gdy postępuje niewłaściwie;
5) chciałbym być dumny z Polski częściej niż jestem.

Dla wszystkich 5 analizowanych pytań procent respondentów wybierających **opcję środkową** „ani się zgadzam, ani nie zgadzam" **lub beztreściową** „TP – trudno powiedzieć" w grupie E1 dla każdego pytania przewyższał znacznie procent wybierających opcję beztreściową w grupie E2 pozbawionej opcji środkowej. Średni procent odpowiedzi uchylających się od zajęcia stanowiska z 5 pytań w grupie w E1 wyniósł **21,6%** w porównaniu z **12,3**% w grupie E2.

---

[190] Wieczorkowska i in., 2009.

[191] Ibidem.

Pozwala to przypuszczać, że respondenci często wybierają **opcję środkową z wygody lub lenistwa**, traktując ją jako wyjście neutralne – nawet wtedy, gdy mają wyrobioną opinię.

Z tego powodu można argumentować, że **należy unikać** umieszczani**a opcji środkowej** na skali, pozostawiając wyłącznie możliwość wyboru odpowiedzi bezteściowej (*TP*) dla tych, którzy rzeczywiście nie są w stanie zająć stanowiska.

Podsumowując: **dostępność** na skali odpowiedzi zarówno **opcji środkowej, jak i beztreściowej**, powoduje **wzrost liczby odpowiedzi nieinterpretowalnych**. Dlatego należy w badaniach unikać odpowiedzi środkowej, ale udostępniać odpowiedź „trudno powiedzieć", którą można następnie przekodować w środek skali odpowiedzi.

P5. Pułapki kolejności pytań

Odpowiadanie na pytania ankiety jest przykładem **dialogu między pytającym i odpowiadającym**. Dla badacza (pytającego) jest ona **zbyt często zbiorem pojedynczych pytań,** respondent zaś prawie zawsze traktuje ją jako **niepodzielną całość**. Bardzo ważne jest to, jak respondenci wnioskują o intencjach pytającego. Zrozumienie intencji pytania zazwyczaj wymaga od odpowiadającego intensywnego wnioskowania na temat „**o co**" i „**po co**" jest pytany. Z popularnych[192] pięciu zasad kooperatywnej konwersacji powinniśmy zapamiętać, że **interpretując pytanie** respondent **odwołuje się do poprzednich pytań**[193].

Jeżeli odpowiadamy na dwa podobne pytania, to drugie będziemy interpretować w ten sposób, aby odpowiedź na nie niosła informację inną niż pierwsza. Niestety, badacze często wielokrotnie pytają o to samo w różny sposób – i wydają się być zaskoczeni, że odpowiedzi badanych nie są spójne. Respondenci starają się traktować podobne (w zamyśle badacza) pytania jako różne – w słusznym przekonaniu, że nikt w pełni władz umysłowych nie zadaje tego samego pytania kilka razy.

Warto uprzedzić respondentów, że niektóre pytania mogą wydawać **się powtórzeniem** tych już zadanych. Można wtedy wyjaśnić, że celowo formułujemy pytania w różny sposób, ponieważ szukamy najlepszego sposobu ich zadania oraz chcemy zwiększyć rzetelność pomiaru.

Odpowiedź na pojedyncze pytanie może być bowiem przypadkowa – np. wynikać z impulsywności, zmęczenia lub chwilowej nieuwagi. Jeśli jednak respondent udziela spójnych odpowiedzi na kilka podobnych pytań, wzrasta prawdopodobieństwo, że pomiar wiernie oddaje jego rzeczywiste przekonania czy postawy.

[192] Grice, 1980.

[193] Skład, Wieczorkowska, 2001.

P#5.1. Wpływ wieku respondentów na efekt kolejności pytań

Ochotników losowo podzielonych na dwie grupy pytano o pogląd na temat dopuszczalności legalnej aborcji z **powodów społecznych** (mężatka nie chce mieć więcej dzieci), wymuszając zajęcie stanowiska poprzez udzielenie odpowiedzi **TAK** lub **NIE**. W grupie **E1** pytanie to zadano jako **pierwsze**. W grupie E2 było ono poprzedzone pytaniem o dopuszczalność legalnej aborcji **z powodów medycznych** (bardzo wysokie prawdopodobieństwo urodzenia chorego dziecka). Uzasadnienie społeczne w zestawieniu z medycznym może wydawać się trywialne. I tak zareagowali respondenci. Liczba odpowiedzi **„tak"** w E2 przy uzasadnieniu społecznym była **istotnie mniejsza**, gdy pojawiało się ono jako DRUGIE **po pytaniu z uzasadnieniem medycznym** – niż wtedy, gdy było zadane jako PIERWSZE. Oznacza to, że kolejność zadawania pytań miała istotny wpływ na otrzymywane wyniki. Gdy porównano[194] poziom zgody na dopuszczalność aborcji w trzech grupach wiekowych, okazało się, że **najsilniej efekt kolejności pytań** wystąpił w grupie **młodych respondentów** (średnia wieku 29,45). W tej grupie wpływ pytania poprzedzającego spowodował **spadek poparcia** dla aborcji z powodów społecznych **o ponad 25%** (z 92,9%, gdy pytanie zadano jako pierwsze, do 67,5%, gdy pojawiło się jako drugie).

Efekt kolejności był znacznie **mniejszy** w grupie **starszych respondentów** (średnia wieku 75,34), którą podzielono dodatkowo na podgrupy według pojemności pamięci operacyjnej mierzonej testem *reading span*. W grupie **starszych osób ze sprawną pamięcią operacyjną** wpływ pytania poprzedzającego spowodował spadek poparcia o **ponad 18%** (z 69,8 do 51,2%). W grupie starszych osób z **niższą sprawnością poznawczą** spadek wyniósł **niecałe 6%** (z 64,1 do 58,5%).

To bardzo czytelna ilustracja faktu, że **wpływ kolejności pytań wymaga sprawnej pamięci operacyjnej** – musimy po prostu pamiętać poprzednie pytanie, by to, które następuje, zyskało nowe znaczenie w jego kontekście. Powinniśmy o tym pamiętać, gdy słyszymy, że w debatach dotyczących ustaw cytuje się dane z badań sondażowych. **Procent poparcia** dla legalizacji aborcji z przyczyn społecznych **zależy od tego, jakie pytania zadano wcześniej.** Tymczasem media podają najczęściej rozkłady odpowiedzi na jedno izolowane pytanie – mimo że kosztowne sondaże na próbach reprezentatywnych zawierają zazwyczaj dużo więcej pytań.

P#5.2. Rozdzielanie dziedzin szczęścia

Pytanie o to, czy jesteśmy szczęśliwi, jest pytaniem wieloznacznym. Czy nasza satysfakcja małżeńska wpływa na poczucie szczęścia?

Możemy uzyskać różne odpowiedzi w zależności od tego, jak sformułujemy pytania.

---

[194] Knäuper i in., 2007, 2004, 2016.

Od ochotników losowo podzielonych na trzy grupy[195], pytanych o to samo – o szczęście małżeńskie i ogólne – otrzymano różne odpowiedzi w zależności od sposobu zadania pytania:

- E1. Chciałabym cię spytać: Czy jesteś szczęśliwy w małżeństwie? Czy jesteś szczęśliwy w życiu w ogóle?
- E2. Chciałabym cię spytać: Czy jesteś szczęśliwy w małżeństwie? Czy jesteś szczęśliwy **w innych dziedzinach życia**?
- E3. Chciałabym cię spytać o **dwie dziedziny życia**: Czy jesteś szczęśliwy w małżeństwie? Czy jesteś szczęśliwy w życiu w ogóle?

Sposób zadania pytania w E2 i E3 „oddziela" szczęście małżeńskie od innych dziedzin życia – i uzyskany związek między satysfakcją małżeńską a ogólnym szczęściem okazał się statystycznie nieistotny (r = 0,18 i r = 0,20 odpowiednio).

Natomiast w grupie E1 uaktywnienie tematu szczęścia małżeńskiego **przed** oceną ogólnego szczęścia spowodowało dodatnią korelację (r = 0,67, im bardziej respondenci byli zadowoleni z małżeństwa, tym bardziej byli zadowoleni z życia w ogóle). Odwrócenie kolejności pytań: najpierw o szczęście w ogóle, potem o szczęście w małżeństwie, spowodowało zanik korelacji.

P#5.3. Ocena asertywności

Ludzie różnią się stopniem, w jakim chcą i potrafią stawiać granice w kontaktach z innymi. Niektórzy z nas myślą o sobie, że brakuje im asertywności. Ale czy potrafimy rzetelnie ocenić naszą asertywność na skali, gdy zostaniemy o to poproszeni? Rzetelnie – to znaczy tak, aby ta ocena nie zależała od tego, o czym myśleliśmy wcześniej. Niestety, odpowiedź na to pytanie brzmi: NIE.

Ochotnicy[196] losowo podzieleni na dwie grupy, zostali poproszeni o przypomnienie sobie 6 (grupa E1) lub 12 (grupa E2) sytuacji, w których zachowali się asertywnie, a następnie oceniali swoją asertywność na skali.

Łatwiej jest przypomnieć sobie 6 sytuacji niż 12. Skoro nie możemy sobie przypomnieć aż 12 przykładów zaczynamy wątpić, czy naprawdę jesteśmy asertywni.

Grupa E1 oceniła swoją asertywność o prawie jeden punkt wyżej (6,3 vs. 5,2) niż grupa E2.

Analogiczne wyniki uzyskano, gdy ocenę asertywności poprzedzono prośbą o przypomnienie sobie 6 lub 12 sytuacji, w których wykazaliśmy się **brakiem** asertywności. Osoby, które miały przypomnieć sobie 12 przykładów braku asertywności oceniały swoją asertywność **wyżej** niż pozostali.

---

[195] Strack i in., 1988; Schwarz i in., 1991b; Schwarz i in., 1991c.

[196] Schwarz i in., 1991a.

P6. Pułapki konstrukcji kafeterii odpowiedzi

Warto pamiętać, że dla respondentów konstrukcja kafeterii odpowiedzi czy podanych przykładów jest informacją, która zmienia ich odpowiedzi, co dobitnie pokazują poniższe wyniki badań.

P#6.1. Wpływ kolejność opcji w kafeterii

Badacze bezrefleksyjnie ustalają porządek w kafeterii prezentowanych opcji (malejący lub rosnący). W kolejnym badaniu sprawdzaliśmy, czy wpływa to na rozkład wyników. Grupa 161 studentów podzielona losowo na dwie części była proszona o odpowiedź na pytanie „Czy masz czas na wypoczynek, telewizję, lektury, kontakty z rodziną, przyjaciółmi?

W grupie E1 kafeteria była uporządkowana w następujący sposób: <1>: Zawsze mam dostatecznie dużo czasu; <2> Zazwyczaj mam czas; <3> Dość często brakuje mi czasu; <4> Często nie mam czasu; <5> Z reguły nie mam czasu. W grupie E2 kolejność opcji w kafeterii była odwrócona od <1> Z reguły nie mam czasu – do <5> Zawsze mam dostatecznie dużo czasu.

Ze względu na losowy przydział do grup E1 i E2 rozkłady odpowiedzi powinny być zbliżone. Tak jednak nie jest. Jeśli chcielibyśmy na podstawie badania określić, jaki procent studentów wybiera opcję „zawsze mam dostatecznie dużo czasu", to w grupie E2 znalazłoby się parokrotnie więcej wskazań niż w grupie E1 (30,7% vs. 5,3%).

P#6.2. Wpływ przykładów w treści pytania

Ochotnicy zostali losowo podzieleni na dwie grupy, którym zadawano pytanie o to, ilu różnych środków przeciwbólowych próbowali. Grupy różniły się zakończeniem pytania: 1? 5? 10? w grupie E1 vs. 1? 2? 3? w grupie E2. Średnia liczba wskazanych środków w grupie E1 była dużo większa niż w grupie E2 (5,2 vs. 3,3). Wniosek? Dokonana przez respondentów **estymacja częstości** używania środków przeciwbólowych zależy od sposobu zadania pytania.

W kolejnym badaniu, również podzielonym losowo na dwie grupy, ochotnikom zadawano pytanie o czas trwania zebrania, w którym uczestniczyli. Średnia w grupie E1 zapytanej **„Jak długo trwało**?", była o 30 minut dłuższa (130 vs. 100 minut) niż średnia w grupie E2 zapytanej **„Jak szybko się skończyło**?". Wniosek? Dokonana przez respondentów **estymacja czasu** trwania zebrania zależy od sposobu zadania pytania.

W następnym badaniu, również z losowym podziałem na dwie grupy, ochotnikom pokazywano zdjęcie koszykarza i proszono o ocenę jego wzrostu. Średnia ocena w grupie E1 zapytanej „Jak wysoki jest ten koszykarz?" była o 26 cm większa (201 vs.

175 cm) niż w grupie E2 zapytanej „Jak niski jest ten koszykarz?". Wniosek? Dokonana przez respondentów **estymacja wzrostu** na podstawie tego samego zdjęcia zależy od sposobu zadania pytania.

P#6.3. Wpływ rozbudowanego zakresu kafeterii

Pytano[197] o częstość oglądania telewizji. Podzielonych losowo na dwie grupy ochotników pytano o czas oglądania telewizji. Grupy różniły się kafeteriami odpowiedzi.

Grupa E1 z kafeterią zawierającą **dużo opcji dla NISKICH wartości** miała do wyboru:

a) mniej niż 0,5 h,
b) od 0,5 do 1 h,
c) od 1 do 1,5 h,
d) od 1,5 do 2 h,
e) od 2 do 2,5 h,
f) więcej niż 2,5 h.

Grupa E2 z kafeterią zawierającą **dużo opcji dla WYSOKICH wartości** miała do wyboru:

a) do 2,5 h,
b) od 2,5 do 3 h,
c) od 3 do 3,5 h,
d) od 3,5 do 4 h,
e) od 4 do 4,5 h,
f) więcej niż 4,5 h.

Respondenci, którzy szacowali czas oglądania TV jako większy niż 2,5 h w grupie E1, mieli do wyboru tylko opcję F. W grupie E2 mieli do wyboru opcje od B do F.

Jeżeli typ kafeterii nie miałby wpływu, to rozkład wyników powinien być zbliżony w dwóch równoważnych grupach E1 i E2. Tak jednak nie jest. Respondentów oglądających TV więcej niż 2,5 h dziennie było **ponad dwa razy więcej** w grupie E2 z kafeterią zawierającą dużo opcji dla DUŻYCH wartości niż w grupie E1 z kafeterią zawierającą dużo opcji dla NISKICH wartości (37,5% vs. 16,2%).

Oznacza to, że podświadomie respondenci zakładają, że **przedstawiona kafeteria odpowiedzi została skonstruowana na podstawie wiedzy badacza o rozkładzie w populacji** i że środkowa wartość w kafeterii odpowiada wartości przeciętnej w populacji, kategorie skrajne odpowiadają zaś ekstremalnym wartościom w populacji. Obie badane grupy pytano także o ocenę (w pytaniu otwartym) średniej w populacji. Grupa E1 oceniła średni czas jako 2,7 h, grupa E2 zaś – 3,2 h.

[197] Schwarz i in., 1985.

Analogiczne wyniki wpływu sposobu konstrukcji kafeterii uzyskano w badaniu eksperymentalnym[198] porównującym wpływ skali odpowiedzi w pytaniu: **„Jaki procent mieszkańców naszego kraju stanowią obecnie Żydzi?”**.

Grupa E1 z kafeterią zawierającą dużo opcji dla WYSOKICH wartości **miała do wyboru:**

a) **mniej niż 2 procent**,
b) 2–9 procent,
c) 10–19 procent,
d) 20–29 procent,
e) 30–49 procent,
f) 50 procent lub więcej,
g) trudno powiedzieć.

Grupa E2 z kafeterią zawierającą dużo opcji dla NISKICH wartości miała do wyboru:

a) **ćwierć procenta lub mniej**,
b) **pół procenta**,
c) **jeden procent**,
d) **półtora procenta**,
e) dwa procent lub więcej,
f) trudno powiedzieć.

Respondenci, którzy szacowali procent jako mniejszy niż 2 w grupie E1, mieli do wyboru tylko opcję A. W grupie E2 mieli do wyboru opcje od A do D.

Jeżeli typ kafeterii nie miałby wpływu, to rozkład wyników powinien być w dwóch równoważnych grupach E1 i E2 zbliżony. Tak jednak nie jest. Respondentów uważających, że Żydów w Polsce jest mniej niż 2% było w grupie E1 z kafeterią zawierającą dużo opcji dla wysokich wartości o wiele MNIEJ niż w grupie E2 z kafeterią zawierającą dużo opcji dla niskich wartości (28% vs. 42% odpowiednio).

P#6.4. Wpływ formy graficznej skali odpowiedzi: drabina vs. piramida

Trzeba pamiętać, że respondenci wydobywają **znaczenie pytania** nie tylko z „nieistotnych” werbalnych elementów pytań, ale także z **graficznego** sposobu prezentacji ich skali czy rozmieszczenia pytań na stronie ankiety.

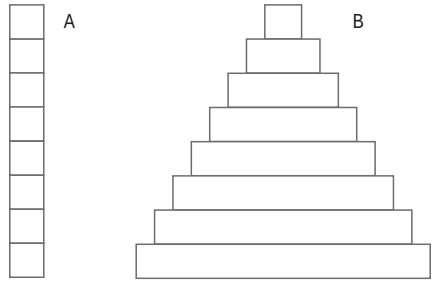
W międzynarodowych badaniach[199], w których proszono o umieszczenie siebie na skali pozycji społecznej, okazało się, że 37% Holendrów lokowało się na jednym z dwóch najniższych szczebli drabiny społecznej w Holandii, podczas gdy w innych krajach takich odpowiedzi było średnio 10%.

---

[198] Sułek, 2001, 2002.

[199] Smith, 1995.

Rysunek 2. Różne skale odpowiedzi użyte w badaniach stratyfikacji społecznej: A (wersja zwykła) vs. B (wersja „holenderska")



Źródło: Wieczorkowska i Wierzbiński, 2013.

Okazało się, że w Holandii skala odpowiedzi została przedstawiona w formie piramidy (rys. 2 B), w pozostałych zaś krajach – w formie drabiny (rys. 2 A). Wpływ formy graficznej skali odpowiedzi wykazały również badania eksperymentalne, w których studenci oceniali niżej swoje dokonania akademickie na skali przedstawionej w formie piramidy niż na skali w formie drabiny.

P7. Pułapki analiz i interpretacji wyników

Trzeba pamiętać, że nawet w sproceduralizowanych, zobiektywizowanych analizach statystycznych badacz musi podjąć szereg decyzji, które mogą doprowadzić do różnych konkluzji[200], dlatego paradygmat metodologiczny WiW zaleca triangulację metod na analizy.

Do tego podstawowym obowiązkiem – i niestety często zaniedbywanym – jest **„czyszczenie" zbioru danych**. W prowadzonym przez mój zespół badaniu stylu jedzenia uzyskaliśmy dostęp do grupy więźniów. Rozdystrybuowana drogą służbową ankieta osiągnęła niemal 100% realizacji próby. Gdy jednak zaczęliśmy czytać odpowiedzi jasne było, że prawie 40% respondentów potraktowało ankiety jako zabawę. Te odpowiedzi zostały zdyskwalifikowane niezależnie przez dwóch „sędziów". Zadanie oceny wiarygodności respondenta stało się proste, ponieważ były to ankiety papierowe. W przypadku komputerowych zbiorów danych nie da się tego ocenić holistycznie. Trzeba próbować zliczać sygnały ostrzegawcze, jak opisano w sekcji pierwszej.

Najłatwiejszym wskaźnikiem jest analiza rozkładu odpowiedzi beztreściowych – zarówno jako cechy pytania, jak i cechy respondenta. Liczbę TP (tj. odpowiedzi typu „trudno powiedzieć") należy analizować zarówno pod kątem konkretnych pytań, jak i indywidualnych respondentów. Pytania z dużą liczbą TP powinny być analizowane

[200] Schweinsberg i in., 2021.

osobno. Respondenci z wysoką liczbą TP powinni być usuwani z analiz – lokalnie lub globalnie (o czym pisaliśmy wcześniej). Po każdej, koniecznie opisanej szczegółowo w dokumentacji modyfikacji zbioru danych, warto przeprowadzić analizy porównawcze.

P#7.1. Nieznani muzycy i politycy

W naszych badaniach dotyczących szerokości obszarów akceptacji[201] w różnych dziedzinach zaobserwowano, że liczba TP przekroczyła 50% przy ocenie polityków oraz 10% przy ocenie albumów muzycznych. W pozostałych badanych dziedzinach odsetek ten nie przekraczał 1–2%. Tego rodzaju wyniki były podstawą do wniosku, że przedstawione respondentom nazwiska polityków oraz tytuły albumów muzycznych zostały źle dobrane, co utrudniało lub wręcz uniemożliwiało udzielenie merytorycznej odpowiedzi.

Analiza liczby odpowiedzi beztreściowych powinna stanowić jeden z pierwszych etapów weryfikacji jakości pytań ankietowych, ponieważ może wskazać na problemy związane z ich treścią, formą lub stopniem trudności dla respondentów. **Sprawdzenie rozkładów odpowiedzi beztreściowych zarówno dla pytań, jak i dla respondentów, powinno być wstępem do kolejnych analiz.**

P#7.2. Styl odpowiadania (wykorzystania skali odpowiedzi/*response style*)

W bardzo wielu badaniach[202] wykazano, że konieczność udzielenia odpowiedzi na dostarczonej skali zmusza respondentów do transformacji ich własnej wewnętrznej skali (np. dwuwartościowej: podoba mi się vs. nie podoba mi się lub bardziej finezyjnej skali dziesięciostopniowej) na zadaną skalę oceny. Efektem tej transformacji są różnice w zakresie wykorzystanej skali oceny, które zaobserwowaliśmy we wszystkich prowadzonych przez nas badaniach. Są one szczególnie zagrażające dla wyników, gdy np. używane do diagnozy kwestionariusze zawierają wyłącznie pytania jednokierunkowe. Wyjaśnimy to na przykładzie.

W badaniu[203] przeanalizowano dane zebrane za pomocą polskiej, komercyjnej wersji *Gallup's Clifton Strengths Finder Assessment*, opartej na **Modelu mocnych stron Gallupa**[204]. Kwestionariusz składa się ze 102 pytań i ma na celu mierzenie 34 talentów pracowników, choć przeszukiwania literatury nie pozwoliły znaleźć żadnych naukowych doniesień uzasadniających założenia, na których opiera się ten model (dlaczego 34 cechy?; dlaczego 4 kategorie?; dlaczego 5 talentów?).

---

[201] Wieczorkowska, 1998.

[202] Np. Wieczorkowska, 1993; Wieczorkowska i in., 2014.

[203] Wieczorkowska i Karczewski, 2019; Karczewski, 2019, 2022.

[204] Clifton i Nelson, 1992; Clifton i Harter, 2003.

W badaniu stosuje się 5-stopniową skalę odpowiedzi opisaną w następujący sposób: (1) **bardzo pasuje**; (2) pasuje; (3) ani tak, ani nie; (4) nie pasuje; (5) **bardzo nie pasuje**.

Gdy zadaniem respondenta był samoopis, na skali odpowiedzi pojawiało się sformułowanie uszczegóławiające „do mnie". Natomiast, gdy opis dotyczył współpracowników, na skali znajdowało się sformułowanie „do niego/do niej".

Konieczność ipsatyzacji zilustrujemy na przykładzie dwóch wymiarów (talentów w terminologii Gallupa), które zgodnie z teorią przedziałowości powinny być skorelowane ujemnie.

Wskaźniki **UPORZĄDKOWANIA** zbudowano, wykorzystując odpowiedzi na 3 pytania: „często koncentruję swoją uwagę na szczegółach"; „lubię tworzyć szczegółowe plany działań"; „często myślę: to potrzeba uporządkować, zaplanować działania krok po kroku".

Wskaźnik **ELASTYCZNOŚCI** zbudowano, wykorzystując odpowiedzi na 3 inne pytania: „łatwo dostosowuję się do zmieniającej się sytuacji"; „lubię robić wiele rzeczy jednocześnie – pojawianie się nowych sytuacji jest dla mnie inspiracją"; „nowe sytuacje są po to, żeby się do nich dostosować – zawsze traktuję zmianę jako coś naturalnego".

Porównajmy wyniki dwóch hipotetycznych osób oceniających się na wymiarach UPORZĄDKOWANIA i ELASTYCZNOŚCI.

Załóżmy, że na wymiarze uporządkowania średni wynik Ewy wyniósł <2>, Adam na <5>.

Na wymiarze ELASTYCZNOŚCI średni wynik Ewy wyniósł <3>, Adam na <4>.

Na podstawie takich surowych (nieprzekształconych) wyników moglibyśmy stwierdzić, że Adam jest zarówno bardziej elastyczny i bardziej uporządkowany niż Ewa.

Sytuacja jednak zmienia się, gdy odkryjemy, że oboje respondenci inaczej wykorzystują skalę odpowiedzi, co zdarza się dość często.

Ewa generalnie preferuje lewy kraniec skali, ponieważ jej średnia odpowiedź na 102 pytania wyniosła **2,5**. Z kolei **Adam** czuje się zgodny z niemal każdym opisem, ponieważ jego średnia wyniosła aż 4,5 na 5-stopniowej skali.

Jeśli, uwzględniając te informacje, dokonamy **ipsatyzacji (centrowania)** wyników surowych, okaże się, że na wymiarze ELASTYCZNOŚCI surowy wynik Adama równy **<4>** zmieni się na 4 – 4,5 = **<–0,5>**, a surowy wynik Ewy równy **<3>** zmieni się na 3 – 2,5= **<+0,5>**.

Porównanie wyników ipsatywnych na wymiarze elastyczności prowadzi do odmiennego wniosku: Ewa z wynikiem **<+0,5>** jest bardziej elastyczna od Adama z wynikiem **<–0,5>** niż porównanie wyników surowych, gdy Adam z wynikiem **<4>** lokował się wyżej niż Ewa z wynikiem **<3>**.

Takie analizy ipsatyzacyjne przeprowadzono na bardzo dużym zbiorze danych. W pierwszym kroku dla **każdego respondenta** obliczono jego indywidualną średnią

z odpowiedzi na 102 pytania. Analizowane samoopisy pochodziły od 1132 pracowników, a pozostałe opisy pochodziły od 4748 współpracowników. Tak, jak we wszystkich analizowanych przez mój zespół badaniach występowało spore zróżnicowanie międzyosobnicze w sposobie użycia skali odpowiedzi (dla dolnych 10% średnia była mniejsza niż 3,36, a dla górnych 10% większa niż 4,11). W związku z tym dokonano ipsatyzacji (centrowania) wszystkich odpowiedzi.

Przedmiotem porównania była jednorodność wskaźników, operacjonalizowana standardowo za pomocą alfy Cronbacha. Wszystkie 4 analizy (2 wymiary × 2 typy ocen) wykazały wzrost alfy Cronbacha (dla ELASTYCZNOŚCI z 0,64 na 0,78 dla samoopisów i z 0,67 na 0,77 dla opisów współpracowników, dla UPORZĄDKOWANIA z 0,71 na 0,85 dla samoopisów i z 0,69 na 0,83 dla opisów współpracowników[205]).

Korelacja między UPORZĄDKOWANIEM a ELASTYCZNOŚCIĄ po ipsatyzacji zmieniła się z –0,1 na –0,36 dla samoopisów. Jeszcze bardziej spektakularna zmiana z dodatniego współczynnika korelacji +0,16 na –0,26 miała miejsce w przypadku opisów współpracowników. Ze względu na wielkość próby wszystkie wyniki są istotne statystycznie, ale zależności znacząco się różnią. Dla samoopisów oczekiwana zależność jest bardzo słaba (r^2 = 0,01) dla danych surowych i blisko 13 razy silniejsza dla danych po ipsatyzacji (r^2 = 0,127). Znak współczynnika korelacji jest zgodny z przewidywaniami teorii.

Dla opisów sporządzonych przez współpracowników zależność również jest silniejsza dla danych po ipsatyzacji, ale co najważniejsze dla danych surowych współczynnik korelacji jest DODATNI, a więc niezgodny z teorią. Widząc istotnie dodatni współczynnik korelacji r = +0,16 moglibyśmy wnioskować, że uporządkowanie sprzyja elastyczności. Dzięki ipsatyzacji wyników otrzymujemy jednak zgodne z teorią konkluzje, co zwiększa nasze zaufanie do pomiaru.

Niebanalne znaczenie, zwłaszcza, gdy chcemy dokonywać podziału grupy według miar pozycyjnych (np. podział medianowy), ma fakt, że po ipsatyzacji zmienna staje się „bardziej ciągła" tzn. przyjmuje więcej wartości.

Z analogicznym problemem mamy do czynienia w 104-elementowym inwentarzu i 13 skalach **Inwentarza Stylów Myślenia Sternberga-Wagnera** (*Sternberg-Wagner Thinking Styles Inventory*). Zadaniem respondenta jest zaznaczenie odpowiedzi na siedmiopunktowej skali (od „w ogóle" do „wyjątkowo dobrze") niezawierającej odpowiedzi beztreściowej typu „trudno powiedzieć". Autor zakłada, że wszystkie style – w tym preferencje opisujące pozornie przeciwstawne wymiary, takie jak lokalność i globalność – są nieskorelowane. W naszych badaniach[206] okazało się jednak, że taki **wynik jest konsekwencją sposobu pomiaru**, ponieważ wszystkie pozycje skali zostały przez autorów ułożone w tym samym kierunku. Dla przykładu skala globalności zawiera wyłącznie opisy osób preferujących globalny styl myślenia, a nie zawiera

[205] Wieczorkowska i Karczewski, 2019.

[206] Eljaszuk, 2001; Wieczorkowska i Eliasz, 2004.

opisów osób unikających takiego przetwarzania informacji. Skala lokalności zbudowana jest odwrotnie. Taka – bardzo często spotykana – jednokierunkowość pozycji testowych uruchamia automatyczną tendencję do potakiwania/zgadzania się wynikającą z omówionej w pierwszej części rafy konfirmacji.

Wpływ tendencji do potakiwania wykazano np. w badaniu eksperymentalnym[207], gdy podzielonym losowo na 2 grupy ochotnikom zadano dwie różne wersje tego samego pytania. Za **nakazem obniżenia temperatury** było **38,3%** w grupie E1 i **29,4%** w grupie E2. Grupa E1 była pytana: „Czy gdyby tej zimy były problemy z dostawami energii, byłbyś za tym, żeby wprowadzić prawo nakazujące ludziom obniżenie temperatury ogrzewania w domach?". Pytanie dla grupy E2 zawierało obie opcje: „Czy gdyby tej zimy były problemy z dostawami energii, byłbyś za tym, żeby wprowadzić prawo nakazujące ludziom obniżenie temperatury ogrzewania w domach, czy też sprzeciwiałbyś się takiemu pomysłowi?".

Styl odpowiadania (wykorzystywania skali odpowiedzi) może zależeć od różnic kulturowych, co jest ważne w porównaniach międzynarodowych. Porównanie danych z internetowych badań własnych włoskiej i japońskiej próby reprezentatywnej i danych zebranych w obu krajach w ramach *International Social Survey Programme* wykazało, że w prawie wszystkich z analizowanych 6 zestawów **pytań** Japończycy udzielali mniej odpowiedzi skrajnych i uchylali się od zajęcia stanowiska istotnie częściej niż Włosi. Można to interpretować jako przejaw różnic w stylach myślenia[208], polegających na dominacji akceptacji sprzeczności (dobre i złe, mocne i słabe itd., istnieje we wszystkim) w kulturze Wschodu i reguły wyłączonego środka (zdanie jest albo prawdziwe, albo fałszywe) w Kulturze Zachodu.

P#7.3. Splątane efekty w badaniach generacyjnych

Zgodnie z rafą konfirmacji widzimy w naszych wynikach to, czego oczekiwaliśmy. Wszystkie wyniki dadzą się zinterpretować na różne sposoby, więc obowiązkiem badacza jest konstruowanie i testowanie alternatywnych interpretacji.

Przykładowo – bardzo popularne są badania porównujące różne generacje np. na rynku pracy. Opisując różnice generacyjne, operuje się pojęciem zbiorowej (prototypowej) jednostki, która została ukształtowana przez dominujące warunki społeczne i kulturowe w okresie formatywnym socjalizacji wtórnej. W takich badaniach najczęściej stosuje się najłatwiejszą strategię porównań poprzecznych (*cross-sectional design*), tj. porównań generacji w jednym punkcie czasowym. Respondenci pytani są o wiek i na tej podstawie przydzielani do kohort generacyjnych. Autorzy takich badań najczęściej nie troszczą się w swoich interpretacjach o fakt, że stosując strategię porównań poprzecznych, nie da się rozdzielić wpływu 3 splątanych efektów

[207] Schuman i Presser, 1996.

[208] Nisbett, 2011.

określanych jako WEK (Wiek, Epoka, Kohorta/Generacja), a w języku angielskim jako APC (od *Age–Period–Cohort*).

Gdy zapytamy laika od czego zależy ocena atrakcyjności np. tatuaży, to bez problemu powie, że ważny jest czas przeprowadzenia badania (teraz tatuaże oceniane są jako bardziej pożądane niż 40 lat temu) i od wieku respondenta (entuzjastów jest zdecydowanie więcej wśród młodszych niż starszych). Oznacza to, że nawet laik jest świadom występowania efektu **epoki** (czasu badania) i efektu **wieku biologicznego**. Rzadko kto jednak wskaże na efekt **generacji**. Nasze sieci neuronowe tworzą wartości prototypowe mające status oczywistości[209] na podstawie obserwacji otoczenia. W **okresie formatywnym socjalizacji wtórnej** (wczesna młodość) w otoczeniu Boomerów było nieporównanie mniej tatuaży niż w otoczeniu Millenialsów.

Niestety, prowadząc badania w jednym punkcie czasowym rozdzielenie wpływu wieku biologicznego (młodsi pracownicy różnią się od starszych, np. siłą mięśni, sprawnością pamięci świeżej od wpływu kohorty/generacji wynikającego z różnic środowiskowych w uprzywilejowanym okresie formatywnym socjalizacji wtórnej) jest niemożliwe, ponieważ wiek respondenta został jednoznacznie określony na podstawie informacji o roku urodzenia (określającym generację) i rok badania.

wiek = rok badania – rok urodzenia

Aby oddzielić wpływ wieku biologicznego od wpływu kohorty konieczne są badania podłużne (*time–lag studies*). Prototypowym przykładem jest badanie[210] *Monitoring the Future* dla uczniów czwartych klas liceum prowadzone od 1976 roku przez Institute for Social Research przy Uniwersytecie Michigan w Ann Arbor. Badane próby są reprezentatywne w skali krajowej, co oznacza, że rozkład podstawowych zmiennych socjodemograficznych (płeć, wiek, rasa, wykształcenie, rejon) w próbie odpowiada rozkładowi tych zmiennych w populacji uczniów amerykańskich szkół. Porównywani respondenci są zawsze w tym samym wieku, więc efekt wieku biologicznego nie ma wpływu na badanie. Przykładowo, wypełniający w 1976 roku ankietę uczniowie z Generacji BB mieli 17–18 lat, tyle samo co uczniowie z Generacji X wypełniający tę ankietę w 1990 roku, Millenialsi w 2005 roku oraz przedstawiciele Generacji Z w 2015 roku.

Dzięki temu możemy zobaczyć np. postępującą inflację stopni.

W takich badaniach można wykluczyć efekt wieku biologicznego, jednakże tego rodzaju dane nie są w stanie oddzielić efektu generacji/kohorty (wynikającego z wpływów kulturowych panujących w trakcie formacyjnego okresu socjalizacji) od efektów epoki (wynikających z aktywnych w trakcie trwania badań wpływów kulturowych oddziaływujących na wszystkie generacje). Z praktycznego punktu widzenia

[209] Wieczorkowska, 2025.

[210] Twenge, 2024.

to czy dana różnica jest wynikiem efektu generacji czy epoki, nie zawsze ma wielkie znaczenie, ponieważ w obu przypadkach mamy do czynienia ze zmianą kulturową.

Jeśli mamy dane pochodzące od wielu generacji na przestrzeni wielu lat, to stosując zaawansowane metody statystyczne można oddzielić efekt wieku biologicznego od efektu epoki i generacji[211].

Temat pułapek w analizach i interpretacjach jest niewyczerpywalny, więc nawet nie będę próbować ich wszystkich omówić. Zasygnalizowałam wybrane trudności, którymi zajmowałam się w mojej pracy naukowej. Wciąż pozostaje bardzo wiele do zrobienia w naszej dyscyplinie, zanim będziemy mogli równać się z postępem, jaki dokonał się w naukach technicznych i przyrodniczych.

Powierzchowne analizy są marnowaniem wysiłku respondentów. Dane należy **„męczyć tak długo aż zaczną zeznawać" – czasem wymaga to przeprowadzenia kolejnych badań nastawionych na testowanie alternatywnych**. Inaczej szkoda nakładów poniesionych na prowadzenie badań i czasu respondentów.

[211] Wieczorkowska i Wilczyńska, 2023.

Zakończenie
Co warto zapamiętać?

W czasie lektury sieci neuronowe Czytelników wyłowiły to, co było dla nich najciekawsze. Skorzystam jednak z przywileju Autorki i wypunktuję, co chciałabym, aby zostało zapamiętane.

Gdy planujemy badania ankietowe:

1. Okazujmy respondentom **szacunek** – dbajmy o ich czas i wysiłek. Nie irytujmy nieprzemyślanymi pytaniami i długością ankiet. Pytajmy tylko o to, co na pewno wykorzystamy w analizach. Wyobraźmy sobie możliwie najdokładniej jak będzie wyglądał nasz „raport z badań", zanim zbierzemy dane. Zazwyczaj mamy tendencję do zadawania zbyt wielu pytań. Zanim będziemy „męczyć" respondentów, **sami** najpierw udzielmy na pytanie odpowiedzi, a potem koniecznie przeprowadźmy pilotaż.
2. Ankieta jest rozmową, respondenci chcą wiedzieć nie tylko, o co są pytani, ale także **po co**. Pamiętajmy, że dla respondentów ankieta jest całością – **poprzednie pytania** wpływają na interpretację kolejnych.
3. Pytania powinny być krótkie, jasne i precyzyjne bez fachowego żargonu, podwójnych przeczeń, ukrytych założeń. Unikajmy pytań **podwójnych** (np. czy zajęcia były przydatne i ciekawe), pytań o **obiekty zbiorowe** (np. ufam współpracownikom, bo respondenci mogą mieć problem z odpowiedzią, gdy ufają bardzo nielicznym) i pytań **z założeniem**.
4. Pamiętajmy, że zadawanie pytań jest informowaniem – np. respondenci **wnioskują z kafeterii** odpowiedzi o tendencji centralnej. Także kolejność opcji w kafeterii wpływa na rozkład odpowiedzi.
5. Pozwólmy przyznać respondentom, że nie znają odpowiedzi na pytania, nie chce im się myśleć… i wybrać opcję beztreściową: „trudno powiedzieć, nie wiem, nie mam zdania". Unikajmy opcji środkowych a opcję **beztreściową** umieszczajmy poza skalą odpowiedzi – nigdy w środku.

6. Pamiętajmy, że **podstawą naukowych twierdzeń jest porównanie** – zanim skonstruujemy ankietę, zastanówmy się z czym porównamy otrzymane wyniki. Jeśli to możliwe, zdecydowanie warto wprowadzić do badania ankietowego losowy podział respondentów na grupy, dzięki czemu możemy umieścić w schemacie badawczym różne wartości zmiennych, których wpływ nas interesuje.

Próby reprezentatywne są konieczne, jeżeli celem badacza jest **estymacja** rozkładów zmiennych w populacji – np. gdy chcemy przewidzieć wyniki wyborów (ale i tu powinniśmy szukać próby reprezentatywnej dla tych, którzy „chodzą" głosować, a nie tych, którzy są uprawnieni do głosowania). Łatwo jest zbadać próbę reprezentatywną przedmiotów nieożywionych, np. wyprodukowanych śrubokrętów, bo one **nie mogą odmówić udziału w badaniach.** Gdy obiektem są ludzie, możemy wylosować próbę spełniającą kryteria, ale nie możemy zagwarantować, że wylosowane osoby zechcą wziąć udział w badaniu. Stopień realizacji próby w badaniach sondażowych spadł w ciągu ostatnich dziesięcioleci wielokrotnie. Co gorsza, mamy do czynienia ze zjawiskiem **fałszywych respondentów** – którzy godząc się na udział w badaniach udzielają odpowiedzi losowych. Dlatego nie ma nic złego w tym, że do testowania hipotez w większości badań używa się prób łatwo dostępnych (*convenience*). Miło byłoby móc wykazać trafność zewnętrzną (możliwość generalizacji na populację) uzyskanych wyników, ale dużo ważniejsza jest trafność wewnętrzna (najpierw trzeba mieć co generalizować), którą można zwiększać za pomocą randomizacji II stopnia prób łatwo dostępnych.

Tak jak pisałam wielokrotnie – czekamy w naszej dyscyplinie na skok technologiczny, który dokona się dzięki wykorzystaniu m.in. biosensorów. Już dziś dane ilościowe – np. informacje o poziomie cukru zbierane przez sensor umieszczony na przedramieniu diabetyka – są skuteczniejszym motywatorem do zmiany zachowania niż większość tradycyjnych oddziaływań motywacyjnych. Głęboko wierzę, że przyszłością naszej dyscypliny są ILOŚCIOWE eksperymentalne studia przypadków.

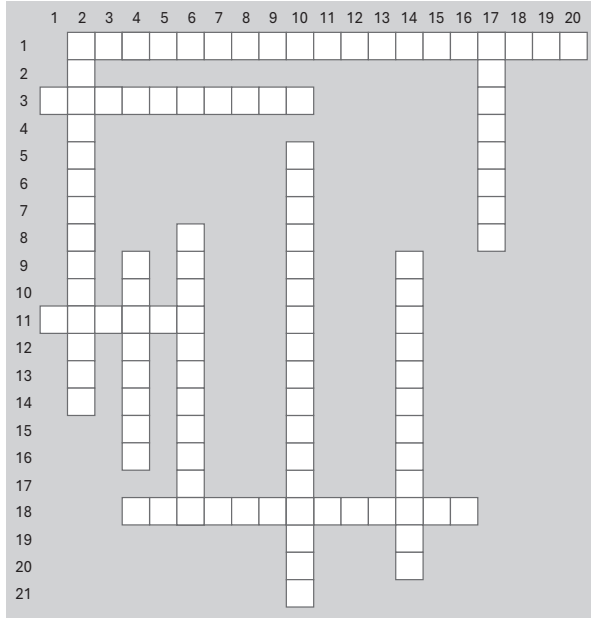
Liczby mają wielką moc – pozwalają na banalnie proste porównania, tworzenie rankingów, wyliczanie bardziej skomplikowanych wskaźników. Niestety liczby mają także ogromny potencjał do rozprzestrzeniania dezinformacji. Tanie chińskie zegarki pięknie wyświetlają ważne parametry życiowe, w tym m.in. ciśnienie krwi, ale – jak miałam okazję niedawno się przekonać – różnica między wynikiem uzyskanym z zegarka a pomiarem wykonanym mankietem wyniosła prawie 30 mm Hg. Takie zafałszowane liczby są większym zagrożeniem niż zafałszowane słowa, ponieważ większość z nas obdarza liczby dużym zaufaniem.

Niepokoi mnie jednak, jeśli traktujemy liczby lepiej niż na to zasługują – np. wyliczany i reklamowany przez lekarzy „indeks zdrowia", w którym sumuje się wartości pochodzące z różnych rozkładów zmiennych – takich jak palenie tytoniu, wskaźnik masy ciała (BMI), aktywność fizyczna czy ciśnienie krwi. Indeks ten opiera się na tym, co łatwo zmierzyć, pomijając wiele równie istotnych, ale znacznie trudniej-

szych do uchwycenia wskaźników – takich jak jakość tkanki mięśniowej czy poziom tłuszczu trzewnego. Czy parafrazując znane powiedzenie „lepsza kiepska maszyna do szycia niż szycie garnituru ręcznie", powiedzieć, że „lepsze kiepskie liczby niż żadne"??? Mam nadzieję, że po lekturze tej książki Czytelnik wie, że takimi kiepskimi liczbami są np. wyniki zdecydowanej większości ankiet ewaluacyjnych, które traktujemy lepiej niż na to zasługują. Trzeba dołożyć wszelkich starań, aby pamiętać o aktualnych słabościach dostępnego pomiaru i robić wszystko, aby ten pomiar doskonalić.

Krzyżówki

#Krzyżówka 0. Wprowadzenie



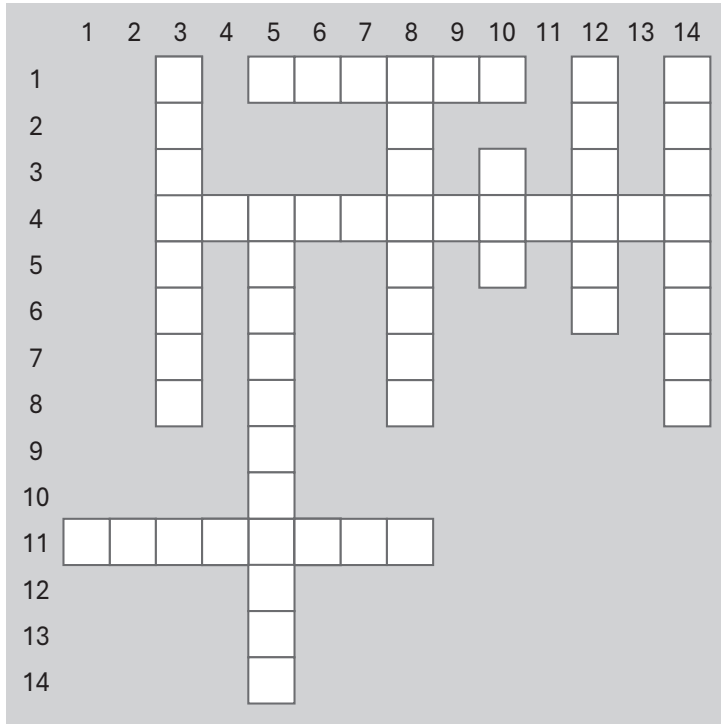
Poziomo

1. Systemów ====== nie da się łatwo rozłożyć na odrębne elementy składowe, które po połączeniu stworzą z powrotem całość.
3. Pod wpływem silnego czynnika stresowego człowiek może zostać ======, co nigdy nie przydarza się śrubokrętom.
11. Przyznanie przed samym sobą, że nasze silne przekonania mogą być po prostu ======, jest dla wielu osób bardzo trudne.
18. ====== oddziaływań psychologicznych oznacza, że TEN SAM efekt można osiągnąć za pomocą RÓŻNYCH oddziaływań (np. zmotywować do zwiększonego wysiłku docenieniem psychologicznym albo obietnicą awansu lub zagrożeniem zabrania premii albo...).

Pionowo

2. Łatwy dostęp do informacji wytworzył usieciowioną ====== – wydaje nam się, że wszystko możemy znaleźć i zrozumieć.
4. Nauki o zarządzaniu wypracowały swoją metodologię, nie różnicując jej ze względu na to, czy dotyczy ona zasobów rzeczowych, czy ======.
6. Główne przesłanie tej monografii mówi, że nie warto bezmyślnie kopiować procedur badania świata ====== do badania ludzi.
10. ====== oznacza, że to samo oddziaływanie psychologiczne może doprowadzić do różnych efektów (np. takie samo zwrócenie uwagi: jednego pracownika zmotywuje do zwiększenia wysiłku, drugiego zniechęci i skłoni do odejścia z pracy).
14. Cechy systemów, które mogą zareagować na czynniki stresowe paradoksalnie – ======.
17. W odróżnieniu od śrubokręta mózg potrafi się czasem sam ======.

#Krzyżówka 1. Rafy metodologiczne w badaniach ludzi



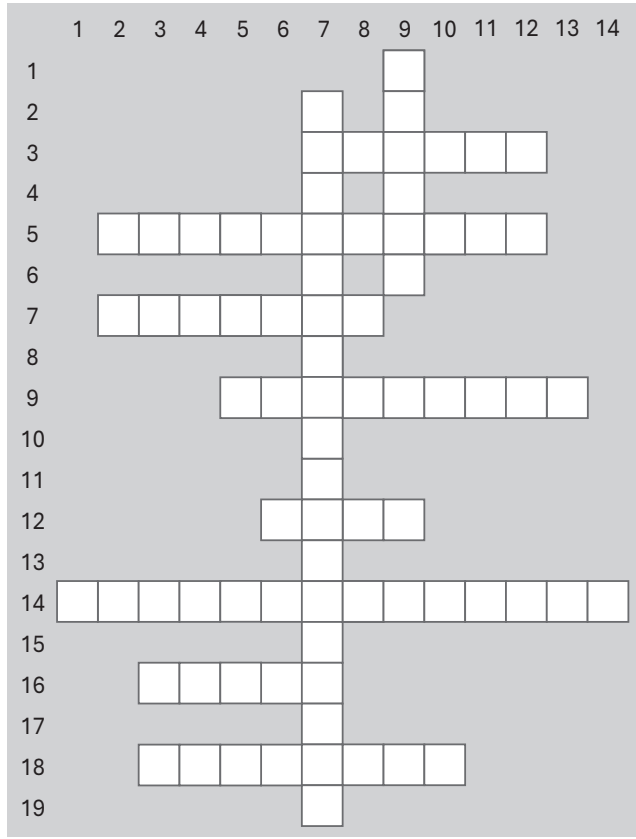
Poziomo

1. ====== doskonale nadają się do roznoszenia ściemy, ponieważ większość ludzi nie czuje się na siłach, aby zakwestionować ich prawdziwość.
4. Gdy respondent zaznacza „trudno powiedzieć" lub „nie wiem", to jest to odpowiedź ======.
11. ====== respondent przeklikujący odpowiedzi losowo.

Pionowo

3. Wykluczenie ======, gdy respondenci „nieuważni w jednym bloku pytań są usuwani z całego badania".
5. Lepszą strategią maksymalizacji trafności ====== niż badanie prób reprezentatywnych jest replikowanie eksperymentów na próbach homogenicznych.
8. W badaniach eksperymentalnych manipulujemy wartościami ====== niezależnej, a nie ludźmi.
10. Akronim dla pytań sprawdzających uwagę respondenta pochodzący od nazwy angielskiej ======.
12. ====== przydział ochotników do grup eksperymentalnych to metoda tworzenia zbiorowych klonów niezbędnych, abyśmy mogli prowadzić wnioskowanie przyczynowe.
14. O dużym realizmie psychologicznym mówimy, gdy sytuacja badawcza ====== uczestników zarówno poznawczo, jak i emocjonalnie.

#Krzyżówka 2. Rafy metodologiczne w diagnozowaniu ludzi



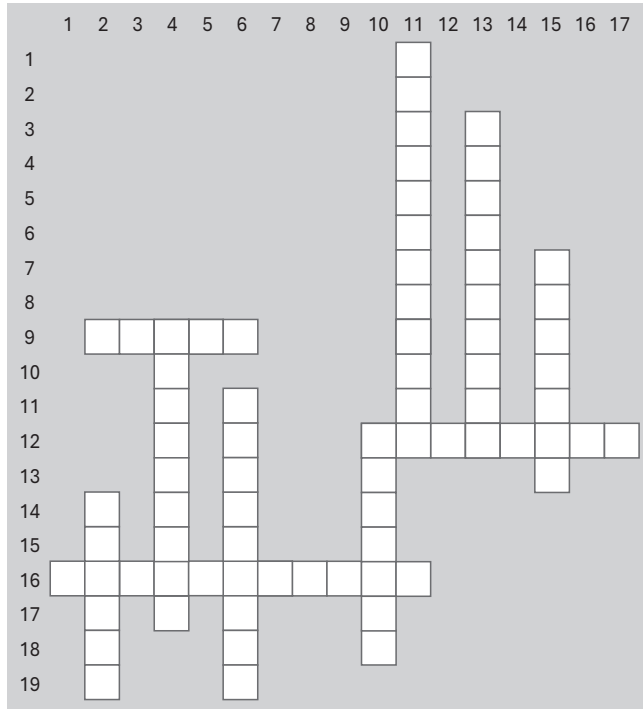
Poziomo

3. ====== postrzegamy dużo trafniej niż siebie.
5. Tendencja do ====== oznacza skłonność respondentów do odpowiadania „prawda”, „zgadzam się”, „tak” niezależnie od treści pytania.
7. Zbiór pytań zamkniętych i otwartych.
9. Nie jesteśmy świadomi mechanizmów, które zniekształcają nasz obraz Ja.
12. Zawiera pytania z prawidłowymi odpowiedziami.
14. Narzędzie diagnostyczne składające się z kilku skal.
16. Zestaw pytań mierzących tę samą zmienną latentną.
18. Skupia się na pomiarze cech, predyspozycji i kompetencji konkretnego pracownika.

Pionowo

7. W badaniach naukowych wykonywanych w paradygmacie *ceteris paribus* wyniki dotyczące pojedynczej osoby są ======.
9. Badanie ankietowe przeprowadzone na dużych próbach.

#Krzyżówka 3. Rafy metodologiczne w ewaluacji ludzi i efektów ich działań



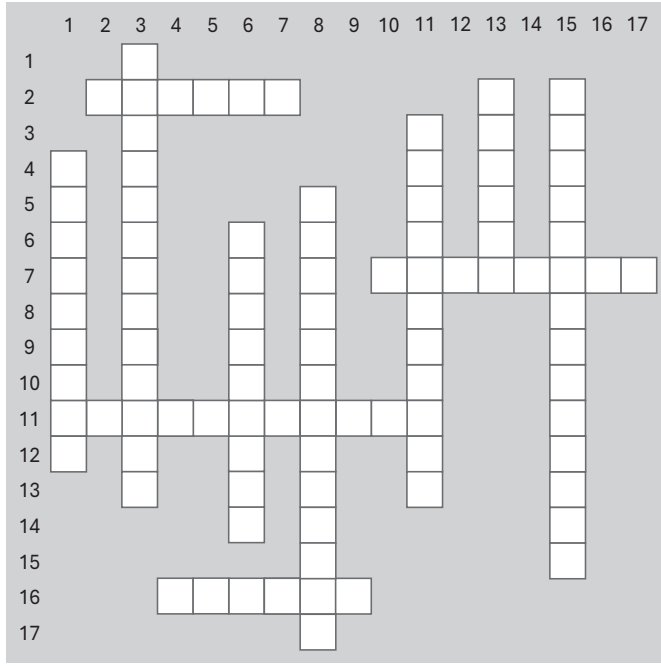
Poziomo

9. Ewaluacje przyjmujące formę słów to numeryczne ======.
12. Wyniki badań wskazują na wyższość ewaluacji ======, w której wszystkie oceniane obiekty (projekty badawcze, artykuły naukowe) są oceniane przez ten sam zespół ewaluatorów.
16. Abyśmy poczuli różnicę w porównaniu z ====== sztabką złota, musiałaby być ona co najmniej o 20 g cięższa, ponieważ wartość ułamka Webera dla ciężaru wynosi 0,02.

Pionowo

2. Ewaluacje przyjmujące formę słów to ======.
4. Osoba oceniająca obiekty podlegające ewaluacji – ======.
6. Sposób ocenienia serii obiektów redukujący efekt HALO ======.
10. Wariant ewaluacji seryjnej, w którym nie można modyfikować wcześniej oceny to wariant ======.
11. Cechą charakterystyczną ewaluatorów prowadzących rozmowę kwalifikacyjną jest ====== swoich umiejętności ewaluacyjnych.
13. Wariant ewaluacji seryjnej, w którym możemy modyfikować wcześniej oceny to wariant ======.
15. Gdy dla każdego ocenianego obiektu, np. projektów grantowych, wybierany jest niezależnie zespół ewaluatorów, to jest to ewaluacja ======.

#Krzyżówka 4. Rafy metodologiczne w ewaluacji rozwoju nauk o zarządzaniu ludźmi



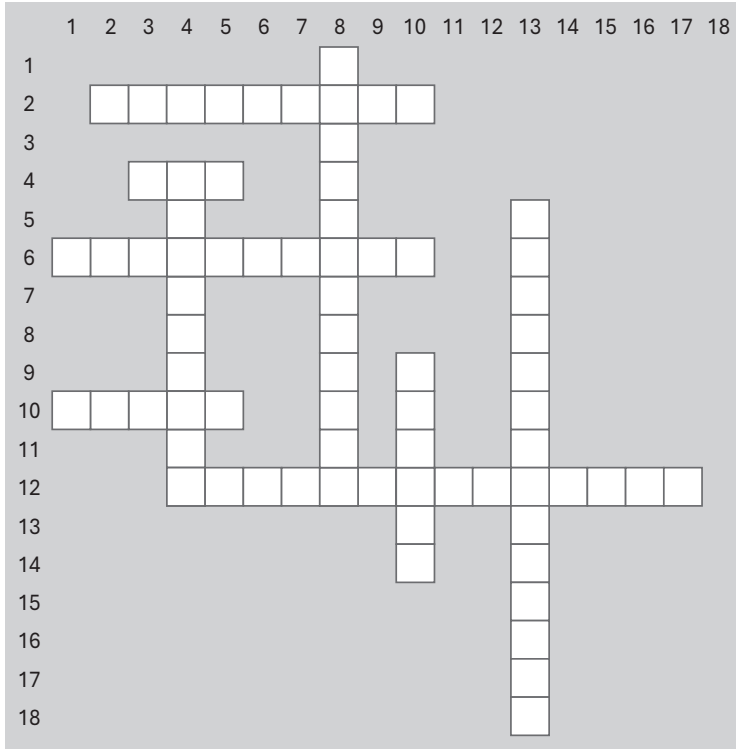
Poziomo

2. Tradycja: uniwersytet to wspólnota ludzi poszukujących ======.
7. Teksty w naukach społecznych, które powierzchownie spełniają naukowe standardy, ale są nastawione na oszukańcze wywarcie wrażenia istotności i głębi nazywane są naukowym ======.
11. Jednostki organizacyjne tego samego uniwersytetu są opisywane jako ====== plemiona walczące o posiadanie jak największego terytorium i przywilejów kosztem swoich sąsiadów.
16. Sfera normatywna zastępowana jest przez regulacje ======.

Pionowo

1. Wzrost bełkotu w szkolnictwie wyższym jest napędzany wiarą we wszechmoc ======.
3. Konfucjusz, odpowiadając na pytanie „Co sądzić o człowieku, którego kochają wszyscy w jego wiosce?”... docenił wagę ====== ocen.
6. Do polityki oświatowej zaadoptowano wolnorynkowe założenie: od każdego według tego, jak wybiera, każdemu według tego, jak jego ======.
8. Prawo mówiące, że miara, która staje się celem, przestaje dobrze mierzyć, podkreśla ====== naturę ilościowych celów.
11. Neoliberalizm: uniwersytet to wspólnota ludzi poszukujących sukcesu ======, który osiąga się przez konkurencję między jednostkami, uczelniami.
13. Bilig opisuje dramatyczny ====== poziomu intelektualnego publikacji.
15. Od nauczycieli akademickich oczekuje się, aby weszli w rolę skutecznych ======.

#Krzyżówka 5. Paradygmat metodologiczny WiW dla badań prowadzonych w obszarze ZZL (HRM)



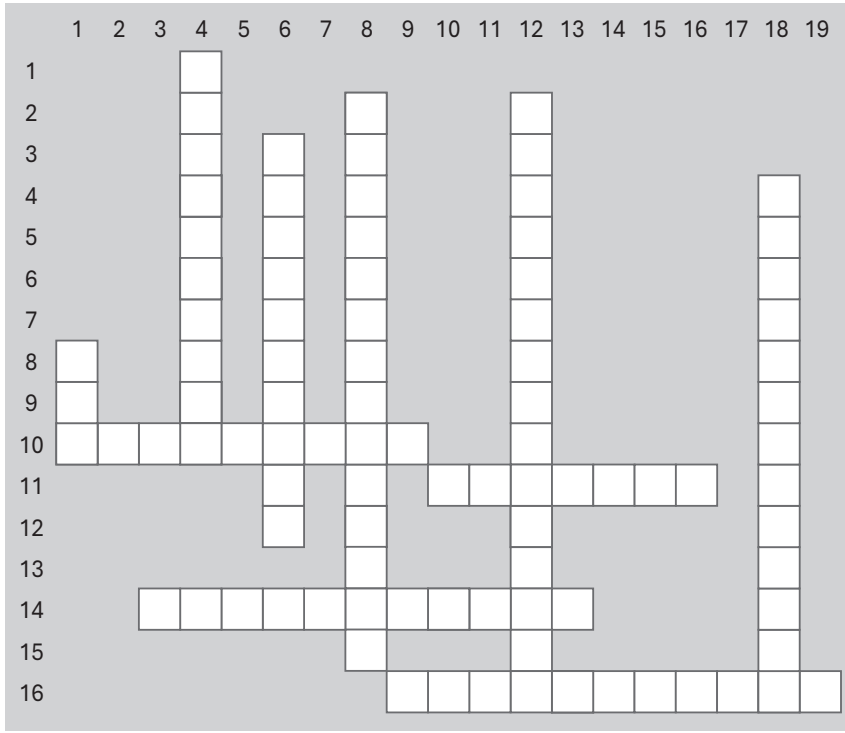
Poziomo

2. Przed przystąpieniem do analizy zbiory danych powinny być starannie oczyszczone z ====== respondentów.
4. Paradygmat metodologiczny uwzględniający specyfikę obiektu badań w ZZL.
6. Analizowane zbiory słów to dane ======.
10. Łączenie w badaniu metod korelacyjnych, eksperymentalnych i jakościowych to triangulacja ======.
12. Nawet w badaniach ankietowych możemy manipulować zmiennymi niezależnymi, czyli prowadzić badania ======, przydzielając ochotników losowo do różnych warunków eksperymentalnych.

Pionowo

4. Analizowane zbiory liczb to dane ======.
8. W medycynie podejście ====== umożliwia identyfikację i klasyfikację różnych schorzeń, a także ułatwia diagnozowanie i leczenie.
10. Gdy testujemy te same hipotezy na różnych zestawach danych to triangulacja ======.
13. ====== metodologiczny nie dopuszcza mieszania metod.

#Krzyżówka 6. Elementy vademecum badacza opinii



Poziomo

10. Zestaw odpowiedzi do wyboru przedstawiany respondentom razem z pytaniem.
11. Wpływ poprzedzających pytań na odpowiedzi jest ====== w grupie seniorów niż młodych respondentów.
14. Pytania ====== umieszczamy na końcu ankiety.
16. Odpowiedzi na zestaw pytań jednokierunkowych są obciążone przesunięciem związanym z tendencją do ======.

Pionowo

1. Akronim oznaczający splątane efekty w badaniach generacyjnych – polska wersja angielskiej: Age-Period-Cohort.
4. Podstawą formułowania wniosków naukowych jest ======.
6. ====== powinien wiedzieć nie tylko to, o co jest pytany, ale także po co jest pytany.
8. W środku skali odpowiedzi nie powinno się umieszczać odpowiedzi ======.
12. Próba jest =======, jeśli socjodemograficzna struktura przebadanej próby odpowiada socjodemograficznej strukturze danej populacji.
18. Zamiast badać bardzo trudno dostępne próby reprezentatywne zwiększamy trafność zewnętrzną (możliwość ====== naszych wyników na populację) poprzez replikowanie badań na wielu różniących się od siebie dostępnych próbach.

Lista haseł do krzyżówek

ACQ
angażuje
ankieta
antykruchość
bełkotem
beztreściowa
beztreściowych
biznesowego
błędne
danych
diagnoza
eksperymentalne
Eksperymentalne
ekwifinalność
ekwipotencjalność
elastyczny
ewaluator
fałszywy
„fałszywych”
fundamentalizm
generalizacji
globalne
ilościowe
innych
jakościowe
kafeteria
kilogramową
konkurujące
korupcjogenną
kwestionariusz
liczby
losowy
ludzkich
marketingowców
metod
metryczkowe
naprawić
nieinterpretowalne
obronnych
oceny
odrębna
opinie
porównanie
potakiwania
potakiwania
prawdy
prawne
przecenianie
przedmiotów
przemądrzałość
psychobiologicznych
publikowanie
reprezentatywna
respondent
seryjnej
skala
słabszy
sondaż
spadek
sygnały
sztywny
test
typologiczne
WEK
WiW
wizerunku
wybierają
wymiarowy
wzmocniony
zewnętrznej
zmiennej
zróżnicowania

Bibliografia

Altbach, P.G., Gumport, P.J., Berdahl, R.O. (red.) (2011). *American higher education in the twenty-first century: social, political, and economic challenges (3rd ed)*. Johns Hopkins University Press.

Altbach, P.G., Reisberg, L. (2018). Global trends and future uncertainties. *Change: The Magazine of Higher Learning, 50*(3–4), 63–67. https://doi.org/10.1080/00091383.2018.1509601

Altbach, P.G., Reisberg, L., Rumbley, L.E. (2010). Tracking a global academic revolution. *Change: The Magazine of Higher Learning*, *42*(2), 30–39. https://doi.org/10.1080/00091381003590845

Alvesson, M., Gabriel, Y., Paulsen, R. (2017). *Return to meaning: A social science with something to say*. Oxford University Press.

Antipov, E.A., Pokryshevskaya, E.B. (2017). Order effects in the results of song contests: Evidence from the Eurovision and the New Wave. *Judgment and Decision Making*, *12*(4), 415–419. https://doi.org/10.1017/S1930297500006288

Anzures, G., Quinn, P.C., Pascalis, O., Slater, A.M., Tanaka, J.W., Lee, K. (2013). Developmental origins of the other-race effect. *Current Directions in Psychological Science*, *22*(3), 173–178. https://doi.org/10.1177/0963721412474459

Ariely, D. (2011). *Potęga irracjonalności: ukryte siły, które wpływają na nasze decyzje*. Wydawnictwo Dolnośląskie.

Ariely, D. (2021). *Payoff. Ukryta logika kształtująca naszą motywację*. Grupa Wydawnicza Relacja.

Ariely, D. (2024). *Nie do wiary. Irracjonane przekonania racjonalnych ludzi*. Smak Słowa.

Armstrong, M. (2011). *Armstrong's handbook of human resource management practice* (12th ed.). Kogan Page.

Aronson, E., Aronson, J. (2020). *Człowiek-istota społeczna*, wyd. nowe. Wydawnictwo Naukowe PWN.

Aronson, E., Wieczorkowska, G. (2001). *Kontrola naszych myli i uczuć*. Santorski & Co.

Baer, R.A., Ballenger, J., Berry, D.T.R., Wetter, M.W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment*, *68*(1), 139–151. https://doi.org/10.1207/s15327752jpa6801_11

Bailenson, J. (2019). *Wirtualna rzeczywistość: doznanie na żądanie*. Helion.

Ball, P. (2007). *Masa krytyczna*. Insignis.

Bauman, Z., Leoncini, T. (2018). *Płynne pokolenie*. Czarna Owca.

Beck, M.F., Albano, A.D., Smith, W.M. (2019). Person-Fit as an Index of Inattentive Responding: A Comparison of Methods Using Polytomous Survey Data. *Applied Psychological Measurement*, *43*(5), 374–387. https://doi.org/10.1177/0146621618798666

Bergstrom, C.T., West, J.D. (2021). *Myśl krytycznie*. JK.

Berinsky, A.J., Margolis, M.F., Sances, M.W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–573. https://doi.org/10.1111/ajps.12081

Bess, T.L., Harvey, R.J. (2002). Bimodal score distributions and the Myers-Briggs Type Indicator: fact or artifact? *Journal of Personality Aassessment*, *78*(1), 176–186. https://doi.org/10.1207/S15327752JPA7801_11

Billig, M. (2013). *Learn to write badly: how to succeed in the social sciences*. Cambridge University Press.

Bokszański, Z. (2007). *Indywidualizm a zmiana społeczna*. Wydawnictwo Naukowe PWN.

Borman, W.C., Hallam, G.L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-differences correlates. *Journal of Applied Psychology*, *76*(1), 11–18. https://doi.org/10.1037/0021-9010.76.1.11.

Bowling, N.A., Gibson, A.M., Houpt, J.W., Brower, C.K. (2020). Will the Questions Ever End? Person-Level Increases in Careless Responding During Questionnaire Completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Breaugh, J.A. (2008). Employee recruitment: Current knowledge and directions for future research. *Human Resource Management Review*, *18*(3), 214–228. https://doi.org/10.1016/j.hrmr.2008.03.003

Brożek, B. (2019). *Myślenie. Podręcznik użytkownika*. Wydawnictwo Naukowe PWN

Bruine de Bruin, W. (2006). Save the last dance II: Unwanted serial position effects in figure skating judgments. *Acta Psychologica*, *123*(3), 299–311. https://doi.org/10.1016/j.actpsy.2006.01.009

Brzezińska, A.I., Brzeziński, J.M. (2011). Skale szacunkowe w badaniach diagnostycznych. W: J. Brzeziński (red.) *Metodologia badań społecznych. Wybór tekstów* (s. 299–399). Zysk i S-ka.

Brzeziński, J. (2019). *Metodologia badań psychologicznych*, wyd. nowe. Wydawnictwo Naukowe PWN.

Burnewicz, J. (2021). *Filozofia i metodologia nauk ekonomicznych*. Wydawnictwo Naukowe PWN.

Carr, N.G. (2013). *Płytki umysł: jak internet wpływa na nasz mózg*. Helion.

Caseras, X., Garner, M., Bradley, B.P., Mogg, K. (2007). Biases in visual orienting to negative and positive scenes in dysphoria: An eye movement study. *Journal of abnormal psychology*, *116*(3), 491. https://doi.org/10.1037/0021-843X.116.3.491

Churchill, G. A. (2002). *Badania marketingowe: podstawy metodologiczne*. Wydawnictwo Naukowe PWN.

Cialdini, R. (2023). *Pre-swazja. Jak w pełni wykorzystać techniki wpływu społecznego*. Gdańskie Wydawnictwo Psychologiczno-Naukowe.

Çiçek, M. (2015). Wearable technologies and its future applications. *International Journal of Electrical, Electronics and Data Communication*, *3*(4), 45–50.

Cichomski, B. (2005). *Polskie Generalne Sondaże Społeczne: skumulowany komputerowy zbiór danych 1992–2005*. Instytut Studiów Społecznych, Uniwersytet Warszawski.

Cichomski, B., Jerzyński, T., Zieliński, M. (2010). *Polskie Generalne Sondaże Społeczne 1992–2010*. Instytut Studiów Społecznych, Uniwersytet Warszawski, Polskie Archiwum Danych Społecznych, Repozytorium Danych Społecznych. https://doi.org/10.18150/G5XB6R, V1

Clarke, I. III. (2000). Extreme Response Style in Cross-Cultural Research: An Empirical Investigation. *Journal of Social Behavior & Personality*, *15*(1), 137–152.

Clifton, D.O., Nelson, P. (1992). *Soar with your strengths*. Delacorte Press.

Clifton, D.O., Harter, J.K. (2003). Investing in Strengths. W: A.K.S. Cameron, B.J.E. Dutton, C.R.E. Quinn (red.) *Positive Organizational Scholarship: Foundations of a New Discipline* (111–121). Berrett-Koehler Publishers, Inc.

Conrad, F.G., Tourangeau, R., Couper, M.P., Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61. https://doi.org/10.18148/srm/2017.v11i1.6304

Corneille, O., Klein, O., Lambert, S., Judd, C. M. (2002). On the role of familiarity with units of measurement in categorical accentuation: Tajfel and Wilkes (1963) revisited and replicated. *Psychological Science*, *13*(4), 380–383. https://doi.org/10.1111/j.0956-7976.2002.00468.x

Costa, P.T., McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five–Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resourses.

Couper, M.P., Tourangeau, R., Conrad, F.G., Crawford, S.D. (2004). What They See Is What We Get: Response Options for Web Surveys. *Social Science Computer Review*, *22*(1), 111–127. https://doi.org/10.1177/0894439303256555

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin*, *8*(2), 214–220. https://doi.org/10.1177/0146167282082005

Curran, P.G., Kotrba, L., Denison, D. (2010). *Careless responding in surveys: applying traditional techniques to organizational settings*. *25th annual conference of Society for Industrial and Organizational Psychology*, Atlanta, USA.

Czakon, W. (2011). *Podstawy metodologii badań w naukach o zarządzaniu*. Oficyna Wolters Kluwer business.

Czapiński, J. (1996). Uziemienie polskiej duszy. W: M. Marody, E. Gucwa-Leśny (red.) *Podstawy życia społecznego w Polsce* (s. 252–275). Instytut Studiów Społecznych Uniwersytetu Warszawskiego.

Damasio, A. (1999/2020). *Błąd Kartezjusza: Emocje, rozum i ludzki mózg*. Rebis.

Damasio, A. (2022). *Odczuwanie i poznawanie*. Kraków Copernicus Center Press.

Dewberry, C., Davies-Muir, A., Newell, S. (2013). Impact and Causes of Rater Severity/Leniency in Appraisals without Postevaluation Communication Between Raters and Ratees. *International Journal of Selection and Assessment*, *21*(3), 286–293. https://doi.org/10.1111/ijsa.12038

Dmochowska, H. (red.) (2014). *Polska 1989–2014*, opracowanie zbiorcze. GUS.

Drucker, P. F. (2006a). *Classic Drucker: essential wisdom of Peter Drucker from the pages of Harvard Business Review*. Harvard Business Press.

Drucker, P. F. (2006b). What executives should remember. *Harvard Business Review*, *84*(2), 144–153.

Durán, J.I., Fernández-Dols, J.-M. (2021). Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. *Emotion*, *21*(7), 1550–1569.

Edelman, B., Larkin, I. (2009). *Demographics, Career Concerns or Social Comparison: Who Games SSRN Download Counts?* Harvard Business School.

Eljaszuk, M. (2001*). Związek przedziałowej strategii aktywności ze stylami myślenia w ujęciu R.J. Sternberga*. Nieopublikowana praca magisterska. Promotor: G. Wieczorkowska.

Fasold, F., Memmert, D., Unkelbach, C. (2012). Extreme judgments depend on the expectation of following judgments: A calibration analysis. *Psychology of Sport and Exercise*, *13*(2), 197–200. https://doi.org/10.1016/j.psychsport.2011.11.004

Feldman Barrett, L. (2022). *Mózg nie służy do myślenia (7 i ½ wywrotowych lekcji o mózgu)*. JK

Galesic, M., Tourangeau, R., Couper, M.P., Conrad, F.G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72*(5), 892–913. https://doi.org/10.1093/poq/nfn059

Ganis, G., Rosenfeld, J.P., Meixner, J., Kievit, R.A., Schendan, H.E. (2011). Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *Neuroimage*, *55*(1), 312–319. https://doi.org/10.1016/j.neuroimage.2010.11.025

Gerrig, R.J., Zimbardo, P.G. (2006). *Psychologia i życie*. Wydawnictwo Naukowe PWN.

Gilbert, D. (2007). *Na tropie szczęścia*. Media Rodzina.

Goldberg, E. (2024). *Dyrygent w Twojej głowie. Płaty czołowe w złożonym świecie*. Wydawnictwo Naukowe PWN.

Gorynia, M. (2025). *Recenzje w postępowaniach naukowych*. Forum akademickie. https://forumakademickie.pl/sprawy-nauki/recenzje-w-postepowaniach-naukowych/

Greenberg, J., Solomon, S., Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. W: M.P. Zanna (red.), *Advances in experimental social psychology*, 29 (s. 61–139). Academic Press.

Greszki, R., Meyer, M., Schoen, H. (2015). Exploring the Effects of Removing 'Too Fast' Responses and Respondents from Web Surveys. *Public Opinion Quarterly*, *79*(2), 471–503. https://doi.org/10.1093/poq/nfu058

Grice, H.P. (1977). Logika i konwersacja. *Przegląd Humanistyczny*, *6*, 85–99.

Grice, H.P. (1980). Logika i konwersacja. W: B. Stanosz (red.) *Język w świetle nauki* (s. 91–114). Czytelnik.

Gromska, J. (2002). Anatomia przemocy. Przemoc a neurobiologia. *Psychiatria w Praktyce Ogólnolekarskiej*, *2*(4), 239–243.

Grudzewski, W.M., Hejduk, I.K., Sankowska, A., Wańtuchowicz, M. (2007). *Zarządzanie zaufaniem w organizacjach wirtualnych*. Difin.

Hallowell, E.M., Ratey, J.J. (2005). *Delivered from distraction: Getting the most out of life with attention deficit disorder*. Ballantine Books.

Hensel, P. (2017). *Legitymizacja badań organizacji*. Wydawnictwo Naukowe PWN.

Hensel, P.G., Kacprzak, A. (2020). Job overload, organizational commitment, and motivation as antecedents of cyberloafing: evidence from employee monitoring software. *European Management Review, 17*(4), 931–942.

Hensel, P.G., Kacprzak, A. (2021). Curbing cyberloafing: studying general and specific deterrence effects with field evidence. *European Journal of Information Systems*, *30*(2), 219–235. https://doi.org/10.1080/0960085X.2020.1756701

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J.L., DeSimone, J.A. (2021). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*, *36*(5), 807–828.

Jerzyński, T. (2008). *Wybrane korelaty liczby odpowiedzi beztreściowych w badaniach sondażowych*. Nieopublikowana rozprawa doktorska. Instytut Socjologii UW. Promotor: J. Wierzbiński.

Jerzyński, T. (2009). Poziom realizacji próby i ważenie poststratyfikacyjne w analizie danych sondażowych. *Problemy Zarządzania*, *7*(4, 26), 196–208.

Jeśka, M. (2016). *Konsekwencje stopnia dopasowania cech pracownika do zadań o różnym poziomie standaryzacji procesów*. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnston, M., Johnston, D. (2023). Badanie typu N = 1. Czego możemy się dowiedzieć, badając jednostkę. W: D. Kwasnicka, G. A. Ten Hoor, K. Knittle, S. Potthoff, A. Cross, J. Olson (red.). *Praktyczna psychologia zdrowia – Przełożenie badań nad zachowaniem na praktykę* (tłum.: Z. Kwissa-Gajewska i E. Gruszczyńska), 1. http://doi.org/10.17605/OSF.IO/M72P5 (Wersja angielska opublikowana w roku 2021).

Jost, J.T., Glaser, J., Sulloway, F.J., Kruglanski, A.W. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. https://doi.org/10.1037/0033-2909.129.3.339

Judge, T.A., Ferris, G.R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal*, *36*(1), 80–105. https://doi.org/10.2307/256513

Jurek, P., Olech, M. (2017). *Siedmiowymiarowy Kwestionariusz Osobowości. Narzędzie do diagnozy osobowościowych predyspozycji zawodowych*. Fundacja Altkom Akademia.

Kabut, M. (2021). False Respondents in Web Human Resource Surveys. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Kahneman, D., Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *2*, 263–291.

Kahneman, D. (2012). *Pułapki myślenia: O myśleniu szybkim i wolnym*. Media Rodzina.

Kahneman, D., Sibony, O., Sunstein, C.R. (2022). *Szum czyli skąd się biorą błędy w naszych decyzjach*. Media Rodzina.

Kane, J.S., Bernardin, H.J., Villanova, P., Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *38*(4), 1036–1051. https://doi.org/10.2307/256619

Karczewski, W. (2019). *Predyspozycje do wykonywania pracy wysoko zrutynizowanej: rekomendacje dla zarządzania zasobami ludzkimi*. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Karczewski, W. (2022). *Predyspozycje do wykonywania pracy wysoko zrutynizowanej w organizacji*. Poltext.

Klein, R.A., Cook, C.L., Ebersole, C.R., Vitiello, C., Nosek, B.A., Hilgard, J., Ratliff, K.A. i in. (2022). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. *Collabra: Psychology*, *8*(1), 35271. https://doi.org/10.31234/osf.io/vef2c

Knäuper, B. (1999). The impact of age and education on response order effects in attitude measurement. *Public Opinion Quarterly*, *63*(3), 347–370. https://doi.org/10.1086/297724

Knäuper, B., Belli, R.F., Hill, D.H., Herzog, A.R. (1997). Question difficulty and respondents' cognitive ability: the impact on data quality. *Journal of Official Statistics*, *13*(2),181–199.

Knäuper, B., Carrière, K., Chamandy, M., Xu, Z., Schwarz, N., Rosen, N.O. (2016). How aging affects self-reports. *European Journal of Ageing*, *13*, 185–193.

Knäuper, B., Schwarz, N., Park, D.C. (2004). Frequency reports across age groups: Differential effects of frequency scales. *Journal of Official Statistics*, *20*, 91–96.

Knäuper, B., Schwarz, N., Park, D., Fritsch, A. (2007). The perils of interpreting age differences in attitude reports: question order effects decrease with age. *Journal of Official Statistics*, *23*, 515–528.

Kociatkiewicz, J., Kostera, M. (2014). Zaangażowane badania jakościowe. *Problemy Zarządzania*, *12*(1, 45), 9–17. https://doi.org/10.7172/1644-9584.45.1

Kołodziej, A., Brzezicka, A. (2015). Jak depresja zmienia sposób patrzenia – przegląd badań dotyczących zmian we wskaźnikach okulograficznych u osób cierpiących na zaburzenia depresyjne. *Neuropsychiatria i Neuropsychologia*, *10*(1), 11–18.

Kosslyn, S., Miller, W.G. (2019). *Górny mózg, dolny mózg*. Copernicus Center Press.

Kowalczyk, K. (2019). Rola recenzji w zarządzaniu środkami na badania naukowe. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Kowalczyk, K.K., Wieczorkowska, G. (2022). Pułapki liczbowych ewaluacji. W: G. Wieczorkowska (red.) *Cztery wyzwania zarządzania* (s. 107–125), Wydział Zarządzania UW.

Kreber, C. (red.) (2009). *The University and its Disciplines*. Routledge.

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J.A., Alwin, D.F. (1989). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*, 526–538.

Król, G., Kowalczyk, K. (2014). Ewaluacja projektów i abstraktów – wpływ indywidualnego stylu ewaluacji na oceny. *Problemy Zarządzania*, *12*(1), 137–155.

Kulczycki, E. (2023). *The evaluation game: How publication metrics shape scholarly communication*. Cambridge University Press.

Kurtz, J.E., Parrish, C.L. (2001). Semantic response consistency and protocol validity in structured personality assessment: the case of the NEO-PI-R. *Journal of Personality Assessment*, *76*(2), 315–332. https://doi.org/10.1207/S15327752JPA7602_12

Kuźmińska, A.O. (2016a). *Psychologiczne następstwa wykonywania funkcji kierowniczych. Konsekwencje dla zarządzania kadrami.* Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Kuźmińska, A.O. (2016b). Third variable effects in management studies. *Problemy Zarządzania*, *14*(2, 60), 147–159.

Kuźmińska, A.O. (2019). *Konsekwencje posiadania władzy i częstego myślenia o pieniądzach*. Wydawnictwa Naukowe WZ UW.

Kuźmińska, A.O., Pazura, D. (2018). The impact of Control Preferences Fit Between Employees and Their Supervisors on Employee Job Satisfaction. *Studia i Materiały*, *29*(2), 18–32.

Kuźmińska, A.O., Schulze, D., Koval, A. (2019). Who Doesn't Want to Share Leadership? The Role of Personality, Control Preferences, and Political Orientation in Preferences for Shared vs. Focused Leadership in Teams. W: A. Kuźmińska (red.) Management Challenges in the Era of Globalization (s. 11–27). University of Warsaw Faculty of Management Press.

Kwasnicka, D., Ten Hoor, G.A., Knittle, K., Potthoff, S., Cross, A., Olson, J. (red.). Praktyczna psychologia zdrowia – Przełożenie badań nad zachowaniem na praktykę (tłumaczenie: Z. Kwissa-Gajewska i E. Gruszczyńska). vol. 1. http://doi.org/10.17605/OSF.IO/M72P5 (Wersja angielska opublikowana w roku 2021).

Lachiewicz, S., Nogalski, A. (red.) (2010). *Osiągnięcia i perspektywy nauk o zarządzaniu*. Oficyna Ekonomiczna Grupa Wolters Kluwer.

Larsen, J.T., Norris, C.J., Cacioppo, J.T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*, 776–785. https://doi.org/10.1111/14698986.00078

Levi, R., Ridberg, R., Akers, M., Seligman, H. (2021). Survey Fraud and the Integrity of Web-Based Survey Research. *American Journal of Health Promotion*, *36*(1), 18–20. https://doi.org/10.1177/08901171211037531

Levitt, S.D., Dubner, S.J. (2009). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow.

Liu, B., Balkwill, A., Reeves, G., Beral, V. (2010). *Body mass index and risk of liver cirrhosis in middle aged UK women: prospective study*. Bmj, 340. https://doi.org/10.1136/bmj.c912

Liu, M., Wronski, L. (2018a). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116–124. https://doi.org/10.1177/0894439317695581

Liu, M., Wronski, L. (2018b). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32–49. https://doi.org/10.1177/1470785317744856

Lloyd, J.B. (2008). Myers-Briggs theory: How true, how necessary? *Journal of Psychological Type*, *68*, 1–8.

Locke, W. (2008). The Academic Profession in England: Still stratified after all these years? W: *The Changing Academic Profession In International Comparative and Quantitative Perspectives* (s. 89–115). Research Institute for Higher Education, Hiroshima University. http://oro.open.ac.uk/11803/

Łazowska, E., Niezgoda, M., Kruszewski, M. (2014). Obciążenie poznawcze kierowców. *Transport samochodowy*, *2*, 73–82.

Mackiewicz, R. (2011). Liczby w decyzjach ekonomicznych: instynkt numeryczny. W: T. Zaleśkiewicz, A. Falkowski (red.) *Psychologia poznawcza w praktyce: Ekonomia, biznes, polityka* (s. 137–186). Wydawnictwo Naukowe PWN.

Maniaci, M.R., Rogge, R.D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Marody, M. (2015). *Jednostka po nowoczesności: perspektywa socjologiczna*. Wydawnictwo Naukowe Scholar.

McKelvey, B. (2017). Model-centered organization science epistemology. W: J. Baum (red.), *The Blackwell companion to organizations* (s. 752–780). https://doi.org/10.1002/9781405164061.ch33

McKibben, W.B., Silvia, P.J. (2017). Evaluating the Distorting Effects of Inattentive Responding and Social Desirability on Self-Report Scales in Creativity and the Arts. *The Journal of Creative Behavior*, *51*(1), 57–69. https://doi.org/10.1002/jocb.86.

Meade, A.W., Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Mehrabian, A. (1971). *Silent messages*. Wadworth Publishing Company.

Michałowicz, B. (2016). Ankiety ewaluacyjne w szkolnictwie wyższym: wpływ wyboru ewaluatorów. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Michałowska, D.A. (2013). *Neoliberalizm i jego (nie)etyczne implikacje edukacyjne.* Adam Mickiewicz University Press.

Monahan, J.L., Murphy, S.T., Zajonc, R.B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science*, *11*(6), 462–466. https://doi.org/10.1111/1467-9280.00289

Myers, B., Myers, P.B. (1995). *Gifts Differing*. CPP Inc.

Nęcka, E. (2005). O roli procesów poznawczych w wyjaśnianiu zjawisk społecznych. W: M. Kossowska, M. Śmieja-Nęcka, S. Śpiewak (red.), *Społeczne ścieżki poznania* (s. 5–11). Gdańskie Wydawnictwo Psychologiczne.

Nęcka, E. (2023). *Psychologia. Wprowadzenie*. Wydawnictwo Naukowe SCHOLAR.

Niemczyk, J. (2011). Metodologia nauk o zarządzaniu. W: W. Czakon (red.) *Podstawy metodologii badań w naukach o zarządzaniu* (s. 19–29). Oficyna Wolters Kluwer.

Nisbett, R.E. (2011). *Geografia myślenia*. Smak Słowa.

Nisbett, R.E. (2016). *MINDWARE: Narzędzia skutecznego myślenia*. Smak Słowa.

Nowak, K. (2019). *Hidden costs of job demands-employee working style misfits*. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska.

Nowak, K. (2021). Konsekwencje braku psychologicznego dopasowania pracownika do wymagań pracy. W: G. Wieczorkowska (red.), *Cztery wyzwania w zarządzaniu ludźmi* (s. 7–28). Wydawnictwa Uniwersytetu Warszawskiego.

Obarska, K., Górowska, M., Jęczalik, K., Gola, M. (2021). Uzależnienia w świecie nowych technologii oraz nowe technologie w niesieniu pomocy osobom uzależnionym. W: Ł. Gaweda (red.) *Kryzys psychiczny w nowoczesnym świecie* (*s.* 185–228). Wydawnictwo Naukowe PWN.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Oppenheim, A.N. (2004). *Kwestionariusze, wywiady, pomiary postaw*. Zysk i S-ka.

Pessoa, L. (2024). *Splątany mózg*. Wydawnictwo Naukowe PWN.

Pigliucci, M. (2019). *Bujda na resorach*. Wydawnictwo Naukowe PWN.

Pittenger, D.J. (2005). Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research*, *57*(3), 210–221. https://doi.org/10.1037/1065-9293.57.3.210

Plous, S. (1993). *The Psychology of Judgment and Decision Making*. McGraw-Hill Inc.

Poldrack, R. (2022). *Nauka czytania w myślach. Co neuroobrazowanie może powiedzieć o naszych umysłach?* Copernicus Center Press.

Poole, G. (2009). Academic disciplines: homes or barricades? W: C. Kreber (red.) The University and its Disciplines (s. 50–57). Routledge.

Pulver, C.A., Kelly, K.R. (2008). Incremental validity of the Myers-Briggs Type Indicator in predicting academic major selection of undecided university students. *Journal of Career Assessment*, *16*(4), 441–455. https://doi.org/10.1177/1069072708318902

Quenk, N.L. (2009). *Essentials of Myers-Briggs type indicator assessment*. John Wiley & Sons.

Ross, M., McFarland, C., Fletcher, G.J. (2001). Wpływ postaw na przypominanie sobie zdarzeń z własnej przeszłości. W: E. Aronson (red.) *Człowiek istota społeczna. Wybór tekstów* (s. 215–229). Wydawnictwo Naukowe PWN.

Russell, J.A., Weiss, A., Mendelsohn, G.A. (1989). Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, *57*(3), 493–502.

Salovey, P., Caruso, D., Mayer, J.D. (2004). Emotional intelligence in practice. W: P. Alex Linley, S. Joseph (red.), *Positive psychology in practice* (s. 411–565). John Wiley & Sons Inc. https://doi.org/10.1002/9780470939338.ch28

Satel, S., Lilienfeld, S.O. (2017). *Pranie mózgu. Uwodzicielska moc (bezmyślnych) neuronauk*. CiS.

Schulz, M. (2002). The role of intuition in decision making. *Journal of Applied Psychology*, *87*(4), 758–767.

Schuman, H., Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Schuster, J.H., Finkelstein, M.J. (2006). *The American faculty: The restructuring of academic work and careers*. JHU Press.

Schwarz, N., Hippler, J., Deutsch, B., Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*, 388–385.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., Simmons, A. (1991a). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology*, *61*, 195–202.

Schwartz, N., Strack, F., Hilton, D.J., Naderer, G. (1991b). Judgmental biases and the logic of conversation. *Social Cognition*, *9*(1), 67–84.

Schwarz, N., Strack, F., Mai, H.P. (1991c). Assimilation and contrast effects in part-whole question sequence: A conversational-logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23. https://doi.org/10.1086/269239

Schwarz, N., Hippler, H., Noelle-Newmann, E. (1992). A cognitive model of response-order effects in survey measurement. W: W. Schwarz, S. Sudman (red.) *Context effects in social and psychological research* (s. 187–201). Springer-Verlag.

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O.R., van Aert, R.C., Van Assen, M.A., Schulte-Mecklenbeck, M. i in. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational behavior and human decision processes*, *165*, 228–249.

Scott-Lichter, D. (2011). Got content, get attention. *Learned Publishing*, *24*(4), 245–246.

Scruton, R. (2012). *Pożytki z pesymizmu i niebezpieczeństwa fałszywej nadziei*. Zysk i s-ka.

Shaffer, F., Ginsberg, J.P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in public health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Shank, D.R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition*, *13*(2), 158–167. https://doi.org/10.3758/BF03197008

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta psychologica*, *68*(1–3), 203–215. https://doi.org/10.1016/0001-6918(88)90056-X

Shaughnessy, J.J., Zechmeister, J.S., Zechmeister, E.B. (2007). *Metody badawcze w psychologii*. Gdańskie Wydawnictwo Psychologiczne.

Schuster, J.H., Finkelstein, M.J., Galaz Fontes, J.F., Liu, M. (2008). *The American faculty: the restructuring of academic work and careers*. Johns Hopkins University Press.

Siarkiewicz, M. (2007). Zastosowanie metody analizy szeregów czasowych do badania zależności pomiędzy składem posiłków a poziomem subiektywnie odczuwanej energii. W: A. Nowak, K. Winkowska-Nowak, A. Rychwalska (red). *Modelowanie matematyczne i symulacje komputerowe w naukach społecznych* (s. 32–44). Wydawnictwo Academica Szkoły Wyższej Psychologii Społecznej.

Sipps, G.J., Alexander, R.A., Friedt, L. (1985). Item analysis of the Myers-Briggs Type Indicator. *Educational and Psychological Measurement*, *45*, 789–796.

Siuta, J. (2006). *Inwentarz Osobowości NEO-PI-R*. Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Skład, M., Wieczorkowska, G. (2001). Sztuka układania ankiet ewaluacyjnych. W: J. Grzelak, M. Lewicka (red.), *Jednostka a społeczeństwo* (s. 250–266). GWP.

Slaughter, S., Rhoades, G. (2004). *Academic capitalism and the new economy: markets, state, and higher education*. Johns Hopkins University Press.

Slaughter, S., Rhoades, G. (2010). The social construction of copyright ethics and values. *Science and engineering ethics*, *16*, 263–293.

Smith, T.W. (1995). Little things matter: A sampler of how differences in questionnaire format can affect survey responses. W: *Proceedings of the American Statistical Association, Survey Research Methods Section*, vol. 1046 (s. 51). American Statistical Association.

Spitzer, M. (2007). *Jak uczy się mózg*. Wydawnictwo Naukowe PWN.

Steedle, J.T., Hong, M., Cheng, Y. (2019). The Effects of Inattentive Responding on Construct Validity Evidence When Measuring Social–Emotional Learning Competencies. *Educational Measurement: Issues and Practice*, *38*(2), 101–111. https://doi.org/https://doi.org/10.1111/emip.12256

Stevens, S.S. (1958). Problems and methods of psychophysics. *Psychological Bulletin*, *55*(4), 177–188. https://doi.org/10.1037/h0044182.

Strack, F., Martin, 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. https://doi.org/10.1002/ejsp.2420180505

Sułek, A. (2001). *Sondaż polski: przygarść rozpraw o badaniach ankietowych*. Wydawnictwo IFiS PAN.

Sułek, A. (2002). *Ogród metodologii socjologicznej*. Wydawnictwo Naukowe Scholar.

Sułkowski, Ł. (2011). Rozwój metodologii w naukach o zarządzaniu. W: W. Czakon (red.), *Podstawy metodologii badań w naukach o zarządzaniu* (s. 30–45). Oficyna Wolters Kluwer.

Szarota, P. (2008). Wielka Piątka-stare problemy, nowe wątpliwości. *Roczniki Psychologiczne*, *11*(1), 127–138.

Śpiewak, S. (2013). *Rozgrzewanie uwagi-wyczerpywanie woli-uległość: Mechanizmy adaptacji umysłu do wysiłku poznawczego*. Wydawnictwo Naukowe Scholar.

Taleb, N. (2012). *Antykruchość: Jak żyć w świecie nieprzewidywalnym*. Helion.

Tatarkiewicz, W. (1951/2014). *Historia filozofii. Filozofia XIX wieku i współczesna*, 3. Wydawnictwo Naukowe PWN.

Trow, M. (2007). Reflections on the transition from elite to mass to universal access: Forms and phases of higher education in modern societies since WWII. W: J. Forest, P. Altbach (red.), *International handbook of higher education* (s. 243–280). Springer Netherlands.

Tsai, Y.F., Viirre, E., Strychacz, C., Chase, B., Jung, T.P. (2007. Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, *78*(5), B176–B185.

Turska, E. (2016). Is Double-Degree Goal Equally Good for All Students? Moderating Impact of Interval Activity Style. *Problemy Zarządzania*, *14*(2, 60), 134–146.

Twenge, J.M. (2024). *Pokolenia*. Smak Słowa.

Tyszka, T. (2010). Decyzje: perspektywa psychologiczna i ekonomiczna. Wydawnictwo Naukowe Scholar.

Unkelbach, C., Memmert, D. (2008). Game management, context effects, and calibration: The case of yellow cards in soccer. *Journal of Sport and Exercise Psychology*, *30*(1), 95–109.

Unkelbach, C., Ostheimer, V., Fasold, F., Memmert, D. (2012). A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes*, *119*(1), 103–113. https://doi.org/10.1016/j.obhdp.2012.06.004

Van Bavel, J.J., Mende-Siedlecki, P., Brady, W.J., Reinero, D.A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459.

Wänke, M., Bless, H., Schwarz, N. (1998). Context effects in product line extensions: Context is not destiny. *Journal of Consumer Psychology*, *7*(4), 299–322.

Wänke, M., Schwarz, N., Noelle-Neumann, E. (1995). Asking comparative questions: The impact of the direction of comparison. *Public Opinion Quarterly*, *59*, 347–372.

Wieczorkowska, G., Eliasz, A. (2004). Rola przedziałowości i temperamentu w adaptacji do zmiany. *Kolokwia psychologiczne*, *10*, 29–51.

Wieczorkowska, G., Gonzalez, R. (2023). *Wprowadzenie do dyskusji panelowej „What next? A necessary shift in empirical social science"*. 18.06.2023 – Wydział Zarządzania UW.

Wieczorkowska, G., Grzelak, J. (1999). Nie wiem, nie myślę, jestem z miasta. Wybrane problemy metodologiczne analizy badań sondażowych. W: B. Wojciszke i M. Jarymowicz (red.), *Psychologia rozumienia zjawisk społecznych* (s. 261–289). Wydawnictwo Naukowe PWN.

Wieczorkowska, G., Karczewski, W. (2019). Employees' predispositions to routinised work: measurement issues. W: A. Kuźmińska (red.), *Management Challenges in the Era of Globalization* (s. 94–106). Wydawnictwo Naukowe Wydziału Zarządzania UW.

Wieczorkowska, G., Kowalczyk, K. (2021). Ensuring Sustainable Evaluation: How to Improve Quality of Evaluating Grant Proposals? *Sustainability*, *13*(5), 28–42. https://doi.org/10.3390/su13052842

Wieczorkowska, G., Król, G., Wierzbiński, J. (2016). Przeszłość, teraźniejszość i przyszłość edukacji akademickiej. *Nauka*, *3*, 87–106.

Wieczorkowska-Wierzbińska, G., Nowak, K., Michałowicz, B., Wilczyńska, K., Król, G. (2018). Analiza związków wyniku okulograficznego pomiaru stopnia koncentracji pracownika z wynikami kwestionariuszowych pomiarów stylu pracy i temperamentu pracownika i innymi pomiarami psychofizjologicznymi. Raport badawczy wykonany w ramach projektu realizowanego przez Topas Sp. z o.o. pt. **„Prace rozwojowe w celu stworzenia innowacyjnej usługi wspomagającej zarządzanie kompetencjami poznawczymi w przedsiębiorstwach"** w ramach Regionalnego Programu Operacyjnego Województwa Mazowieckiego na lata 2014–2020 (RPO WM 2014–2020), Działanie 1.2. Działalność badawczo-rozwojowa przedsiębiorstw. – typ projektu: Bony na innowacje.

Wieczorkowska, G., Wierzbiński, J. (2013). *Statystyka: od teorii do praktyki*. Wydawnictwo Naukowe Scholar.

Wieczorkowska, G., Wierzbiński, J., Król, G. (2015). W: M. Kostera (red.) *Metody badawcze w zarządzaniu humanistycznym* (s. 169–180). Wydawnictwo Akademickie SEDNO.

Wieczorkowska, G., Wierzbiński, J., Siarkiewicz, M. (2009). Wybrane problemy metodologiczne analitycznych badań sondażowych. W: P. Radkiewicz, R. Siemieńska-Żochowska (red.), *Społeczeństwo w czasach zmiany: badania Polskiego Generalnego Sondażu Społecznego 1992–2009* (s. 443–465). Wydawnictwo Naukowe Scholar.

Wieczorkowska, G., Wilczyńska, K. (2023). Generacje internetowe na rynku pracy. *Studia i Materiały*, *2*, 27–40.

Wieczorkowska-Nejtardt, G. (1998*). Inteligencja motywacyjna: mądre strategie wyboru celu i sposobu działania*. Instytut Studiów Społecznych.

Wieczorkowska-Wierzbińska, G. (2011). *Psychologiczne ograniczenia*. Wydawnictwo Naukowe Wydziału Zarządzania UW.

Wieczorkowska-Wierzbińska, G. (2025). *Rafy zarządzania*. *Ludzie różnią się od śrubokrętów*. Wydawnictwo Naukowe Wydziału Zarządzania UW.

Wieczorkowska, G. (1993). Pułapki statystyczne. W: Z. Smoleńska (red.), *Badania nad rozwojem w okresie dorastania* (s. 211–234). Wydawnictwo: Instytut Psychologii PAN.

Wieczorkowska-Wierzbińska, G., Wierzbiński, J., Kuźmińska, A.O. (2014). Porównywalność danych sondażowych zebranych w różnych krajach. *Psychologia Społeczna*, *2*(29), 128–143.

Wieczorkowska-Wierzbińska, G. (2018). *The future of managerial education*. Economic and Social Development: Book of Proceedings, 48–54.

Wieczorkowska-Wierzbińska, G., Wilczyńska, K. (2019). *The future of managerial research*. Economic and Social Development: Book of Proceedings, 369–376.

Wierzbiński, J. (2009). *Badanie zaufania do organizacji: problemy metodologiczne*. Wydawnictwo Naukowe Wydziału Zarządzania UW.

Wilczyńska, K. (2023). *Trafność stereotypów generacyjnych – rekomendacje dla HRM. Nieopublikowana rozprawa doktorska*. Wydział Zarządzania UW. Promotor: G. Wieczorkowska. Promotor pomocniczy: K. Nowak.

Woźniak, J. (2013). *Rekrutacja: Teoria i praktyka*. Wydawnictwo Naukowe PWN.

Yousefpour, A. (2021). *Generational Differences in Human and Work Values in Iran and Poland*. Nieopublikowana rozprawa doktorska. Wydział Zarządzania UW. Promotor: G. Wieczorkowska i A. Khorakian.

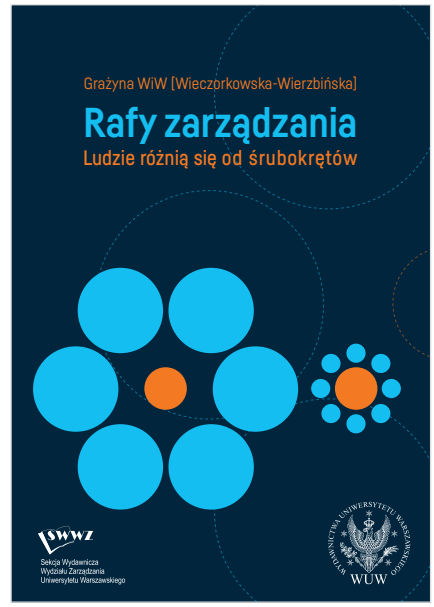
Zimbardo, P.G., Gerrig, R.J. (2012). *Psychologia i życie*. Wydawnictwo Naukowe PWN.

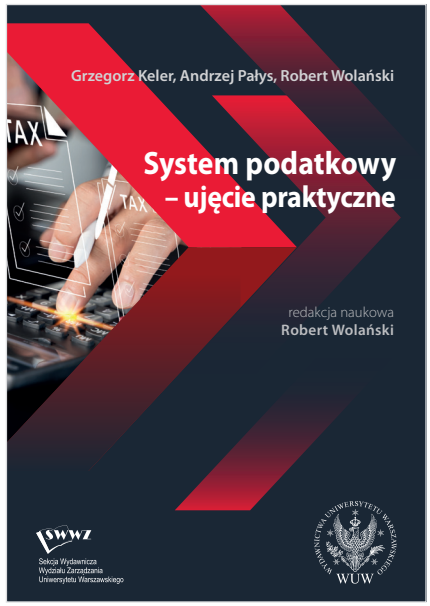
Zimbardo, P., Coulombe, N. (2015). *Gdzie Ci mężczyźni*. Wydawnictwo Naukowe PWN.

Zimbardo, P.G., Haney, C. Banks, W.C., Jaffe, D. (1971). The Stanford prison experiment. https://web.stanford.edu/dept/spec_coll/uarch/exhibits/Narration.pdf

Legenda:

* Pozycje oznaczone gwiazdką były cytowane w przypisach za pomocą pierwszego członu nazwiska.





Wydawnictwa Uniwersytetu Warszawskiego
ul. Smyczkowa 5/7, 02-678 Warszawa
tel. 22 55 31 333
www.wuw.pl

SWWZ
Sekcja Wydawnicza
Wydziału Zarządzania
Uniwersytetu Warszawskiego

Grażyna WiW [Wieczorkowska-Wierzbińska] jest magistrem matematyki i profesorem belwederskim psychologii od 25 lat. Od kilkunastu lat jej praca koncentruje się na wykorzystywaniu wiedzy akademickiej w praktyce. W wolnych chwilach prowadzi sesje coachingowo-mentoringowe. Jej ostatnią pasją zawodową jest diagnoza psychologiczna wykorzystująca metodę ilościowych, eksperymentalnych studiów przypadku.

Była założycielem i pierwszym dyrektorem COME UW (Centrum Otwartej i Multimedialnej Edukacji 1997–2005), kierowała Instytutem Studiów Społecznych im. prof. Roberta Zajonca. Przez kilkanaście lat współpracowała z Altkom Akademia, prowadziła wykłady i program International Business Management w SWPS, była kierownikiem pięcioletniego projektu *Zarządzanie wielokulturowe w dobie globalizacji* finansowanego z funduszy strukturalnych UE. Kierowała pięcioma dużymi programami badawczymi, w tym projektami realizowanymi wspólnie z University of Michigan. W latach 2021–2022 prowadziła badania porównawcze zaangażowania w polskich zakładach pracy. Do września 2024 r. była kierownikiem studiów doktoranckich na Wydziale Zarządzania UW. Wypromowała 6 doktorów psychologii i 15 doktorów nauk o zarządzaniu, jest promotorem 22. rozprawy doktorskiej dotyczącej badań zaangażowania pracowników.

Jest autorką kilkunastu książek (m.in. *Skalowanie wielowymiarowe jako metoda badania percepcji; Punktowe i przedziałowe reprezentacje celu; Inteligencja motywacyjna: mądre strategie wyboru celu i sposobu działania; Kontrola naszych myśli i uczuć* – z Elliotem Aronsonem; *Statystyka: od teorii do praktyki* – z Jerzym Wierzbińskim; *Kierowanie motywacją: rola myśli, emocji i zachowania; Psychologiczne ograniczenia*). Publikowała także artykuły naukowe we współpracy z wybitnymi psychologami amerykańskimi (m.in. z Robertem Zajoncem, Davidem Hamiltonem czy Eugene Burnsteinem) w takich renomowanych czasopismach, jak „Psychological Science", „Journal of Personality and Social Psychology", „Personality and Social Psychology Bulletin". Pełniła funkcję Associate Editor „European Journal of Management".

Autorka *zamieszcza w pracy opisy najczęściej stosowanych psychologicznych narzędzi diagnostycznych [...] podkreśla wielość narzędzi o niesprawdzonej wartości [...] i wpływ błędów poznawczych na ocenę pracowników i procesy rekrutacyjne; [...] proponuje autorski paradygmat WiW jako elastyczne, pragmatyczne podejście do metod badawczych, dostosowane do specyfiki badanego obiektu.*

prof. dr hab. Irena Heszen-Celińska

„Rafy" w tytule oznaczają zagrożenia czy niebezpieczeństwa, na które może natknąć się podróżnik przemieszczający się po rozległym obszarze współczesnych nauk o zarządzaniu [...]. Co chroni przed wpadnięciem na rafy? W świecie żeglugi morskiej – sonary. Czy takie „sonary" możemy wykorzystać, żeglując po rozległych obszarach współczesnych nauk o zarządzaniu? Recenzowana książka stanowi odpowiednik sonaru, ostrzegający przed zbliżaniem się do ryzykownej identyfikacji czy interpretacji problemu. Autorka [...] podejmuje rolę eksperta identyfikującego aktualne i ***przyszłe problemy nauk o zarządzaniu*** *z podziwu godną intuicją i przenikliwością.*

prof. dr hab. Wiktor Adamus

ISBN 978-83-235-7063-9