

Cztery metodologiczne zagrożenia w naukach o zarządzaniu¹

Grażyna Wieczorkowska*, Grzegorz Król**, Jerzy Wierziński***

W tekście omówiono cztery metodologiczne zagrożenia pojawiające się w badaniach ilościowych z zakresu nauk o zarządzaniu: (1) Zbyt mała liczba eksperymentów; (2) Słabość modeli teoretycznych; (3) Słabość pomiaru; (4) Słabość analiz. Większość tych zagrożeń dotyczy także badań psychologicznych i socjologicznych. W podsumowanie znajdują się odpowiedzi na potencjalnie nieuzasadnione zarzuty recenzentów dotyczące próby i braku hipotez.

Słowa kluczowe: metody ilościowe, zagrożenia metodologiczne.

Nadesłany: 20.11.2016 | Zaakceptowany do druku: 28.12.2016

Four Methodological Threats in Management Science

The text discusses four threats in the analysis of data in management science: (1) Too small number of experimental studies; (2) Weakness of theoretical models; (3) Weakness of measurement; (4) Weakness of analysis. Most of these pitfalls apply also to psychological and sociological research. The summary covers answers to reviewers' potentially unjustified objections concerning samples and lack of hypotheses.

Keywords: quantitative methods, methodological threats.

Submitted: 20.11.2016 | Accepted: 28.12.2016

JEL: C18

* **Grażyna Wieczorkowska** – prof. zw. dr hab., Katedra Psychologii i Socjologii Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski.

Adres do korespondencji: Katedra Psychologii i Socjologii Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski, ul. Szturmowa 1/3, 02-678 Warszawa; e-mail: gw@uw.edu.pl.

** **Grzegorz Król** – dr, Katedra Psychologii i Socjologii Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski.

Adres do korespondencji: Katedra Psychologii i Socjologii Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski, ul. Szturmowa 1/3, 02-678 Warszawa; e-mail: grzegorz.krol@uw.edu.pl.

*** **Jerzy Wierziński** – prof. dr hab., Samodzielny Zakład Metod Matematycznych i Statystycznych Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski.

Adres do korespondencji: Samodzielny Zakład Metod Matematycznych i Statystycznych Zarządzania, Wydział Zarządzania, Uniwersytet Warszawski, ul. Szturmowa 1/3, 02-678 Warszawa; e-mail: wierzinski@wz.uw.edu.pl.

1. Wstęp

Od lat spędzamy wiele godzin tygodniowo, analizując dane z badań, nadzorując analizy wykonywane przez studentów, doktorantów, czy pisząc recenzje prac doktorskich, habilitacyjnych i artykułów naukowych z psychologii, socjologii i zarządzania. Kiedyś analizy statystyczne wykonywali specjaliści; rozwój pakietów statystycznych i łatwość ich obsługi powoduje, że teraz nawet skomplikowane analizy można wyklikać w ciągu paru minut. Łatwość wykonywania zaawansowanych analiz nie idzie jednak w parze z odpowiednim poziomem refleksji. Dlatego postanowiliśmy zwrócić uwagę na cztery często pojawiające się słabości analiz, które spotykamy w recenzowanych przez nas pracach.

2. Typ badań: za mało eksperymentowania

W naukach o zarządzaniu najczęstszym typem badań ilościowych są badania **korelacyjne**. Przykładowo, jeśli chcemy zbadać związek **satisfakcji pracowników** (Y – zmienna zależna/objaśniana) z **częstością** uzyskiwania przez nich **pochwał** od bezpośredniego przełożonego (X – zmienna niezależna/objaśniająca), to w badaniach korelacyjnych mierzymy zarówno X (pochwały), jak i Y (satisfakcja). Badacz jednak nie ma wpływu na wystąpienie zmiennej niezależnej (pochwała), co oznacza, że nie jest w stanie wykluczyć wpływu „trzecich” zmiennych (np. kto pochwalił, kiedy, za co, w jakich warunkach). Z tego względu immanentną cechą badań korelacyjnych jest ich **niska trafność wewnętrzna**. W badaniach eksperymentalnych **manipulujemy** pochwałą (wprowadzamy ją w ściśle określonych warunkach, w odmienny sposób w różnych grupach) – dlatego **trafność wewnętrzna** badania może być bardzo **wysoka**. W języku potocznym pojęcie manipulacji ma negatywną konotację, ponieważ wiąże się z ukrytym sposobem wywierania na nas wpływu – wbrew naszej woli i/lub interesowi. W metodologii **możliwość manipulowania wartościami zmiennej niezależnej** jest bardzo pozytywną cechą badania. Jeśli potrafimy manipulować daną zmienną, nie musimy czekać, aż dana wartość zmiennej wystąpi w sposób naturalny, ale wprowadzamy ją sami i sprawdzamy, jakie są tego skutki. W badaniach mani-

pulujemy **zmiennymi**, a nie ludźmi, choć efekty tego z definicji powinny wpływać na zachowanie osób uczestniczących w badaniach. Reakcja na pochwałę może zależeć od wielu zmiennych – samooceny, poziomu aspiracji, neurotyzmu, doświadczeń przeszłości itd. Ludzie różnią się między sobą na wielu wymiarach mogących mieć wpływ na naszą zmienną zależną (satisfakcja), niezależnie od wartości zmiennej niezależnej (pochwała), ale jeśli **losowo** podzielimy badaną próbę na 2 grupy, to średni poziom samooceny, neurotyzmu itd. 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 nielosową 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 (Aronson i Wieszorkowska, 2001).

Przyrost naszej wiedzy budują przede wszystkim wyniki badań eksperymentalnych, a nie korelacyjnych. Ciągłe jednak ten fakt nie może się przebić w naukach o zarządzaniu, choć rozwój ekonomii behawioralnej i **Nagroda Nobla w ekonomii w 2002 roku dla Vernona Smitha** za ustanowienie eksperymentów laboratoryjnych narzędziem empirycznej analizy oraz Daniela **Kahnemana** za zastosowanie narzędzi psychologicznych w badaniach ekonomicznych, ze szczególnym uwzględnieniem teorii perspektywy, już dawno powinny zmienić nastawienie badaczy. Krytycy badań eksperymentalnych mówią, że gromadzimy wiedzę o studentach, bo to oni najczęściej biorą udział w naszych badaniach. O ile badanie ankietowe można przeprowadzić na wielu tysiącach osób, eksperymenty prowadzi się na małych grupach, bo ich przygotowanie wymaga wielu godzin pracy i większej inwestycji czasowej ze strony badanych. Pojawia się więc pytanie, **na ile można uogólnić wyniki uzyskane** np. w badaniu 30 studentów – innymi słowy, jaka jest **trafność zewnętrzna** takiego badania. Przecież nie interesują nas

jedynie reakcje uczestników, chcemy orzekać o populacji, z której ta próba została wybrana. Jeśli interesują nas preferencje wyborcze Polaków, to wylosowana do badań próba powinna być reprezentatywna dla dorosłych Polaków. Próby reprezentatywne są niesłychanie ważne, jeżeli 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 dużo mniej przydatne, gdy chcemy testować zależności między zmiennymi. Skuteczna statystyczna kontrola ważnych zmiennych (choć nieistotnych ze względu na cel 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, ale **stwierdzenie czy różnica w poziomie satysfakcji między badanymi grupami jest istotna statystycznie**. Do tego nie potrzebujemy próby reprezentatywnej. Od badania prób reprezentatywnych dużo lepszą strategią maksymalizacji trafności zewnętrznej jest replikowanie badań na próbach homogenicznych – osobno badamy rolników, osobno pracowników naukowych, osobno studentów itd. Wyniki badań trzydziestu osób losowo podzielonych na warunki eksperymentalne mają większą wartość poznawczą niż trzech tysięcy uczestniczących w badaniach korelacyjnych. Randomizacja drugiego stopnia (losowy przydział do grup eksperymentalnych) zapewnia nam wysoką trafność wewnętrzną, jakiej nigdy nie osiągniemy w badaniach korelacyjnych. Problemem jest to, że ludzie niechętnie biorą udział w badaniach eksperymentalnych, co poważnie 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 (Aronson i Wieszorkowska, 2001). O **realizmie sytuacyjnym** mówimy, jeśli sytuacja w laboratorium jest podobna do zdarzeń, które często spotykają ludzi w świecie zewnętrznym. Jednak o wiele ważniejsze jest to, czy stworzona przez badacza sytuacja angażuje badanego, skłania go do traktowania

poważnie tego, co się dzieje. Tę własność sytuacji określa się jako **realizm psychologiczny**. Dobre sytuacje eksperymentalne, gdy np. uczestnicy grają w angażującą grę, powodują, że po pewnym czasie sytuacja staje się bardzo realna. Osiągnięte wyniki, pochwały, które uzyskują, mają realną wartość. Badani są o wiele bardziej zaangażowani niż przy wypełnianiu ankiet. Niechęć wobec eksperymentów budzi fakt, że często jego **prawdziwy cel musi pozostać ukryty na czas trwania badania**. Gdyby badani wiedzący, że interesuje nas np. ich uległość wobec autorytetu, byli poproszeni o opisanie, jak zachowaliby się pod naciskiem autorytetu, dowiedzielibyśmy się, jacy chcieliby być, a nie jacy byłiby faktycznie. W jednym z badań (Zimbardo i Gerrig, 2012) przeprowadzonych w warunkach naturalnych pielęgniarki otrzymywały od nieznanego im lekarza, prowadzącego pacjenta przebywającego w szpitalu na ich oddziale, telefoniczne polecenie: „Przyjadę za godzinę, ale proszę już teraz podać pacjentowi 20 mg estrogenu”. Na opakowaniu zawarta była informacja, że maksymalna dawka wynosi 10 mg, typowa zaś 5 mg. Polecenie dotyczyło 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 ma na nas sytuacja, więc nasze wyobrażenia na temat naszych zachowań nie mają dużej wartości badawczej.

3. Słabość modeli teoretycznych

Słabością nauk społecznych jest **brak systematyzacji pojęć**. Nawet publikując w fachowych czasopismach, unikamy często używania „żargonu” naukowego, co było przez lata wzmacniane obowiązkiem publikacji prac habilitacyjnych w formie książki dostępnej dla szerszej klasy czytelników. Prawdą jest, że posługujemy się **pojęciami naturalnymi**, więc ich definiowanie w sposób klasyczny nie jest możliwe i na szczęście prace przeglądowe różnych definicji zniknęły z pola naszych zainteresowań, choć można spotkać przeglądy zawierające kilkadziesiąt definicji danego pojęcia, np. zaufania (Grudzewski i Hejduk, 2004).

Porównania wniosków muszą być jednak prowadzone na poziomie **operacjonalizacji**, których niestety wystandaryzować się nie da. Manipulacja zagrożeniem musi być inna, gdy jest stosowana w grupie studentów, niż w grupie pracowników firmy. Także rozkłady wyników uzyskiwane za pomocą standardowych technik pomiaru zależą od próby. Przykładem może być pozycja z kwestionariusza przedziałowości „Jestem nonszalancki w traktowaniu szczegółów”, która zmieniła diametralnie swoją korelację ze skalą w grupie pracowników Wydziału Farmacji. Nonszalanca w traktowaniu szczegółów dla farmaceuty ma zupełnie inne znaczenie niż dla studenta.

Tworząc modele teoretyczne, ciągle zapominamy, że nauki społeczne są wysokokontekstowe. Zależności uzyskiwane w modelu *ceteris paribus* są bardzo podatne na efekt „trzeciej zmiennej”. Dlatego powinniśmy zdecydowanie zmniejszyć poziom ogólności naszych teorii. Zazdrościmy fizykom i chcielibyśmy opracować tak proste reguły jak trzy zasady dynamiki Newtona, ignorując w ten sposób fakt, że nasz umysł rozwija się w odpowiedzi na pojawiające się w otoczeniu wyzwania – jest więc raczej zbiorem modułów wyuczonych reakcji na różne klasy(!) bodźców. Poważnym problemem jest też rozdział między badaniami naukowymi stosującymi głównie podejście **nomotetyczne** i praktyką, która wymaga wiedzy **idiograficznej**. Duże przyrosty naszej wiedzy nie mają jak dotąd przełożenia na zwiększenie skuteczności oddziaływań praktycznych. Dużo więcej wiemy, ale nadal niewiele więcej potrafimy. I tu znów czekamy na rozwój technologii, która umożliwi wielokrotny pomiar tej samej osoby (**ilościowe porównawcze studia przypadków** – Siarkiewicz, 2007) bez wywoływania jej irytacji.

W sporze jednoteoretycznych z eklektykami uważamy, że najważniejsze jest otrzymanie interesujących poznawczo wyników, więc nie mamy nic przeciwko łączeniu różnych szkół teoretycznych. Czterdzieści lat temu mówienie o procesach nieświadomych było totalnie odrzucane przez naukę zdominowaną przez podejście poznawcze, dziś nikt nie ma wątpliwości, że przetwarzanie informacji odbywa się w przeważającej części poza świadomością.

Za dużo mamy w naukach społecznych twórczości (ciągłego wprowadzania nowych pojęć), za mało rzemiosła. Nie jest

tajemnicą, że najłatwiej osiągnąć „sukces”, wprowadzając nowe pojęcie i publikując narzędzie pomiaru. Jeżeli mamy szczęście, wiele osób zacznie korelować nowe narzędzie z istniejącymi, aż fala zainteresowania opadnie, bo wynik tych usiłowań można łatwo przewidzieć. Na wykładach z metodologii powtarzam do znudzenia: replikować, replikować i jeszcze raz replikować, bo prawie żadna udana replikacja w naukach społecznych nie jest banalna (Aronson i Wieszorkowska, 2001). Studenci najpierw powinni nauczyć się uzyskiwać przewidywalne na podstawie innych badań wyniki, a dopiero potem samodzielnie wychodzić dalej. W ostatnich latach mamy do czynienia z kryzysem replikowalności (*a crisis in science in general with the question of replicability*). Pokazuje on, z jak wielozmiennością mamy do czynienia materią (por. np. Van Bavel i in., 2016). Takie replikacyjne porażki powinny być impulsem do poszukiwania warunków brzegowych dla opisanych wcześniej zależności. Informacje na temat: „Kiedy to nie działa” mogą budować naszą wiedzę.

4. Słabość pomiaru

Nie możemy zapominać, że **obiekt naszych badań** jest dużo bardziej **skomplikowany** i **reaktywny** niż np. w naukach biologicznych czy fizycznych, na których wzorujemy naszą metodologię. Taleb (2012) porównuje nauki społeczne, usiłujące stosować metodologię wykorzystywaną przez fizyków do krów, które usiłują latać. Aby zilustrować to studentom, na wykładzie z medycyny psychosomatycznej przeprowadziłam (GW) „**trening relaksacyjny**”. Prawie 300 osób siedzących na sali proszono o podążanie za instrukcjami podawanymi w tle relaksującej muzyki. Nie tylko obserwowałam reakcję, 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 moim oczekiwaniem – 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).

Oznacza to, że modele nieuwzględniające różnic indywidualnych i kontekstu sytuacyjnego nie mogą wyjaśniać wiele wariacji analizowanych zmiennych. Koncentrując się w naszych analizach statystycznych na miarach tendencji centralnej, nie dostrzegamy często efektów kontrastu.

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 obdarzone, niestety, pamięcią.

Zadawanie tego samego pytania wielokrotnie jest źródłem 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 wypicia filiżanki kawy na wynik testu, nie będziemy wiedzieć, co zdarzy się kilka godzin później, kiedy osoby badane opuszczają już laboratorium. Jeżeli nawet warunki środowiskowe pozostaną bez zmian, samo powtórzenie działania bodźca może zmienić reakcję badanego. Zmiany tego typu są opisywane w kontekście trzech efektów: ekspozycji, habituacji i przesytu. **Efekt ekspozycji** polega na tym, że obiekty początkowo neutralne są lepiej oceniane po większej liczbie ekspozycji. Stąd bierze się dążenie polityków do częstego występowania w mediach. Zapominają jednak przy tym, że jeżeli początkowa reakcja była negatywna, kolejne ekspozycje jedynie wzmocnią siłę negatywnej reakcji. Jeżeli zapytamy drugi czy trzeci raz, co respondent sądzi o tatuażach, ryzykujemy, iż nawet jeśli początkowa ocena była neutralna, możemy otrzymywać coraz bardziej pozytywne oceny przy kolejnych pomiarach. Jeżeli eksperymentalnie zbadamy wpływ pochwał od przełożonych na motywację do pracy, nie uda nam się uchwycić wpływu drugiej, trzeciej czy piętnastej pochwały. Psychologia pokazuje, że każda następna pochwała ma niższą wartość nagradzającą – przy zachowaniu innych stałych warunków. **Emocje pozytywne zwykle błędą przy częstym kontakcie z obiektem, który je wywołuje** – szybko przyzwyczajamy się do dobrego. To zjawisko nazywane jest **habituacją hedonistyczną**. Nawet najcenniejsza pochwała traci swoją wartość, kiedy słyszymy ją zbyt

często. W pewnym momencie następuje **nasycenie**, a następnie **przesyt**. To, co na początku nas fascynowało, zaczyna irytować. Chociaż wiadomo, że wiele interesujących zjawisk, jak motywacja czy szczęście, ma **dynamikę podobną do fal**, nie jesteśmy w stanie mierzyć tej dynamiki, o ile nie potrafimy utrzymywać naszych badanych w nieświadomości procesu pomiaru. Na szczęście, rozwój technologii przynosi powoli zmiany w tym obszarze i umożliwia nowe sposoby pomiaru.

W naukach fizycznych czy ekonomicznych przedmiotem pomiaru są **byty realne** (rzeczywiste wielkości, np. długość położonych linii światłowodu). Jest ustalony wzorzec, np. metra, i możemy obliczyć, jak dobrze nasz pomiar odpowiada wzorcowi. **W naukach społecznych** przedmiotem pomiaru są **byty wirtualne** (konstrukty teoretyczne, np. reprezentacje poznawcze, postawy, cechy, stany afektywne). Istnieje podstawowa różnica między pomiarem wielkości fizycznych a pomiarem konstruktywów teoretycznych. Modele pomiarowe wypracowane w naukach fizycznych nie przystają w pełni do pomiaru w naukach społecznych. Nie mamy wzorca postawy, inteligencji..., z którym moglibyśmy wyniki naszego pomiaru porównać. Ekonomista jest zainteresowany ilością pieniędzy dostępnych na rynku – więc będzie starał się mierzyć obiektywną wielkość zarobków. W badaniach społecznych, nawet jeśli prosimy o podanie wielkości zarobków w złotych, to tak naprawdę interesujemy się reprezentacją poznawczą zarobków w umyśle respondenta. Dlatego nie traktujemy tak poważnie, jak w naukach fizycznych, oryginalnej jednostki pomiaru, dokonując często przekształceń powodujących jej zmianę, np. logarytmizowanie zarobków na skali ilorazowej dla zmniejszenia skośności rozkładu. **Rygoryści metodologiczni** twierdzą też, że **skale szacunkowe** (ang. *rating scales*), np. w pytaniu o poczucie szczęścia, nie są skalami ilościowymi, ponieważ nie mają stałej jednostki pomiaru: np. odległość odpowiedzi 3 = raczej szczęśliwy od 4 = szczęśliwy i od 2 = raczej nieszczęśliwy nie jest taka sama, choć wynosi 1 punkt na skali. Ci sami badacze zapominają jednak o wymaganiach stałości jednostki pomiaru, licząc średnią z ocen szkolnych. Nie ma żadnego dowodu na to, że różnice między oceną dobrą i bardzo dobrą oraz dostateczną są takie same. **Niedoskonałe narzędzia, jakimi są**

skale szacunkowe, sprawiają nam kłopot, ale są lepsze niż brak narzędzi, tak jak kiepska maszyna do szycia jest lepsza niż szycie garnituru ręcznie (Gilbert, 2007). Jeśli zdajemy sobie sprawę z niedoskonałości naszego pomiaru, to możemy dokonywać odpowiednich korekt. Wyniki analiz mogą być zniekształcone, jeśli z powodu braku spełnienia założeń nie możemy być pewni, że wyliczona przez nas statystyka ma wymagany rozkład. Ważniejsze jest, aby we wszystkich porównywanych badaniach zniekształcenie związane np. z typem skali było takie samo. Jeśli wszyscy stosują testy parametryczne, to użycie przez nas testu nieparametrycznego – choć zgodne z zasadami – uniemożliwi porównanie wyników uzyskiwanych w różnych laboratoriach. A już kuriozalne wydaje nam się używanie nieparametrycznych testów istotności różnicy średnich, w sytuacji, gdy te same dane chwilę później wykorzystujemy do modelowania strukturalnego.

Dobrze znane powiedzenie Gordona Allporta: „**Jeżeli chcemy wiedzieć, jak ludzie się czują – dlaczego ich o to nie zapytamy?**” zapoczątkowało wysyp samoopisowych technik pomiaru najróżniejszych cech. Tworząc różne narzędzia bazujące na samoopisie, zapominamy, że w seriach badań pokazano (por. przykłady w Aronson i Wierczkowska, 2001), że często jesteśmy nieświadomi wpływów, jakim podlegamy, co jednak nie powstrzymuje nas od 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 na ten temat nie myśleli, więc ich odpowiedzi powstają *online*. Jeden z moich znajomych ma wpisany w dowodzie kolor oczu: piwny. Gdy wyraziłam (GW) 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) i 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, niestety, zadaniem sztucznym i zupełnie niepotrzebnym. W efekcie **nie jeste-**

śmy świadomi mechanizmów obronnych, jakie zniekształcają nasz obraz Ja. Innych spostrzegamy dużo trafniej (por. Wierczkowska-Wierzbńska, 2011). Pożytek z analizy samoopisów zależy od wielu czynników. Odpowiedzi mogą być relatywnie bardziej wiarygodne, jeśli pytania odwołują się do **postaw jawnych**², a diagnozowana osoba nie jest zainteresowana robieniem dobrego wrażenia, jest zmotywowana do uzyskania wglądu w swoje predyspozycje i ma skłonność do samoobserwacji i autorefleksji. Osoby poddające się diagnozie – dokonując często kreacji swojego wizerunku – prezentują się w taki sposób, aby chronić samoocenę i spełnić spostrzegane oczekiwania pracodawcy. Jeden z moich znajomych aplikujących o 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 ma duży **błąd pomiaru** – dlatego w fizyce, chemii czy biologii dokonuje się wielu powtórzeń pomiaru. Niestety, jest to zabieg możliwy w badaniu bytów nieożywionych – a dokładniej – dowolnych pobawionych pamięci i wolnej woli. Ludzie pamiętają, że przed chwilą byli o coś pytani i powtarzanie pytań wywoła ich irytację. Od ankietera zadającego pytania oczekujemy przestrzegania opisanych przez Paula Grice’a (1977) zasad komunikacji kooperatywnej. Dlatego zamiast zadawać wielokrotnie to samo pytanie, pytamy o **różne powiązane kwestie**. Nawet przy wypełnianiu dokumentów czy podań w urzędach pada często 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-kilkaście pytań dotyczących pewnego zjawiska lub cechy, najczęściej dokonujemy dodatkowo weryfikacji jakości takiego wskaźnika, poprzez policzenie **alfy Cronbacha**. Jest to najpopularniejsza miara rzetelności (jednorodności) wskaźnika syntetycznego, która pozwala na określenie spójności pozycji wchodzących w jego skład.

Przykładem może być bardzo popularny kwestionariusz oparty na Pięcio-

czynnikowym Modelu Osobowości (Costa i McCrae, za: Siuta, 2006), który powstał na bazie badań leksykalnych. Skrócona wersja NEO_FFI (**NEO-Five Factor Inventory**) składa się z 60 pytań³, pozwalających opisać respondenta na 5 wymiarach: neurotyczności, ekstrawersji, otwartości na doświadczenie, ugodowości i sumienności, dla których alfa Cronbacha dla ponad 1000-osobowej grupy studentów wyniosła odpowiednio 0,842; 0,704; 0,69; 0,807; 0,712. Żadna z tych 12-itemowych skal nie jest jedno-czynnikowa. Ekstrawersja, otwartość na doświadczenie wykazują aż 4 czynniki. Nie dziwi to, gdy przyjrzymy się treści pytań. Przykładowo, aby uzyskać maksymalny wynik, np. w skali otwartości na doświadczenie, trzeba się zgodzić z 5 itemami i nie zgodzić z pozostałymi.

Tabela 1. Pytania z kwestionariusza NEO_FFI

	Pytania w skali otwartości na doświadczenie NEO_FFI
1.	Rzadko dostrzegam nastroje lub uczucia płynące z otoczenia
2.	Uważam, że pozwalanie uczniom na słuchanie kontrowersyjnych poglądów może im zamieszać w głowach i wprowadzić w błąd
3.	Sądzę, że powinniśmy się odwoływać do autorytetów religijnych przy podejmowaniu decyzji w sprawach moralności
4.	Mało interesuje mnie dociekanie natury wszechświata i natury ludzkiej
5.	Mam duże potrzeby intelektualne
6.	Często sprawia mi dużą przyjemność zajmowanie się teoretycznymi rozważaniami lub abstrakcyjnymi problemami
7.	Intrygują mnie formy, które odkrywam w sztuce i naturze
8.	Poezja działa na mnie słabo albo wcale
9.	Czasami, gdy czytam poezję lub dzieło sztuki, czuję „dreszczyk emocji” i falę podniecenia
10.	Gdy raz znajdę właściwy sposób na robienie czegoś, trzymam się go
11.	Nie lubię tracić czasu na marzenia
12.	Często próbuję nowych i egzotycznych potraw

Odmienne procesy poznawcze uruchamiane przy zgodzie i zaprzeczaniu (por. np. badania nad asymetrią Wanke i Schwarz, 1995) bardzo często powodują, że itemy pozytywne (wymagające zgody) i negatywne (wymagające zaprzeczenia) w analizie czynnikowej są separowane w osobne czynniki. Największym problemem jest jednak **niejednorodność konstruktów teoretycznych**. W otwartości na doświadczenie (zob. tabela 1) aż ¼ pytań dotyczy zainteresowania sztuką/poezją. W sumienności zawarta jest zarówno potrzeba osiągnięć, odpowiedzialność, jak i pedantyczność. Ta niejednorodność utrudnia wyobrażenie sobie osoby, która uzyskała wysoki/niski wynik na skali. Na ten problem zwracał już uwagę Allport (1937 za: Szarota, 2008), twierdząc, że **uzyskane w analizie czynniki stanowią wymiary osobowości uśrednionej, która jest całkowitą abstrakcją, nieprzydatną dla psychologa pragnącego badać osobowości konkretnych osób**. Na pewno warto podkreślić, że mierząc w ten sposób własności osoby **rozrywamy związek z poziomem obserwacji zachowań, analizując 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, 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 częstość jedzenia słodczy, picia piwa, jedzenie w nocy. Wszystkie te aktywności mogą prowadzić do nadwagi, występując samodzielnie. **Elementy wskaźnika syntetycznego nie muszą być skorelowane**, abyśmy mogli je zliczać lub uśredniać, gdy pytania dotyczyły częstości. Podobnie zbudowany został przez Janusza Czapińskiego (1996) syntetyczny wskaźnik aktywności ekonomicznej. 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 widzimy, 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 konstruktów, 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.

Niestety można sądzić, że narzędzia psychometryczne oparte na samoopisie okazały się w nauce ślepą uliczką, czego dowodem jest **brak postępu w psychometrii w ciągu ostatnich kilkudziesięciu lat**, ale możemy powtórzyć po raz kolejny: **niedoskonałe narzędzie pomiaru jest jednak lepsze niż brak narzędzi** (Gilbert, 2007). 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.

5. Słabość analiz

O ile badania eksperymentalne mają dobre i oczywiste standardy analizy, o tyle w badaniach korelacyjnych, jak to zauważa coraz więcej osób (np. Brzeziński, 2012), publikujemy sporo wyników o wątpliwej wartości. Wybitny psycholog prof. Robert Nisbett ogłasza nawet *The Crusade Against Multiple Regression Analysis* (2016). Problem wynika przede wszystkim z niskiej trafności wewnętrznej będącej immanentną własnością badań korelacyjnych i przejawiającej się w splątaniu (*confounding*) wpływów różnych zmiennych (por. przykłady w tekście Kuźmińskiej, 2016). **Wielozmiennowe analizy danych korelacyjnych są**

koniecznością, bo wszystkiego eksperymentalnie zbadać się nie da. Problem jednak tkwi w ich **wrażliwości na trafność modelu.** Jak to powtarzamy do znudzenia studentom: „współczynniki regresji zależą od towarzystwa”. Tylko nierozumiejący istoty tych analiz mogą stosować krokową analizę regresji (pozostawiając decyzje merytoryczne algorytmom statystycznym), pisać z pełnym przekonaniem, że analiza czynnikowa, analiza skupień dowiodła istnienia k czynników czy skupień. Zbyt często usunięcie jednej zmiennej zmienia strukturę macierzy korelacji. Na studentach wielkie wrażenie robi analiza przykładu (Wieczorkowska i Wierziński, 2011), w którym stwierdzają na tych samych danych różne zależności między X a Y w zależności od tego, które zmienne dodatkowe zostały kontrolowane. Oczywiście tam, gdzie jest jasna teoria, lista zmiennych jest wyznaczona jednoznacznie. W analizach danych sondażowych na próbach reprezentatywnych, gdy kontrolowanie zmiennych socjodemograficznych jest koniecznością często zapomina się o wprowadzaniu do modelu efektów interakcyjnych. Przykładowo, sprawdzając socjodemograficzne predyktory poziomu wykształcenia (operacjonalizowanego przez lata nauki) stwierdzimy brak istotnego związku z płcią, bo 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 (Wierziński, 2009). Ale to wymagało wprowadzenia do modelu regresyjnego efektów interakcyjnych, czego często badacze nie robią.

Prawdą jest jednak, że wielu zagadnień nie da się zbadać eksperymentalnie i powinniśmy bardzo intensywnie pracować nad nowym modelem analizy badań korelacyjnych. Zadanie jest to tym ważniejsze, że istnieją **ogromne zbiory ogólnodostępnych badań prób reprezentatywnych** zgromadzonych olbrzymim nakładem kosztów i w niewielkim stopniu wykorzystywanych poznawczo.

O ile testy istotności (oparte przecież na przedziałach ufności) biorą poprawkę na wielkość próby, o tyle sama wartość współczynnika determinacji wielkości próby nie uwzględnia. Dla ilustracji prosimy studentów, aby losowali z całej próby PGSS (N = 15776) podpróby o wielkości

1% (około 150 osób), 2% itd. Mogą się wtedy przekonać, że w małych 150-osobowych próbach można uzyskać dużo większe (ale także i mniejsze) współczynniki determinacji, niż by to wynikało z zależności w „populacji”. Pokazuje to, że nasze zaufanie do statystyki zwiększa się wraz ze wzrostem liczebności próby, więc próba nie może być zbyt duża, choć sądzimy, że duże próby najlepiej wykorzystywać do **bootstrapowej** (np. Davison i Hinkley, 1997) metody analizy, a nie testowania prostych zależności. Nie wiemy, dlaczego nie jest stosowane zalecenie, aby równanie regresji ustalać na połowie dostępnych danych, a testować je na reszcie (Pedhazur, 1997). Potrzebujemy nowych modeli testowania wielkich zbiorów danych. W naszej ocenie **moda na modelowanie strukturalne** przynosi jak dotąd więcej szkody niż pożytku, ponieważ owocuje wieloma publikacjami niewiele wnoszącymi do naszej wiedzy: zawsze można jakiś model do danych dopasować, a umysł badacza zawsze potrafi dopasować do danych nową teorię. Rygorysta metodologiczny Wiliam Starbuck (2016, s. 171) w swoim bardzo interesującym esej (choć nie ze wszystkimi jego postulatami się zgadzamy) piętnuje dwie powszechnie występujące (niestety) praktyki zwiększające szansę powodzenia w badaniach:

- HARKersi (Hypothesizing After Results are Known) najpierw zbierają dane, wykonują analizy statystyczne, potem formułują hipotezy, a na koniec wyszukują teorie i wcześniejsze badania, które wspierają lub przeczą nowo wymyślonym hipotezom.
- P-hakerzy wykorzystują wiele obliczeń i manipulacji danymi w poszukiwaniu równania lub systemu klasyfikacji, który uchwyci silne wzorce.

Innym problemem jest bezrefleksyjne porównywanie procentu wyjaśnionej zmienności w analizach **wskaźników jednostkowych** (opisujących respondenta) i **grupowych** (opisujących obiekty złożone np. kraj). Przykład: w międzynarodowym sondażu European Social Survey respondenci na pytanie „How happy are you?” odpowiadają na 11-stopniowej skali. O ile rozkład wskaźnika jednostkowego szczęścia jest zbliżony do normalnego, o tyle rozkład wskaźnika grupowego (średnia dla kraju) jest jednostajny – nie ma dwóch krajów o tej samej średniej (Wieczorkowska, Wierzbński i Król, 2015). O ile wskaźnik

jednostkowy można uznać za zmienną dyskretną, o tyle wskaźnik grupowy jest zmienną ciągłą. W konsekwencji inne wyniki uzyskamy, testując tę samą hipotezę dotyczącą np. związku stopnia prawicowości–lewicowości poglądów politycznych z poczuciem szczęścia (Napier i Jost, 2008) za pomocą grupowych vs jednostkowych wskaźników szczęścia. Procent wyjaśnionej wariacji jest nieporównanie mniejszy, gdy analizujemy wskaźniki jednostkowe niż grupowe. Nie może więc dziwić, że **ekonomiści**, analizując głównie **wskaźniki grupowe**, uzyskują **wyższe procenty dopasowania** swoich modeli niż analizujący **wskaźniki jednostkowe badacze społeczni**.

6. Podsumowanie: nieuzasadnione zarzuty recenzentów

Stosowanie metod ilościowych w naukach o zarządzaniu wiąże się często z **przyjmowaniem szkodliwych założeń**, które owocują często bezzasadnymi zarzutami metodologicznymi dotyczącymi najczęściej: (1) **braku reprezentatywności próby**, (2) **braku hipotez**. Zapomina się przy tym, że 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 dwukrotnie. Co gorsza, mamy do czynienia ze zjawiskiem **falszywych 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 wewnętrzną (najpierw trzeba mieć co generalizować), którą można zwiększać za pomocą randomizacji

II stopnia prób łatwo dostępnych. Drugi zarzut bierze się z fałszywego przekonania, że praca naukowa musi być zawsze prowadzona w modelu dedukcyjnym, co w szczególności w zarządzaniu nie jest prawdą. Niektóre prace w naukach o zarządzaniu stanowią próbę syntezy wiedzy, tworzenie modelu, który nie jest poddawany testom empirycznym. Dopisywanie wtedy hipotez nie ma żadnego uzasadnienia (dopisywane na „siłę” często mają postać kabaretową). Warto jednak pamiętać, że jeśli nie testujemy eksperymentalnie zestawu hipotez, tylko zbieramy dane w celach eksploracyjnych, to należy „męczyć je tak długo, aż zaczną zeznawać”. Inaczej szkoda nakładów poniesionych na prowadzenie badań.

Przypisy

- 1 Artykuł jest efektem wielogodzinnych dyskusji trójki autorów, której podsumowanie w angielskiej wersji zostało opublikowane przez dwójkę pierwszych autorów w *Problemach Zarządzania*, 2016, tom 2.
- 2 Na nasze odpowiedzi wpływają także pozawerbalne zapisy doświadczenia (połączenia tworzone w wyniku rejestracji częstości współwystępowania bodźców w otoczeniu). Do tych zapisów nasza świadomość mam dużo gorszy dostęp niż do zapisów informacji werbalnych (w tym naszych przemyśleń). Te drugie są podłożem do tworzenia się postaw jawnych, te pierwsze postaw ukrytych.
- 3 Oryginalna wersja zawierająca 240 pytań składała się z trzydziestu czynników pierwszego rzędu, pogrupowanych po 6 w 5 czynników drugiego rzędu. Powszechnie używana jest jednak jej skrócona wersja.

Bibliografia

Allport, G.W. (1937). *Personality: A psychological interpretation*. New York: Holt.

Aronson, E. i Wieszorkowska, G. (2001). *Kontrola naszych myśli i uczuć. Rozdział 4. Jak możemy zdobyć odpowiedź na nurtujące nas pytania?* Warszawa: Santorski.

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

Costa, P.T., McCrae, R.R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.

Czapiński, J. (1996). Uziemienie polskiej duszy. W: M. Marody i E. Gucwa-Leśny (red.), *Podstawy życia społecznego w Polsce* (s. 252–275). Warszawa: Wydawnictwo Instytutu Studiów Społecznych Uniwersytetu Warszawskiego.

Davison, A.C. i Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

ESS Round 7: European Social Survey Round 7 Data (2014). Data file edition 2.0. NSD – Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

Gilbert, D.T., Tafarodi, R.W. i Malone, P.S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221–233.

Gilbert, D. (2007). *Stumbling on Happiness*. New York, NY: Vintage Books.

Grice, P. (1977). Logika i konwersacja. *Przegląd Humanistyczny*, 6.

Grudzewski, W.M. i Hejduk, I.K. (2004). *Zarządzanie wiedzą w przedsiębiorstwach*. Warszawa: Difin S.A.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67, 371–378.

Napier, J.L. i Jost, J.T. (2008). Why are conservatives happier than liberals? *Psychological Science*, 19(6), 565–572.

Nisbett, R. (2016). *The Crusade Against Multiple Regression Analysis*. Pobrano z https://www.edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis

Pedhazur, E.J. (1997). *Multiple regression in behavioral research*. Belmont: Wadsworth Publishing.

Piliavin, I.M., Rodin, J. i Piliavin, J. (1969). Good Samaritanism: an underground phenomenon? *Journal of Personality and Social Psychology*, 13(4), 289–299.

Siarkiewicz, M. (2007). Zastosowanie metody analizy szeregów czasowych do badania zależności pomiędzy składem posiłków a subiektywnie odczuwanym poziomem energii. W: K. Winkowska-Nowak, A. Nowak i A. Rychwalska (red.), *Modelowanie matematyczne i symulacje komputerowe w naukach społecznych* (s. 15–20). Warszawa: Wydawnictwo Academica SWPS.

Siuta, J. (2006). *Inwentarz Osobowo ci NEO-PI-R Paula T. Costy i Roberta R. McCrae. Adaptacja polska. Podręcznik*. Warszawa: Pracownia Testów Psychologicznych.

Starbuck, W.H. (2016). 60th Anniversary Essay: How Journals Could Improve Research Practices in Social Science. *Administrative Science Quarterly*, 2, 165–183, <http://10.1177/0001839216629644> asq.sagepub.com

- Szarota, P. (2008). Wielka Piątka – stare problemy, nowe wątpliwości. *Roczniki Psychologiczne*, 11(1), 127–138.
- Taleb, N. (2012). *Antifragile: Things That Gain from Disorder*. New York: Random House.
- Thorngate, W. i Chunyun, M. (2016). Wiggles and curves: The analysis of ordinal patterns. *Problemy Zarządzania*, 2.
- Turska, E. (2014). *Kapitał kariery ludzi młodych. Uwarunkowania i konsekwencje*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.
- Van Bavel, J.J., Mende-Siedlecki, P., Brady, W.J. i Reinero, D.A. (2016). *Contextual sensitivity in scientific reproducibility*. Proceedings of the National Academy of Sciences. Social Sciences – PNAS, 23 6454–6459, <http://dx.doi.org/10.1073/pnas.1521897113>
- Wänke, M., Schwarz, N. i Noelle-Neumann, E. (1995). Asking comparative questions: The impact of the direction of comparison. *Public Opinion Quarterly*, 59, 347–372.
- Wierzbński, J. (2009). *Badanie zaufania do organizacji: problemy metodologiczne*. Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania UW.
- Wieczorkowska-Wierzbńska, G. (2011). *Psychologiczne ograniczenia*. Warszawa: Wydawnictwo Naukowe Wydziału Zarządzania UW.
- Wieczorkowska-Wierzbńska, G. (2012). Zrozumieć dane i starać się rozwiązać problem. *Roczniki Psychologiczne*, 3.
- Wieczorkowska, G. i Wierzbński, J. (2011). *Statystyka: od teorii do praktyki*. Warszawa: Wydawnictwo Naukowe Scholar.
- Wieczorkowska, G., Wierzbński, J. i Król, G. (2015). Metody ilościowe. W: *Metody badawcze w zarządzaniu humanistycznym*. Warszawa: Wydawnictwo Akademickie Sedno.
- Wieczorkowska-Wierzbńska, G., Król, G. i Wierzbński, J. (2015). Przeszłość, teraźniejszość i przyszłość edukacji akademickiej. W: *Gospodarka na rozdrożu – XXI wiek*. Warszawa: Wyd. Naukowe Wydziału Zarządzania UW.
- Zimbardo, P.G. i Gerrig, R.J. (2012). *Psychologia i życie*. Warszawa: Wydawnictwo Naukowe PWN.