Trafność pomiaru w praktyce psychometrycznej

Paweł Kleka

Trafność pomiaru w praktyce psychometrycznej

Badania empiryczne

Wydawnictwo Nauk Społecznych i Humanistycznych UAM
Poznań 2021

Projekt okładki
Robert Domurat

Redakcja, korekta i łamanie
Izabela Baran

Wydanie I

Publikacja finansowana z funduszy Wydziału Psychologii i Kognitywistyki UAM

978-83-66983-02-1
978-83-7589-085-3

Wydawnictwo Nauk Społecznych i Humanistycznych
Uniwersytetu im. Adama Mickiewicza w Poznaniu
60-568 Poznań, ul. Szamarzewskiego 89c
www.wnsh.amu.edu.pl, wnsh@amu.edu.pl, tel. (61) 829 22 54

Wydawnictwo Fundacji Humaniora
60-682 Poznań, ul. Biegańskiego 30A
www.funhum.home.amu.edu.pl, drozd@amu.edu.pl, tel. 519 340 555

Druk
Drukarnia Scriptor Gniezno
62-600 Gniezno, ul. Poprzeczna 6A

SPIS TREŚCI

WSTĘP

> Opis osoby bez odwołania się do innych może być fragmentem literatury, biografii lub powieści. Ale nauki? Nie.
>
> Meyer, 1926, s. 271[1]

Tematyka podejmowana w tej pracy jest odzwierciedleniem trwającej obecnie na świecie w środowisku psychometrów i psychologów dyskusji na temat jakości psychologii jako nauki (w 2018 roku powołano dwie organizacje zrzeszające dyskutujących o tym psychologów i psychometrów: Psychological Science Accelerator, Society for the Improvements of Psychological Science). Problem naukowości psychologii nie został rozwiązany przez stulecie jej istnienia, a ostatnio przybrał na sile w związku z tzw. kryzysem replikowalności psychologicznych badań naukowych (Turner, 2019). Spadek zaufania do wyników badań nie tylko wśród społeczeństwa, ale także wśród samych naukowców sprawił, że pojawiła się potrzeba uporządkowania problematyki prowadzenia badań i oceny ich jakości. W tej sytuacji przybierają na sile głosy krytyczne, które obecne są na każdym etapie rozwoju nauki, ale często były zagłuszane kolejnymi 'nowinkami' inkorporowanymi na grunt psychologii z ościennych dziedzin: socjologii, fizyki, matematyki czy ostatnio informatyki oraz neuronauki. Zachwyt nad nowymi metodami prowadzenia badań czy też analizą ich wyników sprawiał, że psychologia 'szła do przodu, nie oglądając się za siebie'. Niestety problemy, które nie zostały dostatecznie zbadane i rozwiązane, wpływają teraz na spadek zaufania do psychologii jako dziedziny nauki.

Psychologia zawsze przejawiała aspiracje do bycia nauką opartą na badaniach empirycznych. Przedmiotem nauki empirycznej są obiekty o mierzalnych właściwościach. Dlatego podstawowym zmartwieniem psychologów badaczy jest udowodnienie, że zarówno mierzone obiekty psychiczne, jak i narzędzia pomiarowe używane w psychologicznej praktyce badawczej są godne zaufania. Problematykę empiryczności pomiaru psychologicznego można zatem sprowadzić

[1] „A description of one individual without reference to others may be a piece of literature, a biography or novel. But science? No".

do odpowiedzi na dwa pytania: 1) co specyficznego mierzy pomiar psychologiczny oraz 2) jak dobrze ten pomiar przebiega.

Podstawowym problemem dla psychologów odpowiadających na pytanie o przedmiot swoich zainteresowań badawczych jest udowodnienie, że obiekty psychiczne są mierzalne w jakikolwiek sposób. W latach trzydziestych ubiegłego wieku brytyjskie stowarzyszenie na rzecz rozwoju nauki (British Association for the Advancement of Science) powołało do życia Komitet pod przewodnictwem Allana Fergusona, by zbadał, jakie psychiczne atrybuty mogą być mierzone w sposób naukowy. Członkowie komitetu pod dużym wpływem Normana Roberta Campbella (zob. Michell, 1999, s. 144) stwierdzili, że ponieważ psychologiczne atrybuty nie poddają się konkatenacji (nie można ich ilościowo porządkować), nie są mierzalne naukowo (Ferguson et al., 1940, za: Newman, 1974, s. 141). Niezgoda z tym oświadczeniem była głównym źródłem powstania operacyjnej teorii pomiaru stworzonej przez Stanleya S. Stevensa (1946)[2], w której zdefiniował on pomiar psychologiczny jako przypisywanie liczb do obserwacji zgodnie z różnymi funkcjami matematycznymi w zależności od charakteru mierzonej (wtedy jeszcze określanej jako *sensory events*) właściwości. Pozwoliło to wyróżnić cztery podstawowe skale pomiarowe, których definicja obowiązuje do dzisiaj i konstytuuje prawie wszystkie pomiary psychologiczne. Skoro zaś właściwości psychiczne okazały się mierzalne, to stało się możliwe zdefiniowanie, czym jest pomiar w psychologii. Wbrew obiegowym opiniom nie jest to jednak etap zakończony i definicja pomiaru w psychologii rozwijana jest nadal. Temu rozwojowi poświęcam rozdział 1 niniejszej pracy.

Do pomiaru potrzebne są mierzone podmioty. Z tej perspektywy przyglądam się pomiarowi w rozdziale 2. Poświęcony jest on problemowi pozyskiwania wyników badań psychologicznych głównie za pomocą metod samoopisowych. W dużym stopniu ufamy informacjom dostarczanym nam bezpośrednio przez osoby badane, które zanim informacji udzielą, konstruują ją w swoich umysłach (Kleka, 2017). Zastanawiam się, czy systematyczny błąd, jaki mogą popełniać badani przez nas ludzie, wystarczy uśredniać przez badanie grupowe i odwoływanie się do centralnego twierdzenia granicznego, czy też powinniśmy zacząć bardziej uwzględniać w naszych narzędziach pomiarowych to, że ludzie różnią się poziomem wiedzy o sobie i mają zróżnicowane ograniczenia w dostępie do tej wiedzy. Głębokość wglądu, poziom samowiedzy czy zdolność do samoopisu – to w moim odczuciu synonimy czynnika rzadko uwzględnianego w badaniach psychologicznych. W rozdziale 2 próbuję dokonać rozstrzygnięcia, czy ten immanentny czynnik badań psychologicznych jest nierozróżnialny od szumu tła (czynników losowych), czy też opierając się na przesłankach

[2] Zapewne nie bez znaczenia był też fakt, że komitet jako przykład potwierdzający tezę o niemierzalności atrybutów psychicznych podał skalę głośności opracowaną właśnie przez Stevensa.

teoretycznych i/lub empirycznych, jesteśmy w stanie go wyodrębnić i powinniśmy go kontrolować.

Dyskusja wokół odpowiedzi na drugie pytanie, czyli o jakość pomiaru, koncentruje się wokół jego rzetelności (precyzji) oraz trafności. Pojęcie r z e t e l n o ś c i towarzyszy problemowi pomiaru od zarania psychologii. Już Charles Spearman w swoich pracach na początku XX wieku (1904) zwrócił uwagę na obecność błędu pomiarowego i konieczność korygowania wyników badań psychologicznych o tę nieznaną wartość. Koncepcja rozwijana przez lata, ze znaczącymi pracami Williama Browna (1910), Joya P. Guilforda (1936, 1954), Frederica Kudera i Marion W. Richardson (1937), Philipa J. Rulona (1939), Cirila Hoyta (1941), Louisa Guttmana (1945), Harolda Guliksena (1950), Lee Cronbacha (1951), Frederica M. Lorda i Melvina R. Novicka (1968), Josa M.F. Ten-Berge i Fritsa E. Zegersa (1978), Williama Revelle'a (1979) oraz Rodericka P. McDonalda (1999), przeprowadziła rzetelność narzędzi badawczych w psychologii od alfy[3] do omegi[4] (szczegółowe porównanie współczynników rzetelności jako miary spójności testu por.: Zinbarg, Revelle, Yovel i Li, 2005; Dunn, Baguley i Brunsden, 2014). Klasyczna teoria testu przybrała postać ugruntowanego modelu wyniku prawdziwego składającego się wraz z błędem pomiarowym na wynik obserwowany, rozszerzona o kontekst badania ma swoje uogólnienie w teorii uniwersalizacji (Cronbach, Gleser, Nanda i Rajaratnam, 1972). W obu teoriach rzetelność definiowana jest jako stosunek wariancji wyników prawdziwych do wyników obserwowanych.

Inaczej jest jednakże z rozwojem koncepcji t r a f n o ś c i narzędzi psychologicznych. Adekwatność w opisie rzeczywistości psychicznej i użyteczność wyników uzyskiwanych w badaniach psychologicznych opiera się w całości na ich trafności, sprawiając, że pojęcie to stało się fundamentem, na którym budowana jest psychologia jako nauka. Pojęcie trafności od odpowiedzi na proste pytanie: „Czy test mierzy to, do czego został stworzony?" wyewoluowało do skomplikowanego tworu[5] obejmującego trzy, pięć lub według niektórych autorów sześć różnych aspektów testu i testowania[6]. Zagadnieniu, jak psychologia przy definiowaniu

[3] Współczynnik rzetelności określający spójność wewnętrzną pozycji testowych, dominujący w badaniach psychologicznych od publikacji Cronbacha w 1951 roku, krytykowany przez różnych badaczy, także przez swego twórcę (por. Cronbach i Shavelson, 2004), głównie za bezrefleksyjne stosowanie przez większość badaczy i omijanie założeń stojących za tym współczynnikiem, np. tau-ekwiwalentności pozycji. Współczynnik alfa nie był pierwszym zaproponowanym, ale pierwszym, który został utożsamiony z rzetelnością.

[4] Współczynnik rzetelności wewnętrznej testów oparty na dekompozycji wariancji w modelu hierarchicznym analizy czynnikowej – pozbawiony ograniczeń współczynnika alfa, ale wymagający bardziej skomplikowanych obliczeń matematycznych.

[5] Specjalnie unikam sformułowania „konstrukt", ponieważ będzie mi ono potrzebne w rozważaniach nad przedmiotem pomiaru w rozdziale 1.

[6] Wyczerpujące zestawienie 151 różnych nazw/rodzajów trafności przedstawili w swojej pracy Newton i Shaw (2014, s. 8).

trafności przeszła od pytania o odwzorowanie treści do traktowania trafności jako właściwości interpretacji wyniku testowego (por. Messick, 1980), przyglądam się bliżej w rozdziale 3.

Rozdział 4 poświęcony jest praktycznemu sprawdzeniu odporności stosowanych powszechnie sposobów walidacji narzędzi badawczych na treść, a właściwie na jej brak w narzędziach będących przedmiotem sprawdzania. Uzyskane wyniki wskazują na raczej niski stopień zaufania do metod statystycznych oderwanych od założeń teoretycznych. Przy dużej dostępności różnych, skomplikowanych metod analitycznych poprawność otrzymanych wyników nie ma żadnego związku lub ma związek niewielki z założeniem o trafności walidowanego narzędzia badawczego – algorytmy matematyczne nie odróżniają automatycznie testów wiarygodnych od niewiarygodnych, ale jedynie dostarczają wyników. A te poddawane są najczęściej subiektywnej interpretacji.

Zanim przejdę do omawiania poszczególnych tematów dotyczących pomiaru i problemów z nim związanych, chciałbym dla jasności wywodu poczynić kilka założeń definicyjnych. I tak terminu t e s t używam w znaczeniu narzędzia badawczego stosowanego do pośredniego mierzenia właściwości psychicznych człowieka. W tej szerokiej definicji mieszczą się kwestionariusze, skale i testy zadań, które są prezentowane osobom uczestniczącym w badaniach psychologicznych w celu gromadzenia danych odzwierciedlających obecność (a często również stopień) danej zdolności psychicznej, postawy lub zachowania. Zastrzegam sobie terminy „podtest", „podskala" i „skala" dla oznaczenia podzbioru poszczególnych elementów testu. Zazwyczaj uważa się, że odzwierciedlają one aspekt danego atrybutu wyższego rzędu.

Psychologowie badają nie tylko jeden z najbardziej złożonych systemów naturalnych – człowieka, ale także procesy zachodzące w psychice, których jeszcze nikt nie zdążył z powodzeniem skonceptualizować. Co więcej, robią to, używając owych procesów. Fakty te powinny przynajmniej nasuwać pytanie, do jakiego stopnia ilościowe, eksperymentalne metody nauk przyrodniczych (fizyki, biologii itp.) mogą być sensownie stosowane w psychologii. Jak bardzo obiekt badania, jakim jest człowiek z perspektywy psychologicznej, sprawia, że samo badanie staje się specyficzne i wymaga specjalnych metod pomiaru? Czy opracowane do tej pory i stosowane powszechnie metody pomiaru są wystarczająco dobre? Czy potrafimy odróżnić wiarygodne metody pomiaru psychologicznego od niewiarygodnych? Postaram się odpowiedzieć na te pytania na kartach niniejszej książki.

1. PROBLEMATYKA POMIARU W PSYCHOLOGII[1]

> Jak już zauważyłem, wielu psychometrów zazwyczaj nie wydaje się być w badaniu zainteresowanymi, czy atrybuty, które mierzą, są naprawdę ilościowe. Zamiast tego są zainteresowani przede wszystkim twierdzeniem, że te atrybuty mierzyć mogą.
>
> Michell, 2014, s. 117[2]

Wraz z czasami nowożytnymi wyodrębniająca się z filozofii psychologia pragnęła uchodzić za naukę empiryczną, ścisłą, poddaną naukowemu rygorowi. W XIX wieku psychologia była jeszcze działem filozofii zajmującym się zjawiskami powiązanymi ze świadomością, głównie w trzech obszarach: introspekcji, asocjacji oraz fizjologii. Głównym powodem oglądania się na nauki ścisłe (w których procedury numeryczne były ówcześnie traktowane jako miara precyzji i obiektywności *per se*) było to, że filozofowie odmawiali psychologii naukowości, np. Immanuel Kant (1724–1804) uznawał psychologię za wiedzę li tylko praktyczną. Jak pisze Cezary W. Domański (2018, s. 31): „odmawiając psychologii miana nauki ścisłej, Kant rzucił wyzwanie, które podjęli w XIX wieku przedstawiciele tak zwanego kierunku mechaniczno-matematycznego […] oraz psychofizycy". W związku z tym psychologia inkorporowała z nauk przyrodniczych metody badawcze, które utwierdzały psychologów w przekonaniu, że zajmują się nauką ścisłą. Zmierzone empirycznie (za pomocą zmysłów) dane, modelowane następnie przy użyciu narzędzi matematycznych stały się głównym przedmiotem zainteresowania psychologów. Proces mierzenia w psychologii był i jest nadal definiowany zwykle jako proces przypisywania liczb do obiektów zgodnie z ustalonymi regułami (Stevens, 1946). Nieco później David H. Krantz, R. Duncan Luce, Patrick Suppes i Amos Tversky (1971) doprecyzowali pomiar jako przypisywanie liczb obiektom „w taki sposób, że właściwości atrybutów są wiernie przedstawiane jako właściwości numeryczne".

---

[1] Fragment tego rozdziału ukazał się w pracy zbiorowej *Mity współczesnej psychologii* pod redakcją Moniki Obrębskiej i Andrzeja Pankali w 2020 roku.

[2] „As I have noted, many psychometricians typically do not seem to be interested in investigating whether the attributes they aspire to measure are really quantitative […]. Instead they are primarily interested in already claiming that they can measure such attributes".

Innymi słowy, dane uzyskane w wyniku pomiaru powinny odzwierciedlać cechy charakterystyczne mierzonego zjawiska nazywanego w psychologii konstruktem. Narzędzia badawcze używane przez psychologów (testy, kwestionariusze, skale itp.) opisują metody obserwacji (pomiaru) konstruktów i dostarczają wyników, które pozwalają nam przekształcać zjawiska w dane.

1.1. Od pomiaru fizjologii do pomiaru konstruktu

Historia psychologii jako dyscypliny naukowej jest głównie opowieścią o pomiarze psychologicznym. Początek przekonania o empiryczności psychologii można umieścić w latach czterdziestych XIX wieku. Nie jest to jakaś szczególna zasługa samej psychologii – w rozwoju różnych dziedzin nauki nastąpił okres, kiedy wydawało się, że wszystko uda się prędzej lub później zmierzyć. Co więcej, niektórzy psychologowie, szczególnie z kręgu amerykańskiego, stwierdzili, że nie trzeba czekać, aż sam pomiar psychologiczny będzie dostatecznie zbadany oraz określony, i zaczęli wdrażać procedury służące ocenianiu różnych cech, postaw i zachowań ludzi, które to procedury są podstawą głównego nurtu w psychologii dzisiaj. Wywarło to ogromny wpływ na rozwój psychologii jako dziedziny nauki, a konsekwencją jest m.in. nasycenie studiów psychologicznych na poziomie akademickim statystyką i psychometrią (Nunnally, 1960).

Charakterystyczne dla psychologii są próby ilościowego przedstawienia atrybutów psychicznych, co niesie za sobą konsekwencję w postaci konieczności udowodnienia, że te ilości stanowią dobre odwzorowanie mierzonego atrybutu. Pierwotnie badaniami z wykorzystaniem pomiarów psychologicznych zajmowano się niezależnie w dwóch krajach: Anglii i Niemczech. W tym pierwszym podwaliny kładli Darwin, Galton i Cattell, zajmując się głównie pomiarami różnic indywidualnych i inteligencji; w tym drugim byli to Herbart, Weber, Fechner i Wundt, zainteresowani zróżnicowaniem psychofizycznym, czym doprowadzili do rozwoju psychologii eksperymentalnej i standaryzowanych testów.

W Niemczech Gustaw Teodor Fechner (1801–1887), fizyk i filozof, interesował się związkiem między pomiarem właściwości fizycznych a stanem fenomenów psychicznych. Jego praca *Elemente der Psychophysik* z roku 1860 ukonstytuowała współczesną psychologię eksperymentalną (Domański, 2018, s. 58) i zakończyła etap psychologii filozoficznej. Wykształcony jako fizyk, pierwotnie interesujący się i badający obwody elektryczne (Winter, 1949) Fechner rozumiał pomiar tak, jak rozumie go fizyka. W swoich eksperymentach zajmował się ludzką percepcją obiektów o zbliżonej wadze, wyciągając wnioski na temat krzywoliniowej (logarytmicznej) zależności między bodźcami a odczuciami (por. prawo Webera–Fechnera). Spierając się ze spadkiem po filozofii, czyli kartezjańskim dualizmem ciała i umysłu, Fechner argumentował, że zarówno zjawiska mentalne, jak i fizyczne podlegają obserwacji czy to przez introspekcję, czy też dzięki

obserwacji z zewnątrz. Monistyczne podstawy pomiarów psychofizycznych wyrażały się w przekonaniu, że ciało i umysł nie mogą należeć do osobnych „przestrzeni", ponieważ wtedy nie byłaby możliwa interakcja między nimi. A skoro interakcje następują, to można mierzyć ciało, by zdobywać wiedzę o umyśle. Badania Fechnera ukonstytuowały naukowy pomiar zjawisk psychicznych i stały się bodźcem do rozwoju psychologii ilościowej dzięki powiązaniu zjawisk psychologicznych z fizycznymi. Dzięki prawu psychofizycznemu dołączył on „rzeczy psychiczne" i ich pomiar do istniejącej fizyki ilościowej. Jak pisał po latach Stefan Błachowski (za: Domański, 2018, s. 60): „badając w sposób ścisły związki, jakie zachodzą między podnietami a wrażeniami", wprowadził w obszar zainteresowań psychologów matematykę. Zbierając materiały potwierdzające jego prawo, Fechner zainicjował eksperymentalne metody badawcze, przekonując sobie współczesnych, że metody te są mierzeniem.

Fechner mylnie sądził, że rozwiązał problem ilościowości pomiaru, a metody psychofizyczne były naśladowane przez kolejne pokolenia psychologów ilościowych. W rzeczywistości teza, że istnieją takie zjawiska psychiczne jak doznania, jest tylko hipotezą, w dodatku niemożliwą do weryfikacji. Jest to teza czysto filozoficzna, a Fechner przyjął za oczywiste, że bezpośrednim przedmiotem świadomości w percepcji osoby jest bodziec, a nie wewnętrzny konstrukt umysłowy. Introspekcyjne sądy badanego informują nas jednak nie o doznaniach, ale raczej o wrażliwości na bodziec fizyczny, jego właściwości i relacje (Michell, 1988).

Za datę narodzin współczesnej psychologii uważa się rok 1879, w którym Wilhelm Wundt (1832–1920) założył laboratorium badawcze do przeprowadzania eksperymentów psychofizycznych nad wrażeniami zmysłowymi, kontynuując fizjologiczne i eksperymentalne podejście Fechnera do psychologii. Sławne laboratorium (urządzone zresztą na wzór chemiczno-fizycznego laboratorium promotora Wundta, profesora chemii Roberta Bunsena) przeprowadzało badanie za badaniem. W roku 1893 pracowało w nim 25 osób (Domański, 2018, s. 66), co nawet jak na dzisiejsze standardy jest dużym zespołem. Z upływem czasu zainteresowanie psychofizyką wygasło, a próby zmierzenia zdolności intelektualnych stały się centralnym punktem zainteresowania psychologii ilościowej w XX wieku.

W Anglii Francis Galton (1822–1911), będący pod wpływem dzieła Darwina *O pochodzeniu gatunków*, zauważył, że różnice między osobami obserwowane w eksperymentach niemieckich psychofizjologów mogą mieć wytłumaczenie w różnicach w obszarze zdolności do inteligentnego osądu badanych osób. Eksperymenty Galtona polegały na przykład na odpytywaniu osób biorących w nich udział o dokładne przypomnienie sobie wcześniejszych doświadczeń – różnice między uczestnikami skłoniły Galtona do poszukiwania źródła tychże różnic w inteligencji. Z kolei inteligencja miała być cechą poprawiającą przystosowanie jednostek do środowiska i podlegającą regułom dziedziczenia.

Rozwój badań nad inteligencją w oderwaniu od fizjologii zawdzięczamy z kolei Alfredowi Binetowi (1857–1911). Bazując na różnicach w odpowiedziach dzieci

i dorosłych, postulował on tworzenie testów, które badają stosowanie bardziej złożonych umiejętności niż zwykła trafność lub szybkość reakcji. Przykładem może tu być stałość uwagi i wyrafinowane posługiwanie się językiem. Testy projektowane przez Bineta stały się prototypami dla wszystkich testów psychologicznych i odpowiadały na potrzeby szkolnictwa w selekcji dzieci do specjalnego traktowania, wojska w selekcji żołnierzy do bardziej wymagających stanowisk, studentów i pracowników w dopasowywaniu do profilów studiów i pracy. Testy te konstruowane były w postaci standaryzowanych wywiadów psychologicznych.

Prekursorem w badaniach nad inteligencją był Charles Spearman (1863–1945), który także nie ustrzegł się błędu popełnionego przez psychofizyków. Jego teoria inteligencji i wyznaczenie czynnika g miały swe źródło w obserwacji, że wyniki różnych testów umysłowych są ze sobą skorelowane. Spearman sądził, że czymś, co tę pozytywną korelację powoduje, jest ukryty czynnik – inteligencja. Zupełnie zignorował konieczność dowiedzenia, że korelowane właściwości są ilościowe i mierzalne. Nawet to, że skale mogły być w jakimś sensie ilościowe, nie oznaczało automatycznie, że ich przyczyna również jest ilościowa. To, co powinno zostać dowiedzione (ilościowy, addytywny charakter mierzonej właściwości), było przyjmowane nie jako hipoteza do przetestowania, tylko jako pewnik, na którym oparto cały dział psychologii.

Głównym źródłem tego błędu był środowiskowy nacisk na pragmatyzm efektów badań psychologicznych, który swoje piętno odcisnął szczególnie na testach inteligencji (wtedy nazywanych testami mentalnymi) w Stanach Zjednoczonych. W Ameryce bardzo szybko zaczęto stosować testy oceniający cechy umysłowe (*mental tests*), a pomiar ilościowy był powszechnie obecny. Omijanie pytania o 'ilościowość' mierzonych konstruktów w psychologii stało się możliwe m.in. dzięki temu, że wyniki badań psychologicznych przedstawiane były przeważnie w postaci liczb – łatwiej było je 'sprzedać' decydentom, gdy wyglądały naukowo dzięki swej formie. Głównymi orędownikami użyteczności psychologii, promującymi testy umysłowe w szkolnictwie, w przemyśle i w wojsku, byli James M. Cattell (1860–1944) i Edward L. Thorndike (1874–1949).

Cattell był pod wpływem Wundta, u którego się doktoryzował. Uważał, że psychologia może tylko wtedy być precyzyjna i ścisła jak nauki fizyczne, gdy będzie budowana na bazie eksperymentów i pomiarów (Cattell, 1890). Według niego naukowe było to, co mierzalne, a każda nauka w miarę rozwoju podąża w stronę metod ilościowych. Z kolei jego uczeń, Thorndike, wprost zrównywał psychologię z naukami ścisłymi:

> Cokolwiek istnieje w ogóle, istnieje w pewnej ilości. Znać coś dokładnie, to znaczy znać ilość i jakość […]. Jest to oczywiście to samo ogólne *credo*, co fizyków, chemików czy fizjologów zaangażowanych w myślenie ilościowe, a w rzeczy samej to, czym współczesna nauka jest w ogóle. I, ogólnie rzecz biorąc, charakter pomiarów edukacyjnych jest taki sam jak wszystkich pomiarów naukowych (1918, ss. 16–17).

Prymat pragmatyzmu przypieczętował Truman Kelley (1884–1961), wpływowy psycholog i statystyk (jako pierwszą dziedzinę studiował matematykę), deklarując:

> Nasze testy psychiczne mierzą coś, co może nas obchodzić lub nie, ale jest to coś, co warto zmierzyć, ponieważ poszerza naszą wiedzę na temat tego, co ludzie mogą robić w przyszłości. Przyrząd do mierzenia jako miara czegoś, co jest pożądane do zmierzenia, jest nadrzędny, a to, czego jest miarą, jest drugorzędne (Kelley i Shen, 1929, s. 866).

Kelley postulował, by problematykę ilościowości pomiaru rozwiązywać w sposób statystyczny, co w połączeniu z wpływem Thorndike'a na środowisko psychologów amerykańskich sprawiło, że ilościowość przestała być problemem logicznym, a stała się statystycznym, czyli nieinteresującym dla psychologów niezajmujących się psychometrią i statystyką[3]. A takich psychologów ówcześnie (i prawdopodobnie nadal) była większość. Nic dziwnego, że w krótkim czasie, nawet jeśli poszczególni psychologowie nie zgadzali się co do tego, co mierzą, to większość zgadzała się co do jednego: jeśli coś mierzymy, to jest to mierzalne. Jednocześnie zmieniła się rola teorii w rozwoju psychologii. Coraz częściej teoria przestawała być źródłem pytań, na które trzeba odpowiedzieć, przeprowadzając kolejne badania, a służyła jako rama do konstruowania odpowiedzi na pytania postawione na podstawie danych empirycznych. Teoria zaczęła służyć wyjaśnianiu wyników, a nie ich predykcji. W ten sposób psychologia weszła na ścieżkę rozwoju, na której testowanie nie odbywa się w sposób popperowski (teorię uważamy za fałszywą, jeśli dane nie potwierdzają przewidywań), ale lakatozjański (testujemy idee w tzw. liniach (programach) badawczych, które ze sobą konkurują; por. Meehl, 1990). Budujemy wiedzę opartą na metaanalizach podsumowujących całe zbiory badań – metaanalizach, w których teoria zupełnie ustępuje miejsca empirii, czyli pomiarowi. Sprowadza się to do sprawnego budowania narracji na podstawie zbiorów zastanych faktów[4], które się już wydarzyły. Przy bogactwie koncepcji, konstruktów, poglądów, modeli w psychologii zawsze można znaleźć taki, który tłumaczy to, co zostało zaobserwowane – lub wymyślić nowy „zestaw".

Rozwijająca się psychologia wymagała zdefiniowania obszaru swoich badań. Samo określenie przedmiotu badania terminem „psychologiczny" było zbyt szerokie i nieprecyzyjne. Dyskusja nad tym, co jest *de facto* przedmiotem pomiaru w psychologii, doprowadziła do zdefiniowania k o n s t r u k t u[5]. Jego definicja na najbardziej podstawowym i ogólnym poziomie mówi o tym, że konstrukt to pojęcie teoretyczne, które jest przedmiotem badań psychologicznych. Mimo

[3] Początek psychometrii jako subdyscypliny naukowej można datować na rok 1935 – rok założenia przez Thurstone'a, Thorndike'a i Guilforda czasopisma *Psychometrika*.

[4] Niektórzy byli w tym tak dobrzy, że nawet nie potrzebowali faktów, tylko sami je fabrykowali – zob. np. skandal związany z oszustwami naukowymi Diederika Stapela.

[5] Ma to konsekwencje dla rozwoju sposobów walidowania narzędzi badawczych używanych przez psychologów, o czym mowa będzie w rozdziale 3.

że dopiero w połowie XX wieku pojawiła się wyraźna definicja pojęcia „konstrukt" i ustalono jego rolę w psychologii, to jednak koncept ten wywodzi się z debat toczonych w ramach filozofii nauki już w latach dwudziestych i trzydziestych XX wieku na temat użyteczności pojęć teoretycznych w nauce w ujęciu bardziej ogólnym (Campbell, 1920; Kelley, 1923; Nagel, 1931; Russell, 1937; Thorndike, 1904; Thurstone, 1937).

Nie istnieje jednoznaczna definicja terminu „konstrukt" – poszczególni naukowcy definiują go z różną precyzją, lub wręcz negują jego użyteczność. Na przykład Jane Loevinger rozróżniała cechy od konstruktów. Twierdziła, że o ile u ludzi istnieją cechy, o tyle konstrukty „istnieją w umysłach i czasopismach psychologów" (Loevinger, 1957, s. 642). Z drugiej strony na przykład Cronbach i Meehl w swoim artykule z 1955 roku stwierdzają, iż: „konstrukt jest jakąś postulowaną cechą ludzi, przyjmuje się, że jest odzwierciedlony w skuteczności badania" (Cronbach i Meehl, 1955, s. 283). Powszechnie konstrukt jest utożsamiany z cechami psychologicznymi czy atrybutami, które są przedmiotem badań psychologicznych (np. Anastasi i Urbina, 1989, s. 189), a ostatecznie oznacza on dość dużą klasę odniesień, w tym atrybuty psychologiczne, cechy, wnioskowane procesy oraz konstrukcje logiczne. Tak szeroka definicja jest wygodna, lecz także niepraktyczna.

Już Cronbach w swojej pracy *Construct Validation After 30 Years* (1989) zauważył, że prawie każdy psycholog piszący o terminie „konstrukt" określa go jako niejasny i mylący. W standardach do stosowania testów w edukacji i psychologii (American Educational Research Association, American Psychological Association i National Council on Measurement in Education, 1999) pojęcie „konstrukt" używane jest w znaczeniu wyniku testowego, tj. tego, co test ma mierzyć. Przegląd najnowszych głosów w dyskusji nad tym pojęciem wskazuje na stale obecny problem z jego definicją i interpretacją (Slaney i Racine, 2013a). Michael Maraun i Stephanie Gabriel (2013) oraz Kathleen Slaney i Timothy Racine (2013b) podkreślają, że z praktycznego punktu widzenia precyzyjna definicja konstruktu w psychologii jest rzadkością. Konstrukt jest często zarówno podmiotem, jak i wskaźnikiem podmiotu, co jest samo w sobie sprzeczne z punktu widzenia logiki. Z tego powodu nie istnieją w psychologii ustalone i stałe konstrukty niezależne od rzeczywistości intrapsychicznej – są one badane (klasyfikowane) w zależności od obserwacji poczynionych zawsze w jakimś kontekście: intraindywidualnym, społecznym, kulturowym itd. Przykładem mogą tu być zmieniające się z wydania na wydanie definicje zaburzeń psychicznych i zaburzeń zachowania w kolejnych edycjach DSM (pełny angielski tytuł książki wydawanej przez APA to Diagnostic and Statistical Manual of Mental Disorders, a doczekała się ona już pięciu wydań).

Rzeczywistość badana w psychologii jest nie tylko subiektywna ze względu na podmiot badania (człowiek i jego stany psychiczne), ale także w dużym stopniu zależy od procesu badawczego, ponieważ konstrukty psychologiczne są obiektami społecznymi. Atrybutów psychicznych, takich jak lęk, emocje czy stres, nie można jednoznacznie przypisać do reakcji fizykochemicznych w organizmie. Ponieważ

przedmiot badań w psychologii znajduje się między pojęciem a rzeczywistością, do badań i analiz w tej dziedzinie niezbędny stał się termin k o n s t r u k t p s y c h o l o g i c z n y. Służy on jako pomost pomiędzy nierozpoznawalną rzeczywistością fizyczną (stan i funkcje mózgu i/lub organizmu) a rzeczywistością społeczną atrybutu psychicznego (opartą na przejawach, a zatem ontologicznie subiektywną). Dlatego w psychologii wszystko to, co możemy badać, opiera się na pomiarze konstruktu psychologicznego.

1.2. Klasyczna teoria pomiaru

Przyjmując definicję tego, co jest przedmiotem zainteresowania psychologii, wróćmy do analizy procesu p o m i a r u konstruktów psychologicznych. Próby pomiaru takich konstruktów psychologicznych, jak zdolności poznawcze, cechy osobowościowe lub postawy, spotkały się z różnego rodzaju krytyką. Na przykład Stephen F. Blinkhorn (1998) wskazywał na ograniczone możliwości zastosowania wyników analiz psychometrycznych w testach praktycznych, James Lumsden (1976) podkreślał niechęć psychometrii do korzystania z nowych koncepcji i koncentrowanie się na starych ideach. Zauważył też u psychologów tendencję do korzystania z małych i nieistotnych innowacji technicznych. Moim zdaniem jednak najpoważniejszy argument dotyczy bezkrytycznej akceptacji przez psychologów (i w znacznej części też psychometrów, choć te dwa zbiory nie są rozłączne) założenia, że badane konstrukty mogą być określane ilościowo, a tym samym mierzone bez obawy o konieczność udowodnienia zasadności tego założenia (Borsboom, 2006; Fraser, 1980; Michell, 2000), tak jakby w psychologii sam pomiar usprawiedliwiał to założenie. Problem ten jest konsekwencją powszechnie przyjętej operacyjnej teorii pomiaru zaproponowanej przez Stanleya S. Stevensa (1946) z jej pozornymi zaletami i poważnymi wadami (choćby taką, że wszystko, poza losowym przypisywaniem liczb do obiektów, jest pomiarem, co implikuje, iż każdy psychologiczny konstrukt można mierzyć). Konsekwentnie ignorowany jest fakt, że tylko ilościowe (policzalne) atrybuty (*quantitative attributes*) mogą być w ten sposób mierzone, oraz to, że ilościowe miary mają określoną strukturę i przed każdym badaczem stoi zadanie empirycznego udowodnienia, iż konstrukt, który ma być przedmiotem badania, również taką strukturę posiada (Michell, 1999, s. xi).

Przyjrzę się teraz temu, w jaki sposób paradygmat o p e r a c j o n a l i z m u zdominował pomiar w psychologii. Pomiar zawsze był uważany za akt klasyfikacji, identyfikacji, opisu lub porównania, a czynności pomiarowe polegały na przypisaniu liczb – są one wygodnym (precyzyjnym i jednoznacznym) denotatem wyobrażeń o rzeczach. Pomiar opisać można zatem jako metodę pozwalającą uzyskać wzajemnie jednoznaczną odpowiedniość między psychologicznym (psychicznym) stanem rzeczy a liczbami. Większość badań empirycznych w psychologii polega na jednym z dwóch sposobów pomiaru. Zgodnie z pierwszym zliczana

jest częstotliwość reakcji osoby uczestniczącej w badaniu na określony zestaw bodźców. Dla każdego bodźca istnieje predefiniowana reakcja poprawna (zgodna z wzorcem) i reakcja (lub reakcje) niepoprawna. Pomiar polega na policzeniu liczby reakcji (częstości) i przypisaniu im liczby bądź kategorii ilościowej. Ten sposób pomiaru wywodzi się od eksperymentów Fechnera zajmującego się wrażliwością sensoryczną (*intensity of sensation*). Każda prezentacja bodźca dawała 'szansę' systemowi poznawczemu badanego na prawidłową reakcję. Ponieważ trudno udowodnić założenie, że intensywność reakcji na bodziec jest analogiczna do intensywności samego bodźca, proces pomiaru polegał na prezentowaniu bodźców w parach i porównywaniu ich przez badanego (dla porządku warto zauważyć, że Fechner sugerował, iż związek między wielkością bodźca a wielkością odpowiedzi układu nerwowego jest logarytmiczny; por. Michell, 1999, s. 81). Dzięki pracom Thurstone'a (1927) w pomiarze psychologicznym pojawiło się odniesienie do rozkładu normalnego (krzywej Gaussa) i opracowany został aparat matematyczny pozwalający oszacować najbardziej prawdopodobną liczbę dla wielu porównań, którą przyjmowało się za miarę mierzonej wrażliwości sensorycznej.

Drugi sposób pomiaru, jak już wspomniałem wcześniej, miał źródło w obserwacji różnic w zakresie zdolności intelektualnych ludzi (Binet, 1903 i Spearman, 1904, za: Michell, 1999, s. 9). Osobom uczestniczącym w badaniu prezentowano (dziś powiedzielibyśmy – wystandaryzowany) zbiór zadań, a miarą intelektu była liczba zadań (dziś powiedzielibyśmy – pozycji testowych) poprawnie rozwiązanych.

Do początku XX wieku filozoficzne podejście do pomiaru opierało się w dużej mierze na zasadzie idealizmu platońskiego zakładającego, że istnieje określona właściwość obiektu, który ma być zmierzony, a pomiar ma na celu przypisanie wartości liczbowej, która dokładnie identyfikuje tę właściwość. Mierzalne właściwości musiały spełniać określone warunki, sformułowane przez Otto Höldera (za: Michell, 1999, s. 52) w jego definicji (struktury) ilości. I tak jeśli a, b oraz c są różnymi poziomami mierzonej właściwości, to zachodzą dla tej właściwości następujące zależności:[6]

1. Zachodzi jedno z równań: $a > b$, $a < b$ lub $a = b$.
2. Dla każdego a istnieje takie b, że $a > b$.
3. Dla każdego a oraz b istnieje takie c, że $c = a + b$.
4. $a + b > a$ oraz $a + b > b$.
5. Dla każdego a oraz b istnieją takie x oraz y, że $a + x = b$ oraz $y + a = b$.
6. $(a + b) + c = a + (b + c)$.
7. Jeżeli $a > b$ oraz $b > c$, to $a > c$.

Należy podkreślić, że symbole =, <, > nie oznaczają relacji między obiektami jako takimi, ale relację między właściwościami tych obiektów.

[6] Niewielka poczyniona tutaj modyfikacja dotyczy 7. aksjomatu. W oryginale dotyczy on rozłącznych przedziałów, tutaj *stricte* poziomów mierzonej właściwości.

Teoretycznymi podstawami pomiarów fizycznych zajmował się m.in. Norman Robert Campbell (1880–1949), który widział pomiar jako demonstrację izomorfizmu[7] między ideą ilości a wielkością mierzonej właściwości (Campbell, 1920). Według tego autora pomiar polegał więc na obserwowaniu relacji pomiędzy obiektami fizycznymi w wyniku wykonania operacji empirycznej. Tylko po zdefiniowaniu operacji służącej testowaniu ilości można było porównywać między sobą mierzone obiekty.

Mimo że siedem aksjomatów uporządkowania i addytywności Höldera było empirycznie testowalnych dla właściwości fizycznych (np. masa – za pomocą wagi), to jednak istnieją takie właściwości fizyczne, których nie można empirycznie sprawdzić, jak gęstość lub twardość – można uporządkować obiekty według tych właściwości, ale nie można już ich dodawać (np. temperatura – nie można uzyskać wiadra wody o temperaturze 50º z dwóch półwiaderek z wodą o temperaturze 25º). Posiadanie addytywności posłużyło Campbellowi do wprowadzenia rozróżnienia na dwie mierzalności (*properties in terms of their measurability*): miary szerokie (*extensive – open to fundamental measurement*) oraz miary wąskie (*intensive – open to derived measurement*), które pozwalało przyjąć tezę, że miary pierwszego rodzaju mogą „stwarzać" te drugie w wyniku operacji matematycznych. Co ważne, relacje między obiektami mierzonymi za pomocą miar szerokich (relacja empiryczna, doświadczalna) muszą być zachowane także pomiędzy miarami wąskimi wyliczonymi z miar szerokich. Campbell wyraźnie uznał, że relacje zachodzą tylko między obiektami fizycznymi, dlatego też próby pomiaru muszą być podejmowane w odniesieniu do możliwych do empirycznego zaobserwowania relacji między obiektami rzeczywistymi[8].

Takie definiowanie pomiaru wyprowadza tylko jeden rodzaj skali – skalę ilościową z jednym zbiorem właściwości. Dany atrybut, aby być mierzalny, musi spełniać je wszystkie, a teoria pomiaru Campbella nie oferowała możliwości rozszerzenia jej o atrybuty psychologiczne. Campbell, będąc członkiem komitetu (pod przewodnictwem fizyka Allana Fergusona) powołanego do zbadania, czy doznania (*sensory events*) mogą być mierzone, wywarł wpływ na przyjęcie rezolucji wykluczającej praktycznie psychologię z nauk empirycznych. Mimo głosów za utworzeniem dla właściwości psychicznych nowej kategorii obiektów do mierzenia (Richardson, 1933, s. 587, za: Michell, 1999, s. 145) w raporcie końcowym (Ferguson et al., 1940) stwierdzono, że:

> Każde prawo mające na celu wyrażenie ilościowego związku między intensywnością wrażeń a intensywnością bodźców jest nie tylko nieprawdziwe, ale

[7] Właściwość zbiorów polegająca na tym, że wyniki odpowiednich działań na przyporządkowanych sobie elementach tych zbiorów są również sobie przyporządkowane.

[8] Chociaż oznacza to, że skale mogą być tylko względne. Aby były absolutne, trzeba sporządzić wzorce, normy będące punktem odniesienia.

> w rzeczywistości wprowadza w błąd, chyba że i do czasu, gdy będzie można nadać znaczenie pojęciu addytywności w odniesieniu do wrażeń[9].

Dyskusja członków komitetu objęła temat różnic zmysłowych (*sense-distances*). I mimo krytyki (głównie Guild, zob. Michell, 1999, s. 147), że: 1) różnice w postrzeganiu nie są tożsame z różnicami między postrzeganymi obiektami; 2) nie istnieje przechodnia i symetryczna relacja między różnicami w postrzeganiu; 3) obserwowanie różnic w postrzeganiu w eksperymentach psychologicznych nie dowodzi ich istnienia, w murze odgradzającym psychologię od nauk empirycznych pozostawiono szczelinę. Dopuszczono, iż obserwacje zmysłowe są możliwe do uporządkowania, a stąd wynikał bezpośredni wniosek, że skoro można je porządkować, to znaczy, że można je mierzyć.

Reakcje na wykluczenie psychologii z grona nauk ścisłych były różne. John Drever (1938, s. 328) stwierdzał, że różnice w postrzeganiu nie muszą mieć jednostek pomiaru. Ponownie pojawiły się głosy, by pomiar psychologiczny nie był utożsamiany z pomiarem fizycznym. Sir Frederic Charles Bartlett posunął się do skrajnego pesymizmu, argumentując, że chociaż być może to, czym zajmuje się psychologia, nie jest pomiarem (dodam, że w sensie ilościowym), ale czymś się jednak zajmuje, więc niech to robi dalej (Michell, 1999, s. 154). Taka bierna postawa wobec ustaleń komitetu Fergusona była charakterystyczna dla angielskich przedstawicieli psychologii. W tamtym okresie silne w Anglii i Europie kontynentalnej były zainteresowania jakościowe, by przytoczyć tu choćby wypowiedź Williama Browna (1938, s. 52): „przedmiotem obserwacji w psychologii są głównie właściwości jakościowe, a nie ilościowe”[10].

Wracając do raportu komitetu Fergusona, dwie reakcje środowiska psychologów są warte odnotowania: jedna, która została zupełnie zapomniana, i druga, która doprowadziła do powstania definicji pomiaru zaproponowanej przez Stevensa. I tak Thomas W. Reese (1943), zamiast odrzucać fizyczną definicję pomiaru, dokonał jej interpretacji, pokazując, że każdy rodzaj pomiaru (i fizyczny, i psychiczny) wymaga wykazania empirycznych relacji porządkowych czy też addytywnych mierzonych obiektów. W tym sensie pomiar fizyczny nie różni się od pomiaru psychicznego, ponieważ w obu dziedzinach nauki wymagane jest jego empiryczne zdefiniowanie.

> Prawdą jest, że żadna subiektywna wielkość nie została zmierzona w sposób fundamentalny. Wiara autora, że mogą one być tak zmierzone, jest hipotezą. Tylko eksperymentowanie może dać odpowiedź (Reese, 1943, s. 49)[11].

---

[9] „Any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact misleading unless and until a meaning can be given to the concept of addition as applied to sensation”.

[10] „[…] the observation of psychology are primarily qualitative, not quantitative”.

[11] „It is true that no subjective magnitude has been measured fundamentally. The belief of the author that they may be so measured is an hypothesis. Only experimentation can give the answer”.

Ten obiecujący głos rozsądku został zaprzepaszczony przez Joya P. Guilforda i Andrew L. Comreya, którzy zaproponowali, by traktować pomiar na podstawie relacji uporządkowania jako jeden koniec kontinuum, na którego drugim końcu są obiekty posiadające właściwości addytywne. W ich rozumieniu pomiar właściwości psychologicznych leży gdzieś pomiędzy tymi krańcami. To, co Komitet Fergusona, w tym Campbell, traktował jako dychotomię między pomiarem a jego brakiem, Guilford i Comrey proponowali traktować jako stopień realizacji idealnego pomiaru. Odwracając kierunek interpretacji znaczenia w pomiarze z: „od obiektów do liczb" na: „od liczb do obiektów", twierdzili, że liczby (uzyskiwane w pomiarze) nabierają znaczenia zgodnie z operacją ich przypisywania obiektom (Guilford i Comrey, 1951):

> podane liczby przypisane do każdego pomiaru są jedynie wyrazem wykonywanych operacji[12].

W tej sytuacji, ignorując reprezentacjonistyczne podejście Campbella do pomiaru, a jednocześnie się na Campbella powołując (o dziwo, rzekomo parafrazując jego definicję, ale jako autora podając bezimiennego „członka komitetu Fergusona", Stevens, 1946, s. 680) Stanley S. Stevens przedstawił swoją szeroką definicję pomiaru, jakby była rozszerzeniem definicji Campbella, utrudniając mu tym samym ewentualną krytykę (Michell, 1999, s. 161). Definicja szybko okazała się najważniejszą definicją pomiaru w psychologii i jeśli nie zakończyła, to mocno ograniczyła teoretyczne rozważania na temat pomiaru w psychologii (zob. Michell, 1999, s. 163).

Pierwszą zaletą definicji Stevensa było ominięcie ograniczenia, zgodnie z którym liczby użyte jako wyniki pomiaru muszą reprezentować empiryczne ilości. Drugą innowacją było wprowadzenie możliwości stosowania innych rodzajów skal (poziomów pomiaru) poza ilościową. Podejście Stevensa polegało na zdefiniowaniu pomiaru jako „przypisywania liczb do obiektów lub zdarzeń zgodnie z regułami" (Stevens, 1946), a następnie na zdefiniowaniu tych reguł. Typ (poziom) pomiaru miał być klasyfikowany zgodnie z empirycznymi operacjami opisującymi elementy pomiaru. W ten sposób Stevens rozszerzył jednoskalowy system Campella na cztery skale: nominalną (operacje równości), porządkową (operacje mniejsze lub większe), interwałową (operacje równości interwałów lub różnic) i proporcjonalną (operacje równości proporcji).

Definicja Stevensa *de facto* nie zniosła ważnego ograniczenia podejścia Campbella – wciąż istnieje rozróżnienie między systemem formalnym (liczby) a system empirycznym (właściwości). Przesunął się po prostu akcent w pomiarze z traktowania relacji empirycznych jako nadrzędnych, na akcentowanie właściwości skali, na której przedstawiony jest wynik pomiaru. Definicja zakłada rozdział między

[12] „[…] the meanings given numbers assigned in any type of measurement are merely an expression of the operations performed".

właściwościami obserwowanych obiektów a właściwościami reprezentującego je systemu formalnego (liczby), przy czym relacja ta musi być izomorficzna – skala pomiarowa musi mieć te same właściwości relacyjne co opisywane obiekty. Określenie skali pomiarowej stało się warunkiem wstępnym pomiaru oderwanym od r e a l n i e m i e r z o n y c h o b i e k t ó w. Zaczęto utożsamiać właściwości skali użytej do pomiaru danego konstruktu z właściwościami konstruktu, co samo w sobie nie jest nieprawidłowe, jeśli kierunek relacji pochodziłby od mierzonego obiektu do skali pomiarowej, a nie od wybranej skali do obiektu. Pierwsza sytuacja ma miejsce w naukach przyrodniczych, w których pomiar jest odkrywaniem stosunku mierzonej właściwości do wzorca. To wymusza zdefiniowanie najpierw wzorca. W psychologii obowiązuje pomiar zdefiniowany przez Stevensa – przypisuje się więc liczby obiektom lub zdarzeniom. Te dwie definicje wykluczają się wzajemnie, co sugeruje, że jedna z nich jest błędna.

Problem z definicją pomiaru Stevensa polega na tym, że jest ona zbyt szeroka i „istnieje ryzyko w pomiarze edukacyjnym i psychologicznym, że prawie każdy może opracować swoje własne reguły przypisywania jakichś liczb do jakichś wyników" (Suen, 1990, za: Michell, 1999, s. 18). Ryzyko błędu dotyczy sytuacji, w której do wyników z konkretnego testu używane mogą być różne reguły kwantyfikujące ilość (poziom, natężenie) konstruktu psychologicznego. Jest to dozwolone z punktu widzenia pomiaru, tak jak jest on rozumiany przez Stevensa, ale mało prawdopodobne ze względów praktycznych. Testy będące w użyciu są albo przygotowane przez badaczy i opublikowane, co oznacza, że wyposażone są w informację o sposobie przeliczania wyników (czyli reguła jest jedna i stała dla testu), albo są testami przygotowanymi *ad hoc* na potrzeby danego badania, a wtedy ryzyko, że zostaną zastosowane różne sposoby przeliczania wyników, jest minimalne, ponieważ nie są stosowane szeroko (przeważnie używane są tylko przez swojego autora). Jeśli test zmienia kategorię drugą (*ad hoc*) na pierwszą (opublikowany), musi zostać wyposażony w informacje o przeliczaniu wyników (American Educational Research Association…, 1999, s. 116)[13].

1.3. Aksjomatyczna teoria pomiaru

Paradoksalnie wprowadzenie (jak wykazałem powyżej, problematycznej) Stevensowskiej klasyfikacji skal pomiarowych konstytuowało psychologię jako naukę empiryczną. Najsłabszym punktem pomiarów psychologicznych było założenie, że reguły stosowane do szacowania wielkości konstruktów w obserwowanych

[13] Najnowsze wydanie polskie z roku 2007 jest tłumaczeniem Standardów dla testów… wydanych w 1999 roku. Istnieje nowsze wydanie, z 2014 roku, nieróżniące się zbytnio od wcześniejszego wydania z 1999 roku. Różnicami między kolejnymi wydaniami Standardów dla testów… zajmuję się w rozdziale 3. Podając strony, będę się odwoływał do polskiej wersji z 2007 roku.

wynikach (np. sumowanie odpowiedzi) są poprawne z punktu widzenia definicji pomiaru ilościowego. W testach składających się z wielu pozycji istnieje duże prawdopodobieństwo, że osoby o tej samej sumie punktów rozwiązały poprawnie inne zestawy zadań, a traktowane są tak, jakby ich poziom konstruktu był taki sam. Nie jest to prawdziwe nawet wtedy, gdy sumowane elementy mają tą samą wagę i tę samą zawartość konstruktu. A problem jest jeszcze bardziej skomplikowany, bo należy wspomnieć o interakcji między konstruktami (wszak człowiek jest 'nosicielem' wielu konstruktów). Udzielone przez badanych odpowiedzi na dane zadanie, nawet identyczne odpowiedzi, są wypadkową całych szeregów skojarzeń i wybranie/wskazanie ich w kwestionariuszu może mieć swoje źródło w zupełnie nieprzystających do siebie stanach osób badanych (Tourangeau, 1984).

Mimo że zakres pomiarów podstawowych był szerszy, niż twierdził Campbell, to jednak dopiero po pojawieniu się teorii pomiaru łącznego (*additive conjoint measurements*) z jej dość prostymi technikami oraz możliwymi zastosowaniami w naukach społecznych i fizycznych – potrzeba rewizji definicji pomiaru została zauważona. Aksjomatyczna teoria pomiaru umożliwiła stworzenie formalnego aparatu dowodzenia istnienia zmiennych mierzalnych i ciągłych. Opracowana została przez R. Duncana Luce'a i Johna Tukeya (1964)[14]. Autorzy ci w swoim artykule udowodnili, że konstrukty niezdolne do konkatenacji mogą być kwantyfikowalne. Tym samym stanowisko komitetu Fergusona jest błędne, a to, czy konstrukt psychologiczny jest wielkością mierzalną, może zostać sprawdzone w sposób empiryczny. Prace kolejnych badaczy (Krantz, 1964; Kruskal, 1964; Scott, 1964; Tversky, 1967) rozszerzyły tę teorię pomiaru na wiele połączonych zmiennych (w nomenklaturze teorii: atrybutów), a pełny opis teorii znalazł się w pierwszym tomie monografii *Podstawy pomiaru* Davida H. Krantza, R. Duncana Luce'a, Patricka Suppesa i Amosa Tversky'ego (Krantz et al., 1971).

W tym miejscu należy zaprezentować podstawy aksjomatyczne teorii pomiaru, głównie korzystając z prac Davida H. Krantza i wsp. (1971), Louisa Narensa i R. Duncana Luce'a (1986) oraz Joela Michella (2014). Wyrażenie a d d y t y w n y p o m i a r ł ą c z n y określa uwzględnianie wpływu wszystkich atrybutów na mierzoną zmienną (zależną). Aksjomatyczna teoria pomiaru (*theory of conjoint measurement*) przez spełnienie jej aksjomatów dowodzi mierzalności atrybutów jako przesłanek oraz mierzalności zmiennej zależnej. I tak niech dwa atrybuty ekstensywne (naturalne lub inaczej obserwowalne) *A* oraz *X* odnoszą się do trzeciego atrybutu *P*. Niech a_1, a_2 i a_3 są trzema niezależnymi poziomami atrybutu *A*; x_1, x_2 i x_3 są analogicznie poziomami atrybutu *X*. Trzeci atrybut *P* składa się z dziewięciu par poziomów *A* oraz *X*, tj.: (a_1, x_1), (a_1, x_2), … (a_3, x_3) (tab. 1.1).

[14] Była to algebraiczna prezentacja założeń teorii, a w tym samym roku topologiczne podstawy teorii opisał ekonomista Gérard Debreau. Historycznie pierwszą pracą przewidującą cechy teorii pomiaru łącznego była praca matematyka Otto Höldera (1901).

Tabela 1.1. Wartości zmiennej *P* w powiązaniu z poziomami atrybutów *A* i *X*[15]

	x_1	x_2	x_3
a_1	(a_1, x_1)	(a_1, x_2)	(a_1, x_3)
a_2	(a_2, x_1)	(a_2, x_2)	(a_2, x_3)
a_3	(a_3, x_1)	(a_3, x_2)	(a_3, x_3)

Ocena mierzalności *A*, *X* i *P* zależy od relacji między poziomami *P*. Te relacje są przedstawione jako aksjomaty w teorii łącznego pomiaru.

1.3.1. Aksjomat niezależności (*single cancellation axiom*)

Relacje między poziomami zmiennej *P* spełniają aksjomat niezależności wtedy i tylko wtedy, gdy dla wszystkich *a* w *A* oraz dowolnego *x* w *X* spełniony jest warunek

$$(a_i, x_k) > (a_i + 1, x_k),$$

gdzie $i = \{1, 2\}$, a $k = \{1, 2, 3\}$. Analogicznie dla wszystkich *x* w *X* oraz dla dowolnego *a* w *A* spełniony jest warunek

$$(a_i, x_k) > (a_i, x_k + 1),$$

gdzie $i = \{1, 2, 3\}$, a $k = \{1, 2\}$. Oznacza to, że jeśli dowolne poziomy *a* w *A* są uporządkowane, to uporządkowanie jest zachowane na każdym poziomie *X*. Analogicznie jest dla poziomów *x* w *A*. Aksjomat ten nazywany jest aksjomatem pojedynczego unieważnienia za względu na to, że dowolny wspólny poziom zmiennej *P* unieważnia się, aby pozostawić to samo uporządkowanie dla pozostałych poziomów, np. dla:

$$(a_1, x_1) > (a_1, x_2)$$

redukcja wspólnego a_1 pozostawia $x_1 > x_2$[16].

Spełnienie aksjomatu niezależności jest warunkiem niezbędnym, ale niewystarczającym do udowodnienia ilościowego charakteru zmiennej *P*. Spełniony aksjomat udowadnia tylko istnienie porządku w poziomach *P*. Stwierdzenie uporządkowania w poziomach *P* jest jednak niewystarczające do pełnego określenia uporządkowania w *A* i *X*. Z aksjomatu niezależności można wyprowadzić (zbadać) tylko zależności „*left-leaning diagonal*", czyli „skośnie-rosnąco w lewo". Część zależności pozostaje niemożliwa do zbadania. Na przykład niech będzie prawdziwe, że:

[15] Źródłem tabel, jeśli nie podano inaczej, są badania własne autora.

[16] Jeśli wyobrazimy sobie jeden atrybut jako wielkość pojemnika, a drugi jako rodzaj substancji wlewanej do pojemnika, to na przykład szacując uporządkowanie gęstości według masy dwóch substancji, możemy pominąć kwestie pojemnika, jeśli są tożsame.

$$(a_1, x_1) > (a_2, x_1)$$

oraz

$$(a_1, x_1) > (a_2, x_2),$$

wtedy przez zasadę przechodniości

$$(a_1, x_1) > (a_2, x_2),$$

ale relacja

$$(a_1, x_2) \ldots (a_2, x_1)$$

na bazie aksjomatu pojedynczego unieważnienia pozostaje nieokreślona.

1.3.2. Aksjomat addytywności (*double cancellation axiom*)

Ponieważ może być prawdziwa tylko jedna relacja:
albo

$$(a_2, x_1) > (a_1, x_2),$$

albo

$$(a_2, x_1) < (a_1, x_2),$$

do rozstrzygnięcia tego problemu stosuje się aksjomat podwójnego unieważniania. Postępowanie wygląda w ten sposób, że dwie nadrzędne nierówności unieważniają się, dowodząc trzeciej nierówności, np. zakładając, że

$$(a_1, x_2) > (a_2, x_1)$$

oraz

$$(a_2, x_3) > (a_3, x_2),$$

to

$$(a_1, x_2) > (a_2, x_1)$$

jest prawdziwe wtedy i tylko wtedy, gdy

$$a + y > b + x,$$

a

$$(a_2, x_3) > (a_3, x_2)$$

jest prawdziwe wtedy i tylko wtedy, gdy

$$a_2 + x_3 > a_3 + x_2.$$

Z tego wynika, że

$$a_1 + x_2 + a_2 + x_3 > a_2 + x_1 + a_3 + x_2.$$

Po obustronnym usunięciu wspólnych poziomów wynik brzmi:

$$(a_1, x_3) > (a_3, x_1).$$

Stąd podwójne unieważnianie może zachodzić wtedy i tylko wtedy, gdy *A* oraz *X* są policzalne.

Podwójne unieważnianie jest prawdziwe wtedy i tylko wtedy, gdy nie stoi w sprzeczności z uprzednimi nierównościami. Na przykład, jeśli jako wniosek z powyższych nierówności wynikałoby, że:

$$(a_1, x_3) < (a_3, x_1)$$

lub

$$(a_1, x_3) = (a_3, x_1),$$

to aksjomat nie byłby spełniony i *A* oraz *X* nie byłyby ilościowymi miarami uporządkowanymi. Aksjomat podwójnego unieważniania pozwala zbadać tzw. „*right-leaning diagonal*" – zależności między poziomami *P*, co jest niemożliwe tylko na bazie aksjomatu pojedynczego unieważniania.

1.3.3. Aksjomaty ciągłości (*solvability and Archimedean axiom*)

Aksjomaty pojedynczego i podwójnego unieważniania nie wystarczają z kolei do udowodnienia ciągłości atrybutów *A* i *X*. W tym celu należy sprawdzić dwa kolejne warunki.

Pierwszy to aksjomat rozwiązywalności, który oznacza, że dla znanych dowolnych trzech elementów z czteroelementowego zbioru a_1, a_2, x_1, x_2 czwarty element jest taki, że równanie

$$a_1 \cdot x_1 = a_2 \cdot x_2$$

jest rozwiązywalne. Inaczej mówiąc, oznacza to warunek, żeby każdy poziom *P* miał odpowiadający mu poziom w *A* oraz w *X*. Ponadto aksjomat rozwiązywalności jest warunkiem, który zachodzi, gdy poziomy *A* i *X* są tej samej gęstości (*dense*) jak liczby rzeczywiste, lub też równoodległe jak liczby całkowite (Krantz et al., 1971).

Aksjomat drugi, archimedejski, jest następujący. Niech *i* będzie zbiorem kolejnych dowolnych liczb całkowitych. Poziomy *a* tworzą ciąg standardowy wtedy i tylko wtedy, gdy istnieje x_1 i x_2 w *X*, gdzie $x_1 \neq x_2$ oraz dla wszystkich liczb całkowitych *i* oraz *i* + 1 w *I* zachodzi równość

$$(a_i, x_1) = (a_i + 1, x_2).$$

Spełnienie tej równości oznacza, że jeśli x_1 jest większe niż x_2, to istnieje poziom *A*, dla którego dwa odpowiednie poziomy *P* są równe (Scott, 1964). Inaczej mówiąc, żadna niezerowa zmiana nie jest 'nieskończenie mała' w porównaniu z jakąkolwiek inną zmianą (Luce i Tukey, 1964), co jest przesłanką wniosku, że *P* jest ciągłe.

1.3.4. Zastosowanie aksjomatycznej teorii pomiaru

Teoria pomiaru aksjomatycznego stanowi niezwykle mocne narzędzie sprawdzające właściwości używanej skali pomiarowej. Prawdopodobnie najbardziej znanym przykładem jej użycia jest udowodnienie prawdziwości teorii perspektywy Kahnemana i Tversky'ego, która wyjaśnia działanie heurystyk w przetwarzaniu informacji probabilistycznych i tłumaczy podejmowanie decyzji (P) w sytuacji ryzyka (A) i niepewności (X). Mimo to teoria pomiaru aksjomatycznego w psychologii spotkała się z umiarkowanym zainteresowaniem. Z pewnością ma na to wpływ wymagany przez nią wysoki poziom znajomości matematyki formalnej oraz konieczność przeprowadzenia wielu analiz – tym większej ich liczby, im więcej poziomów mają atrybuty. Na przykład do sprawdzenia prawdziwości aksjomatu podwójnego unieważniania dla zmiennych A i X posiadających po 3 poziomy trzeba przeprowadzić $3! \cdot 3! = 36$ porównań. Co prawda Michell (1988) wykazał, że jeśli prawdziwe są aksjomaty pojedynczego unieważniania, to z 36 aksjomatów podwójnego unieważniania aż 30 jest w oczywisty sposób prawdziwych, a pozostałych 6 jest ze sobą w relacji takiej, że jeśli jeden jest prawdziwy, to wszystkie pozostałe są prawdziwe, co znacząco redukuje złożoność obliczeniową. Michell nazwał ten warunek przypadkiem Luce'a–Tukeya. Podobnie, dla sprawdzania ciągłości istnieją przypadki redukujące liczbę koniecznych do przeprowadzenia porównań (Scott, 1964)[17]. Niestety w praktyce sprawdzanie aksjomatów ciągłości, choć możliwe, jest nieekonomiczne ze względu na czas potrzebny na testowanie nierówności na realnych danych (Karabatsos, 2018; Michell, 1990, s. 79).

1.3.4.1. Przykład analizy: czy inteligencja mierzona testem Omnibus jest konstruktem ilościowym?

Warte uwagi z praktycznego punktu widzenia są aplikacje do badania wyników testów w świetle aksjomatycznej teorii pomiaru – jedna, oparta na sprawdzaniu zestawu równań na bazie uzyskanych wyników, zaimplementowana jest w środowisku R, druga – w środowisku Matlab. Pierwsze podejście stosuje analizy symulacyjne z wykorzystaniem łańcuchów Markowa (Domingue, 2014; Heene, Kyngdon i Sckopke, 2016; Kyngdon, 2011) do oszacowania odsetka przypadków w macierzy wyników niespełniających dwóch pierwszych aksjomatów. Przykład jej zastosowania prezentuję poniżej. Druga, wykorzystująca przybliżone analizy bayesowskie (*approximate Bayesian computation*) (Karabatsos, 2006, 2018; Turner i Van Zandt, 2012), pozwala także testować prawdziwość aksjomatów wyższego rzędu (*triple*, *quadruple*) na dużych zbiorach danych w rozsądnym czasie.

[17] Na przykład dla A i P posiadających po 3 kategorie wystarczy udowodnić jeden aksjomat rozwiązywalności, zaś dla $4 \cdot 4$ kategorii wystarczy, by był spełniony jeden aksjomat trzeciego rzędu (*triple cancellation axiom*).

W przykładzie wykorzystam dane zebrane podczas badań przeprowadzonych na grupie studentów Instytutu Psychologii w roku 2011. W badaniu użyto testu Omnibus badającego inteligencję (Jaworowska, Matczak i Ciechanowicz, 2002). Test składa się z 60 zadań z pięcioma możliwymi do wyboru odpowiedziami[18]. Zadania dotyczą wyboru poprawnych antonimów, sylogizmów, analogii werbalnych, uzupełnienia szeregów liczb i wyrażeń frazeologicznych. Każdy z pięciu typów zadań występuje 12 razy, w zestawach po 3 zadania. Trudność zadań jest narastająca, ale mogą być one wykonywane w dowolnej kolejności. Poprawne odpowiedzi punktowane są taką samą wagą, a wynik w teście stanowi suma punktów. Poza wynikiem ogólnym na podstawie analizy czynnikowej autorki testu wyróżniły dwa czynniki: wiedzy oraz rozumowania (po 25 zadań). Rzetelność (spójność wewnętrzna) waha się od 0,78 dla czynnika wiedzy w normalizacyjnej grupie studentów do 0,93 dla wyniku ogólnego w grupie osób dorosłych. W badaniu wykorzystano 117 kompletnych wyników (stosunek płci M/K: 0,35). Badani byli w wieku od 18 do 40 lat (*Md* = 22, M = 22,8, *SD* = 3,93).

W aplikacji Domingue'a (2014) w pierwszym kroku wyznaczana jest macierz wyników względem sum osiągniętych przez osoby uczestniczące w badaniu (kolumny), które są zagregowane według sumy wyników (rzędy). Następnie wyznaczana jest macierz poprawnych odpowiedzi[19]. W kolejnym kroku kolumny macierzy są sortowane według trudności poszczególnych zadań, co pozwala uzyskać wejściową macierz uporządkowaną zgodnie z założeniami aksjomatycznej teorii pomiaru wzdłuż wierszy (rosnące wyniki osób) oraz kolumn (coraz łatwiejsze zadania). Analizując macierz (jej fragment przedstawia tab. 1.2), widzimy, że zadanie 10 było trudniejsze niż 9, podobnie 9 było trudniejsze niż 18 itd. Wśród osób, które uzyskały przykładowo sumaryczny wynik 33 punkty, cztery odpowiedziały na zadanie 10, sześć na zadanie 9, 18, 29 itd. Z analizy pełnej macierzy wynika, że najłatwiejsze było zadanie 45, najtrudniejsze zaś zadanie 50.

Tabela 1.2. Fragment macierzy z częstością udzielonych odpowiedzi na dane zadanie wśród osób badanych względem sumy punktów

Suma	zad10	zad9	zad18	zad29	zad26	zad21	zad5	zad3	zad13	zad14	zad23
31	0	2	2	0	2	2	2	0	0	2	2
32	2	2	4	4	4	4	4	2	2	2	0
33	4	6	6	6	6	4	4	6	4	6	6
34	2	4	4	4	0	4	2	2	2	2	4
36	2	4	4	2	4	4	2	4	0	4	4

[18] Zadania w postaci sylogizmów mają po trzy możliwe odpowiedzi.

[19] Prezentowana metoda jest opracowana tylko dla odpowiedzi dychotomicznych, a jej autor nie rozwija jej niestety w kierunku stosowania do wyników wielokategorialnych.

Suma	zad10	zad9	zad18	zad29	zad26	zad21	zad5	zad3	zad13	zad14	zad23
37	2	2	0	2	2	0	2	0	2	2	2
38	2	2	0	2	2	2	2	0	0	2	0
40	2	2	0	2	2	2	2	2	0	2	2
41	2	0	2	2	2	2	2	2	2	2	0

W pierwszej kolumnie znajduje się osiągnięta suma punktów. Zadania (kolumny) posortowane są od łatwiejszych do trudniejszych.

Tak przygotowaną macierz przekształca się na macierz prawdopodobieństwa, dzieląc liczbę poprawnych odpowiedzi przez liczbę osób, które uzyskały poszczególne sumy punktów w całym teście. Fragment przykładowej macierzy przedstawia tab. 1.3.

Tabela 1.3. Fragment macierzy z prawdopodobieństwem udzielenia poprawnych wyników

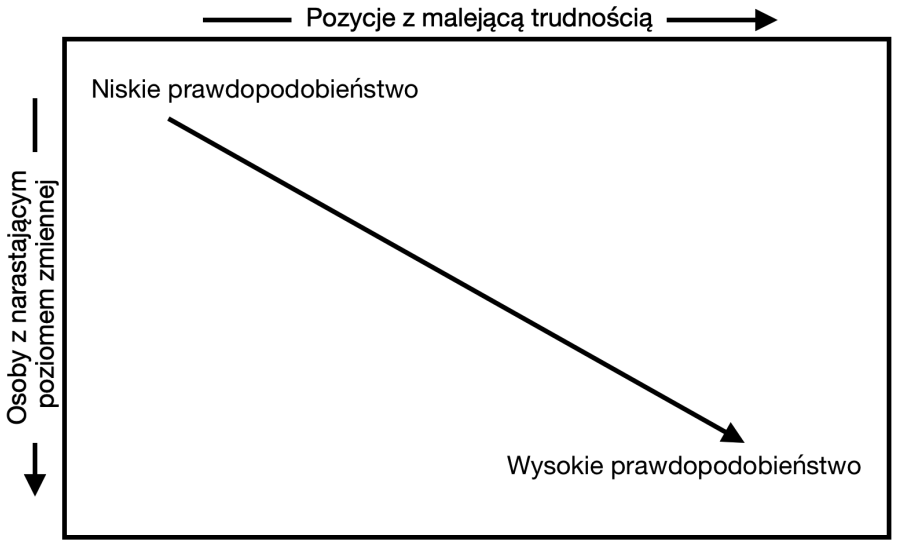
Suma	zad10	zad9	zad18	zad29	zad26	zad21	zad5	zad3	zad13	zad14	zad23
31	0	1,00	1,00	0	1,00	1,00	1,00	0	0	1,00	1,00
32	0,50	0,50	1,00	1,00	1,00	1,00	1,00	0,50	0,50	0,50	0
33	0,67	1,00	1,00	1,00	1,00	0,67	0,67	1,00	0,67	1,00	1,00
34	0,50	1,00	1,00	1,00	0	1,00	0,50	0,50	0,50	0,50	1,00
36	0,50	1,00	1,00	0,50	1,00	1,00	0,50	1,00	0,00	1,00	1,00
37	1,00	1,00	0	1,00	1,00	0	1,00	0	1,00	1,00	1,00
38	1,00	1,00	0	1,00	1,00	1,00	1,00	0	0	1,00	0
40	1,00	1,00	0	1,00	1,00	1,00	1,00	1,00	0	1,00	1,00
41	1,00	0	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0

W pierwszej kolumnie wskazano osiągniętą sumę punktów.

Prawdopodobieństwa przy spełnieniu założeń aksjomatycznej teorii pomiaru powinny układać się zgodnie z porządkiem przedstawionym na rys. 1.1.

W kolejnym kroku, ze względu na złożoność obliczeniową[20], z pełnej macierzy prawdopodobieństw losowane są macierze o wymiarach $3 \cdot 3$, w których przeprowadzane jest sprawdzanie prawdziwości aksjomatu pojedynczego i podwójnego unieważniania. Jeśli któraś z komórek (w jednym z dziewięciu pól wylosowanej macierzy $3 \cdot 3$) nie spełnia pierwszego aksjomatu, jest to odnotowywane. Po przeprowadzeniu wyczerpującej (czyli wszystkie możliwe macierze z przylegających

[20] Przeliczenie tego przykładu przy zadanej liczbie 2000 losowych macierzy komputerowi z 2,4-gigahercowym procesorem Intel Core i5 i z pamięcią 8 GB RAM zajmowało około 26 minut.



Rysunek 1.1. Teoretyczne uporządkowanie prawdopodobieństw udzielenia odpowiedzi w zależności od trudności pozycji i natężenia konstruktu u badanej osoby

wierszy i kolumn) lub losowej (czyli losowe *n* z wszystkich możliwych macierzy[22]) liczby analiz obliczany jest średni ważony procent przypadków złamania założeń. Zgodnie z przyjętą praktyką (Domingue, 2014; Karabatsos, 2001; Kyngdon, 2011) liczba ta nie powinna przekraczać 5% dla poszczególnych zadań testu i testu ogółem (por. tab. 1.4).

Tabela 1.4. Procent przypadków, gdy dana pozycja testu Omnibus (60 pozycji) nie spełniała aksjomatów uporządkowania

	_1	_2	_3	_4	_5	_6	_7	_8	_9	_10
0_	0	0	66,7	33,3	0	35,1	34,8	28,7	40,0	47,2
1_	37,5	29,4	29,8	41,9	30,5	49,2	24,4	38,3	15,7	42,5
2_	42,2	38,3	33,3	28,1	35,7	26,7	22,9	41,3	33,9	42,3
3_	22,0	22,6	29,5	45,4	33,9	34,0	29,4	33,9	34,3	19,8
4_	38,0	27,2	38,5	32,5	27,3	37,1	53,8	31,6	34,0	37,7
5_	40,2	34,5	50,0	12,5	37,4	19,3	14,8	20,4	30,0	1,3

Pozycje przedstawione są kolejno w wierszach – np. liczba na skrzyżowaniu wiersza 1_ oraz kolumny _2 to 12. pozycja testu Omnibus.

[21] Źródłem rysunków, jeśli nie podano inaczej, są badania własne autora.

[22] W analizowanym przypadku liczbę wszystkich możliwych macierzy można obliczyć z iloczynu: (28! / (3! · 25!)) · (60! / (3! · 57!)) i wynosi ona 112 104 720 (Dominique, 2014, pkt 3.2).

W analizowanym przypadku ważony (ponieważ nie testujemy wszystkich przypadków, ale tylko losowy ich zbiór o $n = 2000$) procent wynosi 14,4, co wskazuje, że konstrukt mierzony za pomocą zadań testu Omnibus prawdopodobnie nie ma charakteru uporządkowanego, a wyniki nie są ilościowe w sensie definiowanym przez aksjomatyczną teorię pomiaru. Oznacza to, że wyniki zebrane w postaci punktów za poprawne odpowiedzi nie powinny być sumowane, ponieważ zarówno osoby z identyczną punktacją mają różne zestawy rozwiązanych poprawnie zadań, jak i zadania o identycznej trudności dla różnych osób są w różnym miejscu teoretycznego kontinuum trudności. Inaczej mówiąc, gdyby ułożyć zadania według poziomu trudności dla poszczególnych osób, porządek ten nie byłby identyczny dla każdej badanej osoby. Porządek dla osoby A – powiedzmy, od najłatwiejszego zadania 5, przez zadanie 9, do trudnego zadania 13 – mógłby być inny u osoby B, dla której zadanie 9 mogło okazać się łatwiejsze niż 13, a najwięcej trudności sprawiło jej zadanie 5.

Najnowsze rozwiązanie problemu testowania ogromnej liczby równań zaproponowane przez George'a Karabatsosa (2018) jest oparte na metodologii przybliżonych obliczeń Bayesowskich, z wykorzystaniem bardziej sofistycznych metod próbkowania i estymowania parametrów rozkładów. W połączeniu ze wskaźnikiem rozbieżności Kullback–Leiblera (Karabatsos, 2018, wzór 11) pozwala zidentyfikować problematyczne pozycje testowe i zbadać hipotezę o ilościowym charakterze badanego konstruktu. Niestety obecnie (24 sierpnia 2018 roku) algorytm jest zaimplementowany tylko w środowisku Matlab, które nie było dla mnie dostępne, by sprawdzić jego działanie w praktyce. Jednakże rozwój oprogramowania daje nadzieję, że zmieni się to w niedługim czasie, a sposoby testowania założeń aksjomatycznej teorii pomiaru będą nadal rozwijane i staną się dostępne dla ogółu badaczy.

1.3.5. Krytyka teorii aksjomatycznej

Aksjomatyczna teoria pomiaru nie spotkała się z dużym zainteresowaniem w środowisku naukowym i nie wszyscy badacze uważają ją za wartościowe narzędzie do testowania ilościowego charakteru pomiaru wśród konstruktów psychologicznych. Norman Cliff (1992) nazwał pomiar aksjomatyczny „rewolucją, która nigdy się nie wydarzyła" w psychometrii, Günter Trendler (2009) czy Moritz Heene (2013) sugerują natomiast, że jest to rewolucja, która „nie może się wydarzyć". Powody są dwojakiego rodzaju. Po pierwsze, natura ludzkiego umysłu, który bada psychologia, nie pozwala na pełną kontrolę pomiaru i każdy badany konstrukt, który będzie czymś bardziej złożonym niż proste reakcje, obciążony jest błędem pomiarowym, z którym teoria ACM sobie nie radzi. Zwiększanie zaś kontroli (np. badania laboratoryjne vs kwestionariuszowe) sprawia, że wyniki przestają oddawać rzeczywistość psychiczną i spada trafność zewnętrzna oraz możliwość

generalizacji wyników badań[23]. Drugim powodem jest „rozpędzona maszyna do produkcji artykułów i zdobywania grantów" (Heene, 2013, s. 2), jaką stała się psychologia. Wskazywane przez różnych badaczy nieścisłości czy też wyłapywane nadużycia (afera Diedericka Stapela!) są ciągle zbyt małym impulsem do autorefleksji całego środowiska badaczy i po chwilowym wstrząsie nie jest wytwarzana wystarczająca masa krytyczna motywacji do trwałej zmiany. Pewną nadzieję w tym obszarze widzę w działaniach podejmowanych przez Open Science Collaboration i inne ruchy promujące otwartą naukę (np. OSF – Open Science Framework, AP – Accelerate Psychology) czy rząd Holandii, który przeznaczył na ten cel spore fundusze i wyznaczył plan wprowadzenia do 2024 roku w stu procentach otwartej nauki finansowanej z publicznych pieniędzy. Wracając na grunt psychologii, można, jak sądzę, zaryzykować stwierdzenie, że aksjomatyczna teoria pomiaru jest zbyt wymagająca, by stać się popularna. Podobnego zdania są też inni badacze (Borsboom, Mellenbergh i Heerden, 2004; Cliff, 1992; Michell, 2014), którzy podkreślają, że dane spełniające warunki niezależności, addytywności i ciągłości są niezmiernie rzadko spotykane w psychologii. Godzimy się na wszechobecny i niemały błąd pomiarowy, nadużywając różnych wskaźników dopasowania modelu w miejsce bezpośredniego testowania jego prawdziwości (Heene, 2013). Stąd nadmierna waga przykładana do poziomu istotności statystycznej, opór środowiska[24] we wprowadzaniu wymogu podawania i interpretowania wielkości efektu. Dopasowanie modeli IRT lub SEM nie świadczy o ilościowości konstruktu, wręcz przeciwnie – ilościowość badanego konstruktu pozwala d o p i e r o na stosowanie tych modeli. Nagminną praktyką jest pomijanie niewygodnych pozycji z testów, korelowanie błędów, stosowanie indeksów modyfikacyjnych i eksploracyjne szukanie dopasowania modelowanych danych do teorii ze szkodą dla tej ostatniej.

Powszechnym rozwiązaniem przez „ucieczkę do przodu" jest stosowanie modeli probabilistycznych (IRT) czy równań strukturalnych (SEM) mimo braku dowodów na ilościowy charakter analizowanych zmiennych przez wprowadzenie pojęcia „zmiennej ukrytej o ciągłym charakterze". Wtedy to, co obserwujemy, jest niedoskonałe i może nie spełniać aksjomatów ACM, a przy stosowaniu bardziej elastycznych, ilościowych metod analizy wyników wnioski z niedokładnych modeli są prawdopodobnie lepsze, niż gdyby posługiwać się tylko miarami jakościowymi. Dlatego możemy uzasadnić supremację modelu pomiaru Rascha nad modelem pomiaru Guttmana, ponieważ ten pierwszy model uwzględnia błąd pomiaru. Chociaż jednak wyższość podejścia probabilistycznego opiera się

[23] Zob. dyskusję na temat słynnych badań eksperymentalnych z lat sześćdziesiątych i siedemdziesiątych XX wieku konstytuujących psychologię, które ostatnio okazały się niereplikowalne, jak słynny „test pianki" („marshmallow test"), czy wcześniej doniesienia organizacji Open Science Collaboration (2015) na temat fiaska replikacji prawie siedemdziesięciu eksperymentów z obszaru psychologii społecznej.

[24] Choć pewnie część tego oporu można złożyć także na karb niewiedzy wynikającej z programów statystyki na akademickich kierunkach społecznych.

na obecności błędu w modelowaniu wyników, musimy pamiętać, że nachylenie ogiwy w modelu Rascha będzie rosło wraz ze zmniejszaniem błędu. Paradoksalnie zwiększanie precyzji pomiaru i zmniejszanie błędu sprawia, że przy braku błędu model Rascha równoważy się z modelem Guttmana („Rasch Paradox", zob. Michell, 2008). Niektórzy uznają to za argument odmawiający modelom IRT wyższości nad modelami klasycznymi (Sijtsma, 2012), pomijając fakt, że wyeliminowanie błędu z pomiaru psychologicznego jest praktycznie niemożliwe. Lepiej zatem stosować metody probabilistyczne, które uwzględniają i adresują wpływ błędu pomiarowego, a w analizach relacji między obserwowanymi danymi używają pojęć z teorii prawdopodobieństwa, niż pozostawać przy metodach klasycznych traktujących błąd jako losowe, nieznane źródło obciążające wyniki.

1.4. Podsumowanie problematyki pomiaru

Rozważania na temat teorii pomiaru (np. porównanie teorii aksjomatycznej z teorią odpowiadania na pozycje testowe IRT w: Perline, Wright i Wainer, 1979 lub Karabatsos, 2001) sprawiły, że podjęto próby stworzenia ogólnej teorii pomiaru, której szczególnymi przypadkami są omówione powyżej teorie. Pierwszymi badaczami, którzy próbowali sformułować taką teorię (nazwaną teorią reprezentacji), byli Dana Scott i Patrick Suppes (1958), choć jej początków można się doszukiwać w pracach Hermana von Helmholtza, Otto Höldera i Bertranda Russella. Szczegółowy opis teorii reprezentacji wykracza poza ramy tej pracy[25], ale w uproszczeniu opiera się ona na dwóch twierdzeniach: 1) istnienia – tu na podstawie aksjomatów dowodzone jest, że skala pomiarowa nie jest pusta; 2) unikatowości – tu definiuje się sposób mapowania istniejących poziomów mierzonego konstruktu, czyli odwzorowania w reprezentacjach (np. liczbach) relacji między mierzonymi obiektami. Należy podkreślić, że nie chodzi o pomiar samych obiektów, tylko o pomiar w ł a ś c i w o ś c i obiektów (Bruschi, 2017, s. 2230).

W myśl teorii reprezentacji pomiar jest możliwy, jeśli prawdziwe są cztery przesłanki: 1) leżący u podstaw empirycznego pomiaru konstrukt jest definiowany jako uporządkowana struktura; 2) istnieją aksjomatyczne ograniczenia relacji w obrębie struktury, które odzwierciedlają obiektywne fakty empiryczne; 3) istnieją operacje na liczbach, które odpowiadają operacjom na elementach empirycznych konstruktu; 4) czwarty dowód dotyczy istnienia struktury zachowującej homomorficzne odwzorowanie o d reprezentacji d o struktury empirycznej. Zbiór wszystkich reguł odwzorowujących określa się jako skalę pomiaru (Narens i Luce, 1986, s. 173). W praktyce oznacza to, że możliwy jest naukowy pomiar konstruktów psychologicznych, które, podobnie jak konstrukty fizyczne, mogą być wyrażone jako iloczyny liczb rzeczywistych i jednostek.

[25] Przystępny opis można znaleźć w: Bruschi, 2017, pkt 5.

Chociaż może się wydawać, że psychologia nie ma możliwości posługiwania się miarami podobnymi do tych, których używają na przykład fizyka czy biologia, to teoria łącznego pomiaru daje nadzieję na odkrywanie konstruktów o silnych ilościowych właściwościach. Przejście od, uważanej przez co bardziej krytycznych badaczy za „nienaukową", teorii pomiaru Stanleya S. Stevensa (Tukey, 1961) do sformalizowanej matematycznie teorii pomiaru aksjomatycznego pozwoli odkrywać ukryte ilościowe struktury przez porządkowe rozstrzygnięcia między poziomami obserwowalnych atrybutów. Wykorzystanie teorii odpowiedzi na zadania testowe (IRT) i rozpatrywanie wyników w kontekście probabilistycznego powiązania między obserwowanym a ukrytym, między empirycznym zdarzeniem a latentnym konstruktem je wywołującym, daje nadzieję na stosowanie mocnego pomiaru w psychologii.

Na podstawie zaprezentowanych rozważań można uznać, że przy spełnieniu warunków wstępnych jest możliwy pomiar atrybutów (zmiennych) psychologicznych, takich jak postawy, przekonania, zdolności poznawcze i afekty. Traktowanie wyniku pomiaru w psychologii jako wyniku łączonego nie tylko ma potencjał wyjaśnienia problemów z rzetelnym i trafnym mierzeniem pojedynczego konstruktu, ale też formalizuje relacje między sprzężonymi konstruktami. W podobny sposób możemy traktować wyniki uzyskane w pojedynczym teście – relacja między porządkowymi (a tak najczęściej skonstruowane są skale odpowiedzi testów psychologicznych) wynikami poszczególnych zadań a leżącym u ich podstaw ciągłym, ilościowym konstruktem jest możliwa do udowodnienia. Teraz postulat Höldera (1901), że liczby są używane w pomiarze, ponieważ mierzone atrybuty mają policzalną strukturę, może być wprost udowodniony.

Czy należy zatem zrezygnować z wykorzystywania definicji pomiaru zaproponowanej przez Stevensa? Jeśli jej używanie zastępuje refleksje nad uzasadnieniem pomiaru, co na przykład według Paula F. Vellemana i Lelanda Wilkinsona (1993) Stevens *implicite* czyni, to tak. Pojawienie się w latach czterdziestych ubiegłego wieku takiej definicji pomiaru miało swoje uzasadnienie. Jak uważał Benjamin Drake Wright (za: Trendler, 2009, s. 593):

> [Definicji pomiaru] według Stevensa sprzyjało mocno pragmatyczne podejście, w którym przydatność wyniku była ważniejsza niż jego poprawność, a nacisk kładziono bardziej na cele pomiarowe niż na sposób ich realizacji[26].

Użyteczność nie powinna być wszakże jedyną miarą adekwatności pomiaru. Skale pomiarowe nie mają ostrych granic i o sposobie wykorzystania wyników powinna decydować refleksja nad ich rozkładami oraz kompatybilnością z modelem. Proces pomiarowy powinien sięgać głębiej i zaczynać się wcześniej,

[26] „Stevens' [definition of measurements] was favored by the presence of strong pragmatic culture in which the usefulness of the result dominated their correctness, and the focus was more on the objectives of measuring than how they were obtained".

niż dzieje się to w ramach praktyki badawczej uwiedzionej definicją operacjonistyczną. Zbyt łatwo akceptujemy wyjaśnienie, że mierzony konstrukt jest tym, co mierzy test, tak jakby sam proces pomiaru uzasadniał nasze działania. Stoję na stanowisku, że tylko teoria zbudowana na obserwacjach empirycznych, co oznacza konieczność badania eksperymentalnego, może uzasadniać istnienie postulowanych konstruktów i powinna definiować ich cechy (podobnie uważa wielu badaczy, por. Trendler, 2009, s. 590). W momentach refleksji psychologii nad samą sobą podnoszone są zarzuty, że zbyt często prace badawcze polegają na statystycznej ‘zabawie’ zbiorem liczb uzyskanych w samoopisowych badaniach kwestionariuszowych mierzących w sposób niekontrolowany sprzężone ze sobą ‘cosie’ (Baumeister, Vohs i Funder, 2007; Open Science Collaboration, 2015). Jeśli jako psychologowie chcemy i musimy mierzyć nieobserwowalne konstrukty, to definicja Stevensa prowadzi nas na manowce przybliżeń, zgadywania i podejmowania arbitralnych decyzji.

Trzeba pamiętać, że w minimalistycznej sytuacji pomiar polegający na przypisywaniu według reguł określonych liczb określonym obiektom, robiony w celu uporządkowania wyników, które to uporządkowanie ma oddawać porządek w nieobserwowalnym konstrukcie pomiaru, jest możliwy, ale liczby nie reprezentują niczego poza porządkiem. Błąd popełniamy wtedy, gdy z faktu pomiaru wnioskujemy o właściwościach konstruktu na poziomie ilościowym, bez udowodnienia tej tezy.

2. WPŁYW METOD SAMOOPISOWYCH NA WYNIKI BADAŃ PSYCHOLOGICZNYCH

> Ludzie, grupa lub kultura nie zachowują się lub nie myślą tak, jak przewidują modele, a co ważniejsze, okazuje się, że nasza świadomość, nasza refleksja nad procesem opisanym przez psychologa zmienia ten proces.
>
> Parker, 2007, s. 1[1]

Pomiar konstruktu dotyczącego ludzi, oprócz matematycznej, ma też perspektywę psychologiczną. By mówić o pomiarze konstruktu psychologicznego, musimy mieć przynajmniej do czynienia z czynnikiem, który występuje w różnym natężeniu u różnych osób i możliwe jest jego empiryczne badanie. Przedstawione w rozdziale 1 formalne (matematyczne) teorie pomiaru wskazują na analogię między obiektem pomiaru w naukach społecznych i w naukach przyrodniczych. Oba obszary znacznie się jednak różnią tym, jak i co jest mierzone w ramach ich paradygmatów. I tak jak w naukach przyrodniczych konstruowanie i stosowanie narzędzi pomiarowych polega głównie na manipulowaniu fizycznymi obiektami, tak nauki społeczne operują w większości w obszarze języka. Gdy w fizyce właściwości obserwatora są pomijalne, w psychologii stają się znaczące dla wyniku pomiaru. Ma to poważne konsekwencje dla obiektywności procesów mierzenia. Na przykład w eksperymentach przeprowadzonych z Moniką Obrębską (Obrębska i Kleka, 2016a) wykazaliśmy, jak zmienia się długość i forma wypowiedzi pod wpływem obecności i cech badacza. Zmiana w obserwowanych wynikach objawia się też nie tylko z powodu nieuświadomionej interakcji z badaczem i narzędziem badawczym. Osoby uczestniczące w badaniu mogą świadomie wprowadzać systematyczne obciążenie spowodowane nastawieniem do badania. Wraz z Jarosławem Grothem (Groth i Kleka, 2018) wykazaliśmy, że proste dodanie do instrukcji badania sformułowania „Proszę przedstawić się w jak najlepszym (a w drugiej

[1] „People, a group or a culture do not behave or think as the models predict, and more importantly, it turns out that our awareness, our reflection on the process described by the psychologist changes the process".

wersji – jak najgorszym) świetle" znacząco i w sposób zróżnicowany wpływa na uzyskiwane wyniki w kwestionariuszu badającym profil psychopatii.

Duże znaczenie ma też kolejność prezentowanych bodźców (zadań, pytań, stwierdzeń). Na przykład Fritz Strack, Leonard L. Martin i Norbert Schwarz (1988) pytali osoby badane o częstotliwość chodzenia na randki i poczucie subiektywnego dobrostanu. Wyniki wskazywały, że częste chodzenie na randki sprzyja dobrostanowi w dość znaczącym stopniu (korelacja $r = 0{,}66$). Gdy jednak badacze odwrócili kolejność pytań, tj. najpierw pytano o dobrostan, a potem o randki, relacja między nimi spadała poniżej zera ($r = -0{,}12$).

W badaniach psychologicznych źródłem zmienności wyników, poza kolejnością i znaczeniem bodźca oraz natężeniem badanej cechy, jest też niewątpliwie kontekst osobowy związany na przykład z poziomem inteligencji, stanem emocjonalnym, motywacją, zdolnością do wglądu czy samowiedzą. Czynniki te tworzą interakcyjny filtr oddziałujący z sytuacją badawczą. Odpowiedzi udzielane przez osoby biorące udział w badaniach psychologicznych zawierają znaczną część sytuacyjnych komponentów. Co więcej, ich wpływ może być zmienny w czasie badania, choć pewne aspekty mogą być bardziej stałe, np. sposób przetwarzania informacji objawiający się stylem odpowiadania lub motywacja wpływająca na dokładność udzielania odpowiedzi. Należy także pamiętać, że sam proces udzielania odpowiedzi jest dość skomplikowaną reakcją na bodziec, która w wielu momentach może „pójść nie tak". Przyjrzę się teraz temu, jak przebiega ten proces.

W psychologii, a szczególnie w psychometrii, powszechnie uznawane jest założenie, że udzielanie odpowiedzi na poszczególne pozycje narzędzia pomiarowego jest uwarunkowane posiadaniem badanej cechy, która ma być przez to narzędzie mierzona. Jeśli udzielanie odpowiedzi przez osobę uczestniczącą w badaniu poddamy analizie z perspektywy poznawczej, to możemy wyróżnić cztery jego fazy (Tourangeau, Rips i Rasinski, 2000, s. 7): przetwarzanie informacji zawartych w pytaniu prowadzące do jego zrozumienia, przywoływanie informacji z pamięci, konstruowanie odpowiedzi na podstawie posiadanych informacji i wreszcie kodowanie ich tak, aby pasowały do wymaganego formatu odpowiedzi (więcej na ten temat piszę w Kleka, 2017).

Faza pierwsza, polegająca na przetwarzaniu informacji zawartych w pytaniu, prowadzi do zrozumienia pytania. Wymaga w jakimś stopniu skupienia uwagi i prześledzenia instrukcji, by rozpoznać sens polecenia. Treść semantyczna musi być odszyfrowana, a następnie skojarzona zgodnie z osobistym słownikiem i jego kulturową przynależnością. Inaczej na przykład przetwarzane będą pytania dotyczące postaw, a inaczej te o zachowanie. W pierwszym przypadku osoba uczestnicząca w badaniu może odwołać się do wcześniejszych ustaleń lub sformułować osąd od zera, w drugim musi zidentyfikować potrzebne zachowanie i przywołać odpowiednie wydarzenia z pamięci autobiograficznej. Na tym etapie mogą pojawić się błędy, jeśli osoba uczestnicząca w badaniu nie zrozumie pytania lub

zrozumie je niezgodnie z intencją badacza. Problemy w udzielaniu odpowiedzi na tym etapie mogą mieć źródło w słabej koncentracji i nieskupianiu uwagi osoby uczestniczącej w badaniu lub być wywołane przez cechy pytania: zbyt trudne zdania złożone, niedopasowane językowo, wieloznaczne.

Faza druga polega na przywoływaniu informacji z pamięci długotrwałej według ustalonej (preferowanej, wyuczonej bądź przypadkowej) strategii i/lub wypełnianiu luk w pamięci. Jeśli osoba uczestnicząca w badaniu nie może przywołać wystarczających albo niezbędnych do udzielenia odpowiedzi danych, może dokonać interpretacji pozycji kwestionariuszowej według własnego uznania, co powoduje ryzyko minięcia się z intencją badania i zróżnicowania między osobami badanymi w sposób niedostępny badaczowi, a tym bardziej niepodlegający jego kontroli.

W trzeciej fazie, polegającej na sformułowaniu własnej opinii, zdarza się, że przywołane z pamięci informacje nie stanowią gotowej odpowiedzi i wymagają dodatkowego przetwarzania, np. pytanie o częstość zakupów w zadanym okresie wymaga, oprócz przypomnienia ich sobie, także zsumowania ich liczby. Faza ta obejmuje wszystkie procesy potrzebne do połączenia lub uzupełnienia informacji – również te, w których wskutek braków przywoływane są stereotypowe opinie lub formułowane nowe na podstawie różnych heurystyk (szerzej na temat wpływu heurystyk na posiadaną wiedzę zob. Brycz et al., 2019).

Ostatnia, czwarta faza obejmuje procesy poznawcze dotyczące dopasowania wygenerowanej informacji do możliwych odpowiedzi przygotowanych przez badacza. Osoba uczestnicząca w badaniu dokonuje wyboru, biorąc pod uwagę treść odpowiedzi, które są podane, i uwzględniając aspekty subiektywne (np. dokonując rozróżnienia pomiędzy kategoriami odpowiedzi „zdecydowanie tak" oraz „tak").

Dlatego dowiedzenie uporządkowania i ciągłości mierzonego konstruktu nie wyczerpuje listy problemów, przed którymi stają badacze. Drugim[2] co do ważkości założeniem przyjmowanym *a priori* podczas badań psychologicznych, a mogącym wpływać na uzyskiwane wyniki w sposób utrudniający lub wręcz uniemożliwiający poprawną interpretację[3], jest założenie o takiej samej z d o l n o ś c i d o introspekcji osób uczestniczących w badaniach.

Istnieją badania nad samoświadomością w znaczeniu metawiedzy („wiem, że wiem"), ale dotyczą one przeważnie świadomości funkcjonowania poznawczego i jego podatności na popełnianie błędów (Bar-Tal, Brycz, Dolinska i Dolinski,

[2] Pierwszym założeniem jest przyjmowanie ilościowości i ciągłości wyników uzyskanych w badaniach kwestionariuszowych bez udowodnienia tego.

[3] Co może być jednym z powodów kryzysu pomiarowego w psychologii i niskiego odsetka replikowalnych wyników badań. Z drugiej strony powód ten nie został poruszony w żadnym znanym mi opracowaniu dotyczącym problemów z replikowalnością wyników, waga problemu może być więc tylko moim subiektywnym odczuciem.

2019; Brycz, Jurek, Pastwa-Wojciechowska, Peplińska i Bidzan, 2014; Brycz, Wyszomirska-Góra, Konarski i Wojciszke, 2018; Metcalfe, Metcalfe i Shimamura, 1994). Mimo niewątpliwej użyteczności tego zjawiska i badań nad nim nieznany jest wpływ samowiedzy[4] na trafność[5] wyników uzyskiwanych w badaniach testowych i kwestionariuszowych. W następnym podrozdziale postaram się odpowiedzieć na pytanie, czy można zbadać i jaki ewentualnie jest stopień oddziaływania samowiedzy na obciążenie wyników badań przeprowadzanych (głównie) w psychologii. Inaczej mówiąc, postaram się odpowiedzieć na pytanie, czy posiadanie dobrego narzędzia jest wystarczające do przeprowadzenia dobrego (trafnego i rzetelnego) pomiaru, gdy pomiarowi poddaje się obiekty mierzone z różną „dokładnością".

2.1. Obciążenie wyników badań zdolnością do introspekcji

Moje badanie miało charakter symulacyjny i polegało na określeniu wpływu obciążenia poziomem zdolności do introspekcji (samowiedzą) wyników hipotetycznego testu. Oszacowanie, czy i jak wyniki są obciążone niekontrolowaną w teście, stałą tendencją osób badanych, przeprowadziłem na podstawie analizy zmian w korelacji między wynikami a kryterium reprezentującym rzeczywistą cechę tych osób. Kontrolowana w symulacji manipulacja wielkością i kształtem obciążenia wyników pozwoliła oszacować potencjalną wielkość obciążenia wyników badań psychologicznych błędem mającym swe źródło w niekontrolowanych cechach i zachowaniach osób badanych.

Zastosowałem podejście oparte na symulacji, która składała się z trzech etapów.

1. W pierwszym etapie losowałem próbę z teoretycznej populacji o znanym rozkładzie pewnej cechy i znanej wielkości korelacji z założonym kryterium.
2. Następnie wyniki tej cechy w wylosowanej próbie poddawałem modyfikacji o czynnik samowiedzy. Modyfikacje te przeprowadziłem według różnych reguł oddających możliwe scenariusze obciążania (opisane w dalszej części rozdziału) przez czynnik samowiedzy wyników hipotetycznego testu.
3. Ostatecznie wyznaczyłem rozkład wielkości współczynników korelacji dla nieobciążonych i obciążonych wyników z przyjętym kryterium, by ocenić stopień, w jakim samowiedza może wpływać na dokładność testu.

---

[4] Termin „samowiedza" w znaczeniu „zdolność do introspekcji", w odróżnieniu od „metawiedzy", czyli wiedzy o swoim funkcjonowaniu poznawczym, będącej przedmiotem zainteresowania psychologów społecznych i poznawczych.

[5] Tutaj termin „trafność" użyty jest jeszcze w ogólnym znaczeniu. Pełniejsza psychometryczna definicja znajduje się w rozdziale 3.

Przeprowadzone symulacje odpowiadają sytuacji badawczej, w której grupa osób bierze udział w badaniu jakimś narzędziem, np. w postaci kwestionariusza, które to narzędzie zakłada, że osoby badane mają zróżnicowaną wiedzę o sobie. Ponieważ nie jest znany kształt funkcji obciążającej wyniki badania czynnikiem samowiedzy – różne narzędzia badawcze mogą być podatne na to w różnym stopniu, co więcej, różne osoby mogą cechować się różnym poziomem samowiedzy – pierwszym problemem w badaniu było przyjęcie, w jaki sposób symulować obciążenie uzyskiwanych wyników. Rozwiązaniem przyjętym przeze mnie było równoległe zbadanie kilku prawdopodobnych modeli zaproponowanych na podstawie przeglądu literatury na temat samowiedzy.

Nie chcę tutaj prowadzić szerokich rozważań *stricte* na temat konstruktu samowiedzy – dość wspomnieć, że interesujemy się tym pojęciem od czasów Platona na gruncie różnych obszarów nauki. Antyczni filozofowie wysoko moralnie cenili nakaz „poznaj siebie", a realizację tego nakazu można znaleźć we współczesnej psychologii zajmującej się trafnością samowiedzy. Poznanie siebie pozwala zrozumieć motywy dokonanych działań oraz przewidywać przyszłe zachowania. Na to, co wiemy o sobie, składa się tożsamość osobista, społeczna oraz fizyczna. Po takim zdefiniowaniu „ja" można stwierdzić, że odzwierciedla ono subiektywne postrzeganie tego, kim się jest. To postrzeganie może być bardziej lub mniej zbieżne z obiektywnymi wskaźnikami, a trafność samowiedzy jest skorelowana z samopoczuciem (Brown, 1991, s. 160). W kontekście psychologii i zdrowia psychicznego wielu badaczy zgadza się, że pewna doza uprzedzeń i zafałszowań na swój temat potrzebna jest do utrzymywania równowagi (Allport, 1943; Erikson, 1950; Fromm, 1955; Maslow, 1950).

Jednocześnie badania pokazują, że samowiedza jest systematycznie obciążona (*bias*) – większość ludzi postrzega siebie jako lepszych od innych. Oceniamy pozytywne cechy bardziej jako nasze niż przeciętnego człowieka, cechy negatywne natomiast bardziej jako typowe dla przeciętnego człowieka niż charakteryzujące nas (Taylor i Brown, 1988, za: Brown, 1991, s. 161). Jak podaje Jonathon D. Brown (1991), budujemy nietrafną wiedzę o sobie przez stosowanie różnych strategii dotyczących informacji na nasz temat docierających z otoczenia.

Pierwsza grupa działań to unikanie negatywnego wzmocnienia. Chętniej i wybiórczo: 1) wystawiamy się na pozytywny feedback lub 2) podejmujemy wysiłek poznawczy, by negatywne informacje zwrotne były dwuznaczne, a tym samym stały się mniej ważne. Druga grupa strategii w obszarze poznawczym skupia się na radzeniu sobie z negatywnym wzmocnieniem. Tutaj używamy: 1) selektywnej uwagi, 2) selektywnej interpretacji, 3) selektywnej pamięci lub 4) wybiórczych atrybucji. Ostatnią, trzecią grupę strategii stanowią działania nakierowane na minimalizowanie wpływu negatywnej informacji zwrotnej. Są to: 1) wybiórcze przypisywanie (nie)ważności źródłom informacji, 2) selektywny konsensus, 3) porównania społeczne w dół, 4) pławienie się w chwale zwycięzców i odcinanie

się od niesławy przegranych, 5) kompensacyjne uwypuklanie innych (w innym obszarze) swoich zalet.

Podstawowym sposobem kontroli obciążenia wyników uzyskiwanych w badaniach psychologicznych jest ukrywanie celu badania przed badanymi. Moim zdaniem stosowanie decepcji nie rozwiązuje wszystkich problemów z obciążeniem wyników zróżnicowaną samowiedzą – oddziałuje ona tylko na dostępne świadomości zaburzanie wyników (np. kwestia aprobaty społecznej – zob. Izdebski, Żbikowska i Kotyśko, 2013). Jak twierdzą Daniel Kahneman i Amos Tverski (1996), przekonanie o sobie nie jest rozłożone równomiernie – samo zjawisko jest powszechne, ale nie jest uniwersalne. Inaczej mówiąc, wszyscy mamy uprzedzenia, ale są one u każdego z nas różne. Między innymi dlatego nie mierzy się ich, ale uznaje za element błędu pomiarowego.

Inspiracją do założenia w przeprowadzonych badaniach symulacyjnych związków między samowiedzą a wynikiem były wnioski z metaanalizy Paula A. Mabe i Stephena G. Westa (1982), które pokazały, że istnieje słaby (r = 0,29) związek między dokładnością samooceny a osiągnięciami. Słaba korelacja między samoocenę a wynikami pokazuje, że ludzie źle oceniają swoje możliwości. Badani, którzy zastanawiali się nad swoimi postawami przed badaniem, wykazywali słabszy związek między postawą a zachowaniem w porównaniu z badanymi, którzy nie mieli chwili refleksji. To pokazuje, że refleksja nad sobą nie tylko może obniżać mierzony poziom cechy, ale także wpływać na dokładność w ocenie swoich cech. Badania Kahnemana i Tversky'ego (1996) wykazują ponadto, że to niedocenianie jest często przeszacowaniem, gdy mamy do czynienia z cechami postrzeganymi jako pożądane.

Założony przeze mnie pierwszy model odwzorowywał to, że osoby o niskim poziomie samowiedzy będą zawyżać swoje odpowiedzi, ale tylko przy niskim poziomie badanej cechy. Efekt ten byłby podobny do wpływu czynnika aprobaty społecznej. Gdy badana jest społecznie pożądana cecha – osoby z niskimi wynikami będą „dodawały" sobie trochę wartości, osoby z wynikami wysokimi nie mają zaś takiej potrzeby, dlatego można założyć, że w tym modelu wraz ze wzrostem poziomu cechy efekt zanika, co ilustruje rys. 2.2.

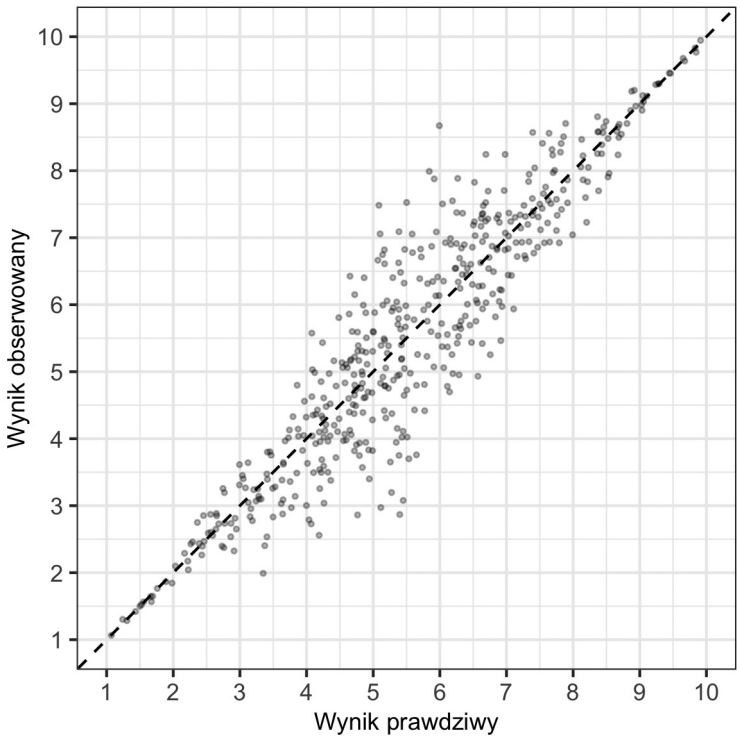
Model pierwszy w założeniu jest modelem liniowym związku między samowiedzą a poziomem badanej cechy. Model drugi (rys. 2.3) ilustruje podobną, aczkolwiek krzywoliniową tendencję. Tutaj zawyżanie wyników (o około 30%) działa w całym zakresie skali, ale dodatkowo występuje efekt sufitowy i maksymalne wyniki w teście uzyskują zarówno osoby z realnie wysokim poziomem cechy, jak i te, które miały dość wysokie wyniki, lecz dodały sobie trochę wartości.

Model trzeci oddaje sytuację, w której zachodzi zarówno efekt podłogowy, jak i sufitowy – osoby o niskim poziomie samowiedzy mają wyniki niedoszacowane, a osoby o wysokim poziomie samowiedzy mają wyniki przeszacowane (rys. 2.4). To sytuacja, w której osoby z niskim poziomem samowiedzy są również

nisko zmotywowane do udzielania odpowiedzi i wybierają odpowiedzi typu „nie wiem", „brak zdania" itp., osoby z wysoką samowiedzą „mieszają się" zaś z osobami aspirującymi do wyników wysokich.

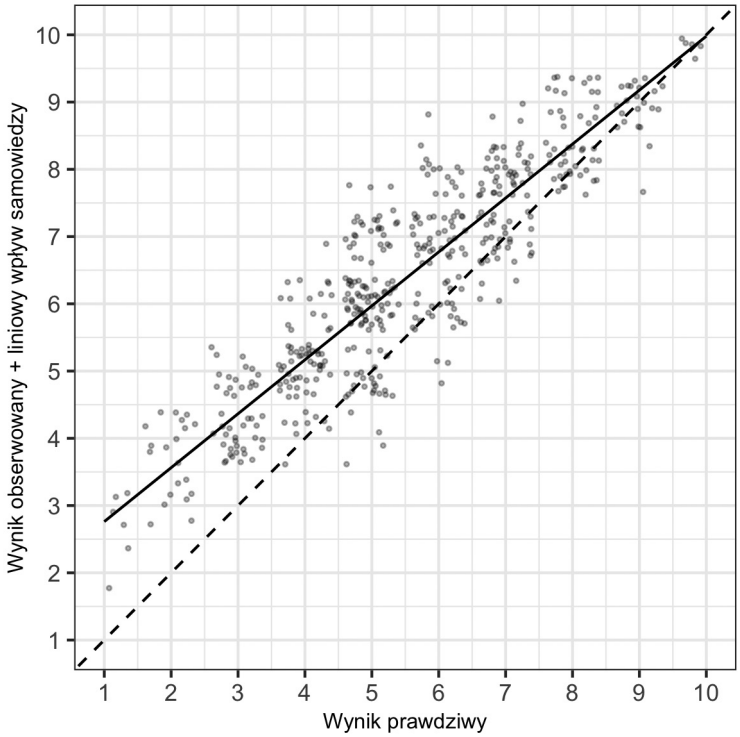
2.1.1. Metodologia i założenia symulacji

Na początku symulacji przygotowałem zbiór danych składający się z liczb wylosowanych z rozkładu normalnego o średniej zero i odchyleniu standardowym jeden ($X \in N(0, 1)$), odpowiadających wynikom hipotetycznego testu osobowości (rys. 2.1). Drugi zbiór liczb ($Y \in N(0, 1)$) reprezentował kryterium, z którym zbiór X był skorelowany z określoną siłą i kierunkiem (ρXY). Następnie oba zbiory liczb przeliczono na wyniki na popularnej standardowej skali stenowej od 1 do 10, ze średnią 5,5 i odchyleniem standardowym 2,0. W ten sposób oba wyniki reprezentowały wyniki przeliczone używane powszechnie w testach psychologicznych.



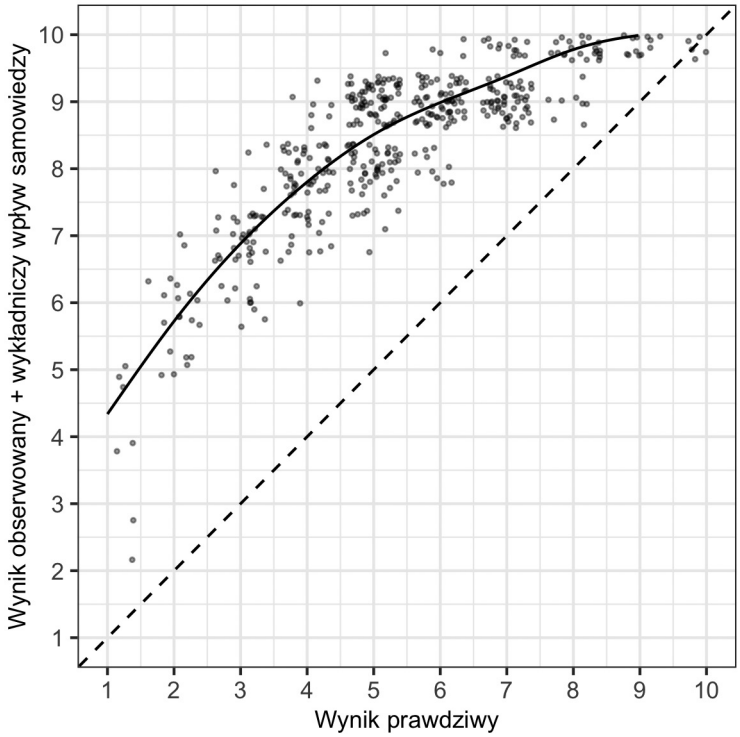
Linia przerywana odwzorowuje idealną korelację między wynikami prawdziwymi i obserwowanymi.

Rysunek 2.1. Rozkład wyników prawdziwych i obserwowanych wykorzystanych w symulacjach – pseudopopulacja (model 0)



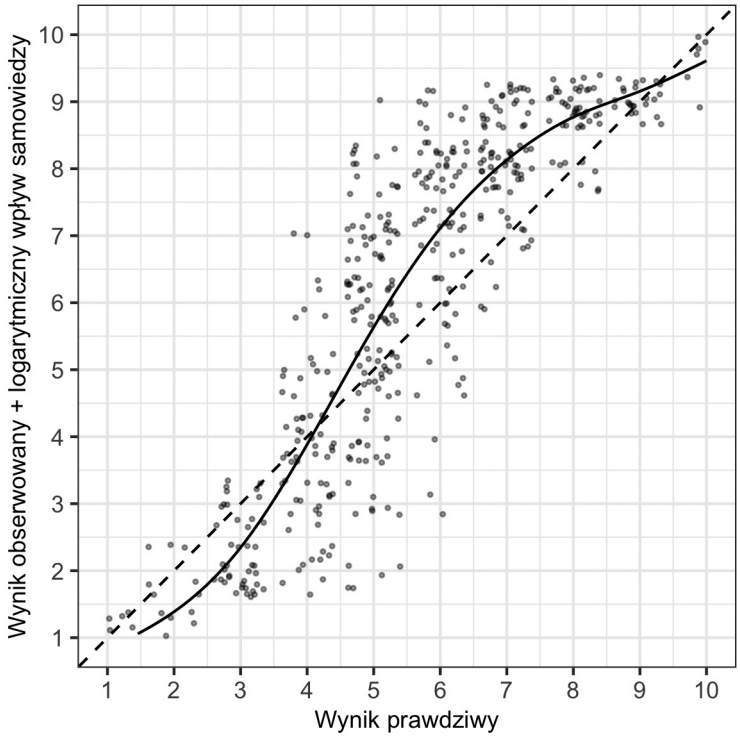
Linia przerywana odwzorowuje idealną korelację między wynikami prawdziwymi i obserwowanymi. Linia ciągła odwzorowuje rzeczywisty związek między wynikami. Wyniki całkowite zostały rozstrzelone na wykresie w celu poprawienia czytelności.

Rysunek 2.2. Model 1 – rozrzut wyników prawdziwych i obserwowanych obciążonych liniowo



Linia przerywana odwzorowuje idealną korelację między wynikami prawdziwymi i obserwowanymi. Linia ciągła odwzorowuje rzeczywisty związek między wynikami.

Rysunek 2.3. Model 2 – rozrzut wyników prawdziwych i obserwowanych obciążonych funkcją wykładniczą



Linia przerywana odwzorowuje idealną korelację między wynikami prawdziwymi i obserwowanymi. Linia ciągła odwzorowuje rzeczywisty związek między wynikami.

Rysunek 2.4. Model 3 – rozrzut wyników prawdziwych i obserwowanych obciążonych logarytmicznie

Punktem wyjścia przy badaniu zmiany w natężeniu związku między zmiennymi *X* i *Y* była korelacja *ρXY* dla danych przeliczonych. W kolejnym kroku zmieniałem wartość zmiennej *X*, dodając pewną wartość wynikającą z przyjętego modelu obciążenia zmiennej aspektem samowiedzy ($X' = X + i$). Analizowane modele obciążenia można scharakteryzować jako liniowe oraz nieliniowe i zostały one zdefiniowane na podstawie następujących założeń:

1. Osoby uzyskujące skrajnie wysokie i skrajnie niskie wyniki w danym teście mają mniej obciążone losowym błędem wyniki niż osoby, które uzyskują wyniki przeciętne (typowe).
2. Zdolność do introspekcji jest stałą właściwością osoby badanej, co oznacza, że wprowadza ona systematyczny błąd do udzielanych odpowiedzi.
3. Wzorując się na badaniach dotyczących wpływu aprobaty społecznej na wyniki kwestionariuszy (Paunonen i LeBel, 2012, s. 162), zakładam, że minimalny wpływ zdolności do introspekcji (samowiedzy) na wyniki obserwowane określa wielkość efektu w granicach $d_{min} = 0{,}5 - d_{max} = 1{,}0$.

W badanej grupie nieznany jest odsetek osób o wysokiej i niskiej zdolności do introspekcji i ma on charakter losowego obciążenia wyników grupowych.

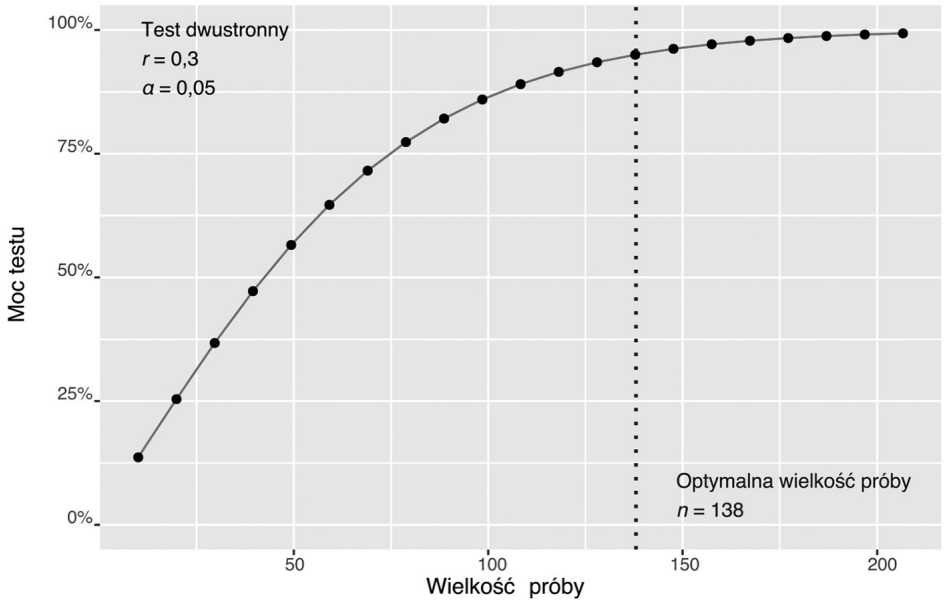
Założyłem, że osoby o wysokiej zdolności do introspekcji odpowiadają na pytania kwestionariuszy w sposób mniej obciążony błędem, bardziej spójny z obrazem własnego ja, natomiast osoby o niskiej zdolności do introspekcji będą bardziej odbiegać od wyników 'prawdziwych', gdyż jakaś część ich odpowiedzi jest zmieniona i zawiera czynnik niezwiązany z przedmiotem badania, np. czynnik aprobaty społecznej lub inny.

2.1.2. Procedura badania i wyniki

Dla każdego analizowanego modelu zastosowano podobną procedurę. W pierwszej kolejności generowano wyniki pseudopopulacji (*X*, $n = 10\,000$) oraz skorelowane z nimi wyniki kryterium (*Y*). W analizie dwa parametry były przedmiotem zainteresowania: 1) wielkość próby, w jakiej szacowano rzeczywistą korelację, oraz 2) wielkość korelacji między zmienną a kryterium przed wprowadzeniem obciążenia zdolnością do introspekcji (samowiedzą). Dla drugiego czynnika przyjąłem jako parametr trzy poziomy korelacji: niska – 0,1, umiarkowana – 0,3 i wysoka – 0,5, opierając się na powszechnie przyjętych poziomach zaproponowanych przez Jacoba Cohena (1988). Ponieważ badanie korelacji między wynikiem a kryterium na tym etapie byłoby trywialne, zgodnie z założeniami 1 oraz 4 do wyników X dodałem losowy błąd według wzoru:

$$X' = X + N(0, 1) \cdot (1 - 0.5 \cdot |X|).$$

Dla pierwszego parametru będącego wielkością prób losowanych z pseudopopulacji przyjąłem wartości na podstawie analizy mocy (rys. 2.5). Zakładając



Rysunek 2.5. Ilustracja graficzna przykładowej analizy mocy wyznaczającej wielkość grupy dla przyjętych wartości parametrów symulacji

wysoką moc $1-\beta = 0{,}95$, dla poszczególnych wielkości korelacji wyników z kryterium przyjąłem stałą wielkość grupy o następujących wielkościach: 1293, 138 oraz 46 obserwacji.

Następnie dla każdego modelu oraz dla każdej wielkości próby przeprowadziłem procedurę wyznaczania współczynnika korelacji polegającą na: 1) wylosowaniu odpowiedniej liczby osób z pseudopopulacji i 2) obliczeniu współczynnika korelacji Pearsona między kryterium a wynikami obciążonymi według różnych modeli. Te dwa kroki były powtórzone 1000 razy dla każdego warunku (3 wielkości próby · 3 modele zależności) w celu wyznaczenia rozkładów wyników. Ponieważ problem badawczy dotyczy oszacowania wpływu obciążenia samowiedzą na współczynnik korelacji, jako wynik wyliczano dla każdej wylosowanej próby różnicę między współczynnikiem korelacji kryterium z „czystą" zmienną X a zmiennymi obciążonymi kolejno samym błędem pomiarowym, następnie zaś błędem i samowiedzą według zakładanych modeli.

Poniżej prezentuję kod w języku R służący do wygenerowania różnic między korelacją wzorcową a korelacjami dla zmiennych obciążonych według założonych modeli:

```
## Funkcja licząca różnice we współczynniku korelacji między
wynikami
## nieobciążonymi, a 1) zaszumionymi,
                    2-4) obciążonymi wg modeli 1-3
przebieg <- function(N, korelacja, popul, iteracje){
   # N - liczebność próbki
   # korelacja - wielkość współczynnika korelacji
   # popul - liczebność pseudopopulacji
   # iteracje - liczba powtórzeń obliczeń dla zadanych parametrów

## Symulacja skorelowanych danych
dane <- as.data.frame(
     sim.correlation(
          matrix(c(1,korelacja,korelacja,1),2,2), popul, T))
colnames(dane) <- c('y', 'x')

## Dane oznaczone jako 'x' modyfikuję o obciążenia:
## Losowy błąd zależny od skali, tzn. mniejszy na krańcach rozkładów
dane$xe <- with(dane,
   x + rnorm(n = popul, mean = 0, sd = 1) * (1 - .25 * abs(x - 5.5)))

## Dodawanie obciążenia wg trzech różnych modeli
dane %>%
     mutate(
# Model 1, Niskie wartości są obciążone bardziej
     x1 = 2 + .8 * xe,
# Model 2, Zależność wykładnicza, efekt sufitowy
     x2 = 10/(1+2 * exp(1) ^(-.5 * xe)),
# Model 3, Zależność logarytmiczna
     x3 = (555 + 10 * xe^4) / (555 + xe ^ 4)
     ) -> dane

## Obliczanie macierzy korelacji
replicate(n = iter, expr = {
     tmp <- psych::corr.test(dane[sample(popul,N),])$r[2:6,1]
# Wyliczanie różnicy z korelacją wzorcową
     tmp[-1] - tmp[1]
})
}
```

Uzyskałem wiele wyników będących różnicami we współczynnikach korelacji. Następnie współczynniki te przeliczyłem na wartości rozkładu normalnego (z), aby wyznaczyć osobno dla każdego modelu statystyki opisowe oraz ich rozkłady. Wyniki przedstawiono w tab. 2.1.

Tabela 2.1. Średnia zmiana współczynników korelacji na skutek obciążenia wyników czynnikiem samowiedzy dla przyjętych modeli

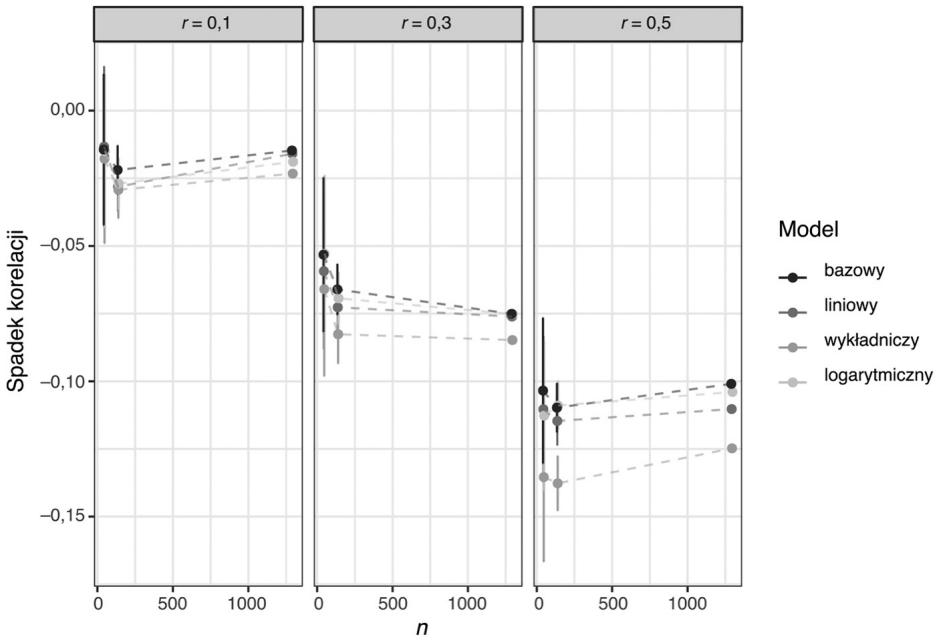
Warunek	M (SD) n = 1293	M (SD) n = 138	M (SD) n = 46
r = 0,3			
model 0	–0,014 (0,09668)	–0,022 (0,05438)	–0,015 (0,01719)
model 1	–0,013 (0,10281)	–0,028 (0,05477)	–0,016 (0,01715)
model 2	–0,018 (0,10841)	–0,029 (0,06344)	–0,023 (0,01940)
model 3	–0,015 (0,09788)	–0,027 (0,05707)	–0,019 (0,01729)
r = 0,5			
model 0	–0,053 (0,09888)	–0,066 (0,05638)	–0,075 (0,01695)
model 1	–0,059 (0,09957)	–0,073 (0,05758)	–0,076 (0,01788)
model 2	–0,066 (0,11131)	–0,083 (0,06514)	–0,085 (0,01918)
model 3	–0,053 (0,10048)	–0,069 (0,05789)	–0,075 (0,01700)
r = 0,8			
model 0	–0,103 (0,09322)	–0,110 (0,05460)	–0,101 (0,0157)
model 1	–0,11 (0,09386)	–0,115 (0,05409)	–0,11 (0,01594)
model 2	–0,135 (0,10839)	–0,138 (0,06105)	–0,125 (0,01762)
model 3	–0,113 (0,09632)	–0,109 (0,05446)	–0,104 (0,01608)

Model obciążenia samowiedzą: 0 – brak; model 1 – liniowe; model 2 – wykładnicze; model 3 – logarytmiczne.

Jeśli przyjrzeć się układowi wyników (rys. 2.6), to widać, że obciążenie wyników „czynnikiem ludzkim" jest proporcjonalne do analizowanej korelacji. Im większa korelacja z kryterium, tym większy spadek – od średnio 0,025 dla słabych związków, przez około 0,05 dla związków średniej siły, po około 0,10 dla silnych korelacji. Drugi wniosek, jaki się nasuwa na podstawie analizy wyników symulacji, to wpływ wielkości próby na możliwość identyfikacji obciążenia. Zakładając hipotetycznie istnienie takiego źródła wariancji przy próbach poniżej 150 osób, nie można zidentyfikować tego obciążenia wyników. Wreszcie trzeci wniosek – model obciążenia jest kwestią drugorzędną. Jedynie przy silnych korelacjach w populacji wykładniczy model obciążenia wyników z efektem sufitowym (brak możliwości zawyżania wyników wysokich ze względu na koniec skali pomiarowej) daje się łatwiej zidentyfikować i osiąga wartości różnicy około 0,10 punktu

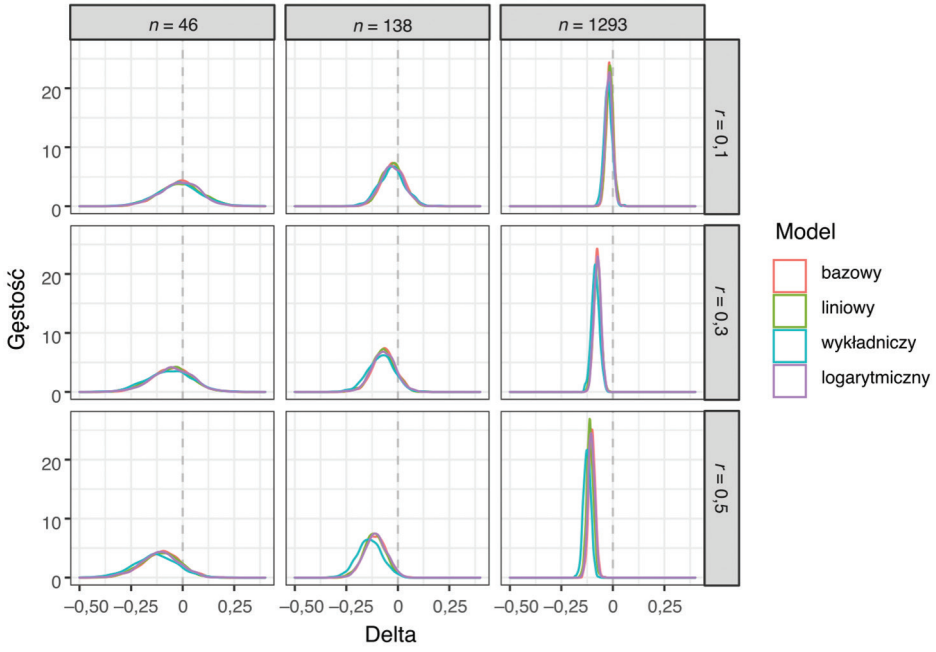
korelacyjnego (dla korelacji z kryterium wynoszącej $r = 0{,}5$). Dla słabych związków i małych grup efekty obciążenia samowiedzą nie są odróżnialne od efektów obciążenia błędem pomiarowym.

Czy efekt obciążenia wyników czynnikiem samowiedzy jest ważny? Dla większości badanych warunków nie różnił się od obciążenia wynikającego z błędu pomiarowego (tzw. model bazowy na rys. 2.6 i 2.7).



Rysunek 2.6. Średnia wielkość (wraz z 95-procentowym przedziałem ufności) obniżenia współczynnika korelacji w zależności od przyjętych warunków

W trakcie interpretacji wyników w znaczeniu wpływu na możliwość stwierdzenia efektu i bycia źródłem błędów we wnioskowaniu ujawnia się jednak destrukcyjny wpływ analizowanego obciążenia. Na rys. 2.7 przedstawiłem rozkład gęstości uzyskanych spadków. I tak jak szerokość tych rozkładów jest funkcją wielkości prób (im większe próby, tym węższe rozkłady), a przesunięcie jest funkcją wyjściowej korelacji (im wyższa korelacja, tym większe spadki), tak zwraca uwagę fakt, że spadki mogą być również wzrostami. Szczególnie dla małych prób (20–60) i niskich korelacji (0,3–0,5), które są bardzo powszechne w badaniach psychologicznych, rozkład wyników obejmuje zarówno wartości ujemne, jak i dodatnie. Inaczej mówiąc, sama wiedza na temat istnienia obciążenia (przy badaniu konstruktu podatnego na taki wpływ) nie pozwala w większości przypadków przewidywać nawet kierunku zmiany.



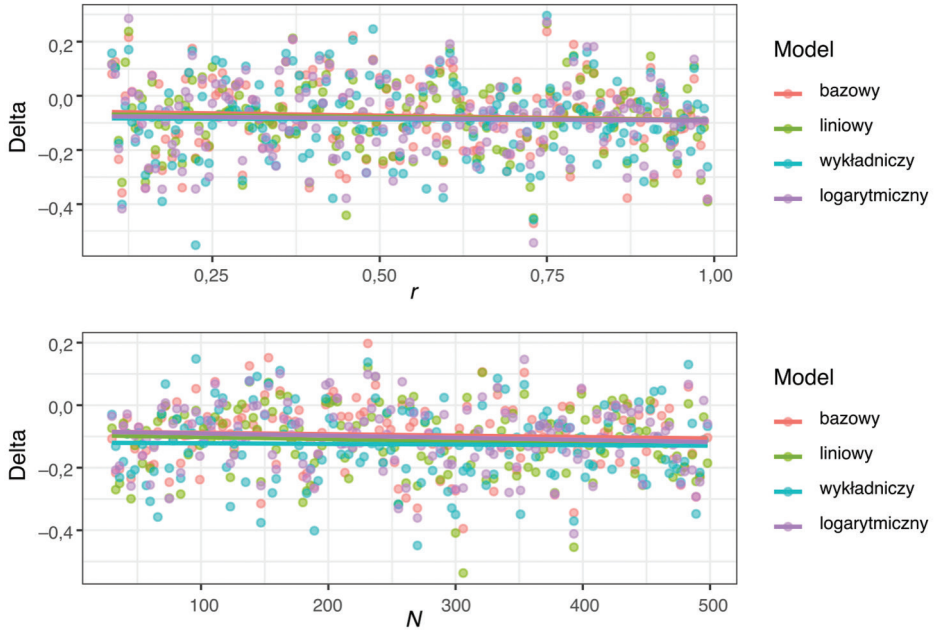
Rysunek 2.7. Rozkłady spadków współczynnika korelacji w zależności od wielkości próby i korelacji w pseudopopulacji

2.2. Podsumowanie problematyki obciążenia wyników

Uzyskane wyniki w znaczeniu wielkości zaobserwowanych efektów nie są znaczące – 10–20% obciążenia wyników najprawdopodobniej zależy od przyjętych przeze mnie parametrów w symulacji[6]. Wniosek, jaki można wyciągnąć na podstawie analizy, dotyczy jednak szerszego kontekstu – obciążenia wyników samowiedzą, która zakłóca liniową zależność między zmiennymi (tak często zakładaną w psychologii, gdy używa się parametrycznych metod analizy wyników). Samowiedza stanowi niewielką składową błędu pomiarowego, a błąd ten – bazowy w mojej symulacji, mający rozkład normalny $\in N(0, 1)$, pogarszał precyzję wyników na tyle silnie, że badane obciążenie nie stanowiło samodzielnego źródła wariancji bez względu na model obciążenia. Jedynym wyjątkiem był

[6] Najsilniej na wielkość różnicy między korelacjami wpływała przyjęta wielkość próby, co podkreśla, jak ważne jest przeprowadzanie badań na licznych grupach badawczych oraz jak ważna jest analiza mocy, która pozwala ustalić wielkość grupy w powiązaniu z wielością spodziewanych efektów.

wykładniczy model obciążenia, który różnił się od pozostałych modeli w niewielkim stopniu (F (3, 36) = 94,16, p < 0, 001, η^2_G = 0,008). Różnica ta jest tym wyraźniejsza, im większe r oraz n badamy (rys. 2.8).



Rysunek 2.8. Przebieg spadku korelacji z kryterium wraz ze wzrostem wielkości korelacji (górny wykres) lub ze wzrostem wielkości próby (dolny wykres)

Czy zwiększanie wielkości prób zabezpiecza przed wpływem obciążenia wyników badania psychologicznego? Rośnie przecież moc takiego badania. Tak, jest to rozwiązanie całkiem sensowne, ponieważ wraz ze wzrostem wielkości próby obciążenie staje się mniej chaotyczne (spada jego wariancja, rys. 2.7). Jeśli jednak osoby badane będą spójnie odpowiadać na zadania w ramach danej grupy tworzącej wskaźnik, a różni badani będą różnie „nasyceni" czynnikiem samowiedzy i w związku z tym pojawią się różnice indywidualne, to i tak dane empiryczne ułożą się w takie wzory, że wskaźniki psychometryczne narzędzia będą potwierdzały jego wysoką spójność.

Czy zatem psychometria powinna się zajmować problemem obciążenia wyników „czynnikiem ludzkim" mimo potencjalnie niewielkiego obciążenia nim realnych wyników? Moim zdaniem warto, by tak było, choć może się wydawać, że nie jest to problem ogromnej wagi, jeśli rozpatrujemy go w wartościach bezwzględnych.

Ważnym problemem według mnie jest to, że manipulacja wynikami, zawyżanie ich czy modyfikowanie kształtu zależności nie jest wykrywalne przy stosowaniu współczynników korelacji. Są one na tyle odporne na łamanie założenia o liniowości związku, że nawet gdy wiem, że on taki nie jest, to nie jestem w stanie tego stwierdzić, ponieważ informacja o tym jest wymieszana z naturalnie występującym błędem pomiarowym. Powszechność, nieodłączność i „wkalkulowanie" błędu pomiarowego w metody pomiarowe w psychologii sprawiają, że jeśli założymy związek liniowy, to stwierdzimy obecność takiego związku, nawet gdy w populacji ma on inny kształt (oczywiście w pewnych granicach). To podkreśla ważność dwóch elementów przeprowadzania badań: 1) teorii jako źródła wątpliwości na temat liniowości związku oraz 2) wizualizacji danych surowych jako etapu, który może badaczy w wątpliwościach utwierdzić. Ufanie wynikom pakietów statystycznych, które są nastawione głównie na podawanie nam wyników w postaci liczb, może sprawić, że popełnimy błąd, sprowadzając potencjalnie ciekawą zależność do trywialnie liniowej, kilka procent słabszej, ale za to łatwiejszej w modelowaniu i interpretacji.

Powyższa analiza nie rozwiązuje w żadnym wypadku problemu dokładności badań samoopisowych, na których opiera się znaczna część współcześnie uprawianej psychologii. Wiele innych czynników, które nie są merytorycznie związane z przedmiotem badania, a zmieniają sposób odpowiadania badanych, zawiera przegląd Norberta Schwarza (1999). Na podstawie badań nad tymi czynnikami wymienia on następujące elementy wpływające na wyniki uzyskiwane w kwestionariuszach:

1. wpływ skrajnych wartości danej skali na odpowiadanie: zmiana skali z zakresu od 0 do 10 na zakres od –5 do +5 sprawiła, że o 11% osób więcej zaznaczyło odpowiedzi maksymalne, wskazujące na udane życie (zmiana z 13% na 34%);
2. wpływ kafeterii odpowiedzi: różnica w towarzyszących odpowiedziach sprawiła, że dolegliwości psychosomatyczne pojawiały się „dwa razy w miesiącu" u 62% badanych osób, gdy skala skonstruowana była od „2 razy lub rzadziej" do „kilka razy lub częściej", a tylko u 39% badanych dla skali od „nigdy" do „dwa razy w miesiącu";
3. wpływ pytań zamkniętych: 61% osób wskazywało, że autonomiczne myślenie ich dzieci jest ważnym zasobem, podczas gdy spontanicznie taką odpowiedź wpisało tylko 4% osób badanych;
4. wpływ możliwej/wymuszonej odpowiedzi: 30% osób określiło, jaki ma stosunek do fikcyjnych przepisów, które nie istnieją;
5. wpływ danych kontekstowych: badania nad przyczynami masowych morderstw afiliowane przy Instytucie Badań nad Osobowością przynosiły odpowiedzi akcentujące uwarunkowania dyspozycyjne, gdy afiliacja przy Instytucie Badań Społecznych wpływała na podkreślanie uwarunkowań społecznych.

Dodatkowym czynnikiem pogarszającym dokładność wyników jest konstrukcja narzędzi psychometrycznych głównie według jednego paradygmatu – większość badaczy, jeśli nie wszyscy, konstruując narzędzia kwestionariuszowe, używają skal likertowskich z arbitralnym ważeniem (poszczególnym odpowiedziom przydzielane są liczby całkowite odpowiadające zakładanemu, takiemu samemu dla każdej kategorii odpowiedzi, natężeniu badanej cechy). Warto jednak pamiętać, że Likert i inni badacze w latach trzydziestych ubiegłego wieku opracowali różne sposoby obliczania wskaźników badanych cech, kładąc akcent na problematykę ważenia odpowiedzi. Na przykład w roku 1928 Gordon W. Allport i Floyd H. Allport opublikowali kwestionariusz do badania wymiaru dominacja–uległość (Studium Reakcji Dominowania i Uległości). Obliczenie wyniku wymagało sumowania wag (ujemnych i dodatnich) uzyskanych przez odniesienie do grup osób, które w odpowiedzi wskazały to samo natężenie zachowania (np. „najczęściej", „czasami", „nigdy"). W roku 1929 Louis Leon Thurstone i Ernest John Cheve opublikowali skalę postaw wobec Kościoła. Założenia teoretyczne tej skali wymagały podczas konstrukcji umiejscowienia użytych twierdzeń na jednym kontinuum. Zastosowanie metody sędziów kompetentnych do budowy skali pozwoliło na ustalenie wag poszczególnych twierdzeń z dokładnością do 0,5 punktu i obliczanie wyniku na podstawie dowolnej liczby wybranych przez osobę badaną twierdzeń z puli dostępnych (osoba badana miała do dyspozycji 22 twierdzenia, z których mogła wskazać dowolne, z którymi się zgadzała). Także Rensis Likert, opisując w 1932 roku konstrukcje skal porządkowych do badania postaw, proponował ważenie odpowiedzi na pytania. Zakładając rozkład normalny cechy w populacji, swe wysiłki kierował na określenie rozkładu poszczególnych kategorii odpowiedzi w stosunku do kontinuum postaw. Opierając się na proporcji osób wybierających dane kategorie odpowiedzi, wyliczano całkowite, dodatnie wagi, które pozwalały na określenie średniego wskaźnika natężenia badanej cechy lub postawy. Przykłady można by mnożyć, ale całe to bogactwo pomysłów uległo zapomnieniu.

Używanie wag pozycji pozwala w większym stopniu zachować addytywny charakter odpowiedzi szczególnie wtedy, gdy kategorie odpowiedzi zawierają różną „ilość" mierzonego czynnika, w przeciwieństwie do arbitralnego przypisywania poszczególnym porządkowym kategoriom kolejnych liczb naturalnych. Mnożenie odpowiedzi przez wagi wydłuża skale, ale też czyni wyniki bardziej precyzyjnymi. Niepopularność takich rozwiązań może mieć przyczynę w zwiększonej trudności konstruowania narzędzi – wymagane są duże grupy badanych i przeprowadzenie stosunkowo pracochłonnych obliczeń. Nie dziwi więc rezygnacja z takich metod na rzecz prostych i szybkich analiz korelacyjnych, które jednak nie skutkują refleksją nad wagami poszczególnych odpowiedzi. Ważenie odpowiedzi zniknęło z obszaru zainteresowań psychologów.

Analiza wyników przedstawiona w tym rozdziale dotyczyła pojedynczego zadania (item). Testy psychologiczne składają się z wielu sprzężonych zadań

tworzących podstawę do wyznaczenia wskaźnika badanej cechy. Natomiast odpowiedź w teście jest wynikiem działania układu wielu wzajemnie powiązanych przyczyn, które są tylko względnie stałe i wśród których badana cecha jest wyłącznie jedną z wielu. Cały system (narzędzie badawcze – osoba badana) jest dynamiczny i układy powiązań mogą zmieniać się w kolejnych odpowiedziach osoby badanej pod wpływem choćby uprzednio udzielonej odpowiedzi. Jeśli na przykład przyznam się w toku pytań do czegoś wstydliwego, to na kolejne pytania mogę zareagować zarówno większą otwartością, jak i z równym prawdopodobieństwem z większym wycofaniem. Psycholog zazwyczaj nie kontroluje i nie jest w stanie kontrolować tego dynamicznego układu. Z tego powodu wydaje mi się uzasadnione stanowisko, że nie potrafimy wyjaśnić związków przyczynowo-skutkowych między badanym konstruktem a udzieloną odpowiedzią. Na próżno szukać tej informacji w podręcznikach testów psychologicznych – nie wiemy, w jaki sposób różnice w mierzonym atrybucie odzwierciedlają się w różnicach w wynikach. Zakładamy tylko, że istnieją w obu równoległych rzeczywistościach – teoretycznej, testowej oraz naturalnej, ludzkiej. I zakładamy, że potrafimy ocenić dopasowanie tych rzeczywistości za pomocą miary trafności, czemu poświęcony jest rozdział 3.

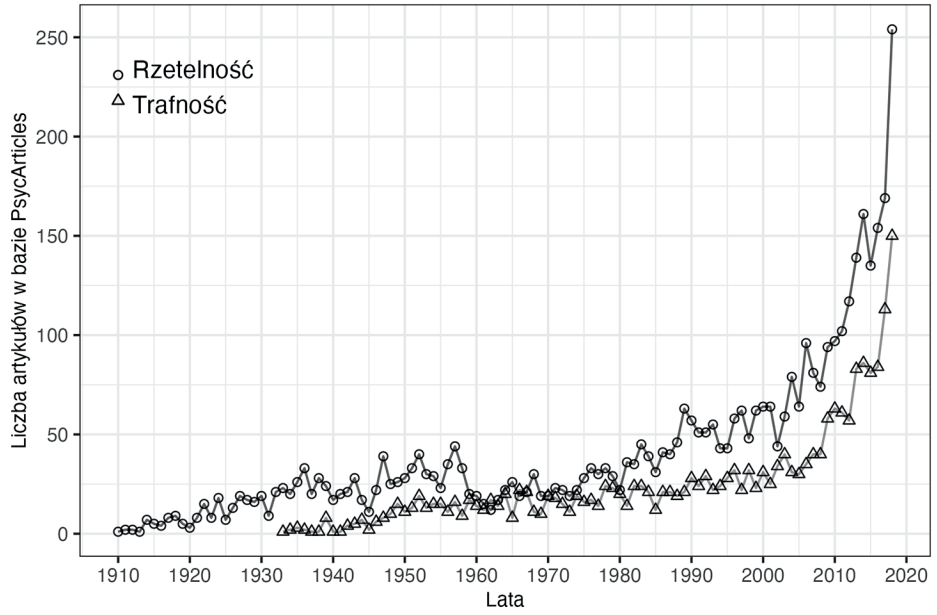
3. JAKOŚĆ POMIARU. OD TRAFNOŚCI JEDNO- DO WIELOASPEKTOWEJ

> Psychologowie wydają się czerpać wielką przyjemność z opracowywania nowych skal pomiarowych (np. istnieje prawie 500 wersji MMPI), ale wykazują małe zainteresowanie w ustaleniu, co oznaczają wyniki w nich uzyskane.
>
> Epstein, 1979, s. 381[1]

Zjawiska badane przez psychologię nie są obserwowalne w sposób bezpośredni. By je badać, opracowuje się narzędzia pomiaru dostarczające danych, na podstawie których można przeprowadzać wnioskowanie o interesujących pierwotnie zjawiskach. Na przykład zaangażowanie w pracę może być mierzone na trzech wymiarach – skuteczności, wypalenia i cynizmu, a każdy z tych wymiarów definiowany jest przez specyficzne postawy i zachowania. Częstość ich występowania pozwala wnioskować o natężeniu tych cech u danej osoby, a sumarycznie w jakiś sposób składają się one na poziom wypalenia zawodowego. Często w takiej sytuacji wynik ogólny jest czymś więcej niż tylko sumą składników (zaprzeczając przy okazji zasadzie addytywności postulowanej przez Normana Roberta Campbella). Uzyskane wyniki w poszczególnych podskalach pozwalają na interpretację struktury badanego konstruktu, gdy jest on (a przeważnie jest) wielowymiarowy lub ma budowę hierarchiczną. Na każdym poziomie agregacji przeprowadzana jest interpretacja obserwowanych wyników, która powinna dostarczać jakichś informacji na temat osoby badanej, a w tym wypadku jej zaangażowania w pracę. To skomplikowany proces uzyskiwania obiektywnej wiedzy, podatny na błędy (Flake i Fried, 2019).

Obserwując tylko liczbę publikacji dotyczących trafności i rzetelności umieszczonych w bazie PsycArticles, można zauważyć rosnące wykorzystanie tych terminów, związane z jednej strony z rozważaniami teoretycznymi nad nimi, ale też z coraz częstszym publikowaniem doniesień dotyczących właściwości nowych

[1] „Psychologists seem to take great pleasure in developing new scales (e.g., the nearly 500 Minesota Multiphasic Personality Inventory scales), but little interest in determining what scores on them mean".



Rysunek 3.1. Liczba publikacji poświęconych trafności i rzetelności na przestrzeni czasu

narzędzi badawczych mających służyć rozwojowi psychologii i nauk społecznych (rys. 3.1).

Badanie konstruktów psychologicznych wymaga przeprowadzenia procesu tworzenia narzędzia pomiarowego od zidentyfikowania celu (mierzonego konstruktu), przez opracowanie teoretyczne struktury konstruktu, wybór sposobu pomiaru, aż do weryfikacji, czy środki prowadzą do celu, czyli do ustalenia, w jakim stopniu zbudowane narzędzie jest trafne (odpowiednie) i rzetelne (precyzyjne). Proces oceny narzędzia polegający na zdobyciu dowodów na to, jak wyniki odzwierciedlają mierzony konstrukt, nazywa się walidacją lub oceną trafności. A jak definiowana jest trafność? Pojęcie to na przestrzeni lat ewoluowało i zanim przedstawię, jak współcześnie rozumiana jest trafność narzędzi wykorzystywanych w badaniach psychologicznych, prześledzę historię rozwoju tego pojęcia.

3.1. Trafność – rys historyczny

Pierwsze oficjalne pojawienie się terminu „trafność" można precyzyjnie umiejscowić w czasie. W 1921 roku organ Standardization Committee of the National Association of Directors of Educational Research przeprowadził ankietę na temat celowości opublikowania oficjalnej listy terminów i procedur. Wyniki ankietyzacji

zostały zaprezentowane i wskazywały, że „[d]wa najważniejsze rodzaje problemów z pomiarem to te związane z określeniem, co test mierzy i jak konsekwentnie mierzy. Pierwszy z nich należy nazwać problemem trafności, drugi – problemem rzetelności" (Buckingham et al., 1921, s. 80).

Na przestrzeni ostatnich 125 lat koncepcja trafności rozwijała się, zaczynając od ogólnej definicji opartej na treści narzędzi, rozszerzając pojęcie najpierw na związki z kryteriami zewnętrznymi wobec narzędzia badawczego, następnie na cechy i czynniki zawarte w teorii badanego konstruktu, i wreszcie wzbogaciła się o konsekwencje stosowania narzędzi pomiarowych. Lata rozwoju i badań nad pojęciem trafności sprawiły, że istnieje bogata literatura, w której rozważa się to pojęcie w różnych kontekstach tak, iż czasami wydaje się, że wyróżniono trafność dla każdego aspektu testu i testowania. Po drodze opracowano wiele specyficznych definicji trafności – do bardziej niezwykłych należy trafność katalityczna (zdolność do reorientacji osób badanych w rzeczywistości po badaniu), jedna ze 151 różnych trafności znalezionych w literaturze przez Paula E. Newtona i Stuarta Shawa (2014).

W miarę podejmowania wysiłków zbudowania spójnej koncepcji trafności różne trendy zdobywały większą uwagę badaczy i teoretyków, co jest wyraźnie widoczne w publikowanych co kilkanaście lat przez APA, AERA i NCME Standardach testów (American Educational Research Association, American Psychological Association oraz National Council on Measurement in Education, 1955, 1974, 1985, 1999, 2014; American Psychological Association, American Educational Research Association i National Council on Measurement in Education, 1966). Zmiany w kolejnych latach i dyskusja wokół pojęcia trafności wzbogaciły naszą wiedzę o niej, ale też spowodowały nadmierną „specjalizację i specyfikację", w miarę jak wprowadzano kolejne rodzaje trafności i rodzaje dowodów na nią.

Rozwój pojęcia trafności można podzielić na cztery historyczne okresy. Pierwszy (od końca XIX wieku do roku 1951) to okres konstytuowania się teorii trafności, następny okres (1952–1974) obejmował definiowanie różnych aspektów trafności, kolejny okres (1975–1999) charakteryzowało ujednolicanie pojęcia, a od roku 2000 trwa okres jego dekonstruowania (Newton i Shaw, 2014).

Początkowy okres dostarczał wielu sprzecznych definicji trafności, które były efektem doprecyzowania pojęcia oraz samego rozwoju pomiaru w psychologii. Pojęcie trafności używane było i nadal jest powszechnie używane w logice, a do psychometrii trafiło jako określenie „precyzji pomiaru atrybutu przy użyciu wyniku testowego" (Newton i Shaw, 2014, s. 1). Jak piszą Newton i Shaw, jest to kraniec kontinuum definiowany w ten sposób także przez Campbella w latach dwudziestych ubiegłego stulecia, podczas gdy po drugiej stronie tego kontinuum jest wymiar etyczny, jako trafność w stosowaniu i interpretowaniu wyniku testu. Opublikowana przez Karla Pearsona formuła współczynnika korelacji pozwala wprost sprawdzać, jak dobrze wyniki testowe powiązane są z mierzonymi atrybutami oszacowanymi w inny sposób. Prawdopodobnie z tego powodu w latach

dwudziestych test był definiowany jako trafny dla pomiaru czegokolwiek, z czym korelował[2]. Przy szacowaniu trafności zaczęto też korzystać z analizy czynnikowej opracowanej przez Spearmana (1904). Guilford (1946) zaproponował i promował nazywanie tych dwóch sposobów szacowania trafności miarą trafności praktycznej (*practical validity*, opartej na korelacji) oraz miarą trafności czynnikowej (*factorial validity*, opartej na analizie czynnikowej).

Do początku lat pięćdziesiątych XX wieku model trafności kryterialnej (praktycznej) był uznawany za standard, ale był niewystarczający do szacowania trafności narzędzi mierzących cechy, dla których kryteria zewnętrzne nie były dostępne. Pojawiały się, co prawda, słabe głosy postulujące rozumienie trafności także od strony teoretycznej, ale nie znalazły się one w centrum zainteresowania psychometrów aż do połowy tej dekady. W 1955 roku kwestia ta wysunęła się na pierwszy plan wraz z publikacją pracy Lee Cronbacha i Paula Meehla definiującej pojęcie trafności teoretycznej (*construct validity*) równolegle z trafnością kryterialną (*criterion validity*) i treściową (*content validity*), co doprowadziło do tzw. trynitarnego spojrzenia na trafność. W drugim wydaniu standardów (American Psychological Association…, 1966) (jeśli liczymy jako pierwsze wydane przez APA w roku 1955 Techniczne Rekomendacje do Testów Osiągnięć) trafność jest rozumiana właśnie w ten potrójny sposób, a trafność teoretyczna jest zdolnością testu do odzwierciedlenia poziomu posiadanej przez badanego właściwości.

Najważniejszym efektem tego okresu było rozróżnienie na trafność dotyczącą pomiaru (Cronbach i Meehl, 1955) i trafność dotyczącą badania (Campbell i Fiske, 1959). Jest to jednocześnie ograniczenie trafności albo do ważności kryterium („jak dobrze wynik testu przewiduje zachowanie poza sytuacją testową?"), albo do ważności treści („jak dobrze do celu badania pasuje treść testu, np. klasyfikacja pracowników na stanowisko?"). Żadne z tych podejść nie definiuje jednak w sposób kompletny trafności testu, a rozróżnienie na dwa rodzaje trafności jest dziś uznawane za rozróżnienie między rodzajami dowodów na hipotezę o trafności, a nie między autentycznie odrębnymi pojęciami.

Definiując trafność, Cronbach i Meehl stwierdzili, że dla określenia precyzji prognozy pomiaru nie jest ważne, co test mierzy, a trafność kryterialna opiera się tylko na empirycznych dowodach korelacji, natomiast trafność pomiaru w znaczeniu trafności teoretycznej i treściowej opiera się na logicznej analizie zawartości testu. Temu celowi służyło wprowadzenie pojęcia sieci nomologicznej – zestawu wzajemnie powiązanych tez stanowiących na przykład próbki zachowań istniejących w umyśle badacza, tworzonych na podstawie doświadczenia, oczekiwań i zdrowego rozsądku. Według badaczy konstrukt teoretyczny budowany jest przez skojarzenia pojęć i obserwacji, które są empirycznie potwierdzane jako wiarygodnie skorelowane z pierwotnym pojęciem (Cronbach i Meehl, 1955, s. 286). Różnica

[2] „[…] a test is valid for anything with which it correlates" (Guilford, 1946, s. 429).

między siecią nomologiczną powiązań teoretycznych między konstruktami a siecią powiązań empirycznych między obserwacjami jest dla autorów miarą trafności konstruktu(ów). Cronbach i Meehl wskazywali, że mamy do czynienia z trafnością teoretyczną bez względu na to, czy wynik testu będzie interpretowany jako miara jakiegoś atrybutu czy też niezoperacjonalizowanej zmiennej (jakościowej) (Cronbach i Meehl, 1955, s. 282)[3]. Konstrukt nie musi być zaś definiowany, ponieważ jest immanentną cechą sieci nomologicznej, w której istnieje (Cronbach i Meehl, 1955, s. 299)[4]. Cronbach i Meehl byli bardziej zainteresowani samym procesem szacowania trafności i silny wpływ na ich koncepcję miał dominujący w latach pięćdziesiątych hipotetyczno-deduktywny model analizy konstrukcji teoretycznych, w których przyjmowano, że teorie to systemy aksjomatyczne. Aksjomaty wykorzystywane są do formułowania i prognozowania obserwowalnych relacji pomiędzy obserwowalnymi zmiennymi, co umożliwia porównanie praw teoretycznych z danymi empirycznymi, i tym samym do dowodzenia słuszności teorii. Odzwierciedlenie rozbicia pojęcia trafności na różne rodzaje znajdziemy w standardach testów opublikowanych przez amerykańskie towarzystwa APA, AERA oraz NCME, które w swoim trzecim wydaniu Standardów dla testów w edukacji i psychologii (American Psychological Association…, 1966) pojęciu trafność nadają liczbę mnogą i piszą o wielu trafnościach. Główny akcent położony jest na wewnętrzną naturę pomiaru i jego użyteczności jako predyktora mającego na celu uporządkowanie i wyjaśnianie danych aspektów rzeczywistości.

W latach pięćdziesiątych sformułowano podstawowe zasady walidacji testów: 1) ustalenie przedmiotu i zakresu interpretacji wyników oraz 2) zapewnienie, że dowody na trafność testu dotyczą proponowanej interpretacji, a jednocześnie wykluczane są interpretacje konkurencyjne. Formalizacją pełnego sposobu walidowania testów była macierz wielu cech wielu metod (*multi-treat, multi-method* – MTMM) opracowana przez Campbella i Fiskego (1959). Była to metoda oparta na analizie wzajemnych związków między wynikami uzyskanymi przy użyciu co najmniej czterech testów: dwóch badających ten sam konstrukt oraz dwóch badających inny. Ponadto w każdej parze (lub grupie, gdy jest ich więcej) testy powinny różnić się metodami badawczymi (np. mogą to być kwestionariusz oraz obserwacja). Aby walidowany test uznać za trafny, powinien on być w większym stopniu skorelowany z innym testem tego samego konstruktu (tzw. aspekt zbieżny trafności) niż z wynikami testów innych konstruktów (aspekt różnicowy trafności).

Biorąc pod uwagę skorelowane testy, przyjmuje się oczywiście, że korelacja dotyczy wyników uzyskanych za ich pomocą. Ważnym warunkiem wstępnym

[3] „[…] construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined' ".

[4] „A construct is defined implicitly by a network of associations or propositions in which it occurs".

była wysoka rzetelność testów używanych do macierzy korelacji. Znacząca i różna od zera korelacja między wynikami pomiaru tego samego konstruktu wskazywałaby na to, jak ta sama badana cecha przewidywana jest za pomocą różnych metod pomiarowych. Ponadto korelacja ta powinna być wyższa od korelacji między innymi konstruktami mierzonymi za pomocą tej samej metody. Wzór tych korelacji powinien być powtarzalny i zauważalny w całej macierzy wielu cech wielu metod. Campbell i Fiske uważali, że nie można oddzielić testu od metody testowania i w analizie trafności poświęcić trzeba uwagę także tej ostatniej – metoda i konstrukt powinny być rozpatrywane jako osobne źródła wariancji wyników testu.

Warto podkreślić za Jerzym M. Brzezińskim (2004, s. 518), że przy opracowywaniu nowego testu aspekt zbieżny i różnicowy trafności powinien być analizowany z kryteriami faktycznie zewnętrznymi, co oznacza, że powinny to być metody nietestowe. Tylko tak można oszacować wpływ metody i uzasadnić wprowadzanie nowego narzędzia badawczego do instrumentarium psychologicznego, które posłuży takim celom, jak budowanie teorii, selekcja czy mierzenie zmian.

Obiecującym podejściem teoretycznym, które skupia się na zrozumieniu wielu źródeł wpływu na wyniki testów, jest teoria generalizacji (Cronbach et al., 1972). Poza cechą i metodą Cronbach ze współpracownikami wskazali także inne źródła błędu: osoby badane, kontekst sytuacyjny oraz badacze, które to czynniki mogą mieć wpływ na wyniki pomiaru.

Teoria generalizacji zapewnia ramy teoretyczne i statystyczne do identyfikacji i oszacowania wielkości tych źródeł, co pozwala zaprojektować dany test w taki sposób, aby zmniejszyć wpływ zidentyfikowanych zakłóceń. Według teorii generalizacji dla każdego testu istnieje uniwersum warunków, które mogą mieć wpływ na wyniki. Przeprowadzenie badań nad skutkami tych warunków nazywane jest badaniami *G*, w odróżnieniu od badań, których wyniki służą do podejmowania realnych decyzji, nazywanych badaniami *D*. Wyniki tych drugich mogą pomóc w modyfikacji warunków testowania zidentyfikowanych w badaniach *G* w celu dostarczenia możliwie najdokładniejszych danych. W teorii generalizacji pojęcie trafności utożsamiane jest ze zdolnością uogólnienia wyników z jednego kontekstu testowego na inny i odpornością narzędzia badawczego na zmiany tychże kontekstów. Niestety teoria generalizacji nie upowszechniła się do dziś, co prawdopodobnie spowodowane jest dużymi wymaganiami w zakresie generowania i testowania różnych źródeł wariancji wyników.

W drugiej połowie XX wieku badacze nie kwestionowali już twierdzenia, że trafność dotyczy pomiaru przez testy psychologiczne konstruktów i głównym zadaniem trafności jest szacowanie adekwatności teoretycznej. Pod koniec lat siedemdziesiątych postulat sformułowany kilkanaście lat wcześniej przez Jane Loevinger (1957, s. 636): „skoro trafność predykcyjna, zbieżna i treściowa są formułowane *ad hoc*, trafność teoretyczna jest całością trafności z naukowego punktu widzenia" staje się coraz powszechniej akceptowany i znajduje

odzwierciedlenie w wydanej w 1985 roku kolejnej wersji Standardów (American Educational Research Association..., 1985), w której pojęcie trafności jest pierwszym omawianym aspektem testów. Trafność teoretyczna nie jest jedną z wielu trafności, ale obejmuje wszystkie możliwe dowody na trafność narzędzia badawczego, włączając w to dowody treściowe i kryterialne, rzetelność i całe mnóstwo metod związanych z teorią testowania (Anastasi i Urbina, 1989; McDonald, 1999; Messick, 1988). Tutaj też pojawia się po raz pierwszy problematyka potencjalnego wpływu procesu testowania na osoby, instytucje i społeczności.

W rezultacie trynitarne ujęcie trafności (równorzędne trzy aspekty) zaczyna słabnąć i zaczyna być dominowane przez trafność teoretyczną. Różne ujęcia trafności są ujednolicane tak, aby trafność była jednym pojęciem obejmującym różne aspekty. W tym okresie trafność teoretyczna nabrała swego dogmatycznego charakteru i uznana została za cechę interpretacji wyników (zyskując asymptotyczny charakter), do której badacze mają dążyć, ale nigdy jej nie osiągną. Duży był tutaj wkład Samuela Messicka (1988), który twierdził, iż „trafność jest jedna". Zmiany w rozumieniu trafności i przesunięcie granic jej definicji dobrze ilustruje zestawienie słownictwa używanego w poszczególnych wydaniach Standardów (tab. 3.1).

Tabela 3.1. Słownictwo używane do określania trafności w kolejnych wydaniach Standardów

Rok	Słownictwo dotyczące trafności	Rodzaje trafności
1955	kategorie	predykcyjna, statusu, merytoryczna, spójna
1966	rodzaje	teoretyczna, kryterialna, treściowa
1974	aspekty	związana z kryteriami, z konstruktem, z treścią
1985	kategorie	związane z kryteriami, z konstruktem, z treścią
1999	źródła dowodów	treść, procesy odpowiadania, struktura wewnętrzna, relacje

Źródło: AERA, APA i NCME, 1955, 1966, 1974, 1985, 1999.

Messick krytykował współcześnie mu opublikowane Standardy za zbyt słabe, jego zdaniem, podkreślenie jedności trafności i propagowanie rozumienia jej w trzech odrębnych obszarach związanych z teorią, treścią i kryterium. Według niego istotą jednorodnej trafności jest to, że adekwatność, sensowność i użyteczność wniosków opartych na rezultatach testowania wspiera się na wiarygodności empirycznie uzasadnialnej zasadności mierzonego konstruktu. Zgodnie z tą definicją trafności to nie test jest walidowany, ale wnioski i decyzje oparte na wynikach testu. Z tego punktu widzenia interpretacja wyników, wykorzystanie testu, społeczne konsekwencje jego stosowania są miarą trafności.

Praca Messicka mimo swej ogromnej wartości nie odcisnęła piętna na powszechnym pojmowaniu pojęcia trafności. Stało się tak głównie przez trudną do zrozumienia formę pracy, w której autor wymaga na przykład integrowania informacji o wszystkich aspektach testu i procedury testowania, aby osiągnąć stan *overall judgment*, niestety nie podaje już, w jaki sposób to zrobić; pisze: „trafność jest zintegrowanym osądem oceniającym stopień, w jakim empiryczne dowody i teoretyczne racjonalne przesłanki potwierdzają adekwatność i stosowność wniosków oraz działań opartych na wynikach testów lub innych efektach oceny" (Messick, 1988, s. 13)[5].

Messick definiował trafność jako proces osądu, w miejsce trafności ocenianego przedmiotu (tu: narzędzia badawczego). Nie pozwalał także na wyjątki w ocenie trafności teoretycznej nawet w przypadku prostych testów, mimo że było to w oczywisty sposób nadmiarowe. Messick podsumował swój pogląd na koncepcję trafności w postaci zagnieżdżonej macierzy (*progressive matrix*), w której jeden wymiar stanowiło źródło oceny, a drugi funkcja lub następstwo testowania. Poszczególne wymiary trafności zawierają narastające znaczenie wraz ze zmianą celu z dostarczania dowodów na konsekwencje testowania oraz wraz ze zmianą z interpretacji na użycie testu (tab. 3.2).

Tabela 3.2. Zunifikowane pojęcie trafności Samuela Messicka

Wymiary trafności	Interpretacja wyników	Użycie wyników
Baza dowodowa	A	A + B
Baza konsekwencji	A + C	A+B+C+D
A – trafność teoretyczna, B – ważność lub użyteczność, C – implikacje wartości, D – konsekwencje społeczne		

Źródło: Messick, 1995, s. 748.

Messick (1988) wprowadził do obszaru rozważań nad trafnością testu etyczną stronę testowania i jego konsekwencji. Ugruntował w ten sposób filozoficzne podstawy teorii trafności, proponując całościowy model oparty na ogólnych zasadach. Koncepcja ta nie spotkała się z pozytywnym odzewem i nie zmieniła od razu myślenia o trafności. Z czasem jednak badacze zaczęli zdawać sobie sprawę, że trafność treściowa polegająca na oszacowaniu, jak dobrze treść uniwersum jest odwzorowana w treści narzędzia, nie zapewnia trafności jego użycia, podobnie sama trafność kryterialna oparta na korelacji z kryterium zewnętrznym nie wnosi dużo do pojęcia trafności narzędzia bez odwołania do znaczenia

[5] „Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment".

i zawartości wyników, na co zwracał uwagę w swych pracach Messick (1980, 1988). Podejście czysto empiryczne jest niewystarczające do oszacowania trafności narzędzia. W miarę jak trafność treściowa i kryterialna zostawały podporządkowywane trafności teoretycznej, traktowanej jako nadrzędna, rozumienie pojęcia trafności zaczęło się zbliżać do koncepcji unitarnej[6], która swój wyraz znalazła w piątym wydaniu Standardów dla testowania w edukacji i psychologii (American Educational Research Association…, 1999). Problem rozczłonkowania trafności autorzy Standardów rozwiązali, przyjmując, że trafność jako cecha testu jest jedna, ale można badać ją na wiele sposobów. Dotychczasowe trafności teoretyczne, kryterialne itp. to nic innego jak dowody w poszczególnych obszarach na trafność testu. Zgodnie z tą rekomendacją mamy więc możliwość gromadzenia i przedstawiania dowodów na trafność narzędzia na podstawie teorii, procesu odpowiadania, wewnętrznej struktury testu, kryteriów zewnętrznych oraz następstw testowania. Główny akcent jest położony na trafność teoretyczną w znaczeniu uzasadniania sposobu konstrukcji i interpretacji wyników testu przez jego użycie.

Choć szacowanie trafności jest nadal uznawane za niekończący się proces gromadzenia naukowych dowodów na poprawność interpretacji wyników[7], to po roku 2000 pojawiają się prace badaczy postulujących przywrócenie pojęciu trafności prostego i praktycznego znaczenia (Goldstein, 2015). Główna zmiana polega na definiowaniu trafności narzędzia badawczego w kontekście tego, jakie ma zdolności do przewidywania przyszłych stanów badanej cechy. Najbardziej pragmatyczne jest tutaj podejście Michaela Kane'a (2013), które powstało jako odpowiedź na potrzebę ułatwienia procesu szacowania trafności. Kane definiuje trafność w ten sam sposób, jak czynił to Messick, proponując do jej szacowania podejście oparte na argumentacji (*argument-based approach*):

> Trafność proponowanej interpretacji lub wykorzystania wyników testów w dowolnym momencie można zdefiniować w kategoriach wiarygodności i stosowności proponowanej interpretacji/użytkowania w tym czasie. […] Trafność jest […] właściwością proponowanych interpretacji i wykorzystania wyników testu (Kane, 2013, ss. 2–3)[8].

---

[6] Nie działo się to bez kontekstu – zdecydowanie wpływ na proces ujednolicania rozumienia koncepcji trafności miały publikowane przez APA, AERA i NCME kolejne Standardy dla Testów Edukacyjnych i Psychologicznych.

[7] „The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (American Educational Research Association…, 1999, s. 9).

[8] „The validity of a proposed interpretation or use of test scores at any point in time can be defined in terms of the plausibility and appropriateness of the proposed interpretation/use at that time. […] Validity is […] a property of the proposed interpretations and uses of the test scores".

Zgadzając się z definicją trafności w ujęciu Messicka, Kane zaproponował dość proste rozwiązanie sposobu jej szacowania oparte na kontekście użycia testu. Według Kane'a do oceny trafności teoretycznej narzędzia badawczego nie jest potrzebna kompletna i wyczerpująca teoria, ale wystarczy argument dotyczący interpretacji bądź użycia testu (*interpretation/use argument* – IUA).

> Po opracowaniu IUA zapewnia on ramy dla walidacji i określa kryteria oceny, czy proponowana interpretacja i zastosowanie zostały odpowiednio zbadane. Jeżeli argument jest spójny i kompletny, a wszystkie jego wnioski i założenia są wysoce wiarygodne […], interpretacja/użytkowanie będzie uznane za wiarygodne lub trafne. Jeżeli jakakolwiek część argumentu nie jest wiarygodna, interpretacja/użycie nie zostałoby uznane za trafne (Kane, 2013, s. 9)[9].

Dzięki takiemu zdefiniowaniu procesu szacowania trafności walidacja nie musi być „niekończącym się procesem" (Cronbach, 1989, s. 151), a i nie jest tak, że „prawie wszystkie informacje zebrane podczas procesu konstruowania i używania testu odnoszą się do jego trafności" (Anastasi, 1986, s. 3). W tym ujęciu badacz przez przedstawienie odpowiednich argumentów powinien uzasadnić, że użycie konkretnego testu jest trafne z punktu widzenia konkretnego celu, np. diagnozy lub prognozy. Najpierw definiuje się, jakie interpretacje i zastosowania wyników testu zostaną użyte, a następnie gromadzi się dowody potwierdzające, że zamierzone i przewidziane zastosowanie testu jest zasadne. Ocena trafności jest zbieraniem i integrowaniem dowodów na poparcie tezy o znaczeniu wyniku testu reprezentującego poziom konstruktu psychologicznego i odbywa się zawsze w kontekście celu (np. diagnoza lub badanie naukowe) i populacji.

3.2. Krytyka trafności opartej na argumentacji

Nie wszyscy badacze zgadzają się jednak z założeniem, że test jest trafny w kontekście celu, dla danej grupy osób badanych oraz w takim stopniu, w jakim osiąga założony cel dla tej konkretnej grupy. Przyjmując to założenie, nie można oddzielić trafności testu od celu pomiaru oraz od grupy, na której pomiar jest przeprowadzany. Jednakże według Denny'ego Borsbooma i współpracowników (Borsboom, 2006; Borsboom, Cramer, Kievit, Scholten i Franić, 2009; Borsboom et al., 2004; Borsboom i Wijsen, 2017; Markus i Borsboom, 2011; 2013) trafne mogą być testy, a nie interpretacje ich wyników. Borsboom ostro skrytykował definiowaną w piątym wydaniu Standardów (American Educational Research Association…, 1999)

[9] „Once the IUA is developed, it provides a framework for validation and it provides criteria for evaluating wheter the proposed interpretation and use have been adequately validated. If argument is coherent and complete and all of its inferences and assumptions are highly plausible […], the interpretation/use would be considered plausible, or valid. If any part of the argument is not plausible, the interpretation/use would not be considered valid".

nadrzędną i ujednolicającą rolę trafności teoretycznej, uważając, że błędem jest przyjęcie idei, iż walidować należy teorie (*construct*), a nie testy.

Ponadto oparcie trafności na sieci nomologicznej, której elementy są ze sobą powiązane, także jest problematyczne, ponieważ w psychologii (i szerzej – w naukach społecznych) wiele rzeczy jest ze sobą skorelowanych i w żaden sposób nie potwierdza to trafności teoretycznej, choć może potwierdzać teorię. Oparcie oceny trafności na korelacji prowadzi też do oceniania jako takich o lepszej trafności testów, które silniej ze sobą korelują, co w ekstremalnym przypadku skutkuje stwierdzeniem, że najbardziej trafny jest ten test, który mierzy to samo co kryterium oceny[10], nawet jeśli jest to korelacja pozorna (Haig, 2003).

> W związku z tym twierdzimy, że tak naprawdę ważne jest to, jak działa test, i z pewnością nie jest to właściwość interpretacji wyników testu, czy nawet wyników testu, ale samego przyrządu pomiarowego (tj. konkretnej, fizycznej rzeczy, którą można upuścić na nogi, a nie jednostki językowej, obiektu teoretycznego lub konstrukcji statystycznej) (Borsboom et al., 2009, s. 149)[11].

Przenoszenie akcentu trafności z właściwości pomiaru na interpretację wyniku pomiaru prowadzi do absurdalnych wniosków. Jeśli test niczego nie mierzy, to i tak może być trafny, jeśli robi to dobrze. Taki wniosek wynika z oparcia oceny trafności na interpretacjach użycia. Trafność powinna być cechą testów jako takich, ale nie oznacza to równocześnie rezygnacji z obecności teorii – pełni ona jednak rolę komponentu przyczynowego w interpretacji wariancji wyników:

> Test jest trafny dla pomiaru atrybutu, jeżeli zmienność atrybutu powoduje zmienność wyników testu (Borsboom et al., 2004, s. 1067)[12].

Według tego podejścia test jest trafny, jeśli spełnione są dwa warunki: 1) istnieje konstrukt, który możemy mierzyć, oraz 2) istnieje przyczynowa relacja między poziomami (natężeniem) konstruktu a wynikami testu. I jeśli u Cronbacha i Meehla sieć nomologiczna jest połączona symetrycznymi związkami korelacyjnymi, to Borsboom i współpracownicy postulują, aby te związki były przyczynowe. Wtedy test jest trafny, jeśli różny stan (poziom, natężenie) badanego konstruktu jest źródłem wariancji wyników testu. Oznacza to, że walidacja powinna dotyczyć bardziej procesów pojawiających się podczas testowania i zachowań, które

---

[10] Wysokość i masa ciała ludzi korelują ze sobą około 0,8 – nie znaczy to jednak, że waga jest trafnym narzędziem do pomiaru wzrostu.

[11] „We submit, therefore, that what really matters in validity is how the test works, and this is certainly not a property of test score interpretations, or even of test scores, but of the measurement instrument itself (i.e., of the concrete, physical thing that you can drop on your feet, rather than of a linguistic entity, set-theoretical object, or statistical construction)".

[12] „A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores".

różnicują osoby, a mniej relacji między badanymi atrybutami połączonymi w sieci nomologiczne. To, co określa trafność, to psychologiczny proces, który oddziałuje na wariancję w wynikach i pozwala na postawienie oraz sprawdzenie hipotez merytorycznych na temat relacji przyczynowo-skutkowych między konstruktem a wynikiem testu. Jednocześnie autorzy podkreślają, że spełnienie tego bez teorii jest awykonalne.

Odnosząc się do definicji trafności Messicka: „Trafność jest zintegrowanym osądem oceniającym" (Messick, 1988, s. 13)[13], Borsboom i współpracownicy twierdzą, że: „Trafność zupełnie nie jest osądem. Jest ocenianą właściwością" (Borsboom et al., 2004, s. 1063)[14].

Dodatkowym argumentem za uznaniem podejścia Borsbooma i przeniesieniem akcentu z oceny przebiegu wnioskowania na właściwość narzędzi badawczych jest nacisk na kontrolę procesu konstruowania testów. Nie jest dobre podejście, zgodnie z którym stworzenie i zastosowanie testu (niezbędne do oceny trafności opartej na uzyskanych wynikach) poprzedza ocenę tego, dla kogo jest on trafny. To już podczas konstruowania testu badacz powinien zadbać o przygotowanie trafnego narzędzia, a nie po jego stworzeniu lub co gorsza – użyciu. Dlatego też to, co jest potrzebne w teorii trafności, to nie tyle modele relacji między mierzonymi konstruktami i kryteriami odniesienia, ile raczej teoria udzielania odpowiedzi w teście.

Argument z testem umiejętności czytania wydrukowanym zbyt małą czcionką i przez to mierzącym też poprawność widzenia nie jest przykładem testu, którego trafność oceny umiejętności czytania jest niska. To przykład testu, który został nietrafnie wykonany. Ten sam test nie może mieć wysokiej trafności dla optometrów do oceny wzroku, ponieważ do tego celu nie trzeba całych zdań, wystarczają pojedyncze litery. Poza tym komponent umiejętności czytania będzie nadal wpływał na wariancję wyników i efekt pomiaru takim testem może być przypadkowy. Nie można powiedzieć, że dla specyficznego celu i specyficznej grupy ten test jest nadal trafny, co ma świadczyć o tym, że trafność to aspekt użycia testu i interpretacji wyników. Po prostu test jest na tyle obciążony wpływem różnych czynników (czytaniem i widzeniem), że ustalenie jednego z tych źródeł wariancji na stałym poziomie (tzn. przeprowadzenie testu na grupie osób dobrze widzących) pozwala stosować go i rozumieć wyniki w kontekście różnicowania czytelnictwa. W praktyce badawczej będziemy jednak raczej drukować test większą czcionką, niż dopisywać kolejne strony w instrukcji dotyczące jego specyficznego zastosowania.

[13] „Validity is an integrated evaluative judgment".

[14] „Validity is not a judgment at all. It is the property being judged".

3.3. Badania własne

Minęło już prawie dziewięćdziesiąt lat badań nad trafnością narzędzi badawczych w psychologii. Regularnie publikowane standardy dotyczące testów wpływają na ujednolicanie koncepcji i rozwiązań praktycznych związanych z tworzeniem narzędzi badawczych i określaniem ich jakości. Chciałbym się zatem przyjrzeć temu, jak w praktyce psychologowie i psychometrzy definiują i badają trafność konstruowanych przez siebie narzędzi. W tym celu przeprowadziłem systematyczny przegląd literatury.

3.3.1. Metoda

Celem systematycznego przeglądu literatury było określenie dominującego modelu badania trafności narzędzi do badań psychologicznych. Badanie jest systematyczne w tym sensie, że włączono w jego obszar wszystkie artykuły naukowe z wybranych źródeł opublikowane w przyjętych ramach czasowych.

3.3.1.1. Kryteria włączania i wykluczania

Zastosowano następujące kryteria włączania artykułów do przeglądu: 1) publikacja musiała dotyczyć walidacji psychometrycznej narzędzia do badań kwestionariuszowych lub testowych; 2) w tytule, abstrakcie, słowach kluczowych lub wstępie musiała być zawarta informacja, a cel (jeden z celów) publikacji dotyczył określenia właściwości psychometrycznych użytych w badaniu narzędzi; 3) artykuł musiał być opublikowany w recenzowanym czasopiśmie; 4) z danego rocznika powinny w bazie znaleźć się minimum trzy pozycje; 5) tekst musiał być opublikowany w języku angielskim.

Wykluczano podręczniki, artykuły traktujące o testach przesiewowych, artykuły teoretyczne, metaanalizy oraz przeglądy literatury. Dodatkowym kryterium wykluczania był autor publikacji i narzędzie badawcze – główny autor oraz przedmiot walidacji mogły być uwzględnione tylko raz, by nie obciążać powstałej próby specyficznym tematem lub preferencjami konkretnej osoby.

3.3.1.2. Strategia przeszukiwania

Jako reprezentatywną próbę opisującą rzeczywiście stosowane metody walidacji narzędzi badawczych postanowiłem przeanalizować 100 artykułów naukowych. Przyjętą przeze mnie datą początkową był rok 1999, tj. ten, w którym opublikowano przedostatnie Standardy (American Educational Research Association…, 1999). Do wyszukania artykułów wykorzystałem wyszukiwarkę Google Scholar oraz platformę EBSCO z dostępem do bazy PsycArticles. Przeglądu literatury

dokonałem w listopadzie 2018 roku. Wyszukiwarka Google Scholar na pytanie zawierające nazwę testu lub kwestionariusza oraz słowa *validation of* lub *validity*, lub *psychometric* wskazała 39 200 rekordów, z których wybrałem 50 artykułów według opisanych wyżej kryteriów. Średnio na 10 wyników 6,5 zawierało link do pliku PDF lub odnośnik DOI, który pozwalał pobrać publikację z dostępnych źródeł. Pozostałe 50 rekordów dobrałem z bazy PsycArticles (3353 rekordy w odpowiedzi na to samo zapytanie), omijając te, które zostały już zaklasyfikowane do próby po użyciu Google Scholar, zawierały tylko abstrakt lub treść dostępna była tylko po opłaceniu dostępu. Średnio 4,8 artykułu na 10 było dostępnych bezpośrednio poprzez platformę EBSCO.

3.3.1.3. Procedura

Dla każdego włączonego do przeglądu artykułu zanotowałem datę publikacji, nazwiska autorów i nazwę walidowanego testu oraz obszar działań badawczych (a więc to, czy opisywane narzędzie było stworzone na potrzeby publikacji, czy była to walidacja już istniejącego narzędzia, ponowna walidacja po wprowadzeniu zmian do narzędzia, jego adaptacji lub po stworzeniu wersji skróconej). Kontrolowałem też: liczbę wymiarów narzędzia i wielkość próby wraz ze średnim wiekiem osób w grupie walidacyjnej. W obszarze trafności notowałem użytą metodę statystyczną i jej najważniejszy wynik: dla czynnikowych analiz eksploracyjnych był to procent wyjaśnionej wariancji, dla konfirmacyjnej analizy czynnikowej miara dopasowania CFI oraz RMSEA (wraz z górną 90-procentową granicą przedziału ufności), dla analiz korelacyjnych wielkość współczynnika korelacji. W obszarze rzetelności notowałem wielkość i rodzaj podanego współczynnika – przeważnie była to alfa Cronbacha lub omega McDonalda, a dla miary stabilności długość odstępu i wielkość współczynnika korelacji test-retest. W przypadku gdy walidowane narzędzie składało się z wielu skal, za wynik przyjmowałem średnią z wyników dla podanych wskaźników. Jeśli wyniki bardzo różniły się między sobą i średnia byłaby wynikiem mocno obciążonym, to za sumaryczny wynik przyjmowałem medianę ze wszystkich podanych w pracy wskaźników. Jeżeli w publikacji opisano kilka narzędzi lub kilka etapów (np. najpierw EFA, a potem CFA przeprowadzone na innych próbach), to dla każdego z nich zanotowałem osobne wartości wskaźników.

W większości przypadków wartości podane były bezpośrednio, a jeśli nie było możliwe ich odczytanie, starałem się uzyskać te informacje pośrednio. Szczególnie często dotyczyło to przeciętnego wieku osób w przebadanej próbie: w 35% przypadków zamiast odchylenia standardowego podawany był zakres wartości. W tych sytuacjach przyjąłem, że zakres obejmuje 6 odchyleń standardowych (3 powyżej średniej oraz 3 poniżej średniej), co pozwoliło estymować wartość odchylenia standardowego dla wieku. Podobną strategię estymowania potrzebnej wartości przyjąłem w przypadku wskaźnika dopasowania modelu RMSEA

i braku przedziałów ufności. Dla 27 przypadków 95-procentowy półprzedział ufności wyznaczyłem ze wzoru:

$$\mathrm{PU} = \sqrt{1/N} \cdot 1{,}96 \cdot \mathrm{RMSEA}).$$

3.3.1.4. Problem

101 zebranych artykułów dostarczyło 178 wyników będących przekrojowym przeglądem sposobów i efektów walidacji właściwości psychometrycznych narzędzi badawczych używanych w badaniach psychologicznych. Artykuły były opublikowane w latach 1999–2018 i przeciętnie opisywały 1,8 (*SD* = 1,07) analizy. Podstawowym celem przeglądu było określenie typowej procedury służącej walidacji narzędzi badawczych w obszarze psychologii: Jakie metody walidacji są najczęściej wykorzystywane? Jakie wskaźniki statystyczne są przedstawiane przez autorów?

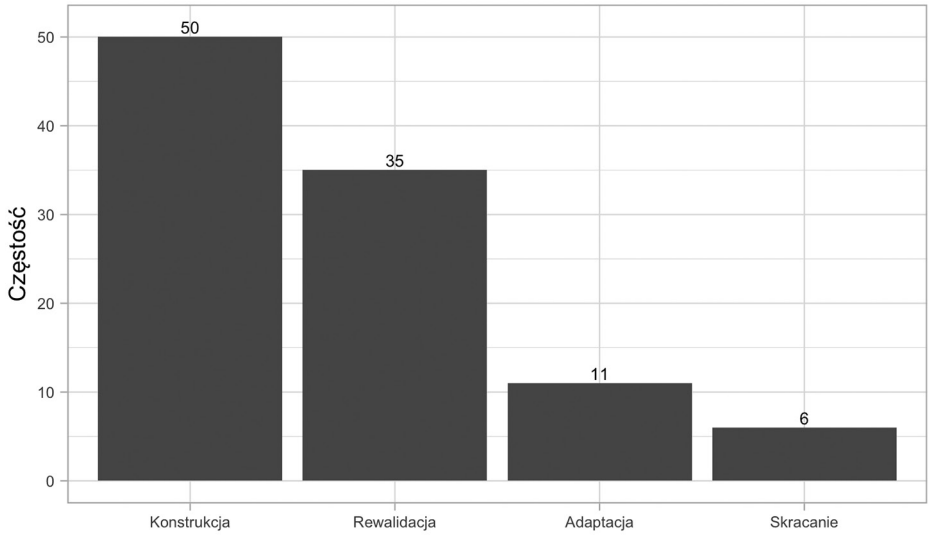
3.3.1.5. Charakterystyka zebranych publikacji

W analizowanym okresie najwięcej (65) publikacji poświęconych walidacji narzędzi badawczych (w 100% testy lub skale kwestionariuszowe) ukazało się w 2018 roku. Prawdopodobnie jest to efekt przedstawiania wyników wyszukiwania przez użyte bazy danych, które najnowsze publikacje przedstawiają na początku listy wyników. Przeciętnie w każdym badanym roku zarejestrowałem 5,5 analizy. Najczęstszym powodem publikacji była konstrukcja nowego narzędzia badawczego (*n* = 50, 49,02%), na kolejnych miejscach były rewalidacje (ponowna ocena właściwości psychometrycznych) i adaptacje (dostosowanie narzędzia badawczego do lokalnego języka), a na końcu tworzenie wersji skróconych (rys. 3.2).

Typowa liczba wymiarów przekładająca się na liczbę skal w danym narzędziu wynosiła 3,5 (od 1 do 30, pierwszy i trzeci kwartyl rozkładu liczebności wymiarów wynosił odpowiednio: *Q*1 = 2, *Q*3 = 5). Natomiast liczba pozycji w skali wynosiła przeciętnie 26 (od 4 do 232, *Q*1 = 17, *Q*3 = 39,75). We wszystkich 178 analizach wzięło udział 89 556 osób badanych – średnia wielkość próby wyniosła *M* = 508,8 (*Md* = 290, *SD* = 805,66, *Q*1 = 174,25, *Q*3 = 512,5). Rozkład wieku w badanych próbach przedstawiono na rys. 3.3.

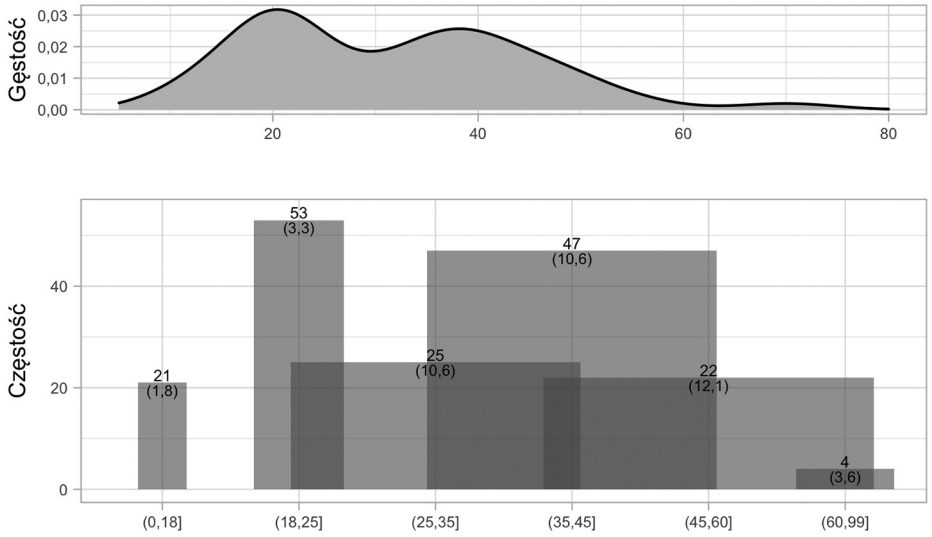
Badania na grupach najmłodszych, do 18. roku życia, stanowiły 12% analizowanych publikacji i tu zróżnicowanie wiekowe było najmniejsze – odchylenie standardowe wynosiło około 1,8 roku. Najwięcej badań (31%) przeprowadzonych było na osobach w wieku od 18 do 25 lat (to najbardziej dostępne grupy dla wielu badaczy, tj. studenci), a poszczególne grupy badawcze pod względem wieku były przeciętnie zróżnicowane o 3,3 roku.

Grupy starsze były o wiele bardziej heterogeniczne – przeciętne odchylenie standardowe było większe niż 10 lat, a dopiero badania ukierunkowane na osoby najstarsze (2% ogółu analizowanych publikacji) cechowały się węższym



Typy działań: konstrukcja – opis procesu tworzenia narzędzia do badań wraz z szacowaniem jego właściwości psychometrycznych; rewalidacja – ponowna ocena właściwości psychometrycznych na nowej grupie lub w nowym zastosowaniu; adaptacja – opis procesu i wyniku adaptacji narzędzia badawczego, w większości przypadków dostosowanie wersji anglojęzycznej do lokalnego języka (w jednym przypadku z francuskiego na angielski); skracanie – opis procesu i wyniku przygotowania nowej, krótszej wersji na bazie oryginału.

Rysunek 3.2. Częstość poszczególnych typów działań walidacyjnych



Górna część rysunku przedstawia gęstość rozkładu wieku. Dolna część przedstawia częstość poszczególnych grup wiekowych w badanych publikacjach – liczby nad słupkami przedstawiają liczbę grup w określonych przedziałach wiekowych, liczby w nawiasach to średnie odchylenie standardowe w poszczególnych grupach. Szerokość słupków jest proporcjonalna do wielkości średniego zróżnicowania.

Rysunek 3.3. Częstość i zróżnicowanie poszczególnych grup wiekowych

rozrzutem wieku. Dominację grupy osób w wieku około 20 lat dobrze widać na górnym wykresie (rys. 3.3) przedstawiającym gęstość rozkładu wieku w analizowanych publikacjach.

3.3.2. Wyniki

Najczęstszą metodą statystyczną stosowaną do określenia trafności narzędzi badawczych w analizowanych publikacjach była konfirmacyjna analiza czynnikowa (CFA) – 37% przypadków w stosunku do 28% przypadków wykorzystywania korelacji i 26% przypadków używania eksploracyjnej analizy czynnikowej (EFA). Te trzy techniki łącznie stanowiły ponad 91% przypadków, co dobrze ilustruje rys. 3.4. Jeśli włączyć analizę głównych składowych (PCA) jako także eksploracyjną analizę czynnikową, to odsetek ten wzrośnie do 93%.

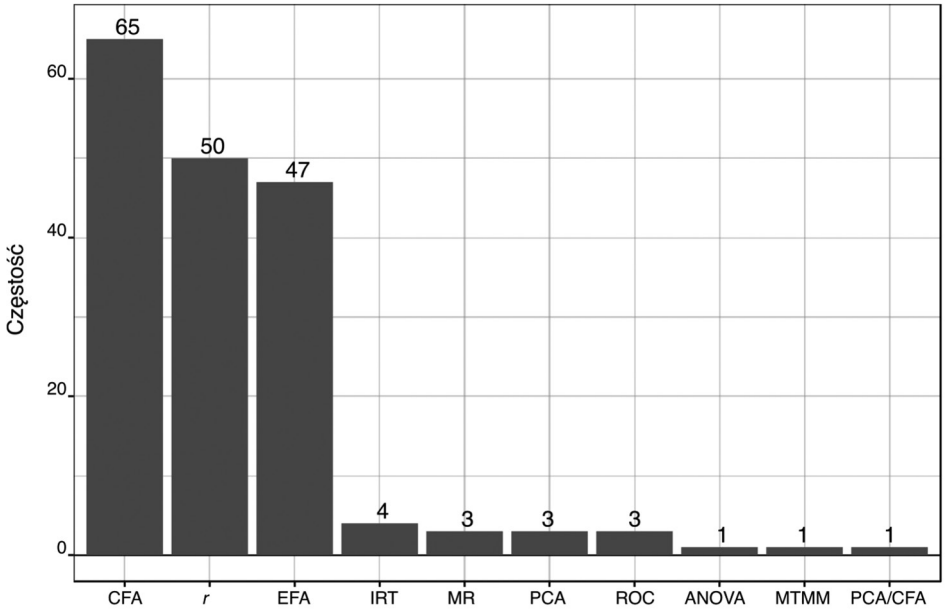
Poza trzema wymienionymi technikami statystycznymi sporadycznie (mniej niż 3%) do walidacji narzędzi wykorzystywane były: teoria odpowiadania na pozycje testowe (IRT), regresja wielokrotna (MR), analiza głównych składowych (PCA)[15], krzywe detekcji sygnału (ROC), analiza wariancji (ANOVA) i tylko raz użyto rekomendowanej przez Campbella i Fiskego (1959) metody wielu cech wielu metod!

Pozostając przy trzech najczęściej wykorzystywanych technikach statystycznych do oceny trafności, dwie z nich – współczynniki korelacji oraz eksploracyjna analiza czynnikowa – nie wymagają założeń teoretycznych i dostarczają wyników tylko na podstawie danych. To pozwala wyciągać badaczom wnioski w sposób subiektywny, ponieważ nie istnieją kryteria mówiące o tym, jak bardzo dany układ wyników wspiera założenia teoretyczne oraz pozwala ukryć, czy teoria nie była mniej lub bardziej świadomie dostosowywana do wyników. Metodą, która dostarcza informacji o relacji między założeniami teoretycznymi a obserwowanymi danymi w postaci wskaźników dopasowania modelu do danych, jest konfirmacyjna analiza czynnikowa. Nie zawsze jest ona jednak wykorzystywana (65 razy na 101 walidowanych narzędzi). Zwracam też uwagę na to, że autorzy 56 publikacji dotyczących oceny właściwości psychometrycznych narzędzia wykorzystali tylko jeden sposób analizy danych (ze wszystkich wymienionych powyżej), 21 użyło dwóch technik (eksploracyjnej i konfirmacyjnej), a 15 autorów użyło trzech technik (eksploracyjnej, konfirmacyjnej i korelacji z wynikami innego narzędzia)[16].

---

[15] W jednym przypadku na tych samych danych przeprowadzono analizę eksploracyjną i konfirmacyjną (co nie jest rekomendowane), stąd przypadek PCA/CFA.

[16] Przypadki użycia większej liczby technik (4 oraz 5) wiązały się z większą liczbą badań w jednej publikacji, np. kilkoma wersjami pośrednimi lub analizami poświęconymi kilku narzędziom.

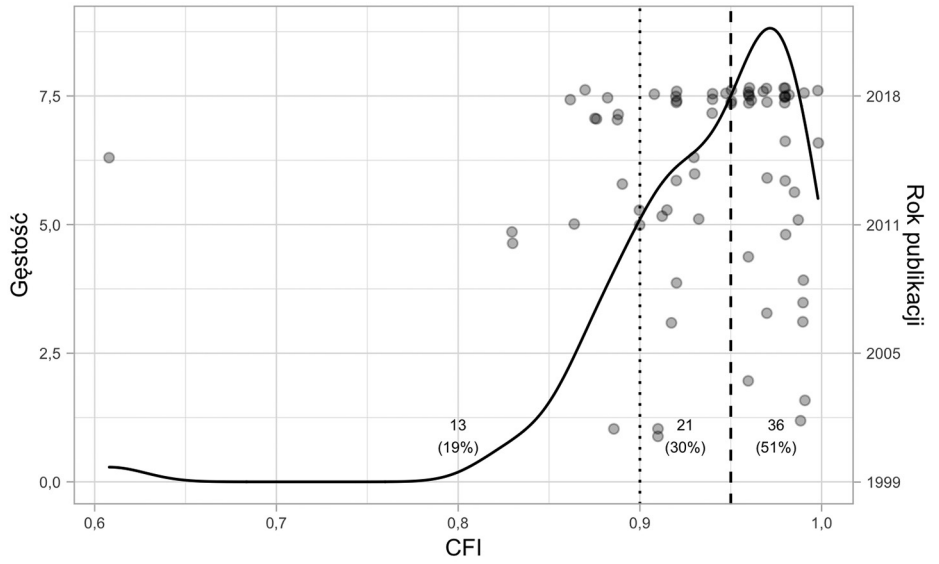


Rysunek 3.4. Częstość wykorzystania poszczególnych metod statystycznych przy określaniu właściwości psychometrycznych narzędzi badawczych (*N* = 178)

Dla każdej z technik statystycznych przyjęte są progowe wartości pozwalające uznać wynik za dowód potwierdzający trafność wyników narzędzia. I tak dla CFA absolutnym minimum jest podanie wskaźnika dopasowania do modelu wysyconego CFI (Comparative Fit Index), który kwantyfikuje rozbieżność między danymi a hipotetycznym modelem, uwzględniając jednocześnie wielkość próby. Ostatnie badania wykazały, że minimalny akceptowalny poziom wskaźnika CFI wynosi 0,90 (a im więcej, tym lepiej) (Hu i Bentler, 1999), zaś model dobrze dopasowany do danych posiada wskaźnik na poziomie większym lub równym 0,95[17].

Jak można zauważyć na rys. 3.5, nie istnieje związek między rokiem publikacji a poziomem wskaźnika (r_s = 0,079), niemniej tylko połowa zamieszczonych wartości CFI w analizowanych publikacjach potwierdza trafność narzędzi badawczych

[17] Wielkość tego wskaźnika zależy od użytego w analizie estymatora – zastąpienie domyślnego estymatora ML (*maximum likelihood*) estymatorem nieparametrycznym, np. DWLS (*diagonal weighted least of squares*) poza wzrostem wymagań co do wielkości próby, zwiększa także wymaganą wielkość wskaźnika do 0,99. Niestety wciąż nieliczni autorzy podają szczegóły dotyczące sposobu obliczania parametrów analizowanych przez siebie modeli.



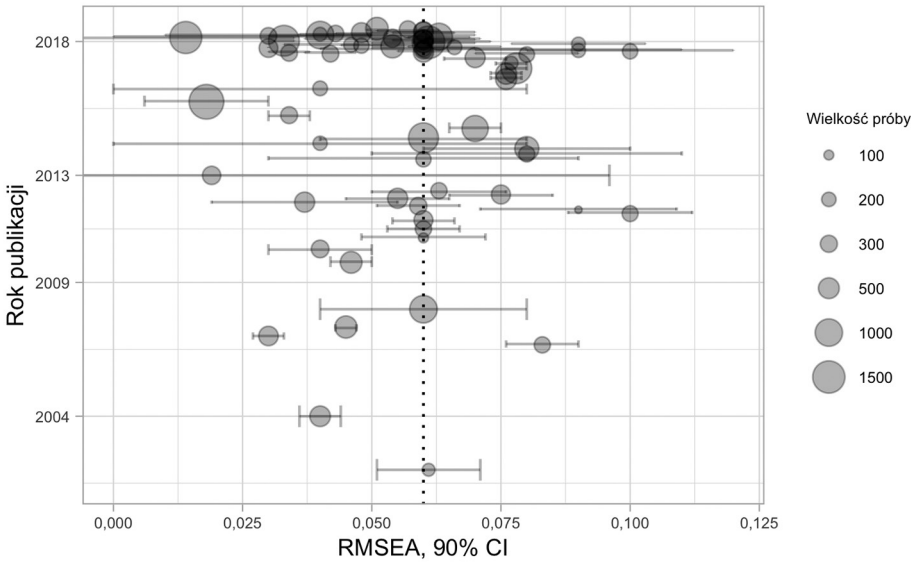
N = 70. Linia ciągła przedstawia gęstość rozkładu wartości wskaźnika CFI, natomiast punkty przedstawiają dokładną jego wartość. Pionowa oś dla roku publikacji jest orientacyjna, a do współrzędnych punktów dodano losową wartość w celu poprawienia czytelności ilustracji.

Rysunek 3.5. Gęstość rozkładu wskaźników dopasowania CFI w analizowanych publikacjach z uwzględnieniem roku publikacji

w znaczeniu dopasowania wyników do teoretycznych założeń. Pozostałe 30% jest w tzw. „szarej strefie", gdzie nie można rozstrzygnąć, czy wyniki potwierdzają założenia teoretyczne co do struktury, a 19% analiz wskazuje na brak dopasowania – wyniki są sprzeczne z oczekiwaniami badaczy.

Drugim wymaganym wskaźnikiem w przypadku wykorzystania CFA do oceny trafności wyników danego narzędzia badawczego jest pierwiastek ze średniokwadratowego błędu aproksymacji (*root mean square error of approximation* – RMSEA), który pozwala oszacować rozbieżność między danymi i modelem, a dokładniej – między macierzą kowariancji teoretyczną i obserwowaną. Zgodnie z propozycją Hu i Bentlera (1999) przyjmuje się, że wartość 0,06[18] lub mniejsza wskazuje na akceptowalne dopasowanie modelu do danych. Zwyczajowo podaje się także 90-procentowy przedział ufności dla tego wskaźnika, który pozwala oszacować, „jak blisko ideału" jest obserwowany model.

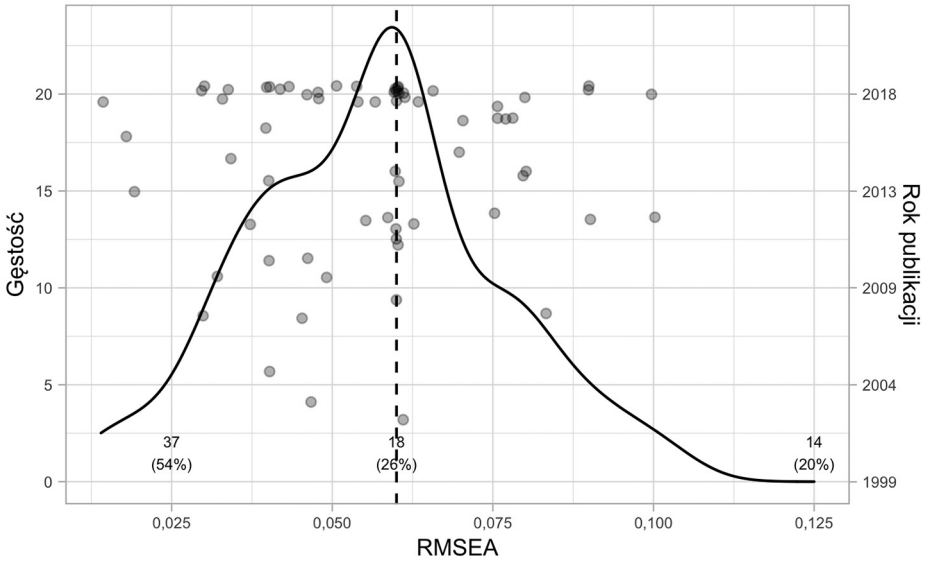
[18] Dla estymatorów nieparametrycznych wartość ta wynosi 0,04.



Wielkość punktów jest proporcjonalna do wielkości próby w danym badaniu. Kropkowana pionowa linia oznacza przyjęty próg minimalnej wartości RMSEA wskazującej na model dopasowany do danych.

Rysunek 3.6. Rozkład wyników wskaźnika RMSEA (punkty) wraz z 90-procentowym przedziałem ufności (linie poziome) w poszczególnych latach

W analizowanych publikacjach (n = 69) nie spostrzegłem zależności wielkości tego wskaźnika od roku publikacji (rys. 3.6) – pozyskane wyniki wydają się losowe, choć analiza gęstości rozkładu (rys. 3.7) wykazuje podwyższoną liczbę wskaźników o granicznej wartości 0,06 (26%). Być może wynika to z zaokrąglania wyników lub z takiego parametryzowania modeli przez badaczy, które umożliwia przekroczenie progu wartości pozwalającego potwierdzić trafność w aspekcie teoretycznym walidowanego narzędzia. Sądzę też, że do roku 2010 może brakować doniesień o wartości wskaźnika RMSEA nieprzekraczającej wymaganego progu 0,06, co można hipotetycznie tłumaczyć uprzedzeniami publikacyjnymi i nieujawnianiem wartości tego wskaźnika, gdy nie był on korzystny dla analizy. Z kolei pojawienie się w publikacjach ostatnich lat wartości z przedziału 0,06–0,10 może być spowodowane tym, że próg, jaki zaproponowali Hu i Bentler, przez kilku badaczy był uznany za zbyt restrykcyjny i pojawiły się w literaturze propozycje interpretacji tego przedziału wartości jako wskaźnika modelu



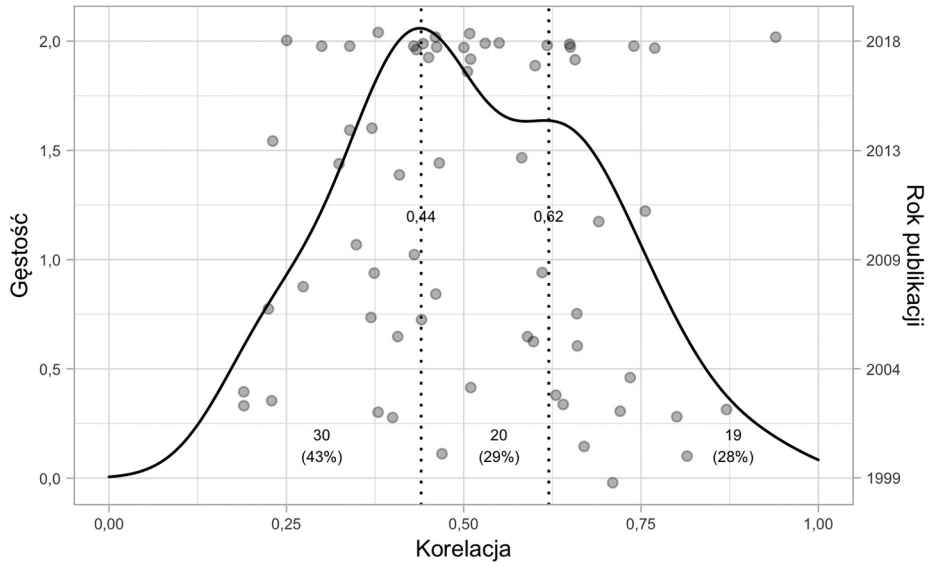
Linia ciągła przedstawia gęstość rozkładu wartości wskaźnika RMSEA, natomiast punkty prezentują dokładną jego wartość. Oś pionowa dla roku publikacji jest orientacyjna, a do współrzędnych punktów dodano losową wartość w celu poprawienia czytelności ilustracji.

Rysunek 3.7. Rozkład wartości wskaźnika dopasowania RMSEA z uwzględnieniem roku publikacji

miernie (*mediocre*) dopasowanego[19]. Wydaje się też, że próby o większej liczebności sprzyjają uzyskiwaniu mniejszych wartości RMSEA.

Wśród analizowanych walidacji znalazła się tylko jedna przeprowadzona za pomocą metody wielu cech wielu metod, ale współczynnik korelacji między wynikami prezentowanego narzędzia a wynikami innych, zarówno w aspekcie zbieżnym, jak i rozbieżnym trafności, przytoczony był w 63 przypadkach (35% wyników). Analizując rozkład wskaźników korelacji, zauważyłem, że ponad 70% wyników jest jednoznacznych – albo silnych korelacji (> 0,60 – wartość lokalnego maksimum rozkładu gęstości wynosi 0,62), świadczących o związku między badanymi konstruktami, albo słabych korelacji (< 0,40 – wartość lokalnego maksimum rozkładu gęstości wynosi 0,44), świadczących na rzecz aspektu rozbieżnego trafności między korelowanymi konstruktami. Aż około

[19] Por. D.A. Kenny, *Measuring Model Fit*, http://davidakenny.net/cm/fit.htm, dostęp: 6 stycznia 2019.



Linia ciągła przedstawia gęstość rozkładu wartości korelacji test-test. Na rysunku zaznaczono hipotetyczne średnie dwóch nakładających się rozkładów – dla korelacji potwierdzających aspekt rozbieżny oraz dla korelacji potwierdzających aspekt zbieżny trafności.

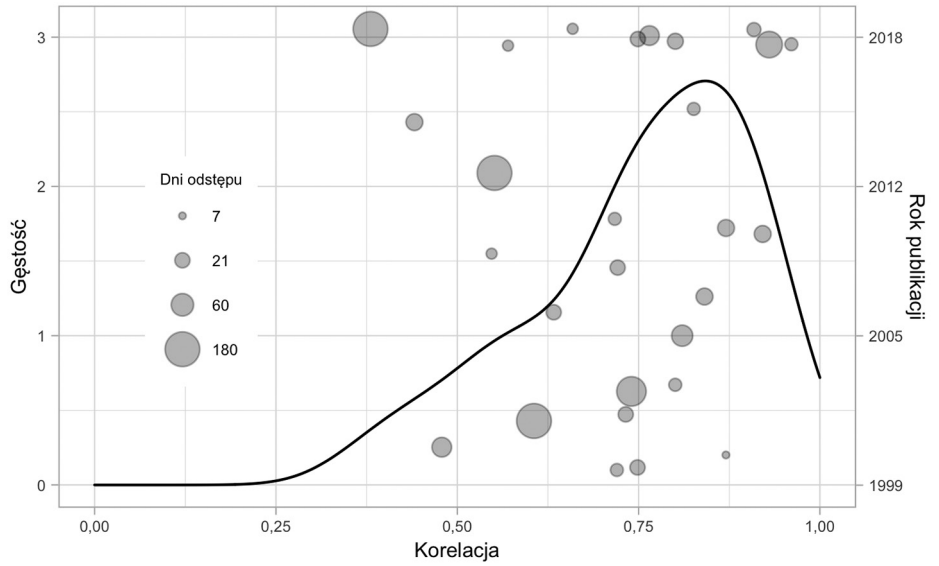
Rysunek 3.8. Rozkład wartości wskaźnika korelacji test-test z uwzględnieniem roku publikacji

29% analizowanych przypadków zawiera jednak wyniki nic lub niewiele mówiące o trafności badanych konstruktów (rys. 3.8).

Oprócz analizy trafności 58% analiz (n = 104) zawierało także oszacowania rzetelności wyników oparte na współczynniku alfa Cronbacha (M = 0,85, Md = 0,88, $Q1$ = 0,81, $Q3$ = 0,91)[20], a 19% analiz (n = 34) ogółem podawało rzetelność w rozumieniu stabilności mierzonej współczynnikiem korelacji test-retest (M_r = 0,75, Md_r = 0,78, $Q1_r$ = 0,66, $Q3_r$ = 0,87), przy odstępie od siedmiu dni do, w skrajnym przypadku, czterech lat, przeciętnie były to trzy tygodnie. Grupę 19% analizowanych badań zawierających stabilność wyników zilustrowałem na rys. 3.9.

Nie jest zaskoczeniem, że dłuższe odstępy sprawiają, iż korelacje są słabsze, ponadto danych jest zbyt mało, a także nie są reprezentatywne dla obszaru stabilności wyników testu (np. nie kontrolowałem treści walidowanych konstruktów), by wyciągać z nich wiążące wnioski.

[20] W pojedynczych przypadkach użyto współczynnika omega McDonalda.



Linia ciągła przedstawia gęstość rozkładu wartości korelacji test-retest. Wielkość punktów reprezentuje odstęp czasowy między badaniami.

Rysunek 3.9. Rozkład wartości wskaźnika korelacji test-retest z uwzględnieniem roku publikacji

3.4. Podsumowanie

Przedstawiony przegląd teoretyczny rozwoju pojęcia trafności pozwala lepiej rozumieć koncepcję trafności i staje się bardziej zrozumiałe, dlaczego nie jest to w naukach społecznych prosta koncepcja. Tradycyjna definicja trafności zakłada stosowanie testu we właściwy sposób, na właściwych osobach badanych i we właściwych warunkach – w takiej sytuacji badany konstrukt można precyzyjnie zmierzyć i podjąć decyzje, które powinny przynieść pozytywny efekt. Pojęcie konsekwencji tych decyzji jest najistotniejsze dla myślenia o trafności, ponieważ cechy latentne nie mogą być obserwowane bezpośrednio, a o ich poziomie (stanie) wnioskuje się na podstawie obserwowanych wskaźników. Dlatego trafność narzędzia badawczego oceniana jest w kontekście procesu wnioskowania. Nie można jednak oddzielić trafności od procesu budowy narzędzia – teoria, która jest podstawą do przygotowania narzędzia badawczego, musi znaleźć swoje miejsce w procesie oceny jego trafności.

W związku z tym ocena trafności musi się odbywać na dwóch wymiarach: wewnętrznym i zewnętrznym oraz teoretycznym i praktycznym. Zaletą takiego myślenia o trafności (tab. 3.3) jest zwrócenie uwagi na jej aspektowość – zbieranie dowodów w jednym wymiarze nie przesądza o jakości narzędzia, choć jest silną przesłanką do takiego stwierdzenia.

Tabela 3.3. Dwuwymiarowość aspektów pojęcia trafności

Wymiary trafności	Wewnętrzny	Zewnętrzny
Teoretyczny	teoretyczna	zbieżna/rozbieżna
Praktyczny	treściowa	prognostyczna + diagnostyczna

Bez względu na to, w jaki sposób zdefiniujemy pojęcie trafności, proces jej walidacji może (za Cronbachem, 1989, s. 162) przyjąć postać słabego lub silnego programu (w sensie rozwoju nauki). Słaby program to zbieranie empirycznych doniesień, w których każda korelacja z inną zmienną jest w jakiś sposób dobra. To droga, na której wytłumaczenie dla obserwacji znajduje się *post factum*, a określenie trafności narzędzia nie ma końca – zadaniem psychometrów jest przeprowadzanie kolejnych badań z ich nieskończonej liczby kombinacji interakcji: narzędzie – osoby badane – kontekst. Natomiast silny program według Cronbacha polega na stawianiu hipotez, szukaniu i odrzucaniu alternatywnych wyjaśnień. W tym sposobie uprawiania psychometrii każde działanie jest celowe i poprzedzone uzasadnieniem jego podjęcia wraz z określeniem miary sukcesu (odrzucenia lub przyjęcia stawianej co do trafności hipotezy).

Przeglądając strategie użyte przez badaczy opublikowane w czołowych czasopismach psychologicznych prezentujących doniesienia na temat walidacji narzędzi badawczych (*Assessment, Educational & Psychological Measurement*, *Journal of Personality Assessment*, *Journal of Psychoeducational Assessment*, *Psychological Assessment*), Robert F. Bornstein (2011) zbadał 486 artykułów. Znakomita większość (91%) wykorzystywała analizę korelacji (wśród nich tylko 24% korzystało z równań strukturalnych), co więcej – prawie 80% z nich opierało się tylko na wynikach narzędzi samoopisowych. Można powiedzieć, że w większości badań trafność narzędzia potwierdzają sami badani podczas pojedynczego badania! W znaczeniu przytoczonej wyżej definicji Cronbacha stosunek słabych badań do silnych wynosi 9 : 1.

Analizując zebrane przeze mnie sposoby i wyniki analiz trafności różnorodnych narzędzi badawczych, daje się zauważyć konsensus co do tego, jak praktycznie powinna wyglądać procedura walidacji. Niezwykłym aspektem tego konsensusu jest to, że pojęcie trafności, z którym zmagają się teoretycy, wydaje się dziwnie nieobecne w praktyce badawczej, kiedy stawia się pytanie o trafność narzędzia. Uzyskany wynik jest realny w przeciwieństwie do pojęcia, które

określa. Obawy Borsbooma i współpracowników (Borsboom, 2006; Borsboom et al., 2009) dotyczące błędnego skupiania się na walidacji procesu szacowania wyników i interpretacji okazują się płonne – znakomita większość badaczy, jeśli nie wszyscy, określa trafność jako właściwość narzędzia.

Najpopularniejszą metodą oceny trafności na etapie eksploracyjnym jest ocena ładunków krzyżowych w wynikach analizy czynnikowej. Dowodem na istnienie struktury potwierdzającej założenia teoretyczne stojące za budową narzędzia jest takie rozłożenie wartości ładunków czynnikowych w macierzy kowariancji, by można było stwierdzić, że każdy czynnik definiuje inny zbiór powiązanych ze sobą pozycji. Występowanie ładunków krzyżowych – wysokich wartości (za takie uważa się wartości 0,32 lub więcej) dla danej pozycji na więcej niż jednym czynniku jest niepożądane (Costello i Osborne, 2005). Analiza ładunków krzyżowych pozwala na oszacowanie trafności w aspekcie różnicowym (*item-level discriminant validity*). Niskie wartości ładunków czynnikowych uzyskuje się, jeśli po usunięciu wariancji dla pożądanego czynnika reszta nie koreluje z pozostałymi czynnikami. Nawet dla niskich kowariancji z czynnikiem pożądanym pozostała kowariancja nie powinna wiązać się z innymi czynnikami, ponieważ jest ona teoretycznie tylko czystą wariancją błędu (Gefen i Straub, 2005, za: Henseler, Ringle i Sarstedt, 2015, s. 118). Kryterium wartości 0,32 jest kryterium bezwzględnym oraz arbitralnym i wynika z przyjęcia progu wspólnej wariancji mniejszej niż 10%.

Niektórzy autorzy proponują względne, bardziej liberalne kryterium oparte na zasadzie większości ładunku na czynniku pożądanym od ładunków na pozostałych czynnikach (Chin, 2010). Ocenę trafności opiera się wtedy na analizie stosunku średniej wydobytej wariancji (*average variance extracted* – AVE) do błędu standardowego. Kryterium AVE-SE Fornella–Larckera (Fornell i Larcker, 1981) wymaga, aby średnia wartość ładunków czynnikowych na wyłonionym czynniku była większa niż największy kwadrat korelacji ładunku z jakimkolwiek innym czynnikiem.

Analizowanie wyników korelacji nie jest złe samo w sobie, ale związki takie dowodzą trafności narzędzi badawczych w psychologii tylko wtedy, gdy badają jasno sformułowaną hipotezę dotyczącą empirycznego związku między reprezentowanymi przez wyniki cechami. Jeśli nie mamy uzasadnionej teoretycznie hipotezy na temat związku, to większość korelacji podawanych jako dowód na trafność łatwiej jest wyjaśnić jako odzwierciedlenie nakładania się definicji badanych cech na poziomie semantycznym.

Dojrzałą metodę oceny trafności opartą na analizie macierzy korelacji zaproponowali już w 1959 roku Campbell i Fiske. Metoda wielu cech wielu metod pozwala oszacować trafność zarówno w aspekcie zbieżnym, jak i rozbieżnym. Ponieważ brak w tej propozycji analizy wskazania jej jednoznacznego wyniku, podejmowane są próby ustalenia prostego kryterium takiej macierzy. Taką propozycją jest iloraz wariancji wielu cech do wariancji cechy (*heterotrait-monotrait ratio* – HTMT).

Licznik wskaźnika HTMT to średnia korelacja z macierzy wielu cech i wielu metod (korelacje wskaźników pomiędzy konstruktami mierzącymi różne zjawiska). Mianownik to średnia geometryczna średnich korelacji dla obu (lub więcej) cech i wielu metod (korelacje wskaźników w ramach tego samego konstruktu). Wartości równe 1 lub więcej świadczą o braku trafności (por. Henseler et al., 2015, ss. 120–122). Jak wykazali autorzy, przy homogenicznych, wysokich ładunkach powszechnie używane metody (analiza ładunków krzyżowych, indeksy modyfikacyjne, kryterium Fornella–Larckera) znacznie rzadziej identyfikują problemy w strukturze kowariancji (do maksymalnie 50% przypadków w badaniach symulacyjnych, gdy wskaźnik HTMT osiąga czułość rzędu 99%). Autorzy proponują próg o wartości 0,85 jako graniczną wartość, poniżej której można uznać „różność" badanych konstruktów. Wyniki powyżej tej wartości świadczą o znacznym nakładaniu się treści i wysokim prawdopodobieństwie, że badane narzędzia mierzą to samo.

***

Obszarem mojego zainteresowania było faktyczne zastosowanie ram teoretycznych do wnioskowań na temat trafności narzędzi pomiarowych. Przeprowadzony systematyczny przegląd literatury w celu wyłonienia typowej procedury walidacji narzędzi badawczych w psychologii pokazał, że badacze w większości świetnie posługują się dostępnymi pakietami statystycznymi i wykorzystują ofertę metod statystycznych dostępnych w naukach humanistycznych. Przytłaczająca popularność metody konfirmacyjnej stawia jej użycie w roli praktycznego panaceum na problemy dowodzenia powiązań teorii z obserwacjami. Co pozytywne, prawie tak samo popularne jest używanie metod eksploracyjnych, lecz negatywnym zjawiskiem jest wciąż bazowanie na współczynnikach korelacji.

Optymalne jest użycie zestawu metod: analizy eksploracyjnej (EFA) w fazie konstruowania narzędzia, a potem analizy konfirmacyjnej (CFA) na etapie jego walidacji. Zaletą takiego zestawu technik jest swoista kompatybilność tych metod. Wyniki EFA łatwo zrozumieć i przełożyć na model, który posłuży do analizy CFA. To, czy dany model jest dobrze wyspecyfikowany i czy nie popełniono błędów podczas fazy eksploracyjnej, błędów wynikających z problemów pomiarowych, nie jest poddawane refleksji. Często teoria jest podstawą do przygotowania zestawu pytań/zadań/bodźców, ale stosowane w powszechnej praktyce szacowanie trafności nie podaje w wątpliwość jakości tego procesu, przenosząc punkt uwagi na to, co można zrobić z wynikami generowanymi przez narzędzie.

Mimo żywej dyskusji między teoretykami reprezentującymi dwa stanowiska – trafność narzędzia vs trafność procesu stosowania narzędzia – w praktyce nie obserwuje się jej odzwierciedlenia. Przyjęty powszechnie jest pewien kanon

raportowania statystycznych dowodów na trafność proponowanych narzędzi do badania zdefiniowanych przez badaczy konstruktów, związany, mam wrażenie, z dostępnością metod obliczeniowych w pakietach statystycznych. Większość z nas wierzy[21], że wyniki tych metod są jednoznacznym dowodem na trafność. Niestety, podobnie jak w pomiarze, samo używanie liczb nie jest dowodem na możliwość pomiaru, tak jak w ocenie trafności używanie macierzy kowariancji (z zastosowaniem jakkolwiek zaawansowanych wskaźników) nie dowodzi istnienia badanego konstruktu, co postaram się pokazać w rozdziale 4.

[21] Wiem, że stwierdzenie obecności „wiary" w nauce może być obrazoburcze, ale niektórzy autorzy składają takie deklaracje w kontekście badania trafności, np. Sireci, 2009, s. 20.

4. WALIDACJA NARZĘDZI BADAWCZYCH W PSYCHOLOGII

> Nie ma żadnych dowodów na to, że testy cokolwiek mierzą lub że atrybuty, które testerzy chcą mierzyć, są mierzalne, więc przedstawianie testów jako instrumentów pomiarowych jest przedstawianiem mitu.
>
> Michell, 2009, s. 112[1]

4.1. Odporność procedury walidacji narzędzi badawczych w psychologii

Nie będzie przesadą stwierdzenie, że metody badawcze w naukach humanistycznych opierają się głównie na kwestionariuszach samoopisowych. Od badań w medycynie, przez edukację, marketing, aż po sąd i wojsko, wszędzie, gdzie w grę wchodzi badanie psychologiczne, występują kwestionariusze i ankiety w postaci pytań odnoszących się do tego, jak bardzo osoba badana zgadza się z twierdzeniem, czy jest jakaś lub ma jakąś cechę. Prosi się ją o wybór jednej z przeważnie pięciu, stopniowanych kategorii, np. od „zupełnie się nie zgadzam" do „zupełnie się zgadzam", od „w pełni mnie nie opisuje" do „w pełni mnie opisuje", czy też od „nigdy", przez „czasami", „rzadko", do „zawsze". Nie tylko w badaniach naukowych, ale także w sytuacjach decydujących o czyimś życiu opieramy się na wynikach takich badań. Dlatego tak ważne jest, by narzędzia używane do pomiaru wszelkich cech i właściwości psychicznych, które służą jako argumenty do podejmowania istotnych decyzji, w dyskusji, albo też wpływają na dystrybucję środków, były jak najwyższej jakości. Jako badacze wypracowaliśmy pewne procedury mające tę jakość potwierdzać. W bieżącym rozdziale chciałbym się przyjrzeć tym procedurom.

---

[1] „There is no evidence either that tests measure anything or that the attributes that testers aspire to measure are measurable, so presenting tests as if they were instruments of measurement is presenting a myth".

Proces konstruowania narzędzia badawczego do pomiaru cech niedostępnych bezpośredniemu doświadczeniu można podzielić na trzy etapy: 1) określanie treści interesującego badacza konstruktu (etap wewnętrzny i teoretyczny), 2) budowanie struktury konstruktu (etap wewnętrzny i praktyczny oraz zewnętrzny i teoretyczny) oraz 3) ocena przydatności pomiaru danego konstruktu (etap zewnętrzny i praktyczny). Celem pierwszego etapu jest identyfikacja obszaru treściowego badanego konstruktu. Kierunek badań na tym etapie może przebiegać mniej lub bardziej autorsko – mogą to być studia lub też wiedza i doświadczenie autorów. Etap ten kończy się na wyborze z uniwersum treści właściwych dla ukrytego, ale już semantycznie zdefiniowanego pojęcia, konkretnych bodźców, zachowań itp., które odpowiadają celowi pomiaru. By zobiektywizować wybór, badacze często posiłkują się metodą sędziów kompetentnych (Hornowska, 2001, ss. 158–169).

Drugi etap skupia się na analizie statystycznej otrzymanych wyników w próbie osób badanych. Od najprostszej analizy rozkładu odpowiedzi, przez szacowanie trudności pozycji, korelacje z wynikiem ogólnym, do analizy czynnikowej i analizy na podstawie teorii odpowiadania na pozycje testowe. Dokonywana jest w ten sposób ocena poszczególnych pozycji i ich miejsca w strukturze pojęcia wyprowadzonego z jego założeń teoretycznych. Etap ten obejmuje też oszacowanie precyzji pomiaru przez dane narzędzie, np. na podstawie współczynników rzetelności, poziomu interkorelacji i korelacji test-retest. Informacje o strukturze narzędzia badawczego (nie konstruktu!) dopełniać powinna analiza odporności wyników na obciążenie różnymi czynnikami ubocznymi, którą przeprowadza się z wykorzystanim oceny niezmienniczości wariancji w analizach różnic grupowych (*measuring of invariance*) lub w IRT za pomocą DIF (*differential item functioning*) (Hornowska, 2001, ss. 169–204).

Konstrukcję narzędzia badawczego zamyka faza trzecia – zewnętrzna. Jest to oszacowanie możliwości diagnostycznych i prognostycznych narzędzia. Te pierwsze można oszacować na przykład na podstawie wspomnianej już analizy wielu cech wielu metod oraz na podstawie różnic międzygrupowych. Odpowiedni układ korelacji uzasadnionych teoretycznie związków z innymi właściwościami psychologicznymi będzie potwierdzeniem trafności w aspekcie diagnostycznym. Natomiast trafność prognostyczną można zbadać na przykład przez analizę różnic międzygrupowych lub na podstawie krzywych ROC (*receiver operating characteristic*) (Hornowska, 2001, ss. 80–127).

Opisany powyżej zbiór działań walidacyjnych narzędzia badawczego w psychologii świadczy o jego dojrzałości psychometrycznej. Nie jest to oczywiście pełna lista działań, które mogą być podjęte, by zweryfikować narzędzie i uznać je za gotowe do zastosowania w praktyce. Opisane elementy są jednak niezbędnym zbiorem, który często spełniają narzędzia oferowane w oficjalnym obiegu (testy sprzedawane do użytku badawczego i diagnostycznego), a bardzo rzadko narzędzia opisywane w publikacjach naukowych. Jak wykazałem w rozdziale 3,

lista analiz wyników dostarczanych przez narzędzia jest bardzo krótka – średnio 1,37 (*SD* = 0,61) analizy na każde z nich. Najczęściej (20 razy) występowało zestawienie w parze analiz EFA i CFA, a na drugim miejscu CFA i test-retest (5 razy). Wniosek z tego jest oczywisty: analizowane przeze mnie doniesienia o walidacji narzędzi badawczych są niekompletne, a mimo to autorzy w większości używają sformułowania: „W świetle przedstawionych wyników można uznać narzędzie X za trafne i rzetelne oraz nadające się do badania Y".

4.2. Badania własne

Biorąc pod uwagę istniejący i opisany do tej pory stan rzeczy oraz silne przekonania badaczy co do tego, że stosowanie procedur statystycznych jest gwarantem odfiltrowania narzędzi dobrych od pozostałych, postanowiłem sprawdzić tę tezę empirycznie. Statystyczne sposoby zbierania dowodów na trafność i rzetelność narzędzi badawczych używanych w psychologii rozwijały się wraz z rozwojem matematycznych metod analitycznych i odpowiadały na zapotrzebowanie badaczy. Prawdopodobnie działa tu swoisty dobór naturalny: stosowane są oraz rozwijane metody silniejsze, lepsze, wypierające zbyt proste i niedostatecznie wrażliwe. Jeśli tak jest, to walidując wyniki uzyskane za pomocą narzędzia wątpliwej jakości, powinniśmy dostać wyniki wskazujące na jego niską wartość. Z analizy przedstawionej w rozdziale 3 wynika, że w obszarze statystycznego testowania trafności i rzetelności testów dominuje analiza korelacji, analiza czynnikowa (eksploracyjna i konfirmacyjna). Celem mojego eksperymentu było sprawdzenie, jak dokładnym filtrem są te metody statystyczne, jeśli chodzi o odróżnienie narzędzi badawczych różnej jakości.

4.2.1. Metoda

Eksperyment polegał na przebadaniu szerokiej próby osób tradycyjnym – w rozumieniu budowy – kwestionariuszem, a następnie poddaniu walidacji uzyskanych wyników tak, aby stwierdzić poziom rzetelności i trafności wykorzystanego narzędzia. Stawiam tezę, że metody walidacji oparte na analizie statystycznej (analiza rzetelności na podstawie alfy Cronbacha, trafność pod względem teoretycznym oparta na analizie eksploracyjnej i konfirmacyjnej czynnikowej oraz trafność pod względem kryterialnym oparta na współczynnikach korelacji) nie pozwolą odrzucić złego narzędzia, lub nawet bezsensownego. W tym celu poprosiłem osoby badane o wypełnienie kwestionariusza GSES służącego do oceny przekonań na temat samoskuteczności (Juczyński, 2001) oraz kwestionariusza IPIP-BFM-20 służącego do określenia cech osobowości na wymiarach Wielkiej Piątki (Topolewska, Skimina, Strus, Cieciuch i Rowiński, 2014). Badanie

odbywało się za pośrednictwem specjalnie przygotowanej strony internetowej pod adresem kleka-badanet.home.amu.edu.pl. Badani zaznaczali odpowiedzi na pytania o wiek, płeć, wykształcenie oraz poszczególne pozycje kwestionariuszy w trzech grupach (stronach). Na początku każdorazowo stawiane były pytania metrykalne, a następnie w sposób zrównoważony prezentowane były albo pytania IPIP-BFM-20, albo GSES. Wymagane było udzielenie odpowiedzi na wszystkie pytania. Osoby uczestniczące w badaniu w każdym momencie mogły przerwać badanie, mogły też do niego wrócić w dowolnym czasie (link w zaproszeniu do badań zawierał indywidualny, anonimowy token), chyba że udzieliły wszystkich odpowiedzi – wtedy dostęp do kwestionariuszy był blokowany i wracającym osobom prezentowana była strona z podziękowaniem za udział w badaniu.

Oprócz kolejności, w jakiej prezentowane były kwestionariusze, manipulacji eksperymentalnej podlegały dwie kwestie. Po pierwsze, kwestionariusz GSES był prezentowany w czterech wersjach, z których tylko jedna pokrywała się z oryginałem w adaptacji Ralfa Schwarzera, Michaela Jerusalema i Zygfryda Juczyńskiego. Pozostałe trzy zawierały taką ingerencję w treść, która pozbawiała sensu prezentowane pytania. Najprostsza modyfikacja polegała na zastąpieniu kluczowych sformułowań słowem „poczuciowość". I tak np. stwierdzenie: „Zazwyczaj jestem w stanie poradzić sobie z tym, co mnie spotyka" zmieniłem na: „Zazwyczaj jestem w stanie poradzić sobie ze swoją poczuciowością", a np. pozycję „Z łatwością potrafię się trzymać swoich celów i je osiągać" zamieniłem na „Z łatwością potrafię się trzymać swojej poczuciowości". Na kolejnym poziomie ingerencji w treść narzędzia badawczego pozycje kwestionariusza zastąpione zostały bezsensownymi zdaniami wygenerowanymi w sposób losowy, np. „Zmartwione doświadczenie projektuje lot" lub „Dlatego chleb podsumowuje opuszczone szkło"[2]. Ostatni etap usuwania sensu z pozycji kwestionariusza GSES polegał na zastąpieniu ich treści przez przypadkowe nowe słowa pozbawione sensu, np. „Pochan enifru midal sapiks trge huze" lub „Ocypy bek aksoza wiku mufy"[3]. Po drugie, zmieniana była liczba możliwych do wyboru odpowiedzi. W oryginalnej wersji badacze przewidzieli cztery stopnie zgody z twierdzeniami: od „zdecydowanie nie", przez „raczej nie" i „raczej tak", do „zdecydowanie tak". W przeprowadzonym badaniu dokonano

[2] Treści poszczególnych pytań brzmiały następująco: *„Matka przyśpiesza plonowanie ziemi"*, *„Prawdziwy ołów uruchamia olej"*, *„Miejsce krytykuje opuszczony protest"*, *„Zmartwione doświadczenie projektuje lot"*, *„Nadmiarowy projekt mówi piosenkę"*, *„Słony komfort słucha wymiany"*, *„Dlatego magenta wygląda nauczanie"*, *„W taki sposób wadliwy ząb biznesu"*, *„Dlatego chleb podsumowuje opuszczone szkło"*, *„Wtedy zaczyna się prosty płacz z powrotem"*.

[3] Było to następujące pseudowyrażenia: *„Ocypy bek aksoza wiku mufy"*, *„Omily mic ogita obogdy dagib hia nem sano"*, *„Ifugma udubna ogoko wof kizyd zadoj joda"*, *„Amefe siu boksid fyko roe ufysy"*, *„Furi ukada myte arallu zyp kila gykse rekis tok"*, *„Pochan enifru midal sapiks trge huze"*, *„Olekse licediwo cimu gil lowsamp asyja cuco"*, *„Kajtu det wyca sasy techach mim amukru owegi"*, *„Dab elakra tagel fybec zej amoksu fan wysub"*, *„Icepa lagaw hed beran zez ufado gok ewawu lei"*.

manipulacji liczbą kategorii, zarówno ją zmniejszając (redukując do dwóch: „nie" oraz „tak"), jak i zwiększając (uzupełniając kafeterię oryginalną o cztery dodatkowe stopnie „nie", „częściowo nie", „częściowo tak" i „tak"). Ponieważ kompletny plan badawczy zawierałby aż 24 warunki (2 kolejności · 4 wersje · 3 długości kafeterii odpowiedzi), ze względów praktycznych ograniczyłem się do 11 warunków, przedstawionych w tab. 4.1.

W ramach pierwszej tury badania, trwającej od 17 do 22 grudnia 2018 roku, wysłano 5635 zaproszeń, z czego 1167 (20,7%) adresowano na błędne/nieaktualne adresy lub trafiły one do adresatów z przepełnionymi skrzynkami i zostały zwrócone przez systemy pocztowe. Na stronę internetową poświęconą badaniu weszło 777 osób, z czego 624 badanie rozpoczęło, a 469 udzieliło odpowiedzi na wszystkie pytania (tab. 4.1).

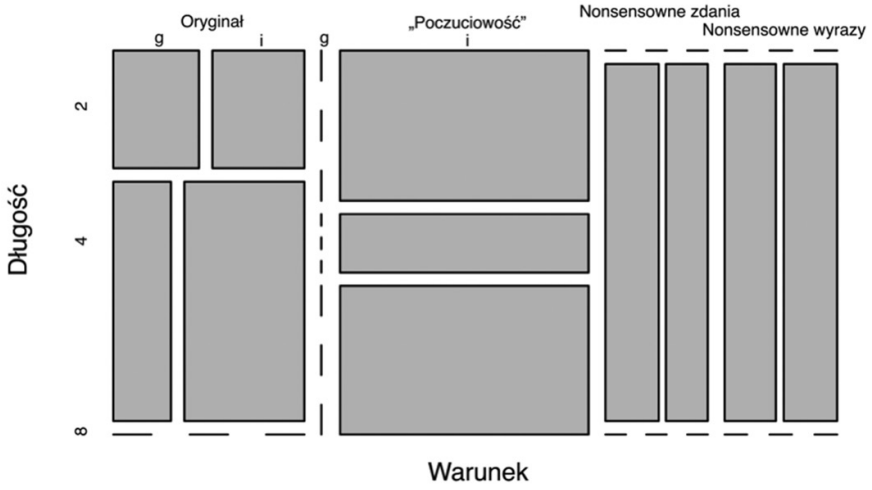
Tabela 4.1. Szczegółowe zestawienie warunków badawczych

Warunek	Kolejność	Wersja	Długość	Np	Nz
o4g	1	oryginalna	4	30	37
o4i	2	oryginalna	4	30	77
o2g	1	oryginalna	2	20	27
o2i	2	oryginalna	2	20	29
p2i	2	z „poczuciowością"	2	70	100
p4i	2	z „poczuciowością"	4	70	39
p8i	2	z „poczuciowością"	8	70	99
n4i	2	bezsensowne zdania	4	30	40
n4g	1	bezsensowne zdania	4	30	51
e4i	2	bezsensowne wyrazy	4	30	51
e4g	1	bezsensowne wyrazy	4	30	49

Oznaczenia: warunek – nazwa warunku; kolejność – który z kolei (pierwszy czy drugi) był badany kwestionariusz; wersja – w jakiej wersji był badany kwestionariusz; długość – liczba poziomów odpowiedzi do wyboru; Np – planowana wielkość próby dla danego warunku; Nz – zrealizowana wielkość próby w warunku.

4.2.2. Analiza przebiegu badania

Badanie przeprowadziłem w okresie od 17 grudnia 2018 do 21 stycznia 2019 roku. Ze względu na mechanizm przydzielania osób badanych do warunków, który losował je proporcjonalnie do liczebności zakładanych grup, niektóre z nich mają mniejszą liczebność (nie zawsze osoba wylosowana przystąpiła do badania

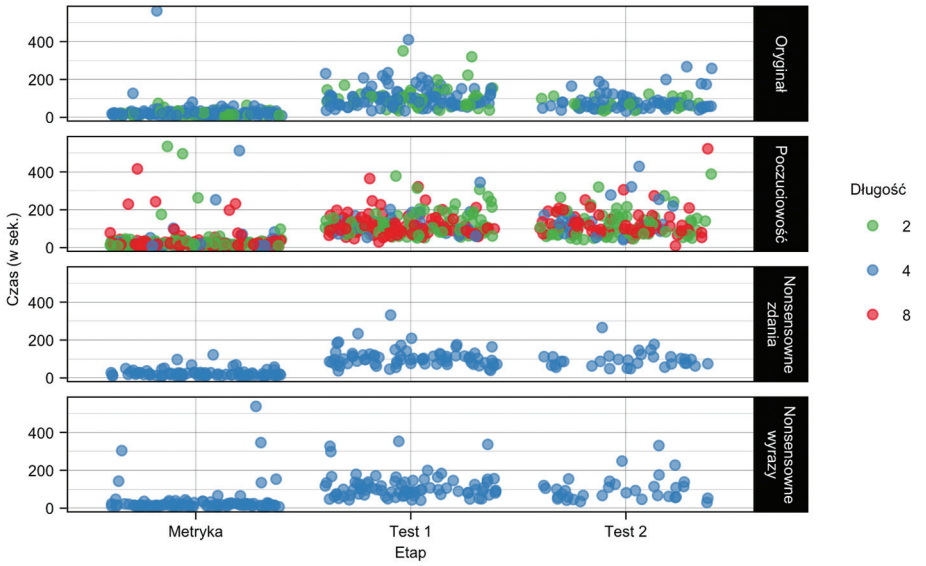


Oznaczenia: g – główny test prezentowany jako pierwszy; i – test osobowości prezentowany jako pierwszy; 2, 4, 8 – liczba kategorii odpowiedzi w teście głównym. Kreski pionowe i poziome oznaczają brak osób w danym warunku.

Rysunek 4.1. Proporcje liczebności grup uzyskanych w badaniu

i je ukończyła). Liczebności te są jednak wystarczające, by przeprowadzić analizę statystyczną wyników. Uzyskane proporcje liczebności grup przedstawia rys. 4.1.

Badanie przeprowadziłem za pomocą dedykowanych stron internetowych. Pierwsza strona z pytaniami, tzw. metryka, zawierała 4 pytania: o wiek, płeć, wykształcenie i zawód (pytanie z opcjonalną, dowolną odpowiedzią). Osoby uczestniczące w badaniach poświęcały na udzielenie odpowiedzi na te pytania średnio 20 sekund (od 16 do 67 – z pominięciem wyników skrajnych, większych niż 3 minuty, które przyjąłem jako wskazania zaburzeń w odpowiadaniu, np. w wyniku pozostawienia otwartej strony w komputerze i zmiany aktywności respondenta). Następnie był prezentowany test osobowości IPIP-BFM-20 składający się z 20 pytań lub zmodyfikowany test samoskuteczności GSES składający się z 10 pytań. Kolejność prezentacji miała wpływ na długość wypełniania IPIP-BFM-20 – była ona dłuższa średnio o 19,5 sekundy (F (1, 1373) = 3,63, p < 0,001), jeśli test ten był pierwszy. Można to interpretować albo jako oznakę zaangażowania – gdy osoba badana widziała „normalny" test po przebrnięciu przez mniej lub bardziej zdeformowane pytania testu GSES, poświęcała mniej uwagi na kolejne pytania, albo jako efekt treningu – test prezentowany jako drugi był



Na rysunku usunięte są czasy odpowiedzi dłuższe niż 10 minut. Kolorem zaznaczono liczbę możliwych do wyboru odpowiedzi w wersji prezentowanej osobie badanej.

Rysunek 4.2. Rozrzut czasów udzielania odpowiedzi w badaniu w zależności od warunków

wypełniany szybciej ze względu na zaznajomienie się z budową strony i opanowanie zasad jej używania (np. nieszukanie przycisku zatwierdzającego odpowiedzi).

Ciekawie przedstawiają się wyniki średnich czasów wypełniania testu będącego przedmiotem badań („Test2" w tab. 4.2) – w wyraźny sposób zależą one od wylosowanej wersji. Najmniej czasu zajmowało osobom uczestniczącym w badaniu wypełnienie sensownej wersji z dwiema możliwymi odpowiedziami („tak" lub „nie"), a czas ten wzrastał nieznacznie, gdy kafeteria odpowiedzi powiększona była do czterech możliwości. Czas wypełniania testu w wersjach bezsensownych był dłuższy, z zachowaniem wydłużającego wpływu większej liczby możliwych odpowiedzi. Natomiast w przypadku wypełniania testu pytającego o wyimaginowaną „poczuciowość" średni czas rósł o kolejne około 20–60 sek. w zależności od liczby kategorii odpowiedzi. Wydłużenie czasu można interpretować jako efekt głębszego przetwarzania treści i poszukiwania sensu w dziwnym słowie, gdy w pozostałych przypadkach identyfikacja „braku treści" pozwalała odpowiadać stosunkowo szybko. Pozostaje tajemnicą, dlaczego prawie tak samo szybko respondenci odpowiadali na pytania bez treści jak w przypadku wersji z treścią – czy przetwarzanie treści w pytaniu zajmuje tyle samo czasu

co samo jego przeczytanie, a dopiero zbicie z pantałyku wymusza dodatkowy czas na przetwarzania informacji? Odpowiedź na to pytanie wymaga zaprojektowania i przeprowadzenia nowych badań, w których osoby uczestniczące będzie można na przykład monitorować przez cały czas badania i precyzyjnie mierzyć proces czytania i odpowiadania (co wykracza poza projekt tej pracy).

Tabela 4.2. Średnie oraz odchylenia standardowe czasów rozwiązywania poszczególnych stron w badaniu

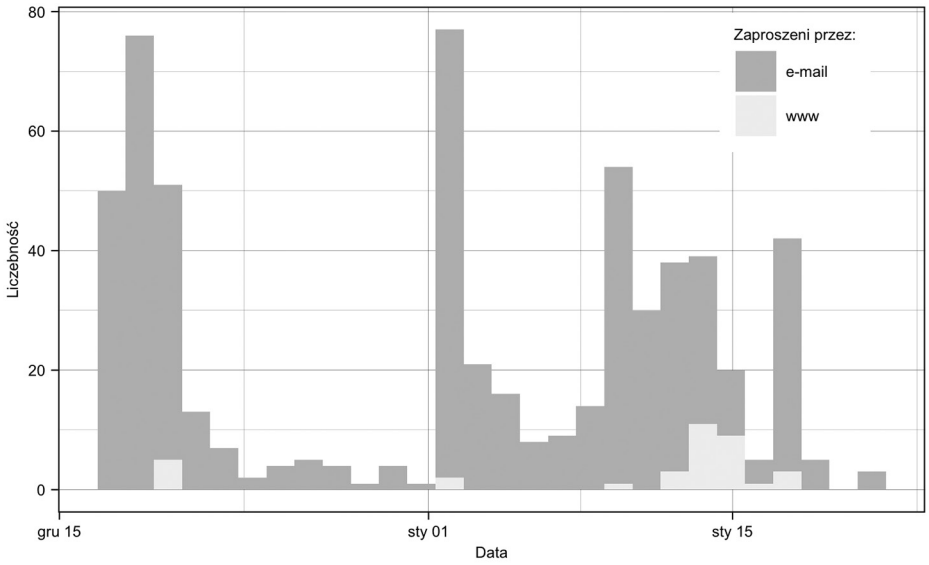
Warunek	Dł.	*N*	Metryka.m	Metryka.sd	T1.m	T1.sd	T2.m	T2.sd
Oryginał	2	26	576,3 sek.	2826,7	134,5 sek.	70,9	74,2	25,9
Oryginał	4	64	23,4 sek.	17,8	138,6 sek.	289,2	83,6	50,4
„Poczuciowość"	2	65	37,1 sek.	125,3	122,0 sek.	122,2	117,5	71,1
„Poczuciowość"	4	22	63,4 sek.	113,3	130,0 sek.	63,6	154,5	95,4
„Poczuciowość"	8	68	25,1 sek.	385,0	112,3 sek.	91,9	121,1	72,7
Nonsensowne zdania	4	32	21,6 sek.	15,9	132,4 sek.	136,2	96,0	43,4
Nonsensowne wyrazy	4	34	16,2 sek.	8,7	117,7 sek.	73,9	101,5	65,3

Oznaczenia: *.m – średni czas; *.sd – odchylenie standardowe. Czas spędzony na rozwiązywaniu testu IPIP-BFM-20 (T1) był o ok. 33 sek. dłuższy, jeśli ten test pojawił się jako pierwszy ($F(1, 419) = 28{,}2$, $p < 0{,}001$).

Analizując częstość odpowiedzi w poszczególnych dniach, zauważyłem, że dni z największą liczbą osób biorących udział w badaniu odpowiadają dniom, w których wysyłane były zaproszenia (rys. 4.3).

Podobnie jak w moich innych badaniach (Kleka, 2017), okazało się, że okres około tygodnia jest sensytywnym okresem, gdy zaproszenie do badania działa motywująco, a najwięcej osób podejmuje decyzję o wzięciu udziału w badaniu w momencie zapoznania się z zaproszeniem. Osoby odkładające udział „na potem" stanowią małą i malejącą z każdym dniem grupę. Jest to wskazówka mogąca sugerować potencjalne obciążenie grup badawczych w badaniach przeprowadzanych za pośrednictwem Internetu pod względem cech osobowości związanych z reaktywnością, samooceną czy innymi cechami związanymi z podejmowaniem decyzji i utrzymaniem motywacji.

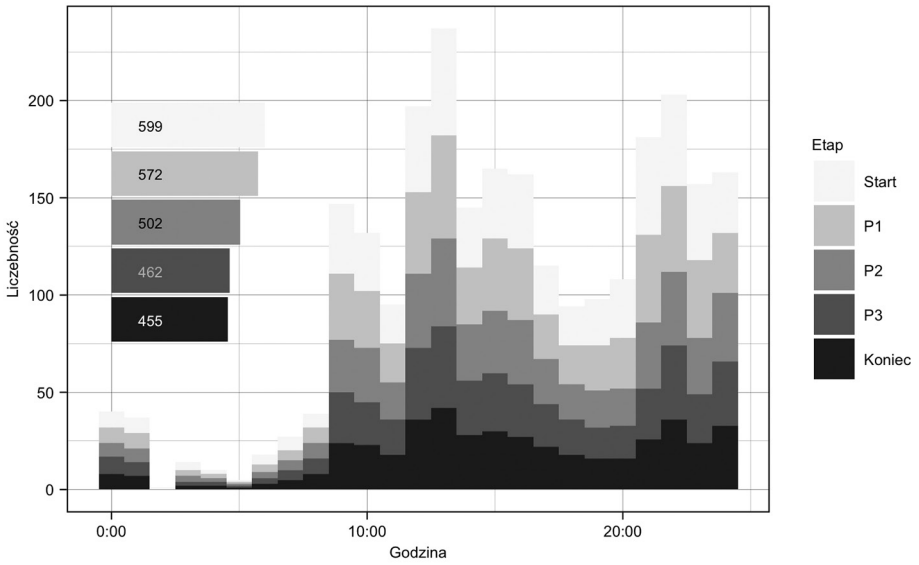
Osoby badane, otrzymując zaproszenie do badań, mogły na nie odpowiedzieć w dowolnym czasie. Analizując godzinę przystępowania do badania, zauważyłem trzy szczyty frekwencji: w godzinach 9.00, 13.00 i 21.00. Związane są one



Rysunek 4.3. Cykl dzienny.
Liczba osób, które przystąpiły do badania danego dnia

z momentem wysyłania zaproszenia do badań (rys. 4.4). Uwzględniając to, można zauważyć wzór dobowej aktywności – spadek aktywności w południe oraz około godziny 20.00, co nie jest żadnym odkryciem, ale może być wskazówką przy wyborze godziny wysyłania zaproszeń.

Liczebność grup osób kończących poszczególne etapy jest coraz mniejsza, co jest cechą badań prowadzonych za pośrednictwem Internetu. Koszty rezygnacji są niskie, w związku z czym zdarza się ona częściej niż w przypadku badań papierowych, choć w porównaniu z nimi wycofanie się z badania nie oznacza utraty wszystkich danych. Mimo że do końca badania dotarła mniejsza grupa, to dzięki rejestracji odpowiedzi możliwe jest przeanalizowanie odpowiedzi na te pytania, które są umieszczone na początku badania. Może się to stać wskazówką do korekty, gdy okaże się, że grupa porzucająca badanie charakteryzuje się specyficznymi cechami (np. gdy będą to osoby starsze, może się okazać, że ankieta nie jest dostosowana do ich potrzeb, albo gdy będą to kobiety, może to wskazywać, że ankieta jest obciążona treściami demotywującymi właśnie tę płeć).



Oznaczenia: Start – moment wejścia na stronę internetową z badaniem; P1, P2 i P3 – czasy przejścia przez kolejne strony badania: metryczkę, pierwszy i drugi kwestionariusz; Koniec – moment zakończenia badania. Duże prostokąty z lewej strony z naniesionymi liczbami przedstawiają proporcjonalne wielkości grup osób, które ukończyły dany etap badania.

Rysunek 4.4. Cykl dobowy.
Liczba osób, które przystąpiły do badania o określonej godzinie

4.2.3. Osoby uczestniczące w badaniu

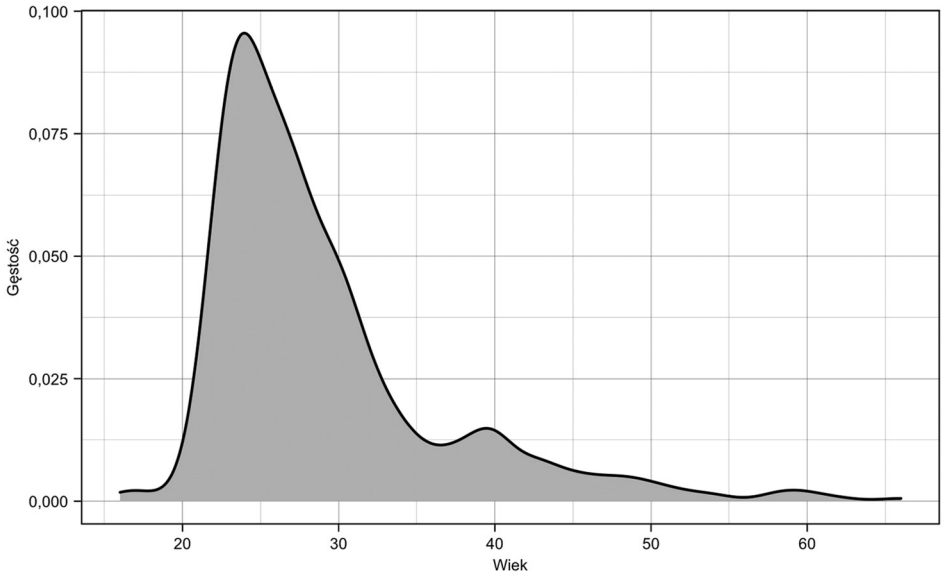
Większość osób, które wzięły udział w badaniu, była w wieku od 24 do 31 lat (odpowiednio pierwszy i trzeci kwartyl). Najstarsza osoba miała 66, a najmłodsza 20 lat. Średni wiek osób badanych, które wpisały wartość w metryczce (N = 541, 90,3% ogółu badanych), wynosił 29,1 (SD = 7,53, Md = 27). Rozkład gęstości wieku przedstawia rys. 4.5. Nie stwierdziłem różnicy w wieku między osobami, które porzuciły badanie, a tymi, które je ukończyły ($t(75, 42) = -1{,}43$, $p = 0{,}157$).

Dominującą płcią była płeć żeńska (N = 440, 73,5% ogółu badanych), co było niezamierzonym i niekontrolowanym efektem. Mężczyźni stanowili mniejszą grupę (N = 122, 20,4% ogółu badanych) i byli przeciętnie starsi o około 2 lata ($\Delta M = -1{,}98$, 95% CI [−3,74, −0,22], $t(162, 27) = -2{,}22$, $p = 0{,}028$). Płeć nie wpływała na decyzję o nieukończeniu badania ($\chi^2(1, n = 599) < 0{,}001$, $p = 0, 989$).

Dominującym wykształceniem respondentów było wykształcenie wyższe (*N* = 351), na drugim miejscu były osoby z wykształceniem wyższym niepełnym (*N* = 115) oraz z wykształceniem średnim (*N* = 93). Szczegółowy rozkład częstości poszczególnych kategorii wykształcenia z uwzględnieniem płci przedstawiają tab. 4.3 i rys. 4.5.

Tabela 4.3. Rozkład częstości wykształcenia osób badanych z uwzględnieniem płci

Wykształcenie	Kobieta *N* (%)	Mężczyzna *N* (%)	Brak odpowiedzi *N* (%)	Razem *N* (%)
Podstawowe	2 (0,5)	7 (5,7)	0	9 (1,5)
Zawodowe	2 (0,5)	0	0	2 (0,3)
Średnie	73 (16,5)	18 (14,8)	2 (5,4)	93 (15,5)
Wyższe niepełne	88 (20,0)	24 (19,7)	3 (8,1)	115 (19,2)
Wyższe	273 (62,0)	73 (59,8)	5 (13,5)	351 (58,6)
Brak odpowiedzi	2 (0,5)	0	27 (73,0)	29 (4,8)
Razem	440 (73,4)	122 (20,4)	37 (6,2)	599 (100)



Rysunek 4.5. Rozkład gęstości wieku osób uczestniczących w badaniu

4.2.4. Wyniki

Analiza uzyskanych wyników obejmuje trzy obszary – analizę korelacji, analizę rzetelności oraz analizę czynnikową. Chcąc udowodnić, że statystyczne metody walidacji są działającym filtrem pozwalającym odróżnić narzędzia mierzące rzetelnie zakładane konstrukty od narzędzi złych i mierzących przypadkowy zlepek treści, zbadałem najpierw współczynniki korelacji, których dotyczyła pierwsza testowana hipoteza. Jeśli treść ma znaczenie, jeśli osoby badane odpowiadają refleksyjnie na zadawane im pytania, to związki wyników z narzędzi mierzących samoskuteczność oraz wymiary osobowości powinny istnieć tylko dla ich sensownych wersji. Macierz korelacji przedstawia tab. 4.4.

Tabela 4.4. Wartości współczynników korelacji Pearsona i Spearmana między wymiarami osobowości a różnymi wersjami badanego narzędzia

Wymiar osobowości	Oryginał	„Poczuciowość"	Nonsensowne zdania	Nonsensowne wyrazy
r Pearsona				
ekstrawersja	0,3147	0,2476	0,0257	0,1962
ugodowość	0,0942	0,0822	0,1722	0,0536
sumienność	–0,2114	–0,3209	0,1656	0,1594
emocjonalność	–0,3788	–0,3335	–0,0785	0,1459
intelekt	0,3987	0,2474	0,2037	0,3145
rho Spearmana				
ekstrawersja	0,2924	0,2856	0,0305	0,2034
ugodowość	0,1259	0,0985	0,1437	0,0546
sumienność	–0,1889	–0,3234	0,2088	0,1893
emocjonalność	–0,3877	–0,3287	–0,0721	0,0817
intelekt	0,3563	0,2705	0,2042	0,2873

Nie będę interpretował znaczenia tych wyników – jeśli istnieje związek między osobowością a samoskutecznością, to jest on analizowany przez wielu innych badaczy. Mnie interesują różnice między korelacjami dla różnych wersji kwestionariusza badającego samoskuteczność – „poczuciowość". Punktem wyjścia do interpretacji wyników na potrzeby pierwszej hipotezy jest pierwsza kolumna wartości współczynników korelacji. Układ korelacji z samoskutecznością: umiarkowanie pozytywne z wymiarami ekstrawersji i intelektu, umiarkowanie negatywne z wymiarem emocjonalnym i słabe z sumiennością oraz

bliskie zera z ugodowością – powinien być wyjątkowy i istnieć tylko dla narzędzia w wersji oryginalnej, mierzącego „prawdziwą" samoskuteczność. Jednakże taki sam układ korelacji możemy obserwować dla wersji narzędzia mierzącego „poczuciowość". Można zatem dostrzec, że zmiana głównego pojęcia w poszczególnych pozycjach kwestionariusza nie wpłynęła na wybór odpowiedzi przez osoby badane!

Dla wersji zawierających nonsensowne zdania lub nonsensowne wyrazy układ korelacji jest na szczęście inny, ale niepokojące jest mimo wszystko istnienie umiarkowanych i małych korelacji na przykład ze skalą intelektu. Te wyniki nic nie znaczą merytorycznie, ale podobne wartości (około 40% wyników analizowanych w rozdziale 3, por. rys. 3.8) są wskazywane jako dowody potwierdzające trafność. Na przykład w podręczniku do GSES, w par. 2.2. Trafność (s. 90), korelacje rzędu 0,25–0,35 z dyspozycyjnym optymizmem Testu Orientacji Życiowej, poczuciem własnej wartości Skali Poczucia Własnej Wartości Rosenberga, wewnętrznym umiejscowieniem kontroli zdrowia z Wielowymiarowej Skali Umiejscowienia Kontroli Zdrowia, zachowań zdrowotnych z Inwentarza Zachowań Zdrowotnych Juczyńskiego są uznane za dowody na trafność skali.

W prezentowanych badaniach wymiary osobowości są czymś innym niż nonsens, ale nie jest to argument wskazujący, jak trafnie jest to coś mierzone. Z tak niskich wartości wskaźników korelacji nie wynika trafność ani jednego, ani drugiego narzędzia.

Wyniki analizy pokazują, jak łatwo jest uzyskać niskie korelacje, i nie powinniśmy interpretować wartości praktycznie bez znaczenia merytorycznego. Uzyskanie około 10% wspólnej wariancji (tj. korelacji wynoszącej 0,3163) nie jest żadnym osiągnięciem i wynika bardziej ze zbieżności metod (tutaj obie metody są kwestionariuszowe) niż ze związku między ich treściami.

Drugi analizowany obszar dotyczył rzetelności. Skoro trafność w aspekcie zewnętrznym nie powinna się opierać na prostych współczynnikach korelacji, to jak wygląda trafność wewnętrzna, tj. spójność pozycji narzędzia jako próby z nomologicznej sieci powiązanych treści? By odpowiedzieć na to pytanie, skorzystałem z najpopularniejszych wskaźników, tj. alfy Cronbacha i lambdy 6 Guttmana, oszacowanych dla poszczególnych badanych wersji narzędzia (tab. 4.5).

Uzyskana wartość rzetelności przewyższa nawet tę z badań adaptacyjnych autorów polskiej wersji narzędzia (0,96 w stosunku do 0,85 dla wersji adaptowanej w Polsce). Co więcej, dla żadnej z wersji, także tych nonsensownych, rzetelność nie spada poniżej powszechnie przyjętego kryterium, jakim jest wartość 0,70. Oznacza to, że uzyskując wyniki takie jak w tab. 4.5, wszystkie wersje uznane byłyby za spełniające warunek spójności wewnętrznej.

Zarówno lambda 6 Guttmana, jak i alfa Cronbacha przyjmują wyższe wartości dla skal z większą liczbą kategorii odpowiedzi, obydwa wskaźniki wskazują też na spójność, mimo że średnia korelacji pozycji (dla wersji nonsensownych) nie przekracza 0,25.

Tabela 4.5. Wartości współczynników spójności wewnętrznej alfa Cronbacha i lambda 6 Guttmana oraz średnia z wartości korelacji dla poszczególnych wersji badanego narzędzia

Wersja	Alfa	Lambda	Średnie *r*
Oryginał	0,9635	0,9653	0,7253
„Poczuciowość" (2)	0,8868	0,9061	0,4393
„Poczuciowość" (4)	0,8579	0,9112	0,3764
„Poczuciowość" (8)	0,9413	0,9522	0,6159
Nonsensowne zdania	0,7613	0,7765	0,2418
Nonsensowne wyrazy	0,7342	0,7580	0,2164

W wierszach są zamieszczone parametry dla poszczególnych wersji badanego narzędzia. Wartości 2, 4, 8 przy wersji „poczuciowości" oznaczają liczbę kategorii odpowiedzi.

Analizując rzetelność jako spójność wewnętrzną pozycji, badacze zazwyczaj podają wartość alfy Cronbacha po usunięciu kolejnych pozycji. Pozwala to wykluczyć pozycje słabiej skorelowane z pozostałymi (teoretycznie słabiej korelujące z cechą ukrytą mierzoną przez narzędzie). Zasadność takich działań jest dyskusyjna (por. Kleka i Soroko, 2018) ze względu na faworyzowanie treści centralnych dla konstruktu kosztem peryferyjnych, niemniej są one praktykowane. Przyjrzyjmy się, jakie wartości towarzyszą badanym wersjom narzędzia (tab. 4.6).

Tabela 4.6. Wartości współczynnika alfa Cronbacha po usunięciu kolejnych pozycji

Pozycja	O	P(2)	P(4)	P(8)	NZ	NW
p1	0,959	0,883	0,849	0,936	0,750	0,707
p2	0,962	0,877	0,847	0,935	0,761	0,714
p3	0,962	0,872	0,857	0,937	0,739	0,728
p4	0,959	0,872	0,837	0,935	0,729	0,723
p5	0,958	0,871	0,837	0,935	0,735	0,712
p6	0,958	0,869	0,831	0,934	0,753	0,707
p7	0,962	0,878	0,849	0,937	0,750	0,688
p8	0,959	0,876	0,847	0,933	0,731	0,719
p9	0,958	0,873	0,841	0,938	0,737	0,724
p10	0,959	0,886	0,849	0,932	0,728	0,706

O – wersja oryginalna, P(2, 4, 8) – wersja z „poczuciowością" licząca 2, 4 lub 8 kategorii odpowiedzi, NZ – wersja z nonsensownymi zdaniami, NW – wersja z nonsensownymi wyrazami.

Podobnie jak w przypadku interpretacji wyników dla całej skali, także tu, analizując wartości współczynnika alfa Cronbacha w dowolnym postępowaniu walidacyjnym, każda wersja badanego narzędzia uznana zostałaby za spójną. „Najsilniejsza" pozycja w wersji nonsensownych wyrazów to stwierdzenie „*Olekse licediwo cimu gil lowsamp asyja cuco*", którego usunięcie obniża rzetelność narzędzia do poziomu alfa = 0,688. Podanie wyjaśnienia tego faktu byłoby w oczywisty sposób czystą spekulacją, choć w większości publikowanych walidacji narzędzi badacze są zdolni *post factum* podać sensownie brzmiące argumenty za pozostawieniem (lub usunięciem) pozycji z podobnymi parametrami.

Przytoczone wyniki wskazują, że analiza spójności wewnętrznej nie jest skuteczną miarą jakości testu i interpretacja uzyskanych w niej wyników w żaden sposób nie chroni przed uznaniem losowych pozycji testu za narzędzie o wysokich właściwościach psychometrycznych.

Według przytoczonych w poprzednim rozdziale badań zdecydowanie najpopularniejszą metodą walidacji narzędzi jest analiza czynnikowa. Dlatego wyniki uzyskane w eksperymencie także poddałem takiej analizie z użyciem algorytmu MinRes. Badany test jest narzędziem mierzącym jednowymiarowy konstrukt. Uzyskane na badanej próbie wyniki potwierdzają to założenie dla wszystkich czterech wersji. W przypadku oryginalnej wersji wyjaśniona wariancja wyniosła 73%, natomiast wersja z „poczuciowością" miała aż 87% wyjaśnionej wariancji. Wersje nonsensowne, wbrew oczekiwaniom, również pozwoliły umiejscowić w hipotetycznym konstrukcie odpowiednio 25% i 23% wspólnej wariancji. I tak jak tylko trochę zaskakuje wynik dla dwóch ostatnich wersji, ponieważ oznacza on, że nawet brak sensu może generować wspólny czynnik, choć na szczęście o niskich wartościach, tak pozór sensu wypada lepiej niż skala rzeczywiście mierząca ugruntowane teoretycznie pojęcie samoskuteczności. Ładunki czynnikowe poszczególnych pozycji przedstawia tab. 4.7.

Analizując wartości ładunków czynnikowych dla oryginalnej wersji narzędzia, można zaobserwować wysokie i bardzo wysokie wyniki. Dla wersji mierzącej „poczuciowość" ładunki są już tylko bardzo wysokie. Wyniki dla wersji z nonsensownymi zdaniami oraz nonsensownymi wyrazami sugerują problem z trzema (innymi) pozycjami (lambda < 0,40). Znając treść (a raczej jej brak) pozycji, wiemy, że wariancja wyjaśniona przez rozwiązanie czynnikowe nie wyjaśnia niczego, ale metoda analizy nie wskazuje tego jednoznacznie.

Na podstawie ładunków czynnikowych takich jak w analizowanym przykładzie podejmuje się decyzje o budowie narzędzia, gdy tymczasem z dużym prawdopodobieństwem mogą one wskazywać na puste rozwiązanie, niemające nic wspólnego z teoretycznie badanym konstruktem. Ale w badaniach EFA wyjaśniamy wyniki takie, jakie są, i nie mamy możliwości narzucenia rozwiązaniom czynnikowym zakotwiczenia w teorii. Dlatego analizy te mogą wskazywać na istnienie czynników kontekstowych, które są silniejsze niż postulowane teoretycznie. Mam tu na myśli na przykład styl odpowiadania czy brak wiedzy na temat badanego

Tabela 4.7. Ładunki czynnikowe w EFA
dla poszczególnych wersji badanego kwestionariusza

Pozycja	Oryginał	„Poczuciowość"	Nonsensowne zdania	Nonsensowne wyrazy
p1	0,855	0,935	0,390	0,530
p2	0,794	0,933	0,304	0,462
p3	0,783	0,924	0,501	0,337
p4	0,863	0,935	0,594	0,345
p5	0,893	0,945	0,556	0,492
p6	0,902	0,953	0,413	0,520
p7	0,798	0,917	0,392	0,671
p8	0,868	0,943	0,597	0,423
p9	0,883	0,915	0,539	0,344
p10	0,878	0,948	0,635	0,527

zjawiska przy jednoczesnym postawieniu osób badanych w sytuacji konieczności wyboru z ograniczonej puli odpowiedzi. Respondenci często nie mają możliwości odmowy udzielenia odpowiedzi lub też celowo wybierają je losowo, natomiast analiza czynnikowa wszystkie wyniki traktuje tak, jakby zawierały taką samą porcję wiedzy na temat osób badanych.

Uwzględnienie założeń teoretycznych w analizie zapewnia konfirmacyjna analiza czynnikowa. Dlatego czwarta hipoteza dotyczyła analiz CFA. Ponieważ analizowane narzędzie w wersji oryginalnej jest jednowymiarowe dla wszystkich wersji, obliczyłem wartości indeksów dopasowania takiego jednoczynnikowego modelu do danych (tab. 4.8).

Tabela 4.8. Indeksy dopasowania modelu
dla poszczególnych wersji badanego narzędzia

Indeks	Oryginał	„Poczuciowość"	Nonsensowne zdania	Nonsensowne wyrazy
CFI	0,953	0,960	0,921	0,831
TLI	0,940	0,948	0,899	0,782
RMSEA	0,118	0,140	0,057	0,080
RMSEA.CI.LOWER	0,092	0,118	0,000	0,023
RMSEA.CI.UPPER	0,143	0,164	0,107	0,124
SRMR	0,028	0,016	0,079	0,083

W przypadku indeksów dopasowania według powszechnie przyjętych kryteriów (Hu i Bentler, 1999) wersje „puste" wypadają bardzo podobnie do wersji oryginalnej. Dla CFI wersja z nonsensownymi wyrazami różni się na niekorzyść, ale wypada lepiej w kryterium RMSEA, gdy wersja oryginalna nie osiąga wymaganych wartości dla danych dobrze dopasowanych do modelu. Porównując wyniki, nietrudno zauważyć różnice. Problemem jest jednak to, że podczas walidacji narzędzia widzi się tylko wyniki dla niego samego. Gdyby skupić się wyłącznie na wynikach wersji z nonsensownymi wyrazami, można dostrzec, że uzyskane wartości indeksów dopasowania stanowią bardzo dobry punkt wyjścia do ich „podkręcenia", tak by uzyskać wymagane progi krytyczne. Wystarczy skorelować ze sobą błędy pozycji albo zmienić metodę obliczania estymatora, aby uzyskać idealne wyniki (tab. 4.9).

Tabela 4.9. Indeksy dopasowania modeli skorygowanych dla nonsensownych wyrazów

Indeks	Estymator ML	Skorelowane błędy	Estymator DWLS
CFI	0,831	0,963	0,969
TLI	0,782	0,948	0,960
RMSEA	0,080	0,039	0,060
RMSEA.CI.LOWER	0,023	0,000	0,000
RMSEA.CI.UPPER	0,124	0,097	0,108
SRMR	0,083	0,068	0,093

Wartości indeksów dopasowania modelu do danych dla wersji narzędzia wykorzystującej w badaniu pozycje składające się z bezsensownych wyrazów, zarówno po skorelowaniu trzech błędów, jak i po zmianie metody estymacji na nieparametryczną DWLS, spełniają wymagane progi, by uznać narzędzie za trafne. Jest tylko jeden problem – to narzędzie niczego nie mierzy!

4.3. Podsumowanie

Jak wykazałem w rozdziale 3, w przypadku kwestionariuszowych narzędzi samoopisowych działania w zakresie walidacji zazwyczaj składają się z trzech etapów: 1) oszacowania ogólnej rzetelności pomiaru; 2) opracowania modelu zmiennej/zmiennych latentnych; 3) oszacowania związków między mierzoną zmienną a zmiennymi zewnętrznymi (niestety przeważnie również mierzonymi kwestionariuszowo).

Zazwyczaj w etapie pierwszym wykorzystywany jest współczynnik alfa Cronbacha, rzadko uzupełniany innymi współczynnikami lub oszacowaniami precyzji

pomiaru. W etapie drugim przeważają analiza czynnikowa (eksploracyjna lub konfirmacyjna) oraz analiza głównych składowych, rzadko pojawia się zaś modelowanie oparte na IRT. Wreszcie trzeci etap na ogół realizowany jest na podstawie analizy macierzy korelacji wyników nowego narzędzia z wynikami narzędzi już istniejących, a przeznaczonych do mierzenia podobnych cech.

Strategie analityczne użyte w przedstawionym w tym rozdziale badaniu zostały wybrane w wyniku wcześniejszej analizy powszechnie stosowanych i zalecanych procedur walidacyjnych. Manipulacja eksperymentalna polegała na tym, że w wybranych przypadkach pozycje kwestionariuszowe zostały tak skonstruowane, by niemożliwe było spójne ich zinterpretowanie przez osoby badane. Biorąc pod uwagę brak teorii psychologicznych dotyczących tego, co te pozycje mierzą i w jaki sposób oddziałują na osoby badane, uzasadnione wydawało się oczekiwanie, że wyniki stosowania procedur walidacji narzędzi powinny wyraźnie wskazywać na braki w tych wersjach, a hipoteza mówiąca, że kwestionariusz mierzy cokolwiek, powinna zostać z łatwością odrzucona. Jednakże procedury, o których mowa, nie doprowadziły do jednoznacznego wniosku, trudno więc polegać na nich jako na skutecznym filtrze pozwalającym odrzucić słabe, a w skrajnych przypadkach pozbawione sensu narzędzie. Analiza czynnikowa oraz analiza rzetelności powinny sugerować problemy z wersjami kwestionariusza niczego niemierzącymi, a zmiana kafeterii odpowiedzi nie powinna wpływać na wskaźniki rzetelności oraz wartości ładunków czynnikowych. I wreszcie korelacje wyników narzędzia niereprezentujących niczego sensownego z wynikami Wielkiej Piątki powinny wskazywać na problemy z analizowanym narzędziem. Niestety same wartości analiz nie wskazują na problem z narzędziem i możliwe są dowolne spekulacje interpretacyjne co do ich znaczenia – nawet takie, które stwierdzą rzetelność, trafność i sensowność kwestionariusza pozbawionego tych przymiotów. Uzyskane przeze mnie w prezentowanym eksperymencie wyniki można by podsumować stwierdzeniem: przez poważne badanie nieistniejących zjawisk uzasadnia się i uwiarygodnia ich istnienie.

Problem jest poważny, a swoje źródło ma w dominującej roli empiryzmu w nauce od połowy XX wieku. Psychometria zbyt często opiera interpretacje obserwacji na wcześniejszych założeniach. Z jednej strony to zrozumiałe – zajmujemy się badaniem nieobserwowalnych zjawisk psychicznych, z drugiej jednak – nie powinniśmy zapominać, że z definicji nie tylko są one nieobserwowalne, lecz najprawdopodobniej także nie istnieją w sposób, w jaki je definiujemy, a to, co powołuje je do istnienia, to tylko nasze słowa.

ZAKOŃCZENIE

> Podzielam niepokój słynnego psychologa społecznego Roya Baumeistera i współautorów [...], jak również polskiego psychologa społecznego Dariusza Dolińskiego [...], iż psychologia [...] odchodzi od badania rzeczywistych zachowań, aby badać [...] „samoopisy" (wyniki wypełnianych przez osoby badane kwestionariuszy) i „ruchy palców" (w domyśle: na klawiaturze komputera).
>
> Brzeziński, 2019, *Wprowadzenie*

> Zbyt często, moim zdaniem, interakcje te są traktowane jako odkrycie, a nie jako dowód braku zrozumienia zasady łączenia miar niezależnych zmiennych.
>
> Luce, 1995, s. 21[1]

Pomiar w psychologii w latach sześćdziesiątych i siedemdziesiątych XX wieku był bezkrytyczny wobec liczb. Ponieważ trudno zaprzeczyć, że liczby są jednoznaczne, obiektywne, pozwalają na klasyfikowanie i porządkowanie oraz poddają się operacjom logicznym i matematycznym, wskutek których nie tracą tych właściwości, to liczby uzyskane w trakcie pomiaru psychologicznego nabierały tych właściwości bez względu na to, jakie były właściwości mierzonego obiektu. Psychologia borykająca się z subiektywnymi ocenami jakościowymi, dla których nie było narzędzi pozwalających na ich bardziej złożoną analizę, oparła się na metodach ilościowych, często inkorporowanych z pobliskich dziedzin. Jak pisał Mieczysław Choynowski (1971, s. 39):

> językowi brak nazw na oznaczanie bardziej zróżnicowanych stopni cech i przedstawianie przebiegu zmian; [język] nie pozwala na określanie błędu obserwacji oraz stanowi w pewnym sensie ślepy zaułek [...] pomiar sprowadza się [...] do umiejętnego manipulowania odpowiednim narzędziem

[1] „All too often, in my opinion, the interactions are treated as a finding and not as evidence of a lack of understanding of the combining rule for measures of the independent variables".

> pomiaru. W ten sposób subiektywne obserwacje zostają zastąpione obiektywnym narzędziem pomiarowym [wynikiem – P.K.], wydawane sądy stają się niezależne od osoby badacza.

W tym dążeniu do obiektywności i precyzji psychologia pominęła jednak badanie relacji między podmiotem badania, czyli człowiekiem z całą jego złożonością wewnętrzną, a reprezentującymi jego stany liczbami. Po uzyskaniu liczb w dowolnym procesie pomiaru człowiek przestawał być potrzebny i liczyły się (*nomen omen*) tylko one. Mając do dyspozycji logikę i matematykę, mogliśmy nie tylko zrozumieć istniejące tu i teraz reguły rządzące powiązaniami między liczbami, ale także zaglądać w przyszłość. Kazimierz Ajdukiewicz, na którego powołuje się Choynowski, napisał wprost: „przez pomiar możemy korzystać z aparatu matematyki dla wyprowadzenia z twierdzeń zdobytych na drodze obserwacji odległych ich konsekwencji" (Choynowski, 1971).

Z tego punktu widzenia nie istnieje opisywany przeze mnie problem pomiaru i jego trafności – odkryte prawa i reguły są prawdziwe. Tyle tylko, że są prawdziwe jedynie w świecie abstrakcji, w świecie liczb, których wartości zostały oderwane od znaczenia. I jak twierdził Marvin L. Minsky: „Psychologia zeszła z właściwej swej drogi, ponieważ starała się naśladować fizykę, która odniosła sukces, odkrywając proste prawa, obowiązujące zawsze i wszędzie" (Minsky i Laske, 1992, s. 33).

Odpowiedź uzyskana w badaniu i zakodowana w postaci liczbowego wyniku jest efektem działania układu wielu wzajemnie powiązanych przyczyn, które są tylko względnie stałe i wśród których badana cecha jest tylko jedną z wielu. System jest dynamiczny i układy powiązań mogą zmieniać się w kolejnych odpowiedziach osoby badanej pod wpływem choćby tylko uprzednio udzielonej już odpowiedzi. Jeśli, o czym była już mowa, przyznam się w toku pytań do czegoś wstydliwego, to na kolejne pytania mogę zareagować zarówno większą otwartością, jak i – z równym – prawdopodobieństwem z większym wycofaniem. Psycholog zazwyczaj nie kontroluje i nie jest w stanie kontrolować tego dynamicznego układu. Dlatego wydaje mi się uzasadnione stanowisko, że nie potrafimy wyjaśnić związków przyczynowo-skutkowych między badanym konstruktem a udzieloną odpowiedzią. To, czym dysponujemy, to tylko uśredniona reakcja. Dlatego bez krytycznej refleksji nad stosowanymi metodami i narzędziami badawczymi nie przekroczymy w psychologii pewnej granicy dokładności i precyzji, powyżej której działają przyrodnicze dziedziny nauki.

Pierwszym źródłem problemów jest odrzucenie refleksji nad skalami pomiarowymi. Każdy student nauk społecznych, a szczególnie psychologii, jest uczony, że istnieją cztery klasy danych: nominalne, porządkowe, interwałowe i proporcje (ilorazowe) (Stevens, 1946). Każdy mierzony w psychologii atrybut musi pasować do jednej z czterech kategorii i najlepiej, by pasował do najwyższej. O klasyfikacji decydują arbitralne decyzje lub statystyczne testy normalności rozkładu,

które nie testują mierzonej cechy, lecz tylko i wyłącznie rozkład zebranych liczb. Wynik (statystycznie istotny lub nie) przesądza nie o sensowności pomiaru, tylko o wyborze między klasą porządkową a ilościową (ilorazowa i interwałowa traktowane są łącznie, czego głównym powodem jest takie ich opisanie w najpopularniejszym pakiecie statystycznym w naukach społecznych – SPSS). To, że liczby pasują do jakiejkolwiek funkcji matematycznej, nie jest w żaden sposób powiązane z manifestacjami mierzonej cechy. Mało tego, decyzja o klasie pomiarowej może być oparta na jeszcze słabszych przesłankach. Na przykład, skoro liczę średnią punktację z pozycji dostarczających dowolnych wyników: od zero-jedynkowych za odpowiedzi „tak"/„nie", po kilkupunktowe za stopień zgody z twierdzeniem, to mam skalę ilościową (swoją drogą, kategoria „skala ilościowa" jest kolejnym uproszczeniem, bo nawet nie muszę definiować położenia zera i zastanawiać się nad względnością wyników). W efekcie obszarem troski jest to, jakie transformacje mogę stosować na wynikach (od liczby dozwolonych transformacji zależy poziom Stevensowskiej skali), a nie stopień, w jakim wyniki oddają rzeczywistość intrapsychiczną badanych osób. Czy sumaryczne „piątki", składające się na przykład z zestawów 1 + 2 + 2 oraz 1 + 1 + 3 lub 5 + 0 + 0 są tożsame na gruncie badanej cechy? Czy zero oznacza jej brak czy tylko poziom odniesienia? Jak zatem definiowany jest ten poziom? Jaka jest minimalna zmienność cechy i jaki jest jej charakter – ilościowy czy jakościowy? Czy natężenie cechy (jeśli zakładamy pomiar natężenia) jest funkcją liniową, ciągłą i monotoniczną? To tylko czubek góry lodowej pytań, których nie zadajemy, zasłaniając się liczbami w pomiarze psychologicznym, a jak słusznie zauważył Fredric M. Lord (1953): „liczby nie wiedzą, skąd pochodzą".

Inni badacze krytykujący typologię Stevensa zwracali też uwagę na kwestię reprezentatywności – te same liczby mogą odwzorowywać różne zasoby, a poziom pomiaru determinujący operacje statystyczne nie jest cechą danych (Velleman i Wilkinson, 1993). Parafrazując przykład podany przez autorów: to, czy zwycięski numer w loterii potraktujemy jako informację o kategorii, do której przynależy (zwycięski vs przegrywający), czy w sposób ilościowy (ile losów sprzedano) bądź porządkujący (mam los o numerze mniejszym czy większym od zwycięskiego) lub interwałowy (o ile losów bliżej zwycięskiego jest jeden z dwóch, jeśli losy sprzedawano według narastającej kolejności), zależy nie od liczb wydrukowanych na losach, ale od pytań badawczych, które sobie stawiamy.

Wracając do kwestii narzędzi – dominujące w psychologii narzędzia badawcze w postaci kwestionariuszy opierają swoje istnienie na wnioskowaniu, że:

1. (jeżeli) konstrukty psychologiczne objawiają się w postaci cech,
2. (jeżeli) cechy te uzewnętrzniają się w zachowaniach,
3. (jeżeli) ludzie potrafią obserwować, rejestrować i komunikować swe zachowania,
4. (to) pytając o zachowania, możemy ustalić natężenie (posiadanie) określonego konstruktu psychologicznego.

Przyjmując, że są to prawdziwe założenia, to na trzech pierwszych poziomach tej struktury powstają wątpliwości, które zaprzeczają możliwości wnioskowania przedstawionego w punkcie czwartym. Jeśli potrzebne nam do predykcji lub diagnozy konstrukty psychologiczne, będące pewnym semantycznym opisem, objawiają się w postaci cech, nie oznacza to automatycznie, że występowanie cech świadczy na rzecz konstruktu. Cechy mogą mieć i mają różne źródła, co więcej – relacja między cechami ze zmienną ukrytą nie jest izomorficzna. Podobnie jest z cechami jako źródłem zachowań. Zachowanie *a* może wynikać z cech *A*, *B* i *C*, mimo że z cechy *A* zawsze wynika zachowanie *a* – relacja ta nie jest jednoznaczna i zwrotna.

Drugą otwartą kwestią pozostaje założenie o zdolności do introspekcji osób badanych. To także jest cecha, która może mieć zróżnicowany rozkład w populacji. Zakładanie, że ludzie tak samo komunikują swoje zachowanie, podczas gdy wiemy, że nawet ci sami ludzie w zmienionym kontekście (np. warunki laboratoryjne vs staranie się o pracę) zmieniają odpowiedzi, popełniają błędy w myśleniu (por. Bar-Tal et al., 2019) czy ulegają złudzeniom, jest bardzo optymistycznym założeniem. Dlatego nie dziwi mnie, że większość efektów w psychologii ma niską wartość, szacowaną na około $r = 0{,}2$ (lub $d = 0{,}4$) (Open Science Collaboration, 2015).

Badając cechy fizjologiczne, np. refleks, czynimy pewne założenia idealizacyjne co do warunków i stanu osoby. Jeśli ulegną one zmianie, np. na skutek zmęczenia badanego, to wiemy, że mają wpływ na wyniki. Natomiast jeśli badamy konstrukty psychologiczne, które z definicji są strukturami o wysokim stopniu skomplikowania i składają się z wielu elementów, to zakładamy, że są one odporne na kontekst sytuacyjny i podmiotowy. Wrażliwość metodologiczną mamy zaślepioną przekonaniem o rozkładzie normalnym wariancji błędów do tego stopnia, że zwolnieni jesteśmy (albo sami się zwalniamy) z konieczności udowodnienia tej odporności.

Sam wniosek w punkcie czwartym także zawiera jeszcze co najmniej dwa ukryte założenia. Pierwsze wynika z faktu, że pytanie staje się bodźcem, na który może zareagować osoba badana dopiero wtedy, gdy zostanie ono przez nią zrozumiane – oczywiste jest, że proces rozumienia jest podatny na różnego rodzaju zakłócenia i różnice indywidualne. Co więcej, nie zawsze reakcja następuje po zrozumieniu – w skrajnych przypadkach może odbywać się nawet bez niego (przetwarzanie bodźców podprogowych). Drugie założenie odwołuje się do problemów opisanych w rozdziale 1: aby ustalić istnienie i natężenie czegokolwiek, musimy najpierw udowodnić mierzalność tegoż. Psychologowie swymi badaniami potwierdzają stanowisko, że mierzenie jest wystarczającym dowodem mierzalności, konstruując kolejne i kolejne narzędzia w postaci kwestionariuszy, skal, ankiet, inwentarzy itd.

Rozumienie trafności w kategoriach wiarygodności interpretacji pomiarów[2]: czy test mierzy cechę, którą ma mierzyć, czy też nie, zakłada, że atrybuty pomiaru

[2] Trafność zgodnie ze standardami testowania edukacyjnego i psychologicznego (American Educational Research Association…, 1999, s. 9) jest definiowana jako „stopień, w jakim dowody i teoria wspierają interpretację wyników testu".

psychologicznego są mierzalne. Jeśli te atrybuty nie mają w rzeczywistości takiej struktury, która pozwala na pomiar, to po prostu nie będzie możliwe ich zmierzenie bez względu na to, ile czasu i wysiłku zostanie zainwestowane w proces opracowywania testu.

Problem leży także w samej procedurze walidacji. Przeniesienie przez Cronbacha i Meehla akcentu badania trafności z narzędzia na interpretację wyników uzyskanych za pomocą tego narzędzia doprowadziło do rozmycia pojęcia trafności oraz występowania kuriozalnych sytuacji. Możemy sobie na przykład wyobrazić, że istnieje test polegający na wyciąganiu słomek. Powtarzana procedura pozwala każdej osobie przypisać od 0 do *n* punktów za wyciągnięcie słomki krótszej od innych. Jedynie sensowna interpretacja zebranych wyników brzmi: test słomkowy niczego nie mierzy. Do tego momentu wszystko jest akceptowalne i intuicyjnie poprawne. Ale zgodnie z przyjętymi standardami, aby ocenić wiarygodność (trafność) tej interpretacji musimy zgromadzić dowody na nią. Oczywiste jest, że w toku analiz walidacyjnych nie stwierdzimy związków z ekstrawersją, inteligencją, różnic międzypłciowych, związanych z wiekiem czy miejscem zamieszkania. Na poziomie oceny trafności interpretacji wyników nasza ocena wskazuje na trafność i rzetelność testu – aspekty rozbieżne trafności potwierdzają istnienie niezależnego czynnika X. Ale z poziomu samych wyników niczego nie mierzymy. Uzyskujemy sprzeczność, która stawia pod znakiem zapytania naszą, jako psychologów i psychometrów, zdolność do odróżnienia rzeczy wartościowych od bezwartościowych.

Źródłem problemów jest też niska świadomość metodologiczna badaczy – w czasach gdy liczą się głównie punkty za publikacje, badacze poświęcają zbyt mało uwagi kwestii trafności używanych narzędzi. Według analizy Flake i Fried (2019, s. 4) wcale nie robi tego nawet 93% badaczy, a zdumiewająca większość (69%) nie odwołuje się do uprzednich badań w obszarze ujętym w publikacji. Trudno zatem oceniać wagę doniesień naukowych, gdy brakuje informacji o narzędziach wykorzystanych do ich odkrycia. Struktura publikacji naukowych jest wbrew pozorom dosyć sztywna – badacze przedstawiają, w jaki sposób odnieśli się do podejmowanego problemu badawczego, raportują środki (w tym narzędzia badawcze), jakie wykorzystali, i to, w jakim stopniu ich praca potwierdza (lub nie potwierdza) istniejące hipotezy. Kultura naukowa wywiera znaczną presję na przedstawianie projektów badawczych jako rozstrzygających narracji, w których nie ma miejsca na wątpliwości lub niejednoznaczne wyniki. Presja ta sprawia, że wyniki badań są zgodne z oczekiwaniami, mimo że jest to sprzeczne z tym, jak *de facto* wygląda pełen wątpliwości proces badawczy. Uzyskane korelacje często pokazują niedoskonały związek między cechami. Aby poradzić sobie z tą niedoskonałością, wprowadza się do pomiaru pojęcie błędu. Do jego kontroli stosuje się albo metody matematyczno-statystyczne (np. zwiększając liczebność próby), albo *post hoc* szuka czynników poza tymi, które były badane, aby nimi wyjaśnić wpływ na uzyskane wyniki. Jest też trzeci sposób, który zdecydowanie powinien

być częściej stosowany: jeśli błędy mają systematyczny charakter, to można zaprojektować eksperymenty, które to potwierdzą i pozwolą na ich wyeliminowanie, co znacznie zwiększy trafność pomiarów pierwotnych cech.

Paradoksalnie, opierając się na szeroko opisanych strategiach walidacji opisanych na przykład w wydaniach Standardów z 1999 roku (tab. 5.1), można by oczekiwać, że w ciągu ponad dwudziestu lat psychologia wykroczy poza korelacyjno-obserwacyjną konceptualizację trafności. Niestety jednak, jak wykazałem w rozdziale 3 i jak na przykład uważa też Bornstein (2011), obecne praktyki w tym zakresie nie zmieniły się.

Tabela 5.1. Typy dowodów trafności z wersji Standardów z 1999 roku

Typ dowodu	Typowa procedura walidacji
Oparte na treści	analiza logiczna i ocena ekspercka zawartości pozycji, dopasowanie pozycji do konstruktu, znaczenie pozycji, pobieranie próbek z uniwersum i kryterium zanieczyszczenia
Oparte na procesie odpowiadania	oparte na wywiadzie i obserwacji analizy odpowiedzi uczestników na elementy lub zadania; porównanie różnic w procesach w poszczególnych grupach; badania procesów decyzyjnych obserwatora/interpretatora
Oparte na wewnętrznej strukturze	analizy czynnikowe, analizy skupień, analizy pozycji, analizy różnicowego funkcjonowania pozycji
Oparte na relacjach z innymi zmiennymi	trafność równoległa i predykcyjna, trafność zbieżna i dyskryminacyjna, uogólnienie trafności, kryteria różnic grupowych, badania oceniające wpływ interwencji/manipulacji na wyniki testów, badania podłużne
Oparte na konsekwencjach testowania	badania oczekiwanych/osiągniętych korzyści z badań; badania niezamierzonych negatywnych skutków

Źródło: Bornstein, 2011, s. 535.

Analizy Thomasa P. Hogana i Jessiki Angello (2004, za: Bornstein, 2011) pokazały, że większość (87%) opublikowanych badań nad trafnością narzędzi badawczych dotyczyła korelacji. Kilka lat później Cizek ze współautorami (Cizek, Rosenberg i Koons, 2008; za: Bornstein, 2011) przebadali 283 zrecenzowane testy i stwierdzili, że tylko niecałe 2% z nich oceniało procesy odpowiadania uczestników badań, a pozostałe dowody trafności opierały się nadal tylko na współczynnikach korelacji. Bornstein szacuje, że liczba walidacji opartych na procedurach eksperymentalnych nie przekracza 10%. Niestety opieranie się psychologii na metodach korelacyjnych w celu oszacowania trafności z definicji prowadzi

do tego, że wynik testu jest ważny dla dowolnego korelującego konstruktu – zbieżnie lub rozbieżnie, prognostycznie lub kryterialnie.

Zestaw analiz składający się z analiz czynnikowych i analiz korelacji jest w zasadzie niereaktywny w stosunku do teorii – wielu (większość?) badaczy uważa owe analizy za uniwersalny zestaw strategii walidacyjnych, których użycie wystarcza w zasadzie, by uznać test za „zatwierdzony", niezależnie od obecności lub braku teorii psychologicznej wyjaśniającej znaczenie zmienności wyników czy ich przyczynowo-skutkowe powiązanie z latentnym konstruktem. Ale czy analiza czynnikowa może być dowodem na zasadność teoretyczną proponowanego modelu? Taka analiza opiera się na różnicach między osobami, ale to nie znaczy, że postulowana zmienna latentna ma podobną strukturę. Taki izomorfizm jest zakładany, czyli jako taki powinien mieć status hipotezy i zostać udowodniony.

Teoria zmiennej latentnej także budzi moje wątpliwości. Często jest to raczej rozwiązanie wygodne, a nie rozwiązanie prawdziwe. Uważam tak dlatego, że wprowadzenie do modelu zmiennej latentnej jest jednocześnie łatwo (w sensie dostępności technik statystycznych) i trudno weryfikowalne (w sensie merytorycznym). A przecież wraz z rozwojem statystyki możliwe stało się stawianie i sprawdzanie hipotez konkurencyjnych wobec założenia, że wariancja wyników ma przyczynę w pojedynczej, niedostępnej pomiarowi zmiennej. Jako przykład weźmy inteligencję – mimo wielu kontrowersji wokół definicji tego pojęcia ma ono ugruntowaną pozycję zarówno w rzeczywistości naukowej, jak i potocznej. Tymczasem w 2006 roku Han L. Van der Mass (2006) ze współpracownikami wykazali, że możliwe jest wytłumaczenie wyników w teście na inteligencję przez model dynamiczny, w którym procesy poznawcze korzystnie oddziałują na siebie podczas rozwoju jednostki. Skutkiem tego jest taki sam pozytywnie skorelowany zbiór wyników jak wtedy, gdy zakładamy istnienie nieobserwowalnego bezpośrednio czynnika *g*. Tym samym nie jest nam on (i pojęcie inteligencji) potrzebny do wyjaśnienia zmienności wyników.

Zgodnie z teorią trafności teoretycznej (*construct validity*) nie można zebrać dowodów na trafność testów, ale jedynie na interpretacje (zastosowania) zabranych za ich pomocą wyników. Jeśli te wyniki mają być interpretowane jako miary cech psychicznych, a trafność jest uważana za dowód na sieć nomologiczną łączącą te wyniki i pozycje testu, to można spostrzec, że trafność teoretyczna jest również właściwością interpretacji wyników (Cronbach i Meehl, 1955). To oznacza, że trafność teoretyczna nie odnosi się do pytania, czy testy rzeczywiście mierzą na przykład inteligencję, tylko do pytania, jak dobrze pewne interpretacje wyników testów mierzących inteligencję poparte są dowodami. Trafność teoretyczna jest oceną precyzji procesu odwzorowania „pojęcia teoretycznego" czysto semantycznego i symbolicznego, które nie jest z definicji mierzalne – w rzeczywistości mierzalnej, która jest przyczyną obserwowanych różnic u badanych osób. Proces ten wymaga zaistnienia pomiaru, mimo że cechy istnieją niezależnie od niego i od konstruktu. Tworzy to sytuację, w której, aby określić trafność, musimy

mierzyć, ale aby mierzyć, powinniśmy mieć trafne narzędzie. Używanie pojęcia konstruktu nie pomaga w rozwiązaniu tego dylematu i dlatego Borsboom (2009, s. 152) proponuje porzucenie terminu konstrukt i wprowadzenie w jego miejsce rozłącznych: „pojęcie teoretyczne" oraz „atrybut psychologiczny"[3].

Uwzględniający propozycję Borsbooma[4] proces budowy narzędzia zaczynałby się od precyzyjnego opisania pojęcia teoretycznego z uwzględnieniem potencjalnych korelatów. Opis ten powinien uwzględniać strukturę, na przykład to, czy jest ona jednowymiarowa czy hierarchiczna. Drugi etap polegający na tworzeniu pozycji do pomiaru atrybutów psychologicznych powinien być wrażliwy na zaplanowany format udzielania odpowiedzi oraz powinien unikać skażenia innymi pojęciami teoretycznymi. Takim błędem jest na przykład używanie sformułowań typu „Czasami…", które są zbyt ogólne, lub „Martwię się o…", które mogą zawierać obciążenie zdolnością do introspekcji lub neurotycznością. Tworzenie zbioru pozycji powinno być dokonywane ze świadomością istnienia paradoksu spójności – z jednej strony pożądane są spójne treściowo pozycje, z drugiej – im są spójniejsze, tym bardziej redundantne. Na etapie analitycznym warto rozważyć zastosowanie IRT, które ma dwie ważne zalety w porównaniu z innymi strategiami wyboru pozycji. Po pierwsze, umożliwia określenie poziomu atrybutu, na którym każda pozycja jest maksymalnie informacyjna. Po drugie, pozwala na oszacowanie poziomu każdego atrybutu bez konieczności administrowania całym zestawem pozycji. Elastyczność ta umożliwia opracowanie komputerowych testów adaptacyjnych (CAT), w których ocena skupia się przede wszystkim na podzbiorze pozycji, mających charakter maksymalnie informacyjny dla każdej osoby uczestniczącej w badaniu. Należy jednak pamiętać, że IRT nie jest panaceum na problem z istnieniem mierzalnego pojęcia teoretycznego – już w 1978 roku Robert Wood wykazał, że losowy rzut monetami jest zgodny z modelem Rascha. Dlatego dopasowanie matematycznego modelu do obserwowanych danych nie jest wystarczającym dowodem na istnienie badanego zjawiska w postaci latentnego konstruktu.

W moim przekonaniu rozwiązaniem problemu jest eksperymentowanie – poszukiwanie takich manipulacji, które oddziałują na psychologiczną zmienną ukrytą, i obserwowanie, czy mierzone zmienne zmieniają się zgodnie z zakładanym modelem powiązań między teorią a empirią. Wymogiem gwarantującym efektywność tego procesu jest stosowanie go na etapie konstruowania narzędzia badawczego, w przeciwieństwie do obecnego nacisku, położonego na analizę wyników badań. Jednocześnie należy pamiętać, że wymierność pojęcia teoretycznego jest

[3] Borsboom (2004, s. 1065) uważa, że równie niepotrzebne jest pojęcie sieć nomologiczna – według niego zastąpienie jej „odniesieniem" i przywrócenie realistycznej perspektywy sprawi, że wiele z tego, co się mówi w teorii trafności teoretycznej, i tak pozostanie spójne i wiarygodne.

[4] A także zgodny z postulatem Loevinger (1957, s. 642): „To prawda, że psychologowie nigdy nie poznają cech bezpośrednio, ale jedynie przez pryzmat swoich konstruktów, lecz dane, które podlegają ocenie, są manifestacją cech, a nie manifestacją konstruktów".

niemożliwa do udowodnienia – pragmatyczne podejście wskazuje na szukanie argumentów mogących pokazać, że możliwy jest za to pomiar atrybutu psychologicznego, i po stronie badacza stoi obowiązek udowodnienia możliwości i jakości tego pomiaru, niekoniecznie tylko ilościowego.

Podzielamy silne przekonanie, że pomiar ilościowy jest niezbędnym elementem badań naukowych, ale zapominamy o tym, że w wielu obszarach nauki, nie tylko psychologii, przełomy dokonywane były bez niego. Jeśli pomyślimy o dziesięciu największych: Skinnerze, Piagecie, Freudzie, Bandurze, Festingerze, Jamesie, Pawłowie, Rogersie, Eriksonie czy Wygotskim, to więcej niż połowa z nich wywarła niezaprzeczalny wpływ na psychologię, opierając się na badaniach jakościowych.

Naukowe teorie wyjaśniają niewyjaśnione: pokazują, jak powiązane są ze sobą różne elementy, i zwiększają zrozumienie otaczającej nas rzeczywistości – w przypadku psychologii także rzeczywistości intrapsychicznej. Psychologia wypracowała sposób narracji przedstawiający uzyskane wyniki w sposób, który potwierdza intuicję odbiorców: „tak, jak o tym pomyślę, to oczywiste". Łączymy kropki na białej przestrzeni i osoba oglądająca rysunek nie wątpi w nasz pomysł ich połączenia, nie kwestionuje, nie sugeruje, że można by połączyć je inaczej. Badacze rzadko muszą przekonywać do słuszności swych interpretacji, do sensowności decyzji i świadomości ich konsekwencji. Wystarczy, że narracja wokół wyników badań empirycznych jest spójna – szczegółowość prac powstających obecnie sprawia, że rzadko kto jest ekspertem w tematyce podjętej przez innego badacza na tyle, by zaproponować konkurencyjne „połączenie kropek".

Jednocześnie nie mam wątpliwości, że psychometria będzie się rozwijać. Podobnie, w klasycznej fizyce kolor był zmienną jakościową, współcześnie zaś na podstawie odkryć optyki definiuje się go jako zmienną ilościową. Obecnie, mając do dyspozycji znane sposoby pomiaru konstruktów (atrybutów psychologicznych), są one, jak się wydaje, w większości jakościowe. Ale wraz z rozwojem sposobów pomiaru zwiększa się nasza wiedza, co może przynieść zmianę myślenia i powstawanie mocniejszych, o większej „rozdzielczości" testów. Wciąż testy psychologiczne – a właściwie wyniki w nich uzyskiwane – często stanowią najuczciwszą i najdokładniejszą metodę podejmowania ważnych decyzji.

BIBLIOGRAFIA

Ajdukiewicz, K. (1961). *Pomiar*. Studia Logica, *11*(1), 223–231.

Allport, G.W. (1943). The ego in contemporary psychology. *Psychological Review, 50*(5), 451.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1955). *Technical Recommendations for Achievement Tests*. Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1974). *Standards for Educational and Psychological Tests*. Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington, DC.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37* (1), 1–16.

Anastasi, A., Urbina, S.P. (1989). *Psychological Testing*. London: Collier Macmillan.

Bar-Tal, Y., Brycz, H., Dolinska, B., Dolinski, D. (2019). When saying that you are biased means that you are acurate? The moderating effect of cognitive structuring on relationship between metacognitive self and confirmation bias use. *Current Psychology, 38*, 1706–1712.

Baumeister, R.F., Vohs, K.D., Funder, D.C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396–403.

Blinkhorn, S.F. (1998). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50*(2), 175–185.

Bornstein, R.F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23*(2), 532.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6.

Borsboom, D., Cramer, A.O., Kievit, R.A., Scholten, A.Z., Franić, S. (2009). The end of construct validity. W: R.W. Lissitz (red.), *The Concept of Validity: Revisions, New Directions and Applications* (ss. 135–170). Charlotte, NC: IAP Information Age Publishing.

Borsboom, D., Mellenbergh, G.J., Heerden, van, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071.

Borsboom, D., Wijsen, L.D. (2017). Psychology's atomic bomb. *Assessment in Education: Principles, Policy & Practice, 24*(3), 440–446.

Brown, J.D. (1991). Accuracy and bias in self-knowledge. W: C.R. Snyder, D.R. Forsyth (red.), *Handbook of Social and Clinical Psychology: The Health Perspective* (t. 162, ss. 158–178). New York, NY: Pergamon Press.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322.

Brown, W. (1938). *Psychological Methods of Healing. An Introduction to Psychotherapy*. London: University of London Press.

Bruschi, A. (2017). Measurement in social research: some misunderstandings. *Quality & Quantity, 51*(5), 2219–2243.

Brycz, H., Jurek, P., Pastwa-Wojciechowska, B., Peplińska, A., Bidzan, M. (2014). Auto-atrybucje metawiedzy Ja w ujęciu teorii Bernarda Weinera. *Przegląd Psychologiczny, 57*(3), 347–355.

Brycz, H., Wyszomirska-Góra, M., Konarski, R., Wojciszke, B. (2018). The metacognitive self fosters the drive for self-knowledge: The role of the metacognitive self in the motivation to search for diagnostic information about the self. *Polish Psychological Bulletin, 49*(1), 66–76.

Brycz, H., Konarski, R., Kleka, P., Wright, R. (2019). The metacognitive self: The role of motivation and an updated measurement tool. *Economics & Sociology, 12*(1), 208–232.

Brzeziński, J.M. (2004). *Metodologia badań psychologicznych*. Warszawa: Wydawnictwo Naukowe PWN.

Buckingham, B.R., McCall, W.A., Otis, A.S., Rugg, H.O., Trabue, M.R., Courtis, S.A. (1921). Report of the standardization committee. *Journal of Educational Research*, *4*(1), 78–80.

Campbell, D.T., Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Campbell, N.R. (1920). *Foundations of Science: The Philosophy of Theory and Experiment*. New York, NY: Dover Publications.

Cattell, J.M. (1890). Mental tests and measurements. *Mind, 15*(59), 373–381.

Chin, W.W. (2010). How to write up and report PLS analyses. W: V. Esposito Vinzi, W.W. Chin, J. Henseler, H. Wang (red.), *Handbook of Partial Least Squares: Concepts, Methods and Applications* (ss. 655–690). Heidelberg–Dordrecht–London–New York: Springer.

Choynowski, M. (1971). Pomiar w psychologii. W: J. Kozielecki (red.), *Problemy psychologii matematycznej* (ss. 15–42). Warszawa: Państwowe Wydawnictwo Naukowe.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*(3), 186–190.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Costello, A.B., Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), 7.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
Cronbach, L.J. (1989). Construct validation after thirty years. *Intelligence: Measurement, Theory, and Public Policy, 3*, 147–171.
Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley & Sons, Inc.
Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–303.
Cronbach, L.J., Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418.
Domański, C.W. (2018). *Historia psychologii w Europie Środkowej. Badacze, inspiracje i koncepcje.* Warszawa: Wydawnictwo Naukowe PWN.
Domingue, B.W. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika, 79*(1), 1–19.
Drever, J. (1938). The quantitative relation between physical stimulus and sensory event. *British Association for the Advancement of Science, 108*, 331–334.
Dunn, T.J., Baguley, T., Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.
Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097.
Erikson, E.H. (1950). Growth and crises of the „healthy personality". W: M.E. Senn (red.), *Symposium on the Healthy Personality* (ss. 91–146). New York, NY: Josiah Macy, Jr. Foundation.
Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., … Tucker, W.S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science, 1,* 311–341.
Flake, J.K., Fried, E.I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv*. https://doi.org/10.31234/osf.io/ hs7wm.
Fornell, C., Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.
Fraser, C.O. (1980). Measurement in psychology. *British Journal of Psychology, 71*(1), 23–34. https://doi.org/10.1111/j.2044-8295.1980.tb02725.x.
Fromm, E. (1955). The present human condition. *The American Scholar, 25*(1), 29–35.
Goldstein, H. (2015). Validity, science and educational measurement. *Assessment in Education: Principles, Policy & Practice, 22*(2), 193–201.
Groth, J., Kleka, P. (2018). Patterns of intentional faking in questionnaire-based study of psychopathy. *Current Issues in Personality Psychology, 6*(4), 305–317. https://doi.org/10.5114/cipp.2018.80199.
Guilford, J.P. (1936, 1954). *Psychometric Methods*. Pt. I. New York, NY: McGraw-Hill.
Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427–438.
Guilford, J.P., Comrey, A.L. (1951). Measurement in psychology. W: H. Helson (red.), *Theoretical Foundations of Psychology* (ss. 506–556). New York, NY: Nostrand Co.
Gulliksen, H. (1950). *Theory of Mental Tests*. New York, NY: Wiley & Sons, Inc.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282.
Haig, B.D. (2003). What is a spurious correlation? *Understanding Statistics, 2*(2), 125–132. https://doi.org/10.1207/S15328031US0202/_03.
Heene, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology, 4,* 246–250.
Heene, M., Kyngdon, A., Sckopke, P. (2016). Detecting violations of unidimensionality by order-restricted inference methods. *Frontiers in Applied Mathematics and Statistics, 2*, 1–13.
Henseler, J., Ringle, C.M., Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43* (1), 115–135. https://doi.org/10.1007/s11747-014-0403-8.
Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. *Mathematisch-Physische Class, 53,* 1–64.
Hornowska, E. (2001). *Testy psychologiczne: teoria i praktyka.* Warszawa: Wydawnictwo Naukowe „Scholar".
Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*(3), 153–160.
Hu, L.-T., Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.
Izdebski, P., Żbikowska, K., Kotyśko, M. (2013). Przegląd teorii aprobaty społecznej. *Acta Universitatis Lodziensis. Folia Psychologica, 17*, 5–20.
Jaworowska, A., Matczak, A., Ciechanowicz, A. (2002). *Omnibus: test inteligencji: podręcznik.* Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.
Juczyński, Z. (2001). *Narzędzia pomiaru w promocji i psychologii zdrowia*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.
Kahneman, D., Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–591. https://doi.org/10.1037/0033-295X.103.3.582.
Kane, M.T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38*(4), 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.
Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. http://dx.doi.org/10.1111/jedm.12000.
Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*(4), 389–423.
Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology, 50*(2), 123–148.
Karabatsos, G. (2018). On Bayesian testing of additive conjoint measurement axioms using synthetic likelihood. *Psychometrika, 83*(2), 321–332.
Kelley, T.L. (1923). The principles and techniques of mental measurement. *American Journal of Psychology, 34*, 408–432.
Kelley, T.L., Shen, E. (1929). The statistical treatment of certain typical problems. W: C. Murchison (red.), *The Foundations of Experimental Psychology* (ss. 855–883). Worcester, MA: Clark University Press.
Kleka, P. (2017). Paradane z kwestionariuszy jako źródło wartościowej informacji o osobach badanych. W: W.J. Paluchowski (red.), *Diagnozowanie – wyzwania i konteksty* (ss. 133–148). Poznań: Wydawnictwo Naukowe Wydziału Nauk Społecznych.

Kleka, P., Soroko, E. (2018). How to Avoid the Sins of Questionnaires Abridgement? *Survey Research Methods, 13*(2), 9999–10012.
Krantz, D.H. (1964). Conjoint measurement: The Luce–Tukey axiomatization and some extensions. *Journal of Mathematical Psychology, 1*(2), 248–277. https://doi.org/10.1016/0022-2496(64)90003-3.
Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A. (1971). *Foundations of Measurement* (t. 1: *Additive and Polynomial Representations*). New York, NY: Academic Press.
Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*(2), 115–129.
Kuder, G.F., Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.
Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology, 64*(3), 478–497.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694.
Lord, F.M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*(4), 517–549.
Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Readings, MA: Addison-Wesley.
Luce, R.D., Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X.
Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*(1), 251–280.
Mabe, P.A., West, S.G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296. https://doi.org/10. 1037/0021-9010.67.3.280.
Maraun, M.D., Gabriel, S.M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology, 31*(1), 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006.
Markus, K.A., Borsboom, D. (2011). The cat came back: Evaluating arguments against psychological measurement. *Theory & Psychology, 22*(4), 452–466.
Markus, K.A., Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology, 31*(1), 54–64. https://doi.org/10.1016/j.newideapsych.2011.02.008.
Maslow, A.H. (1950). Self-actualizing people: A study of psychological health. *Personality, 1*, 11–34.
McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ–London: Psychology Press.
Meehl, P.E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.
Messick, S. (1988). Meaning and values in test validation: The science and ethics of assessment. *Research Report Series, 2*, i–28. https://doi.org/10.1002/j.2330-8516.1988.tb00303.x.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Metcalfe, J., Shimamura, A.P. (1994). *Metacognition: Knowing about Knowing. Cambridge*, MA: MIT Press. https://books.google.pl/books?id= Ci0TDgAAQBAJ.
Meyer, M. (1926). Special reviews. *Psychological Bulletin, 23*, 261–276.
Michell, J. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology, 32*(4), 466–473. https: //doi.org/10.1016/0022-2496(88)90024-7.
Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. New York, NY: Psychology Press.
Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept* (t. 53). Cambridge: Cambridge University Press.
Michell, J. (2000). Normal science, pathological science and psychometric. *Theory & Psychology, 10*(5), 639–667.
Michell, J. (2008). Conjoint measurement and the Rasch paradox: A response to Kyngdon. *Theory & Psychology, 18*(1), 119–124.
Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma. *Theory & Psychology, 24*(1), 111–123. https: //doi.org/10.1177/0959354313517524.
Minsky, M.L., Laske, O. (1992). A conversation with Marvin Minsky. *AI Magazine, 13*(3), 31. https://doi.org/10.1609/aimag.v13i3.1009.
Nagel, E. (1931). Measurement. *Erkenntnis, 2*, 313–333.
Narens, L., Luce, R.D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99*(2), 166–180.
Newman, E.B. (1974). On the origin of „Scales of Measurement". W: H.R. Moskowitz, B. Scharf, J.C. Stevens (red.), *Sensation and Measurement: Papers in Honor of Stevens* (ss. 137–145). Dordrecht–Boston: D. Reidel Publishing Company.
Newton, P.E., Shaw, S. (2014). *Validity in Educational and Psychological Assesment*. London: Sage Publications.
Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20*(4), 641–650. https://doi.org/10.1177/001316446002000401.
Obrębska, M., Kleka, P. (2016a). Gadatliwość kobiet: prawda czy stereotyp? W: J. Mampe, H. Makurat, Ł. Owczinnikowa, F. Marzouk (red.), *Socjolingwistyczne badania w teorii i w praktyce. Ujęcie interdyscyplinarne* (ss. 77–86). Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.
Obrębska, M., Kleka, P. (2016b). Wpływ aktywizacji schematu płci i potrzeby dostosowania interpersonalnego na wybory leksykalne kobiet i mężczyzn. *Psychologia Społeczna, 11*(37), 170–182.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251), aac4716.
Parker, I. (2007). Critical psychology: What it is and what it is not. *Social and Personality Psychology Compass, 1*(1), 1–15.
Paunonen, S.V., LeBel, E.P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology, 103*(1), 158–175. https://doi.org/10.1037/a0028165.
Perline, R., Wright, B.D., Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237–255. https://doi.org/10.1177/014662167900300213.

Reese, T. (1943). The application of the theory of physical measurement to the measurement of psychological magnitudes, with three experimental examples. *Psychological Monographs, 55*, 1–89.
Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*(1), 57–74.
Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*(1), 99–103.
Russell, B. (1937). *Principles of Mathematics*. Cambridge: Cambridge University Press.
Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93.
Scott, D. (1964). Measurement structures and linear inequalities. *Journal of Mathematical Psychology, 1*(2), 233–247. https://doi.org/10.1016/0022-2496(64)90002-1.
Scott, D., Suppes, P. (1958). Foundational aspects of theories of measurement 1. *The Journal of Symbolic Logic, 23*(2), 113–128.
Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology, 22*(6), 786–809.
Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats it-self again. W: R.W. Lissitz (red.), *The Concept of Validity: Revisions, New Directions and Applications* (ss. 19–37). Charlotte, NC: IAP Information Age Publishing.
Slaney, K.L., Racine, T.P. (2013a). Constructing an understanding of constructs. *New Ideas in Psychology, 31*(1), 1–3. https://doi.org/10.1016/j.newideapsych.2011.02. 010.
Slaney, K.L., Racine, T.P. (2013b). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*(1), 4–12. https://doi.org/10.1016/j.newideapsych. 2011.02.003.
Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.
Stevens, S.S. (1946). On the theory of scales measurement. *Science, 103* (2684), 677– 680.
Strack, F., Martin, L.L., Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology, 18*(5), 429–442.
Ten Berge, J.M., Zegers, F.E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43*(4), 575–579.
Thorndike, E.L. (1904). *Theory of Mental and Social Measurement*. New York, NY: Science Press.
Thorndike, E.L. (1918). The nature, purposes, and general methods of measurements of educational products. W: G.M. Whipple (red.), *Seventeenth Yearbook of the National Society for the Study of Education* (t. 2, ss. 16–24). Bloomington, IL: Public School Publishing.
Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.
Thurstone, L.L. (1937). Psychology as a quantitative rational science. *Science, 85*, 228–232.
Topolewska, E., Skimina, E., Strus, W., Cieciuch, J., Rowiński, T. (2014). Krótki kwestionariusz do pomiaru Wielkiej Piątki IPIP-BFM-20. *Roczniki Psychologiczne, 17*(2), 367–384.
Tourangeau, Roger. (1984). Cognitive aspects of survey methodology: Building a bridge between disciplines. W: T.B. Jabine, M.L. Straf, J.M. Tanur, R. Tourangeau (red.) *Cognitive aspects of survey methods* (ss. 73–100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L.J., Rasinski, K. (2000). *Psychology of Survey Response*. London: Cambridge University Press.
Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology, 19*(5), 579–599.
Tukey, J.W. (1961). Data analysis and behavioural science or learning to bear the quantitative man's burden by shunning badmandments. W: L.V. Jones (red.), *The Collected Works of John W. Tukey* (t. 3, 1986, ss. 391–484). Belmont, CA: Wadsworth.
Turner, B.M., Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56*(2), 69–85.
Turner, M.D. (2019). An Orientation to the Replication Crisis in Psychology. Jacksonville, FL: Conference: *Southeastern Psychological Association 65th Annual Meeting*.
Tversky, A. (1967). A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology, 4*(1), 1–20. https://doi.org/10.1016/0022-2496(67)90039-9.
Van Der Maas, H.L., Dolan, C.V., Grasman, R.P., Wicherts, J.M., Huizenga, H.M., Raijmakers, M.E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842.
Velleman, P.F., Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*(1), 65–72.
Winter, H. (1949). The work of G.T. Fechner on the Galvanic circuit. *Annals of Science, 6*(2), 197–205.
Zinbarg, R.E., Revelle, W., Yovel, I., Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

SPIS TABEL

SPIS RYSUNKÓW