dr Artur Pokropek
Instytut Badań Edukacyjnych

Porównywalność w badaniach międzynarodowych. Przykład wskaźnika motywacji do nauki w badaniu PISA 2006

Wstęp

Od momentu powstania nowoczesnych nauk społecznych porównania między krajami, kulturami czy społecznościami były ich konstytutywną częścią. Trudno wyobrazić sobie prace Durkheima, Webera czy Marksa bez elementów analizy porównawczej. Poszukiwanie podobieństw i różnic między krajami czy kulturami pomaga odkryć mechanizmy rządzące działaniem ludzi i społeczeństw, niezależnie od tego, czy badanie ukierunkowane jest na odkrycie uniwersalnych praw (jak u Durkheima i Marksa), czy raczej na pokazanie różnic (jak u Webera). Durkheim twierdził, że w socjologii systematyczne porównania są analogią do badań eksperymentalnych przeprowadzanych w naukach przyrodniczych, dlatego w zasadzie cała socjologia ma naturę porównawczą (Durkheim, 1895 [2000]). Dla nowoczesnych badań społecznych kwestia porównywalności, poszukiwania podobieństw i różnic w różnych systemach społecznych stała się motywem przewodnim (por. Trieman, 1977; Tyree, Semyonov i Godge, 1979; Heath, 1981; Erikson i Goldthrope, 1992).

Korzyści płynące z badań porównawczych szybko dostrzeżone zostały w epidemiologii, ekonomii czy w badaniach edukacji. Badania porównawcze zyskały też przychylną opinię w oczach polityków i w ogóle społeczeństw, gdyż wypływające z nich wnioski są mniej hermetyczne, łatwiejsze do komunikacji i zrozumienia. Co oznacza, że 70% Polaków jest szczęśliwych w życiu, albo że polscy uczniowie zdobyli 496 punktów z testu mierzącego umiejętność czytania ze zrozumieniem? Trudno powiedzieć. Ale już komunikat, że Polacy są, średnio rzecz biorąc, bardziej szczęśliwym narodem od Szwedów, lub że polscy uczniowie wypadają lepiej niż uczniowie niemieccy, ale gorzej niż fińscy, przedstawiają konkretne, łatwiejsze do interpretacji fakty.

Zainteresowanie dokonywaniem porównywań między społeczeństwami doprowadziło na przełomie XX i XXI wieku do stworzenia nowego typu badań: badań zaprojektowanych dla porównań, które umożliwiły badaczom społecznym rozpoczęcie szerokiej fali badań porównawczych. Niemniej pomimo wysiłków wkładanych w przygotowywanie tego typu badań, w zapewnienie międzynarodowej porównywalności wyników, założenie o porównywalności powinno zawsze zostać zweryfikowane empirycznie, na podstawie zebranych danych. W artykule tym przedstawione zostaną teoretyczne podstawy takich weryfikacji i statystyczne narzędzia, które używane są do ich przeprowadzenia. Zbiór danych, który wykorzystany zostanie w celu prezentacji problematyki, to dane z badania PISA 2006. Problemem, który zostanie podjęty, to porównywalność wskaźnika instrumentalnej motywacji do nauki przedmiotów przyrodniczych.

Badania zaprojektowane dla porównań

Badania, które próbowały dokonywać porównań między społeczeństwami na podstawie różnych danych zastanych, np. oficjalnych statystyk cenzusowych, były w socjologii znane od dawna. Korzystanie z zastanych źródeł danych wiąże się jednak z poważnymi trudnościami metodologicznymi. W różnych krajach i w różnych biurach statystycznych badane zjawiska mogą być różnie definiowane i operacjonalizowane. Brak kontroli lingwistycznej sprawia, że znaczenia odpowiedzi respondentów nie mogą być ze sobą bezpośrednio porównywalne, a różne definicje populacji i schematy doboru próby w badaniach narodowych stanowią istotną trudność dla badaczy chcących dokonać porównań (por. Rokkan, 1969; Verba, 1969). Te fundamentalne problemy związane z trafnością porównań mogły zostać przezwyciężone dopiero po wprowadzeniu badań zaprojektowanych specjalnie do porównań międzynarodowych.

Pierwsze udokumentowane użycie badania zaprojektowanego do porównań przeprowadzono podczas drugiej wojny światowej i dotyczyło ono wpływu efektów bombardowań na morale cywilów w Niemczech, Wielkiej Brytanii i Japonii (Mohler i Timothy, 2010: 17). Badanie zostało zrealizowane w ramach amerykańskiego programu badawczego dotyczącego skutków bombardowań (*the United States Strategic Bombing Survey*) w sekcji odpowiedzialnej za uwarunkowania społeczne nalotów (Morale Division). Przedsięwzięcie to wymagało technik wykraczających poza krajowe doświadczenia. Po raz pierwszy projekt badawczy od początku (od samego pomysłu) poprzez konstrukcję wywiadów, realizację i prezentacje wyników skupiał się na międzynarodowych porównaniach. Zadbano o to, żeby pytania zadane w trzech krajach miały podobne znaczenie, co uzyskano głównie przez analizy lingwistyczne.

Jak zwykle w przypadku pionierskich badań standardy przyjęte w *Strategic Bombing Survey* odbiegają znacząco od dzisiejszych, jednak stanowiły one pierwszy krok na drodze rozwoju metodologii, pozwalającej uzyskiwać coraz bardziej trafne wyniki.

Jako pierwsze z nurtu nowoczesnych badań porównawczych wymienia się badanie *The civic Culture: Politica; Attitudes an Democracy In Five Nations* (Almond i Verba, 1963). Mierzono w nim poziom politycznego zaangażowania w Stanach Zjednoczonych, Meksyku, Wielkiej Brytanii, Niemczech i we Włoszech. Zbieranie danych zostało zaprojektowane w ten sposób, by odpowiedzi można było zamknąć we wcześniej ustalonych porządkowych i przedziałowych skalach. Starano się zagwarantować porównywalność mierzonych konstruktów. Dobrano duże jak na tamte czasy, 1000-osobowe ekwiwalentne losowe próby respondentów (z pewnymi odstępstwami w przypadku Meksyku). Jakkolwiek krytykowane za niektóre aspekty metodologiczne (Kavanagh, 1980) badanie wyznaczyło standardy projektów porównawczych na kolejne lata i stało się wzorem dla kolejnych projektów, z których można wymienić choćby kilka największych: *International Social Survey Programme* (ISSP), *World Mental Health Initiative Survey* (WMH) *World Fertility Survey* (WFS). *European Social Survey* (ESS), *Trends in International Mathematics and Science Study* (TIMSS), *Progress in International Reading Literacy Study* (PIRLS), *Programme for International Student Assessment* (PISA), *World Value Survey* (WVS), *Programme for the International Assessment of Adult Competencies* (PIAAC).

To, co cechuje badania zaprojektowane dla porównań, to między innymi międzynarodowo uzgadniane schematy badań, wieloetapowe pilotaże, konstrukcja kwestionariuszy uzyskiwana w drodze konsensusu, wieloetapowy proces tłumaczeń, określone standardy co do minimalnej stopy realizacji. Cały proces, obejmujący standaryzację narzędzi, określenie populacji docelowej oraz sposobu losowania i ważenia próby, wreszcie zbieranie danych, podporządkowany jest zapewnieniu możliwości dokonywania szczegółowych porównań różnych krajów czy społeczeństw. Zebranie danych w badaniach zaprojektowanych do porównań to jednak dopiero początek drogi badawczej. Wyniki uzyskane nawet z najlepiej zaprojektowanych badań wymagają weryfikacji i potwierdzenia tego, czy pomiar był trafny, a co za tym idzie – porównywalny.

Problem pomiaru i problem porównywalności

Współczesne, rozszerzone rozumienie trafności pomiaru (Messick, 1989, 1996a, 1996b) definiuje ją w kategoriach zdolności formułowania, na podstawie wyników pomiaru, wniosków dotyczących konkretnie sformułowanych funkcji pomiaru. Trafność pomiaru w tym ujęciu to charakterystyka, która odnosi się do relacji między narzędziem pomiarowym a celami, jakie są przed nim stawiane. W kontekście badań porównawczych głównym celem badania jest dokonanie wiążących porównań między konstruktem mierzonym w różnych kontekstach społecznych, a trafny pomiar zapewnia osiągnięcie tego celu.

Jakość pomiaru, a co za tym idzie porównań między grupami, zależy od trafności wskaźników wykorzystanych w pomiarze, a dokładniej od ich ekwiwalentności w różnych kontekstach społecznych. Ekwiwalentność wskaźników nie odnosi się przy tym do dosłownego znaczenia leksykalnego, ale definiowana jest przez równoważność funkcji, jaką pełnią one w pomiarze. Wskaźnik, jak pisze Scheuch (1968), nie powinien być definiowany przez swoje leksykalne znaczenie, ale poprzez probabilistyczną relację, która wiąże go z mierzonym konstruktem. Posiadanie kawałków kolorowego papieru, posiadanie kolorowych kamyków, posiadanie okrągłych kawałków metalu to trzy wskaźniki, które mają różne leksykalne znaczenia, lecz w odpowiednim kontekście mogą być traktowane jako ekwiwalentne wyznaczniki bogactwa. Wskaźniki są wtórne wobec funkcji, jaką pełnią, a przez to są względem nich wymienne (por. Scheuch, 1968).

Leksykalna tożsamość wskaźników jest więc sytuacją pożądaną, lecz nie jest warunkiem koniecznym uzyskania ekwiwalentnego pomiaru. Na przykład w wielu językach nie ma dobrego odpowiednika dla angielskiego wyrażenia *fair play*, a w języku japońskim nie ma jednego słowa, które definiowałoby *męża*. W tych wypadkach trzeba każdorazowo szukać wskaźników, które będą funkcjonalnie, a nie leksykalnie ekwiwalentne. Przy tłumaczeniu skali dystansu społecznego Bogardusa na niemiecki odkryto, że termin: sąsiedztwo, *neighbourhood* w języku angielskim, odnosi się w dużej mierze do charakteru przestrzennego i oznacza raczej miejsce niż dystans społeczny. Niemiecki leksykalny odpowiednik tego słowa, *Nachbarschaft*, posiada silne konotacje społeczne, dlatego autorzy adaptacji zdecydowali się na opisowe ujęcie tego terminu. Na Bliskim Wschodzie skala Bogardusa okazywała się zbyt „krótka" i nie oddawała całego przedziału wrogości. Postanowiono tam dodać nowe

pytanie: *Czy pragnąłbyś, aby ktoś pozabijał wszystkich tych ludzi?* Tylko to pytanie spełniało funkcje pomiaru dużego dystansu (Scheuch, 1968).

Problem porównywalności jest złożony i wymaga nie tylko wnikliwej analizy treści, ale również analizy danych. W kolejnych częściach artykułu przedstawiona zostanie klasyczne podejście statystyczne, służące do badania porównywalności wskaźników.

Badanie porównywalności

Podstawową i jedną z pierwszych metod opracowanych dla porównań ekwiwalentności międzygrupowej jest wielogrupowa konfirmacyjna analiza czynnikowa (Joreskog, 1971). Jest ona oparta na modelu analizy czynnikowej, której jednowymiarowy wariant opisany jest przez równanie (1). Analiza wielogrupowa polega na jednoczesnej estymacji modelu czynnikowego dla kilku bądź kilkunastu grup. Model ten można zapisać jako:

$$y_i = \tau_i + \lambda_i \theta + e_i \quad (1)$$

gdzie y_i oznacza wartość i-tego wskaźnika dla respondenta (dla przejrzystości pomijany jest tutaj indeks identyfikujący respondenta; w pełnym zapisie: y_{ij}, czyli wskaźnik *i*, dla respondenta *j*). Wartość wskaźnika opisana jest przez równanie liniowe, analogiczne do klasycznego modelu regresji, w którym τ_i jest stałą czynnikową (analogiczną do wyrazu wolnego w równaniu regresji), a λ_i oznacza ładunek czynnikowy (analogiczny do współczynnika kierunkowego w regresji). Wartość λ_i informuje o tym, o ile zmieni się przewidywana wartość y_i, gdy wartość zmiennej zależnej wzrośnie o jedną jednostkę (w tym wypadku zmienną zależną jest nieobserwowalny konstrukt θ). e_i to efekt losowy, inaczej wyraz błędu wyrażający probabilistyczny charakter modelu.

Podstawowy model wielogrupowy estymuje osobny zestaw parametrów dla każdej z grup (estymacja jest łączna dla wszystkich grup, ale parametry osobne). Możliwe jest jednak nałożenie pewnych ograniczeń na wartości parametrów estymowanego modelu. Główne ograniczenia to (gdzie *g* oznacza indeks grupy):

1. $\tau_{ig} = \tau_i$
2. $\lambda_{ig} = \lambda_i$
3. var (e_{ig}) = var (e_i)

Pierwsze ograniczenie pozwala na szacowanie jednej, takiej samej dla wszystkich grup, wartości stałej czynnikowej i-tego wskaźnika. Drugie i trzecie ograniczenie są analogiczne do pierwszego z tym, że odnoszą się do ładunku czynnikowego i wariancji swoistej (efektu losowego) wskaźnika.

W ramach liniowego modelu czynnikowego relacje między mierzonymi konstruktami i wskaźnikami mogą różnić się w odmiennych kontekstach społecznych pod względem:

1. struktury czynnikowej,
2. ładunków czynnikowych,
3. stałych czynnikowych,
4. efektów losowych (wariancji swoistej wskaźników).

Wymienione różnice, odnoszące się do parametrów modelu statystycznego, mają praktyczne implikacje. Spełnienie założeń modelu pomiarowego co do zgodności różnych zestawów parametrów pozwala na przeprowadzanie trafnych porównań. Dla niektórych porównań spełnienie wszystkich założeń jest konieczne, dla innych nie wszystkie parametry muszą być zgodne. W niektórych analizach wymaga się, aby parametry dla wszystkich wskaźników były ekwiwalentne między grupami, w innych analizach części wskaźników pozwala się nie spełniać tych założeń. Poziom porównywalności jest stopniowalny, rodzaj zgodności określa nam, jakie porównania są uprawomocnione, a jakie będą obarczone błędami.

Wstępnym krokiem do rozpoczęcia jakichkolwiek porównań jest ustalenie, czy w dwóch społeczeństwach istnieją dwa porównywalne konstrukty (*construct invariance*). Jest to najbardziej podstawowy rodzaj zgodności, który jest pierwszym warunkiem porównywalności. Kolejnym stopniem zgodności jest zgodność co do struktury (*configural invariance*), empirycznie określana za pomocą eksploracyjnej i konfirmacyjnej analizy czynnikowej. Kolejne rodzaje zgodności korespondują z konfirmacyjnym modelem czynnikowym i relacjami opisanymi wcześniej.

Sytuacja, w której w badanych grupach mamy do czynienia z takimi samymi wartościami ładunków czynnikowych (wraz ze zgodnością co do struktury) określana jest jako „słaba zgodność" – *weak invariance* lub terminem „zgodność co do metryki" (*metric invariance*). Oznacza ona, że powstała na podstawie pomiaru skala będzie miała w różnych krajach jednakowe jednostki pomiaru, ale różne osadzenie skali, tak jak w przypadku temperatury mierzonej stopniami Celsjusza i Kelwina. Taki rodzaj zgodności skal pozwala na nieobarczone błędami porównania międzygrupowe współzmienności między wskaźnikami (np. porównanie współczynników korelacji między dwoma konstruktami w różnych krajach). Nie pozwala jednak na porównywanie średnich z różnych krajów, gdyż nie gwarantuje takiego samego osadzenia skal w każdym z porównywanych społeczeństw.

Jeżeli mamy do czynienia ze słabą zgodnością, a dodatkowo w badanych grupach wartości stałych czynnikowych pozostają takie same, sytuacja zmienia się na „silną zgodność" (*strong invariance*), nazywaną też „zgodnością skalarną" (*scalar invariance*). W tym wypadku uzasadnione będzie również porównywanie średnich wskaźników między badanymi krajami. Jeżeli dodamy do tego spełnienie warunku dotyczącego błędów losowych (takiej samej wariancji swoistej), będziemy mieć do czynienia z „pełną zgodnością" (*strict invariance*, nazywaną niekiedy *full invariance*). W tej sytuacji możliwe jest dokonywanie wszelkiego typu porównań.

Brak pełnej zgodności oznacza, iż rzetelność skal w różnych grupach jest różna. Nie będzie to miało znaczenia przy punktowym oszacowaniu średniej, lecz może prowadzić do błędów w oszacowaniach błędów standardowych średniej. W niektórych analizach, na przykład korelacji, niska rzetelność zmiennych wpływa na niedoszacowanie parametrów (*attenuation bias*): im niższa rzetelność, tym niedoszacowanie większe. Różnice w rzetelności mogą zatem wpływać na trafność porównań. Niemniej jednak istnieje wiele metod statystycznych, pozwalających

w takiej sytuacji uzyskać nieobarczone oszacowania parametrów, których wartości chcemy porównywać. Istnieją znane analityczne korekty rzetelności, np. jedną z najprostszych korekt dla korelacji jest podzielenie jej przez wartość estymowanej rzetelności skali (np. wartość Alfy Cronbacha). Istnieją też analogiczne korekty dla wielu modeli regresyjnych (Maddala, 1983). Można również stosować bardziej wyrafinowane i precyzyjniejsze metody: modele strukturalne (Kaplan, 2009) czy wykorzystanie metodologii plausible values (Wu, 2005) znosi problemy związane z różnicami rzetelności.

Statystyczna identyczność zarówno ładunków czynnikowych, jak i stałych czynnikowych dla wszystkich pytań we wszystkich porównywanych grupach jest rzadkością, a w porównaniach międzynarodowych jest praktycznie niespotykanym typem idealnym. Zgodność poszczególnych elementów modelu pomiarowego traktuje się raczej jako cechę stopniowalną, a nie zerojedynkową. Stąd też wprowadzenie terminu „częściowej porównywalności", który przeciwstawia się idealnej sytuacji „pełnej porównywalności". Byrne wskazuje, iż w całym modelu pomiarowym wystarczy jedno pytanie zachowujące zgodność między grupami, aby porównania międzygrupowe były uzasadnione (Byrne i inni, 1989). Według Widamana co najmniej połowa pytań w jednej grupie powinna zachować zgodność (Widaman i inni, 1993). Oczywiście im więcej porównywalnych pytań, tym porównania stają się pewniejsze; im mniej ekwiwalentnych pytań w modelu pomiarowym, tym wnioski wyciągane z takich analiz powinny być bardziej ostrożne.

Jak łatwo zauważyć, w modelu analizy czynnikowej poprzez dobranie odpowiednich ograniczeń można estymować modele, które będą odpowiadały podstawowym rodzajom zgodności. Nałożenie 1. ograniczenia dla wszystkich wskaźników prowadzi do estymacji modelu przy założeniu słabej zgodności, ograniczeń 1. i 2. prowadzi do estymacji modelu przy założeniu silnej zgodności wskaźników, a narzucenie wszystkich trzech ograniczeń równocześnie skutkuje estymacją modelu przy założeniu pełnej zgodności. Ocena zgodności wskaźników polega na estymacji modeli nakładających kolejne ograniczenia na wartości parametrów i porównywaniu wartości indeksów dopasowań tych modeli.

Motywacja do nauki przedmiotów przyrodniczych

W badaniu PISA 2006 mierzono kilka różnych typów motywacji do nauki przedmiotów przyrodniczych. W tym tekście skupimy się na instrumentalnej motywacji do nauki przedmiotów przyrodniczych (INSTSCIE). Wskaźnik skonstruowany został na podstawie pięciu pytań zadanych uczniom. Uczniowie mieli ustosunkować się do następujących twierdzeń:

1. Warto włożyć wysiłek w naukę biologii, chemii lub fizyki, bo to mi pomoże w pracy, którą chcę wykonywać w przyszłości.
2. To, czego się uczę na biologii, chemii lub fizyce, jest dla mnie ważne, ponieważ będzie mi potrzebne w dalszej nauce.
3. Uczę się biologii, chemii i fizyki, ponieważ wiem, że są dla mnie użyteczne.
4. Warto się uczyć biologii, chemii i fizyki, bo to, czego się nauczę, zwiększy w przyszłości moje szanse zawodowe.
5. Na biologii, chemii i fizyce nauczę się wielu rzeczy, które pomogą mi dostać pracę.

Wykorzystując cztery możliwe odpowiedzi:

1. *Zdecydowanie się zgadzam.*
2. *Zgadzam się.*
3. *Nie zgadzam się.*
4. *Zdecydowanie się nie zgadzam.*

Szczegóły metodologiczne konstrukcji tego wskaźnika można znaleźć w raporcie technicznym badania PISA 2006 (OECD 2009). Analizy przeprowadzone przez ekspertów OECD zapewniają, iż w przypadku tego wskaźnika w badanych społeczeństwach istnieją porównywalne konstrukty (*construct invariance*) oraz iż istnieje zgodność co do struktury (*configural invariance*), empirycznie określana za pomocą eksploracyjnej i konfirmacyjnej analizy czynnikowej. Nie przeprowadzono natomiast analiz, które pozwoliłyby orzec o porównywalności wskaźników. Zakłada się z góry, iż są to wskaźniki porównywalne. Analizy, dzięki którym można orzec, iż takie założenie jest prawdziwe, przedstawione zostały w tym artykule.

Zastosowana została tutaj wielogrupowa analiza czynnikowa dla zmiennych porządkowych. Różni się ona od prezentowanego wcześniej modelu przede wszystkim tym, iż nie specyfikuje się dla niej efektu losowego oraz jedno pytanie może mieć kilka stałych czynnikowych (szczegółowy opis można znaleźć między innymi w: Skrondal i Rabe-Hesketh, 2004). W prezentowanym modelu nałożone zostały restrykcje, aby w każdej grupie ładunki czynnikowe oraz stałe czynnikowe dla wszystkich pytań były sobie równe. Do oceny porównywalności poszczególnych pytań użyto indeksu modyfikacji (*Modification Index*) opartego na teście chi-kwadrat. W kontekście analizy wielogrupowej indeks ten mówi, o ile zmieni się statystyka dopasowania chi-kwadrat dla całego modelu po zniesieniu ograniczenia dla danego wskaźnika w danej grupie. Wartości indeksu modyfikacji przekraczające 100 będą oznaczały, iż odpowiedzi uczniów w danym kraju nie są porównywalne z odpowiedziami uczniów w innych krajach. Przyjęcie wartości 100 dla wartości indeksu modyfikacji nie jest posunięciem arbitralnym, przyjęcie takiej wartości dla licznych prób prowadzi do wyników, które będą porównywalne z innymi statystycznymi metodami wykorzystywanymi przy orzekaniu o porównywalności (por. Pokropek, 2012).

W tabeli 1 przedstawiono wartości wskaźników modyfikacji dla każdego pytania i każdego kraju biorącego udział w badaniu PISA 2006. Każde pytanie charakteryzowane jest przez kilka wskaźników modyfikacji (dla stałych czynnikowych i ładunku czynnikowego) w tabeli przedstawiono jedną liczbę charakteryzującą maksymalną wartość wskaźnika modyfikacji dla pytania. Wartości przekraczające 100 i wskazujące na problem porównywalności zaznaczone zostały pogrubionym tekstem.

Tabela 1. Wartości indeksów modyfikacji dla wielogrupowej analizy czynnikowej dla danych PISA 2006 i pytań tworzących wskaźnik instrumentalnej motywacji do nauki przedmiotów przyrodniczych

Kraj	Pytanie					Liczba porównywalnych pytań	Liczebność próby
	1	2	3	4	5		
Argentyna	1 268	109	0	81	569	2	4 339
Australia	1 131	839	4	2	806	2	14 170
Austria	145	92	229	983	57	2	4 927
Azerbejdżan	54	7	11	1	19	5	5 184
Belgia	789	26	162	186	33	2	8 857
Bułgaria	184	6	29	31	0	4	4 498
Brazylia	89	10	25	12	0	5	9 295
Kanada	308	481	83	115	831	1	22 646
Szwajcaria	710	36	402	48	4	3	12 192
Chile	164	81	7	11	13	4	5 233
Kolumbia	38	7	2	60	4	5	4 478
Czechy	237	2	26	1	10	4	5 932
Niemcy	76	197	325	11	4	3	4 891
Dania	74	12	0	0	8	5	4 532
Hiszpania	255	150	18	0	141	2	19 604
Estonia	29	5	77	6	50	5	4 865
Finlandia	81	8	13	0	27	5	4 714
Francja	54	56	117	323	96	3	4 716
Wielka Brytania	224	1831	70	63	153	2	13,152
Grecja	130	73	41	20	4	4	4,873
Hongkong, Chiny	871	0	388	313	0	2	4,645
Chorwacja	318	66	8	6	144	3	5,213
Węgry	231	2	19	357	39	3	4,490
Indonezja	59	42	1	4	5	5	10,647
Irlandia	71	814	1	61	92	4	4,585
Islandia	25	135	1	3	2	4	3,789
Izrael	779	418	61	530	542	1	4,584
Włochy	632	42	341	540	40	2	21,773
Jordania	373	55	3	2	5	4	6,509
Japonia	94	598	72	154	161	2	5,952
Kirgistan	252	20	84	35	22	4	5,904
Korea	49	12	90	13	38	5	5,176
Liechtenstein	12	2	0	0	0	5	339
Litwa	56	16	73	186	22	4	4,744
Luksemburg	56	25	49	26	32	5	4,567
Łotwa	100	134	106	245	0	2	4,719

Macao-Chiny	349	130	40	8	114	2	4,760
Meksyk	104	3	7	10	5	4	30,971
Czarnogóra	89	5	5	28	15	5	4,455
Niderlandy	128	0	1	163	19	3	4,871
Norwegia	68	5	0	4	113	4	4,692
Nowa Zelandia	35	99	46	156	1	4	4,823
Polska	178	137	11	122	19	2	5,547
Portugalia	116	0	18	27	350	3	5,109
Katar	481	293	41	177	0	2	6,265
Rumunia	31	7	6	4	4	5	5,118
Rosja	65	111	75	100	20	4	5,799
Serbia	40	8	204	6	4	4	4,798
Republika Słowacka	58	173	0	39	19	4	4,731
Słowenia	103	57	1	112	102	2	6,595
Szwecja	178	196	16	35	0	3	4,443
Tajwan, Chiny	190	29	5	5	12	4	8,815
Tajlandia	371	12	3	0	9	4	6,192
Tunezja	363	5	1	43	13	4	4,640
Turcja	153	0	125	19	3	3	4,942
Urugwaj	36	14	5	34	3	5	4,839
Stany Zjednoczone	92	8	0	0	3	5	5,611

Jak pokazują wyniki przedstawione w tabeli 1, problem porównywalności jest niebagatelny. Pełną porównywalnością wskaźników charakteryzuje się tylko 14 krajów z puli 57. Cztery porównywalne wskaźniki można znaleźć w 18 krajach. Pozostałe kraje charakteryzują się mniejszą liczbą porównywalnych wskaźników (szczegóły w tabeli 2).

Tabela 2. Rozkład liczby porównywalnych wskaźników w grupie badanych krajów

Liczba porównywalnych pytań	Liczba krajów	Procent krajów
1	2	3,51
2	14	24,56
3	9	15,79
4	18	31,58
5	14	24,56

Na przedstawione wyniki można patrzeć dwojako. Po pierwsze pokazują one, iż wskaźnik motywacji do nauki konstruowany przez ekspertów prowadzących badanie PISA 2006 nie jest wskaźnikiem porównywalnym wśród wszystkich krajów. W badaniu PISA 2006 (ale także w późniejszych edycjach) nie zastosowano żadnych korekt, które umożliwiałyby skonstruowanie porównywalnego wskaźnika. Jakość wniosków porównawczych dla analiz przeprowadzanych z wykorzystaniem tego wskaźnika przez różnych badaczy może zostać poddana w wątpliwość. Drugi sposób patrzenia na wyniki jest bardziej optymistyczny. Ponad 96% krajów z badanej próby ma przynajmniej dwa porównywalne

wskaźniki, prawie 72% ma aż trzy porównywalne wskaźniki. Otwiera to możliwość zastosowania modelu zgodności częściowej do konstrukcji porównywalnych wskaźników. Innymi słowy, wskaźniki dystrybuowane przez OECD nie są porównywalne (w tym wypadku), ale badacze, podejmując pewien intelektualny i techniczny wysiłek, są w stanie stworzyć lepsze wskaźniki.

Podsumowanie i dyskusja

Porównywalność badanych wskaźników w każdym badaniu porównawczym jest kwestią fundamentalną i trudną do osiągnięcia. Nigdy nie należy jej brać jako pewnik, a zawsze trzeba ją traktować jako kwestię otwartą na weryfikację empiryczną. Jak zostało pokazane nawet w dużych i renomowanych badaniach porównywalności nie jest oczywistością. W artykule tym pokazana została analiza odnosząca się do jednej skali, podobne analizy mogą zostać przeprowadzone dla innych skal budowanych w tym badaniu, jak również w innych badaniach edukacyjnych skonstruowanych w celach porównawczych, takich jak PIRLS, TIMSS czy PIAAC. Twórcy tych badań zadbali o pełną porównywalność skal mierzących umiejętności, lecz taka sytuacja nie odnosi się do innych skal tworzonych w tych badaniach. Znaczenie poszczególnych pytań w niektórych krajach mimo szczegółowych kontroli lingwistycznych okazuje się różne i wymaga kontroli statystycznej.

Bibliografia

1. Durkheim, E (2000). *Zasady metody Socjologicznej*, PWN, Warszawa, *Les Règles de la méthode sociologique 1895.*
2. Joreskog, K. G. (1971). *Simultaneous factor analysis in several populations.* Psychometrika, 36, 409-426.
3. Kaplan, D. (2008). S*tructural equation modeling: Foundations and extensions.* Sage Publications, Inc.
4. Maddala G. S (1989) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press Cambridge.
5. Messick, S. (1989). *Validity. In R.L. Linn* (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.
6. Messick, S. (1996a). *Standards-based score interpretation: Establishing valid grounds for valid inferences. Proceedings of the joint conference on standard setting for large scale assessments,* Sponsored by National Assessment Governing Board and The National Center for Education Statistics. Washington, DC: Government Printing Office.
7. Messick, S. (1996b). *Validity of Performance Assessment.* In Philips, G. (1996). *Technical Issues in Large-Scale Performance Assessment.* Washington, DC: National Center for Educational Statistics.
8. OECD (2009). PISA 2006 *Technical Report*, PISA. OECD Publishing.
9. Pokropek A. (2012) *Porównania międzynarodowe* [w:] H. Domański (red.), *Metodologia Badań nad Stratyfikacją Społeczną*, Wydawnictwo Naukowe Scholar, Warszawa.

10. Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.
11. Wu, A.D, Z. Li, B.D. Zumbo *Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data Practical Assessment, Research & Evaluation*. Volume 12, Number 3, February 2007.