dr hab. Artur Pokropek, prof. IFiS PAN
Polska Akademia Nauk
Polskie Towarzystwo Diagnostyki Edukacyjnej

Badania edukacyjne, porównywalność, niezmienność pomiarowa i metody jej testowania

Abstrakt

Porównawcze badania edukacyjne od wielu lat są cenionym źródłem wiedzy dla badaczy, polityków, nauczycieli, a w ostatnich latach także dla rodziców i opinii publicznej. Porównania nastręczają jednak pewnych problemów. Nie każde porównanie musi być uprawnione i nie każdy wniosek płynący z porównań musi być trafny. W ostatnich latach pierwszoplanową rolę w rozwoju metodologicznym badań porównawczych odgrywa analiza niezmienności pomiarowej ilościowych narzędzi wykorzystywanych w badaniach. W tym artykule omówiona została relacja między porównywalnością a niezmiennością pomiarową. Przedstawione zostały klasyczne i nowoczesne metody testowania niezmienności pomiarowej oraz omówione pojęcia częściowej i przybliżonej niezmienności. Metodologiczne i statystyczne rozważania dopełnione są przykładem empirycznym, w którym analizowania jest niezmienność pomiarowa skali „zainteresowania nauką" z badania PISA 2006 i 2015. Niezmienność pomiarowa jest tutaj analizowana nie tylko w kontekście międzynarodowym, lecz także w kontekście niezmienności pomiarowej w czasie.

Dlaczego potrzebne są nam badania porównawcze?

Badania edukacyjne są szczególnie trudne. Potencjalne wyniki kształtujące efektywność edukacyjną są zazwyczaj niewielkie, a ich wpływ rozkłada się zazwyczaj na lata. Elementów mogących wpływać na proces kształcenia jest bardzo wiele i najprawdopodobniej tworzą ze sobą sieć skomplikowanych interakcji. Dodatkowo badania szkół, uczniów i nauczycieli nie są etycznie neutralne, a środowisko badania jest trudne do kontrolowania. Badania eksperymentalne, złoty standard badawczy, są w tym kontekście bardzo skomplikowane, a potencjalna generalizacja wniosków może być wątpliwa. Poprawnie przeprowadzone badania eksperymentalne powinny być długotrwałe, a co za tym idzie są kosztowne i trudne w realizacji. Prawidłowa kontrola potencjalnych czynników zakłócających w zróżnicowanym środowisku szkolnym i rodzinnym jest niemałym wyzwaniem, a implementacja czynników eksperymentalnych w heterogenicznych środowiskach szkolnych jest bardzo złożona.

Popularną alternatywą dla badań eksperymentalnych w edukacji są porównawcze badania międzynarodowe. Porównania międzynarodowe, mimo iż nie są wolne od licznych metodologicznych pułapek, mogą zapewnić niezbędny dystans i kontekst, dostarczają istotnych informacji na temat badanego systemu edukacyjnego. Wyniki uzyskane na podstawie badań

porównawczych nigdy nie będą tak precyzyjne i jednoznaczne jak badania eksperymentalne, jednak poprawnie przeprowadzone porównania mogą znacznie pogłębić zrozumienie naszej własnej edukacji i mieć istotne znaczenie dla decydentów, nauczycieli oraz tych wszystkich, którzy kształcić chcą lepiej (Noah, 1986). Gdy porównania międzynarodowe wzbogacone zostają dodatkowo o wymiar czasowy – czyli gdy mamy możliwość porównań nie tylko między krajami, lecz także zmian między różnymi punktami czasu – wartość informacyjna takich badań staje się naprawdę wysoka.

Problemy z porównaniami

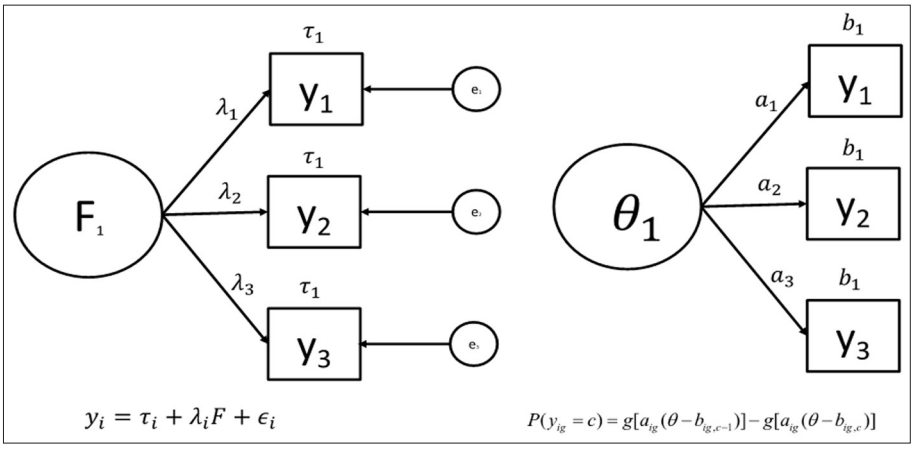
Międzynarodowe badania porównawcze to przede wszystkim ilościowe badania kwestionariuszowe, dopełnione niekiedy testem umiejętności lub wiadomości. Zbieranie danych odbywa się za pomocą papierowego lub elektronicznego testu i/lub papierowego bądź elektronicznego kwestionariusza. W badaniach tych do analizy wykorzystuje się zazwyczaj konstrukty mierzone przez kilka wskaźników (pytań lub zadań). Tak samo jak wiedzę matematyczną ucznia zazwyczaj mierzymy za pomocą przynajmniej kilku zadań, tak i inne cechy ucznia (motywację, wartości, zaangażowanie, zainteresowania) mierzymy za pomocą kilku pytań po to, aby uzyskać bardziej rzetelne informacje. Za pomocą różnych modeli skalowania wskaźniki wykorzystywane są do budowania jednej skali, która ma odzwierciedlać intensywność badanego zjawiska. Badane zjawisko można następnie porównywać między różnymi grupami. Najczęściej z krajami. Czasami porównania mogą być jednak problematyczne.

Dobrym przykładem jest skala zasobów gospodarstwa domowego (*Home Possesion Index*) wykorzystywana w badaniu PISA. Skala ta budowana jest na podstawie kilkunastu pytań odnoszących się do posiadania różnego rodzaju dóbr, między innymi telefonów komórkowych i samochodów (OECD, 2012). Wykorzystywanie takich pytań do określenia pozycji ekonomicznej w kontekście międzynarodowym wydaje się jednak problematyczne. Podczas gdy w Stanach Zjednoczonych posiadanie samochodu jest oczywiste dla wszystkich (odległości między miejscowościami są ogromne, a koszty utrzymania samochodu niskie), w Japonii posiadanie samochodu jest mniej powszechne nawet w dosyć zamożnych rodzinach (odległości między miejscowościami są niewielkie, transport publiczny jest dobrze rozwinięty i można na nim polegać, a koszty utrzymania samochodu są wysokie). Podobny problem odnosi się do porównań między badaniami z różnych punktów czasowych. O ile w roku 2000 posiadanie telefonu komórkowego w gospodarstwie domowym wydawało się dobrym wskaźnikiem zamożności, to w roku 2012 jest to wątpliwe. Mimo opisanych wyżej problemów w modelu skalowania w badaniu PISA zakłada się pełną porównywalność na przestrzeni czasu (informacja o liczbie telefonów w 2000 roku i w roku 2012 ma mieć takie samo znaczenie dla statusu społeczno-ekonomicznego) oraz pełną porównywalność międzynarodową (liczba samochodów w gospodarstwie domowym w Japonii ma być takim samym wyznacznikiem statusu społeczno-ekonomicznego jak w Stanach Zjednoczonych) (Pokropek, Borgonovi i McCormick, 2017).

Inny przykład pochodzi z Europejskiego Sondażu Społecznego (European Social Survey, ESS), gdzie skonstruowano skalę mierzącą poziom religijności. Nie jest to przykład edukacyjny, ale bardzo pouczający. Skala składa się z kilku pytań dotyczących ogólnej religijności i praktyk religijnych. Turcja okazała się krajem o najniższej średniej wartości tego wskaźnika. Szczególnie niski wskaźnik obserwowany był wśród kobiet. Jedno z pytań dotyczyło częstości odwiedzania kościoła/meczetu. Z przyczyn kulturowych w Turcji kobiety nie chodzą regularnie do meczetów. Ten wskaźnik religijności okazał się nieporównywalny między Turcją i innymi krajami i przyczynił się do znacznych błędów w ocenie średniej religijności w Turcji (Billiet, 2013). Inny przykład z badania ESS dotyczy imigrantów. W badaniu zadano serię pytań o stosunek do imigrantów. Danii w 2002 roku odnotowano szczególnie niski wskaźnik „niechęci" do imigrantów. W kolejnych latach wskaźnik ten był jednak wysoki. Problem, jak się okazało, tkwił w tłumaczeniu. Sformułowanie, które powinno odnosić się do *zbrodni*, przetłumaczono jako *obrazę*. Podobnie jak w przykładzie z religijnością błąd ten doprowadził do nietrafnych porównań.

Mimo iż badania porównawcze mogą poszczycić się długą tradycją, to do tej pory problemy konstrukcji międzynarodowo porównywalnych wskaźników stanowią poważną przeszkodę stojącą na drodze do trafności dociekań substantywnych (van de Vijver, 2011). Na szczęście badaniom porównawczym z pomocą przyszły różnorodne metody statystyczne. Szczególnie metody powiązane z modelowaniem cech ukrytych.

Do szacowania intensywności występowania badanej cechy u uczniów lub wyników uzyskanych przez uczniów na testach wykorzystywane są zazwyczaj modele cech ukrytych, w których obserwowalne wskaźniki łączone są za pomocą metod statystycznych z nieobserwowalną bezpośrednio mierzoną cechą. Gdy w analizach przyjmujemy, że wskaźniki mają charakter ciągły – najczęściej wykorzystywane są modele konfirmacyjnej analizy czynnikowej (Confirmatory Factor Analysis, CFA).



Rysunek 1. Model konfirmacyjnej analizy czynnikowej (lewa strona) oraz model IRT (prawa strona)

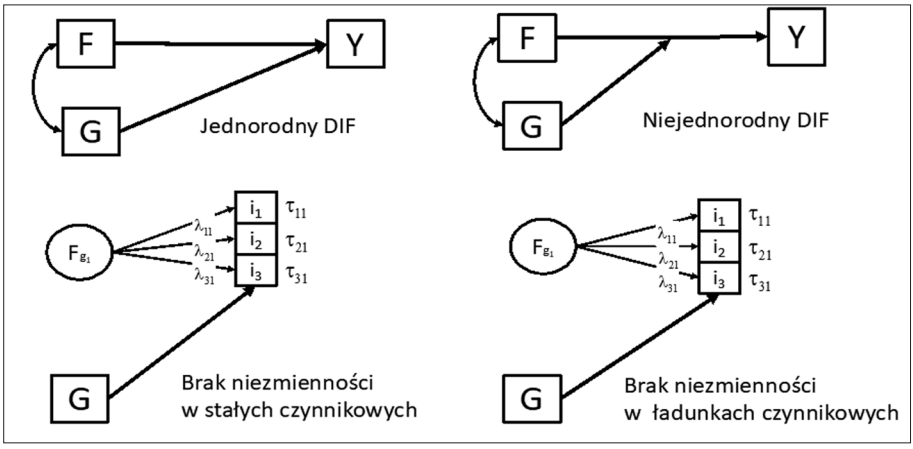
Gdy wskaźniki mają charakter kategorialny – stosowane są modele IRT (obydwa modele w języku polskim opisane szczegółowo w: Pokropek, 2015). Modele w graficzny sposób przedstawione zostały na rysunku 1.

Problem testowania niezmienności pomiarowej sprowadza się w zasadzie do testowania, czy poszczególne parametry modelu pomiarowego (ładunki czynnikowe, stałe czynnikowe w modelach CFA lub parametry dyskryminacji *a* i trudności *b* w modelach IRT) w różnych badanych grupach (krajach, punktach czasowych) są takie same, czy różne.

Formalnie niezmienność pomiarowa oznacza spełnienie założenia *lokalnej niezależności* Mellenbergh (1989):

$$f(Y \mid F = f, G = g) = f(Y \mid F = f)$$

gdzie *F* to badana cecha, *Y* to wskaźnik, za pomocą którego mierzymy daną cechę, a *G* oznacza zmienną grupującą. W uproszczeniu, z niezmiennością pomiarową mamy do czynienia, gdy czynnik grupujący nie zakłóca pomiaru – ścieżki prowadzącej od wskaźnika do badanej cechy, jak pokazane zostało to na rysunku 2. Niezmienność pomiarowa jest bezpośrednio związana ze zjawiskiem zróżnicowanego funkcjonowania wskaźników/zadań (*Differential Item Functioning*, DIF) i do pewnego stopnia jest z nią tożsama. Jednorodny DIF oznacza brak niezmienności pomiarowej w stałych czynnikowych, a niejednorodny DIF brak niezmienności pomiarowej w ładunkach czynnikowych.



Rysunek 2. Zróżnicowane funkcjonowanie wskaźnika *(Differential Item Functioning,* DIF) i niezmienność pomiarowa w modelu CFA

Występowanie DIF nie przekreśla jednak uzyskania słabszych form niezmienności pomiarowych, które wciąż dają możliwość przeprowadzania poprawnych porównań. Najbardziej klasycznie podejście do niezmienności wyróżnia trzy (a w niektórych ujęciach cztery) formy niezmienności. Pierwsza forma niezmienności pomiarowej to *niezmienność konfiguracyjna* (*configural invariance*) i odnosi się do sytuacji, w której porównywane konstrukty mają taką samą strukturę czynnikową, czyli badana cecha ma tyle samo wymiarów.

Zaistnienie tego typu niezmienności nie pozwala na bezpośrednie porównania, ale stanowi pierwszy warunek konieczny do podjęcia porównań. Druga forma niezmienności pomiarowej to niezmienność metryczna (*metric invariance*). Warunkiem koniecznym do uzyskania takiej niezmienności jest równość ładunków czynnikowych dla danego wskaźnika we wszystkich analizowanych grupach. Zachowanie niezmienności metrycznej, z metodologicznego punktu widzenia, pozwala na uprawnione porównania związków między badanymi konstruktami (np. współczynników regresji liniowej), nie pozwala natomiast na bezpośrednie porównywanie średnich poziomów danej cechy między grupami (Davidov i in., 2014). Aby porównywać średnie poziomy badanych cech między grupami, niezbędna jest niezmienność skalarna (*scalar invariance*) – ze statystycznego punktu widzenia sytuacja, w której zarówno ładunki czynnikowe (lub parametry dyskryminacji), jak i stałe czynnikowe (lub parametry trudności) są takie same we wszystkich grupach.

Empirycznie poziomy klasycznej niezmienności testuje się, porównując zagnieżdżone modele pomiarowe. Jeśli modele metryczne lub skalarne nie są istotnie gorzej dopasowane do danych niż model konfiguracyjny, przyjmuje się, że kolejne stopnie niezmienności zostały uzyskane. Najczęściej porównania te odbywają się za pomocą miar dopasowania. Formalne testy statystyczne nie są optymalnym rozwiązaniem, gdyż w przypadku dużych zbiorów danych zbyt często wskazują na istotne statystycznie różnice, które z praktycznego punktu widzenia są bez większego znaczenia. W przypadku modeli dla zmiennych ciągłych zazwyczaj wykorzystywane są miary CFI, RMSEA i SRMR. Zbyt duże osłabienie dopasowania modelu w porównaniu z modelem konfiguracyjnym oznacza odrzucenie kolejnych typów niezmienności. W badaniach często powołuje się na badania symulacyjne Chena (2007), który wskazywał, że dobrymi progami dla tych miar są wartości: 0,01 dla CFI, 0,15 RMSEA i 0,15 dla SRMR w przypadku testowania niezmienności skalarnej i 0,030 w przypadku testowania niezmienności metrycznej.

Problem z klasycznym ujęciem polega na tym, że jest ono bardzo restrykcyjne i nie pozwala na jednostkowe odstępstwa dla niektórych wskaźników. Możliwa jest bowiem sytuacja, w której tylko jeden wskaźnik w jednej grupie będzie wyraźnie problematyczny, tak jak to pokazywały przykłady przywołane z ESS. Wtedy jeden problematyczny wskaźnik potrafi przekreślić całą skalę. Jest to wysoce nieefektywne. Podejście, które umożliwia budowę skal, nawet gdy część wskaźników nie jest porównywalnych, nazywane jest częściową niezmiennością (*partial equivalence*) (Byrne i in., 1989). W modelach częściowej niezmienności wymaga się, aby tylko część parametrów wskaźników była niezmienna. Nowe badania symulacyjne (Pokropek i in., 2019) wskazują, iż do zachowania częściowej niezmienności całej skali wystarczy nawet bardzo niewielka liczba niezmiennych wskaźników. W określonych sytuacjach nawet jeden niezmienny wskaźnik wystarczy na to, by zachować dobre właściwości skali. Problemem pozostaje oczywiście to, jak ustalić, które wskaźniki są niezmienne, a które nie. Literatura tu jest bardzo bogata, a zaproponowanych metod wiele. Problem w tym, że żadna z metod nie zapewnia 100% skuteczności. Wstępne badania symulacyjne wskazują jednak, iż iteracyjna metoda zaproponowana przez Asparouhova i Muthéna (2014) przy okazji rozwoju algorytmu *wyrównywania* (*allignement*) daje bardzo dobre wyniki w kontekście wielu grup. Tą metodą posługiwać się będziemy w przykładzie empirycznym.

W ostatnich latach rozwinięto również inne metody, które można wspólnie określić jako metody przybliżonej niezmienności (*approximate invariance*). Należą do nich: wielopoziomowa analiza czynnikowa (*Multilevel Confirmatory Factor Analysis*) (Fox, 2010; Hox, de Leeuw i Brinkhuis, 2010), Bayesowskie modele strukturalne (*Bayesian Structural Equating Modeling*) (De Jong, Steenkamp i Fox, 2007; Muthén i Asparouhov, 2012) oraz metody optymalizacji wyrównującej (*Alignment optimization method*) (Asparouhov i Muthén, 2014). Metody te łączy założenie, iż niewielkie różnice w parametrach między grupami, szczególnie w badaniach międzynarodowych, są raczej regułą niż wyjątkiem. Niewielkie różnice muszą występować, gdyż sam fakt różnic językowych oraz kulturowych nie pozwala na przetłumaczenie pytań w sposób jednoznaczny. Te niewielkie różnice można jednak modelować i uwzględniać w analizach tak, aby uzyskać zadowalający poziom niezmienności.

Różne typy modelowania niezmienności odpowiadają różnym stanom faktycznym. Zatem aby przeprowadzić pełną analizę niezmienności, należy testować je wszystkie i ustalić, który model najlepiej odwzorowuje rzeczywistość. Rozsądnie jest zacząć od koncepcji najbardziej restrykcyjnej, czyli pełnej zgodności. Jeśli pełna zgodność skalarna zostanie osiągnięta, na tym można poprzestać i przejść do substantywnych analiz. Gdy niezmienność skalarna nie zostanie potwierdzona, rozsądną strategią jest przejście do testowania częściowej niezmienności. Innymi słowy, należy rozpocząć poszukiwania tych wskaźników, które odpowiadać mogą za brak niezgodności. Po testowaniu częściowej niezgodności warto przejść do testowania przybliżonej niezmienności dla tych wszystkich parametrów, które we wcześniejszej analizie uznane zostały za niezmienne (czy też bez detekcji DIF). Jeśli wykryta zostanie przybliżona niezmienność, wtedy do czynienia mamy z częściową przybliżoną niezmiennością. Jeśli przybliżona niezmienność nie została wykryta, model częściowej niezmienności pozostaje modelem, za pomocą którego analizy porównawcze mogą być przeprowadzone. Przykład takiej analizy przedstawiony zostanie w dalszej części tekstu.

Zanim przejdziemy do tego przykładu, podkreślić należy, iż niezmienność pomiarowa nie gwarantuje porównywalności. Jest jedynie warunkiem koniecznym. Teoretycznie można wyobrazić sobie sytuację, w której parametry wskaźników są takie same między grupami, a badane konstrukty inne. Takie sytuacje mogą się zdarzyć wtedy, gdy porównujemy niewielką liczbę grup, a do pomiaru cechy ukrytej używamy niewielkiej liczby wskaźników. Statystyczna analiza nie zastąpi zdrowego rozsądku i substantywnej znajomości badanej problematyki.

Przykład: Zainteresowanie nauką w badaniach PISA 2006 i 2015

Do empirycznego przykładu wykorzystana zostanie skala zainteresowania nauką z badania PISA 2006 i 2015. Do analiz wykorzystane zostaną jedynie kraje OECD. Skala ta skonstruowana jest z pięciu wskaźników, które powstały na podstawie pięciu pytań: *Jak często robisz wskazane rzeczy? 1) Oglądasz programy naukowe w telewizji; 2) Pożyczasz lub kupujesz książki popularnonaukowe; 3) Odwiedzasz strony internetowe o nauce; 4) Czytasz artykuły naukowe w magazynach naukowych lub innych gazetach; 5) Uczęszczasz na spotkania klubu naukowego.* Kategorie odpowiedzi dla wymienionych pytań to: a*) Bardzo często [3]; b) Regularnie [2];*

c) Czasami [1]; d) Nigdy albo prawie nigdy [0]. Wskaźniki traktowane są tutaj jako zmienne ciągłe i zostały wystandaryzowane w taki sposób, aby średnia wskaźnika dla wszystkich krajów wynosiła zero, a odchylenie standardowe jeden.

W tabeli 1 umieszczone zostały wyniki niezmienności w czasie.

Tabela 1. Klasyczna niezmienność pomiarowa. Porównania między rokiem 2006 i 2015 dla każdego kraju OECD uczestniczącego w obydwu falach badania

Kraj	Niezmienność metryczna			Niezmienność skalarna			Czy niezmienność	
	$_{r}$CFI	$_{\Delta}$RMSE	$_{\Delta}$SRMR	$_{r}$CFI	$_{\Delta}$RMSE	$_{\Delta}$SRMR	metryczna?	skalarna?
AUS	0.028	-0.030	-0.059	0.050	-0.038	-0.066	Nie	Nie
AUT	0.036	-0.028	-0.062	0.087	-0.051	-0.077	Nie	Nie
BEL	0.031	-0.034	-0.059	0.062	-0.048	-0.068	Nie	Nie
CAN	0.027	-0.031	-0.062	0.056	-0.047	-0.072	Nie	Nie
CHE	0.015	-0.025	-0.037	0.051	-0.055	-0.047	Nie	Nie
CZE	0.014	-0.016	-0.036	0.052	-0.045	-0.051	Nie	Nie
DEU	0.008	-0.004	-0.024	0.055	-0.041	-0.043	Tak	Nie
DNK	0.022	-0.005	-0.043	0.066	-0.023	-0.054	Nie	Nie
ESP	0.006	-0.006	-0.018	0.040	-0.040	-0.031	Tak	Nie
FIN	0.048	-0.038	-0.075	0.108	-0.063	-0.088	Nie	Nie
FRA	0.030	-0.023	-0.055	0.119	-0.071	-0.078	Nie	Nie
GBR	0.015	-0.020	-0.039	0.030	-0.029	-0.046	Nie	Nie
GRC	0.001	0.010	-0.006	0.075	-0.069	-0.053	Tak	Nie
HUN	0.003	0.006	-0.012	0.055	-0.047	-0.043	Tak	Nie
IRL	0.004	0.000	-0.014	0.024	-0.019	-0.024	Tak	Nie
ISL	0.031	-0.012	-0.055	0.119	-0.053	-0.075	Nie	Nie
ITA	0.018	-0.015	-0.035	0.118	-0.077	-0.062	Nie	Nie
JPN	0.009	-0.007	-0.026	0.033	-0.030	-0.037	Tak	Nie
KOR	0.011	-0.022	-0.036	0.060	-0.076	-0.062	Nie	Nie
LUX	0.015	-0.013	-0.040	0.043	-0.033	-0.051	Nie	Nie
MEX	0.004	0.000	-0.010	0.043	-0.032	-0.022	Tak	Nie
NLD	0.023	-0.005	-0.041	0.034	0.001	-0.045	Nie	Nie
NOR	0.016	-0.001	-0.040	0.037	-0.006	-0.048	Nie	Nie
NZL	0.031	-0.036	-0.060	0.059	-0.049	-0.068	Nie	Nie
POL	0.002	0.006	-0.004	0.063	-0.047	-0.034	Tak	Nie
PRT	0.008	0.007	-0.022	0.029	-0.003	-0.035	Tak	Nie
SVK	0.022	-0.023	-0.046	0.100	-0.076	-0.071	Nie	Nie
SWE	0.023	-0.007	-0.044	0.060	-0.021	-0.055	Nie	Nie
TUR	0.006	-0.001	-0.021	0.072	-0.067	-0.058	Tak	Nie
USA	0.013	-0.015	-0.041	0.038	-0.035	-0.052	Nie	Nie

Uwaga: Nazwy krajów oznaczone wg standardów ISO.

W każdym kraju sprawdzone zostało, czy pełna niezmienność pomiarowa między rokiem 2006 i 2015 została zachowana. Testowana jest tu niezmienność metryczna i niezmienność skalarna. Do oceny niezmienności posłużono się trzema wskaźnikami dopasowania. W tabeli podano różnice między wartością wskaźnika dopasowania dla modelu konfiguracyjnego a modelem metrycznym bądź skalarnym. Kryteria oceny przyjęte zostały za Chenem (2007).

Jak pokazują analizy, pełna niezmienność nie jest tylko problemem porównań między krajami. Jak się okazuje, pełna niezmienność w przypadku badanej skali nie została zachowana do porównań w tych samych krajach, ale w różnych punktach czasowych. Tylko w garstce krajów odnajdujemy niezmienność metryczną. W żadnym przypadku nie potwierdzona została niezmienność skalarna.

W kolejnym kroku przechodzimy do analizy częściowej niezmienności, gdzie niezmienne parametry oznaczamy za pomocą procedury porównań parami w optymalizacji wyrównującej (Asparouhov i Muthén, 2014). Dzięki tej metodzie możemy ocenić, iloma niezmiennymi parametrami (w przypadku CFA ładunków i stałych czynnikowych) dysponujemy w danym porównaniu. Wyniki te przedstawione zostały w tabeli 2.

Tabela 2. Częściowa i przybliżona niezmienność pomiarowa. Porównania między rokiem 2006 i 2015 dla każdego kraju OECD uczestniczącego w obydwu falach badania

Kraj	Niezmienność parametrów		Przybliżona niezmienność (wariancja parametrów)		Dalsze analizy
	Ładunki	Stałe	Ładunki	Stałe	
AUS	80%	20%	0.000	0.000	Tak
AUT	60%	40%	0.000	0.000	Tak
BEL	80%	40%	0.001	0.001	Tak
CAN	60%	20%	0.001	0.000	Nie
CHE	60%	20%	0.000	0.000	Nie
CZE	60%	40%	0.001	0.001	Tak
DEU	60%	40%	0.000	0.000	Tak
DNK	80%	40%	0.001	0.001	Tak
ESP	60%	40%	0.000	0.001	Tak
FIN	60%	0%	0.000	---	Nie
FRA	80%	20%	0.001	0.000	Tak
GBR	80%	40%	0.001	0.000	Tak
GRC	100%	40%	0.001	0.000	Tak
HUN	80%	20%	0.000	0.000	Tak
IRL	80%	20%	0.001	0.000	Tak
ISL	40%	20%	0.001	0.000	Tak
ITA	60%	20%	0.000	0.000	Tak
JPN	60%	40%	0.000	0.001	Tak
KOR	60%	20%	0.001	0.000	Tak
LUX	60%	20%	0.001	0.000	Tak

Kraj	Niezmienność parametrów		Przybliżona niezmienność (wariancja parametrów)		Dalsze analizy
	Ładunki	**Stałe**	**Ładunki**	**Stałe**	
MEX	80%	40%	0.005	0.000	Tak
NLD	80%	60%	0.001	0.001	Tak
NOR	80%	20%	0.001	0.000	Tak
NZL	80%	40%	0.000	0.000	Tak
POL	80%	40%	0.001	0.000	Tak
PRT	80%	40%	0.001	0.000	Tak
SVK	60%	20%	0.001	0.000	Tak
SWE	60%	0%	0.001	---	Nie
TUR	40%	0%	0.000	---	Nie
USA	80%	20%	0.001	0.000	Tak

Uwaga: Nazwy krajów oznaczone wg standardów ISO.

Jak widać, w zdecydowanej większości krajów większość ładunków czynnikowych pozostaje niezmienna. Inaczej sprawa ma się ze stałymi czynnikowymi. W większości krajów mniejszość ładunków czynnikowych zachowuje niezmienność. W niektórych krajach zastosowana metoda nie wykryła nawet jednego niezmiennego ładunku czynnikowego. Kraje te zostaną usunięte z kolejnych analiz. Brak nawet jednej niezmiennej stałej czynnikowej wskazuje na brak możliwości dokonania kolejnych porównań.

W kolejnym kroku przeanalizowana została przybliżona niezmienność. Do tej analizy wykorzystany został model BSEM, gdzie kolejno testowane były modele międzygrupowych wariancji parametrów: 0; 0,001; 0,005; 0,01 i 0,02. Dopasowanie modelu testowane było za pomocą miary DIC. Przyjęte zostało, iż DIC>3 oznacza istotną poprawę dopasowania modelu. Mniejsze wartości wskazują na nieistotną zamianę dopasowania modelu. Modele testowane były sekwencyjne, poczynając od najniższych wariancji i kończąc testowanie w momencie, gdy zwiększanie wariancji nie poprawiało dopasowania modelu. Wyniki przedstawione zostały w tabeli 2. Wartości wariancji są małe i rzadko przekraczają wartość 0,001, co pozwala uznać je za nieznaczące (Pokropek i in., 2019). Podsumowując ten etap analiz, w zdecydowanej większości krajów odnajdujemy częściową niezmienność, stanowiącą warunek konieczny dla porównań wyników tego samego kraju w dwóch punktach czasowych. W trzech krajach brak niezmiennych stałych czynnikowych przekreśla możliwość dokonywania porównań średnich między dwoma punktami czasowymi. Jako że naszym celem jest dokonanie porównań zarówno między dwoma punktami czasowymi, jak i między krajami, owe trzy kraje zostały wykluczone z dalszych analiz.

W kolejnym kroku analiz sprawdzamy, czy pełna niezmienność pomiarowa może zostać uzyskana, gdy analizujemy kraje w różnych punktach czasowych, traktując je jako oddzielne grupy. Analizujemy tym samym łącznie 54 grupy (27 krajów w dwóch punktach czasowych). Nie jest zaskoczeniem, że w tej konfiguracji pełna niezgodność nie została osiągnięta, co przedstawiono w tabeli 3.

Tabela 3. Klasyczna niezmienność pomiarowa. Wartości miar dopasowania modeli dla 27 krajów w dwóch punktach czasowych rozpatrywanych jako 54 grupy

Model:	**CFI**	**RMSEA**	**SRMR**
Konfiguracyjny	0.9831	0.0743	0.017
Metryczny	0.9491	0.0964	0.071
Skalarny	0.8505	0.1377	0.1043
Różnice:	**$_{\Delta}$CFI**	**$_{\Delta}$RMSEA**	**$_{\Delta}$SRMR**
Konfiguracyjny – Metryczny	0.034	-0.022	-0.054
Konfiguracyjny – Skalarny	0.1326	-0.063	-0.087

Kolejny etap analiz polegał na testowaniu modelu częściowej niezmienności. W tym kroku analizy przeprowadzane były na 54 grupach (27 krajów w dwóch punktach czasowych). W tabeli 4 przedstawiono odsetek niezmiennych parametrów dla wszystkich grup. Wyniki są podobne do tych odnoszących się do porównywalności w czasie. Większość ładunków czynnikowych (ogółem 65%) jest niezmiennych pomiarowo. W jednej grupie żaden z ładunków nie został oznaczony jako niezmienny. Jeśli chodzi o stałe czynnikowe, średnio rzecz biorąc, sytuacja jest nieznacznie gorsza (57%), lecz aż w 5 grupach żaden ładunek czynnikowy nie został oznaczony jako niezmienny. Oznacza to, że 6 grup, a co za tym idzie również krajów, zostaje usuniętych z dalszych analiz.

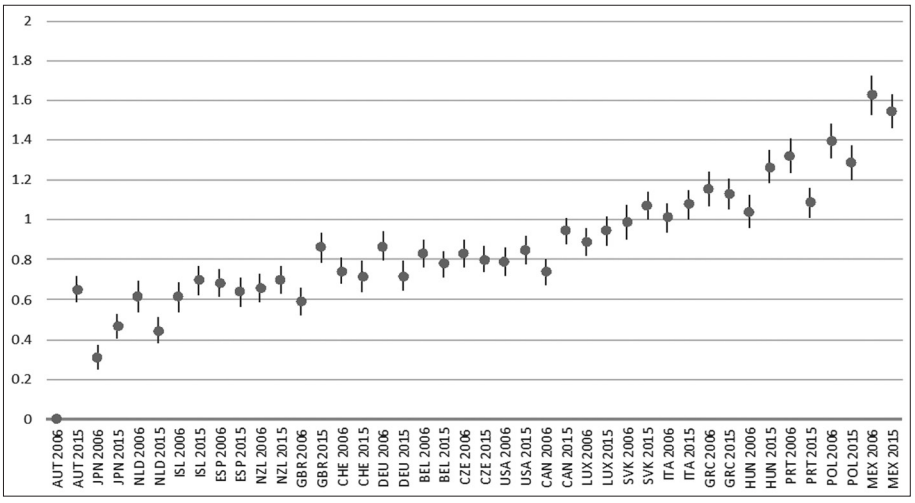
Tabela 4. Procent niezmiennych ładunków i stałych czynnikowych dla 27 krajów w dwóch punktach czasowych rozpatrywanych jako 54 grupy

Kraj	Rok 2006		Rok 2015		Dalsze analizy
	Ładunki	**Stałe**	**Ładunki**	**Stałe**	
AUS	60%	0%	40%	20%	Nie
AUT	60%	60%	40%	60%	Tak
BEL	60%	80%	60%	20%	Tak
CAN	80%	80%	40%	20%	Tak
CHE	100%	100%	60%	20%	Tak
CZE	40%	20%	80%	60%	Tak
DEU	80%	80%	80%	40%	Tak
DNK	20%	60%	40%	0%	Nie
ESP	100%	100%	40%	20%	Tak
FRA	80%	80%	40%	0%	Nie
GBR	100%	100%	60%	20%	Tak
GRC	60%	40%	20%	40%	Tak
HUN	60%	80%	20%	40%	Tak
IRL	80%	100%	0%	20%	Nie
ISL	60%	100%	20%	60%	Tak
ITA	80%	100%	40%	60%	Tak
JPN	80%	80%	60%	20%	Tak
KOR	100%	100%	60%	0%	Nie

Kraj	Rok 2006		Rok 2015		Dalsze analizy
	Ładunki	**Stałe**	**Ładunki**	**Stałe**	
LUX	80%	100%	40%	20%	Tak
MEX	100%	100%	80%	20%	Tak
NLD	100%	100%	40%	60%	Tak
NOR	100%	100%	40%	0%	Nie
NZL	100%	100%	60%	20%	Tak
POL	100%	100%	60%	20%	Tak
PRT	100%	100%	60%	60%	Tak
SVK	100%	100%	60%	60%	Tak
USA	100%	100%	80%	20%	Tak

Uwaga: Nazwy krajów oznaczone wg standardów ISO.

W kolejnym kroku testowany był model częściowej przybliżonej niezmienności. Model ten nie pokazywał lepszego dopasowania do danych od modelu częściowej niezmienności, który został uznany za wystarczająco dopasowany. Model częściowej niezmienności został oszacowany po raz kolejny po usunięciu 6 krajów, w których brakowało parametrów niezmiennych pomiarowo. Wyniki z tego modelu przedstawiono na rysunku 3.



Rysunek 3. Średnie oraz 95% przedział ufności wskaźnika zainteresowania nauką dla 21 krajów OECD w dwóch punktach czasowych. Średnie estymowane na podstawie modelu CFA częściowej niezmienności

Uwaga: Nazwy krajów oznaczone wg standardów ISO.

W tym miejscu kończą się metodologiczne rozważania. Zaczynają się sprawy trudniejsze. Wydaje się, że zainteresowanie nauką jest w miarę stałe między 2006 a 2015 rokiem w badanych krajach. Ciekawe jest to, że w krajach mniej zamożnych zainteresowanie nauką jest większe. Dlaczego? To już temat na inny artykuł.

Dodatkowe informacje

Tekst ten został napisany w ramach grantu Sonata 8 finansowanego przez Narodowe Centrum Nauki (NCN) pt. *Porównywalność skal w międzynarodowych badaniach sondażowych* (UMO-2014/15/D/HS6/04934).

This text has been prepared under the Scales Comparability in Large Scale Cross-country Surveys Project, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934).

Bibliografia

Asparouhov, T. i Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508.

Billiet J. 2013. *Quantitative methods with survey data in comparative research*. [W:] A *Handbook of Comparative Social Policy*, (red.) P Kennett, ss. 264–300. Cheltenham, UK: Edward Elgar. 2nd ed.

Noah, H. 1986. *The Use and Abuse of Comparative Education*. [W:] *New Approaches to Comparative Education*, [red.] P. G. Altbach and G. P. Kelly, 153–166. Chicago, IL: University of Chicago Press.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the Cross-Country Comparability of Indicators of Socioeconomic Resources in PISA. *Applied Measurement in Education*, *30*(4), 243–258.

van de Vijver, F. J. (2011). *Capturing bias in structural equation modeling*. [W:] E. Davidov, P. Schmidt i J. Billiet [red.], *Cross-cultural analysis. Methods and applications* (pp. 3–34). New York: Routledge.

Pokropek, A. (2015). Modele cech ukrytych w psychologii, socjologii i badaniach edukacyjnych. *Teoria i zastosowania.* Warszawa: IBE.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P. i Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology* (40), 55–75.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Byrne, B. M., Shavelson, R. J. i Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–21.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*: Springer.

Hox, J. J., de Leeuw, E. D. i Brinkhuis, M. J. (2010). Analysis models for comparative surveys. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell i T. W. Smith (red.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (s. 395–418). Willey: Hoboken, NJ.

Muthén, B. i Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods, 17*(3), 313.

De Jong, M. G., Steenkamp, J. B. E. i Fox, J. P. (2007). Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model. *Journal of consumer research, 34*(2), 260–278.