Polskie Forum Psychologiczne, 2018, tom 23, numer 2, s. 330-356
DOI: 10.14656/PFP20180207

# ZASTOSOWANIE ANALIZY RÓWNOWAŻNOŚCI POMIAROWEJ W BADANIACH PSYCHOLOGICZNYCH

Katarzyna Lubiewska, Karolina Głogowska

Instytut Psychologii, Uniwersytet Kazimierza Wielkiego
Institute of Psychology, Kazimierz Wielki University in Bydgoszcz

IMPLEMENTATION OF MEASUREMENT EQUIVALENCE ANALYSIS IN PSYCHOLOGICAL RESEARCH

**Summary**. Analysis of measurement invariance is slowly but consistently becoming the standard of quantitative analysis in psychological research. It tests whether a measuring instrument (e.g. a scale) has the same measurement characteristics in groups under the study (e.g. differing in terms of age, culture, or type of research). Establishing the measurement invariance of instruments used in the study allows, at the initial stage of the data analysis, for further testing of the main research hypotheses addressing the relations between constructs or the mean differences. The paper provides an overview of strategies and problems related with the invariance analysis aiming to introduce more widely this method to researchers in Poland. To this end, we first describe levels of measurement invariance and decisions' criteria related with its establishing. We also describe strategies to cope with noninvariance, briefly introducing bayesian methods. Finally, we provide an example guiding step-by-step through the invariance analysis with the use of R environment 'lavaan' package. Summarizing, we highlight the importance of invariance analysis in psychological research postulating that this analysis does not limit researchers but increases the precision of conclusions derived in psychological research.
**Key words**: measurement invariance, psychometrics, cross-cultural research, comparative psychology

Analiza równoważności skal pomiarowych, czyli testowanie uprawnień do dokonywania porównań i badania relacji pomiędzy wynikami pomiaru w badanych grupach, jest jednym ze sposobów, który pozwala na zdobycie przez badacza (względnej) pewności, że uzyskane przez niego wyniki analiz danych przedstawiają niezafałszowany obraz badanego zjawiska uprawniając do dalszego testowania hipotez głównych o związkach pomiędzy zmiennymi czy różnicach średnich. Inne

Adres do korespondencji: Katarzyna Lubiewska, e-mail: lubkat@ukw.edu.pl

pomocne w tym zakresie analizy, np. błędu wspólnej metody (odsyłamy do tekstu Razmus, Mielniczuk, 2018) czy natury braków danych (Pokropek, 2018), są omówione w innych artykułach włączonych do tego numeru czasopisma. W niniejszym tekście zajmiemy się dokładniejszą analizą tego, czym jest równoważność pomiarowa, jakie są konsekwencje zignorowania tej analizy w badaniach, jak przeprowadzić samą analizę krok po kroku, oraz jakie korzyści ona przynosi.

Analiza równoważności pomiarowej, choć promowana już w latach osiemdziesiątych XX wieku (np. Hui, Triandis, 1985) w psychologii przyjmuje się powoli. Wyjątkiem wydaje się być psychologia międzykulturowa, gdzie analizy te są powszechne. Czołowe czasopisma tego nurtu nie dopuszczają do druku artykułów porównujących różne grupy kulturowe, jeżeli analiza równoważności pomiarowej nie jest przeprowadzona na wstępnym etapie analiz. Świadomość ta nieco wolniej, aniżeli w psychologii międzykulturowej (np. Cieciuch i in., 2014), dociera również do innych subdyscyplin psychologii takich, jak np. psychologia rozwojowa (np. Knight, Zerr, 2010a, 2010b; Haltigan i in., 2014) czy organizacji (np. Vandenberg, Lance, 2000; Cheung, 2007). Niezależnie od subdyscypliny psychologii równoważność bada się też, sprawdzając uprawnienia do porównywania wyników grup badanych w różnych warunkach, np. pomiędzy grupą eksperymentalną a kontrolną (np. Pentz, Chou, 1994) czy grupą wypełniającą test w formie papier-ołówek a grupą badaną komputerową wersją testu (np. van de Vijver, Harsveld, 1994; De Beuckelaer, Lievens, 2009; Cieciuch, Davidov, 2015).

Biorąc jednak pod uwagę, że znaczna mniejszość badań dotychczas opublikowanych zawiera raport dotyczący równoważności pomiarowej skal wykorzystanych do analiz związków czy różnic wyników analizowanych grup może pojawić się wątpliwość dotycząca zakresu, w jakim możemy raportom z dotychczasowych badań ufać. Pytanie to wynika nie tylko z braku w większości z dotychczasowych raportów z badań analizy równoważności pomiarowej, ale również z braku innych analiz (np. analizy wzorca braku danych), które zapewniają jak najlepszą „diagnozę" danych przed przystąpieniem do analiz głównych. Zapewne problemy z replikacją w badaniach psychologicznych (np. Open Science Collaboration, 2015) mogą być, przynajmniej po części, związane z brakiem porównywalności wyników i badanych konstruktów w analizowanych w różnych badaniach grupach.

## Równoważność pomiarowa

W literaturze anglojęzycznej używa się zamiennie pojęć *ekwiwalencja pomiarowa* (*measurement equivalence*) i *równoważność pomiarowa* (*measurement invariance*). Celem tej złożonej z kilku poziomów analizy jest sprawdzenie, a najlepiej wykazanie, że przedstawiciele różnych badanych przez nas grup tak samo rozumieją badany konstrukt i taką samą wagę przypisują pozycjom testowym skali odpowiadając na nie, przez co odpowiedzi uzyskane w obu grupach są ze sobą porównywalne.

Analizę równoważności przeprowadza się po to, żeby na wstępnym etapie analiz sprawdzić, czy w naszym badaniu jest ona potwierdzona, co uprawnia do przy-

stąpienia do testowania sformułowanych w badaniu hipotez głównych. Jeżeli ten pożądany stan rzeczy nie jest osiągnięty, wnioskowanie dotyczące naszych hipotez głównych (gdybyśmy przystąpili do ich testowania) dostarczy wyniki, które nie są wiarygodne i mogą doprowadzić do błędnych wniosków. Brak potwierdzenia równoważności pomiarowej w dwóch badanych grupach można porównać do próby zestawienia ze sobą nie dwóch jabłek, a jabłek z pomarańczami, których natura jest zupełnie różna. Tego błędu można uniknąć. Przykładowo, w jednym z badań (Lubiewska, van de Vijver, 2015) porównując wyniki pomiaru niepokoju przywiązaniowego w grupie nastolatków, ich matek i babć wykazaliśmy, że gdybyśmy nie kontrolowali równoważności pomiarowej, nasz wniosek z badań postulowałby, że nastolatkowie mają wyższy poziom niepokoju aniżeli ich mamy i babcie. Wynik ten okazał się jednak uwarunkowany brakiem równoważności pomiarowej odpowiedzi testowych kilku twierdzeń skali. Ich wyłączenie z puli twierdzeń badających niepokój ujawniło brak różnic w poziomie niepokoju we wszystkich badanych grupach wiekowych.

Warto przy tym zaznaczyć, że równoważność nie opisuje tylko właściwości danego narzędzia pomiarowego, np. skali, a raczej opisuje właściwości skali oraz różnic pomiędzy grupami, w których się to narzędzie zastosowało (van de Vijver, Leung, 1997). Inaczej rzecz ujmując, skala nigdy nie funkcjonuje tak samo niezależnie od grupy, w której jest stosowana. Dlatego też, przygotowując raport z badań należy przed przystąpieniem do analiz głównych przetestować na własnych danych strukturę skali, jej rzetelność oraz sprawdzić jaki zakres równoważności jest potwierdzony jeżeli nasza próba nie jest homogeniczna (np. pod względem kultury, płci, wieku czy warunków eksperymentalnych).

## Poziomy równoważności pomiarowej

Jest wiele poziomów równoważności pomiarowej, które mogą być testowane w badaniach. Najbardziej interesujące dla badaczy i uważane za krytyczne w badaniach psychologicznych są zwykle równoważność: konstruktu, metryczna i skalarna (van de Vijver, Leung, 1997). Niemniej poza wymienionymi rodzajami ekwiwalencji można badać też równoważność: błędów pomiaru zmiennych obserwowalnych modelu pomiarowego (poziom uznawany za zbyt restrykcyjny i rzadko testowany), wariancji czy kowariancji zmiennych latentnych, czyli związków pomiędzy czynnikami latentnymi, którą badamy, kiedy skala ma kilka podskal (np. Schmitt, Kuljanin, 2008).

Poniżej opiszemy trzy najważniejsze wymiary ekwiwalencji pomiarowej: równoważność konstruktu, metryczną i skalarną, które są niezbędne do przystąpienia do dalszego testowania hipotez o związkach pomiędzy zmiennymi i różnicach średnich poziomów wyników skali (tabela 1). Te trzy wymiary równoważności mają strukturę hierarchiczną i potwierdzenie niższego (łatwiejszego do udowodnienia) poziomu równoważności zezwala na przejście do testowania poziomu wyższego (trudniejszego).

Tabela 1. Podstawowe poziomy (etapy) testowania równoważności pomiarowej

| Poziom równoważności pomiarowej | Rodzaj restrykcji równości sprawdzany w grupach | Kroki przeprowadzenia analiz | Poziom trudności potwierdzenia | Uprawnienia i ograniczenia, jeżeli poziom nie jest (częściowo lub całkowicie) potwierdzony |
|---|---|---|---|---|
| Równoważność konstruktu | Równość struktury czynnikowej (ilość czynników i wzorzec ładunków czynnikowych) | 1. Stworzenie modelu konfiguralnego.<br>2. Oszacowanie dobroci dopasowania modelu konfiguralnego (bazowego). | Najniższy (najłatwiejszy) | Nie można testować równoważności metrycznej. |
| Równoważność metryczna[a] | Równość ładunków czynnikowych | 1. Narzucenie restrykcji równości ładunków czynnikowych we wszystkich grupach.<br>2. Sprawdzenie istotności statystycznej różnicy ($\Delta \chi^2$) pomiędzy wartościami $\chi^2$ modelu konfirmacyjnego i $\chi^2$ modelu konfirmacyjnego z narzuconą równością ładunków czynnikowych lub zastosowanie kryteriów odcięcia do oceny. 3. Jeżeli $\Delta \chi^2$ nie jest istotna statystycznie lub parametr dopasowania modelu spełnia kryteria odcięcia można przejść do testowania równo-ważności skalarnej. Jeżeli tak nie jest można sprawdzić warunki częściowej równoważności lub zaprzestać na tym poziomie. | Średni | Nie można testować równoważności skalarnej. Można analizować związki pomiędzy zmiennymi. |

cd. tabeli 1

| | | | | |
|---|---|---|---|---|
| Równoważność skalarna | Równość stałych regresji | 1. Dodatkowe narzucenie restrykcji równości stałych regresji (poza restrykcjami równości ładunków czynnikowych) we wszystkich grupach.<br>2. Sprawdzenie istotności statystycznej różnicy ($\Delta \chi^2$) pomiędzy wartościami $\chi^2$ modelu konfirmacyjnego z narzuconą równością ładunków czynnikowych i $\chi^2$ modelu konfirmacyjnego z narzuconą równością ładunków czynnikowych i stałych regresji. | Najwyższy (najtrudniejszy) | Nadal można analizować związki pomiędzy zmiennymi, ale nie można testować różnic średnich latentnych w grupach. |

[a] Jeżeli struktura skali składa się z więcej, aniżeli jednego czynnika, po analizie równoważności metrycznej możemy sprawdzić poziom uzyskania równoważności strukturalnej, gdzie narzucone są restrykcje równości pomiędzy latentnymi czynnikami skali.

Najniższym poziomem jest potwierdzenie równoważności konstruktu. Poziomem wyższym jest osiągnięcie równoważności metrycznej. Poziomem najwyższym i najtrudniejszym jest potwierdzenie równoważności skalarnej. Jeżeli nasze analizy wykazują, że mamy osiągnięty poziom równoważności metrycznej możemy przystąpić do testowania głównych hipotez badawczych analizujących związki pomiędzy zmiennymi. Jeżeli uda nam się również potwierdzić równoważność skalarną, możemy przystąpić do testowania hipotez o różnicach średnich.

**Równoważność konstruktu** (*construct* lub *functional invariance*) określana jest też jako równoważność konfiguralna, jeżeli mierzonych jest kilka konstrunktów, np. podwymiarów skali (*configural invariance*). Analiza tego poziomu równoważności odpowiada na pytanie, czy w różnych grupach mierzymy ten sam bądź inny konstrukt (van de Vijver, Leung, 1997). Potwierdzenie tego poziomu równoważności dostarcza empirycznych dowodów na to, że mamy w każdej z grup tę samą liczbę czynników latentnych (podskal skali) wskazywanych przez te same wskaźniki obserwowalne (zmienne/twierdzenia skali) o tym samym wzorcu związków pomiędzy nimi. W szerszym znaczeniu potwierdzenie tego poziomu równoważności wskazuje na to, że badani w analizowanych grupach nadają to samo znaczenie konstruktowi przez nas badanemu jako całości oraz, że konstrukt ten ma tę samą strukturę w analizowanych grupach lub w różnych punktach pomiarowych (Welkenhuysen-Gybels, van de Vijver, 2001). Pomimo tego, że jest to najłatwiejszy do potwierdzenia poziom równoważności, można i na tym etapie napotkać problemy, jeżeli analizowane przez nas grupy bardzo się różnią.

Przykładowo, w jednym z badań pierwszej autorki i współpracowników (Lubiewska i in., w recenzji), w którym analizowaliśmy związek wymiarów zachowań rodzicielskich matek z ufnością przywiązaniową nastolatków w 14 (upraszczając) zachodnich i niezachodnich krajach, napotkaliśmy na problem braku równoważności konstruktu dotyczący skali kontroli rodzicielskiej (skrócona wersja PARQ; np. Rohner, Rohner, Roll, 1980). Dalsze analizy wykazały, że skala ta w niektórych krajach niezachodnich (np. w Indonezji czy w Indiach) tworzy jeden spójny konstrukt, podczas gdy w krajach zachodnich, postkomunistycznych, Bliskiego Wschodu oraz w Chinach konstrukt ten składał się z dwóch umiarkowanie skorelowanych ze sobą czynników – kontroli psychologicznej (w wymiarze przymusu emocjonalnego ograniczającego autonomię dziecka) i behawioralnej. Niestety, z powodu niskiej spójności wewnętrznej czynnika kontroli behawioralnej w krajach niezachodnich zdecydowaliśmy się na testowanie hipotez głównych tylko z użyciem czynnika kontroli psychologicznej – zmieniając zamierzenie autora skali co do jej stosowania. Przy tym rozwiązaniu udało nam się potwierdzić równoważność konstruktu, co pozwoliło na testowanie następnych poziomów równoważności. Niemniej z powodu nieakceptowalnego poziomu spójności wewnętrznej (rzetelności) musieliśmy z dalszych porównań dodatkowo wyłączyć dane z czterech krajów.

**Równoważność metryczna** (*metric invariance*) lub równoważność jednostki pomiarowej (*measurement unit invariance*). Nieco trudniejszym do uzyskania poziomem

ekwiwalencji pomiarowej jest równoważność metryczna, czyli tzw. porównywalność jednostek pomiaru. Najbardziej trafnym przykładem tego poziomu analizy jest porównanie ze sobą pomiaru temperatury za pomocą stopni Celsjusza i Kelvina. Pomimo tego, że zeru Kelwina odpowiada -273 stopni Celsjusza, zmianie o jeden kelwin odpowiada zmiana również o jeden stopień w skali Celsjusza, wskazując na to, że interwały pomiarowe są równe (co prawda źródła skali różne).

Podobnie jest np. ze skalą Likerta. Sprawdzając ten poziom równoważności, pytamy o to, czy jednostka pomiarowa w analizowanych grupach jest porównywalna (van de Vijver, Leung, 1997), przykładowo, czy odległość odpowiedzi w skali Likerta, np. od (1) *zdecydowanie się nie zgadzam* do odpowiedzi (2) *raczej się nie zgadzam* jest tożsama w analizowanych grupach. Aby to sprawdzić badamy, czy wartości ładunków czynnikowych są równe w badanych grupach, co jest równoznaczne z równością interwałów skali narzędzia. Należy tutaj jednak pamiętać, że na tym poziomie sprawdzamy tylko równoważność jednostki pomiarowej, nie zaś źródło skali. Źródło skali, czyli punkt startowy pomiaru wskazywany w analizach przez stałą regresji (*intercept*), może pozostawać w tym przypadku różny tak, jak w sytuacji, kiedy zeru kelwina odpowiada -273 stopni w skali Celsjusza.

Warto tutaj ponownie zaznaczyć, że potwierdzenie równoważności metrycznej nie oznacza jeszcze, że możemy porównać średnie wyników w analizowanych grupach. Udowadniając, że mamy równą jednostkę pomiarową, nie udowodniliśmy jeszcze, że wyniki średnie są porównywalne ze sobą. Może być tak, że pomimo równych jednostek pomiaru średnia w jednej grupie może być podwojona z powodu innego poziomu startowego pomiaru w porównaniu do średniej w drugiej grupie. Aby mieć uprawnienia do porównywania średnich poziomów skali musimy przetestować najwyższy poziom ekwiwalencji pomiarowej – równoważność skalarną.

Należy dodać, że brak potwierdzenia tak równoważności konstruktu, jak i równoważności metrycznej może wskazywać na stronniczość metody (*method bias*) lub stronniczość twierdzeń skali (*item bias*), czyli ich odmienne funkcjonowanie w grupach (*differentia item functioning*, DIF analizowany często w ramach Teorii Odpowiedzi Testowych poza Klasyczną Teorią Testów; np. w: Reynolds, Suzuki, 2013). Stronniczość metody może wynikać z wielu źródeł będących przyczyną różnic pomiędzy grupami, np. braku odporności metody na styl odpowiedzi testowych respondentów (*response style*)[1], tendencji badanych do ukazywania się w określonym świetle (*social desirability*) czy niedostosowania metody pomiaru do np. poziomu edukacji badanych (Zawadzki, 2006). Stronniczość twierdzeń zaś może wynikać

[1] Styl odpowiedzi testowych odnosi się do typowego dla danej grupy (często kulturowej) stylu odpowiadania na pytania testowe. Wśród rodzajów stylu odpowiedzi są: ugodowość [tendencja badanych do częstszego wybieranie odpowiedzi (4) *zgadzam się* aniżeli (2) *nie zgadzam się* niezależnie od rodzaju pytania czy poziomu badanej cechy]; preferencja do wybierania środkowych wartości skali [(3) *ani tak, ani nie*] czy preferencja do udzielania ekstremalnych odpowiedzi [(5) *zdecydowanie zgadzam się*/ (1) *zdecydowanie nie zgadzam się*].

z nietrafnego ich tłumaczenia[2], nieadekwatności ekologicznej twierdzenia (np. gdy pytamy o oglądanie telewizji tam, gdzie nie ma jej w domu) czy z niezrozumiałego dla badanych ich sformułowania (van de Vijver, Leung, 1997). Należy jednak pamiętać, że stronniczość metody czy twierdzeń skali sama w sobie mówi nam dużo o różnicach pomiędzy analizowanymi grupami, choć niekoniecznie o różnicach międzygrupowych w analizowanym przez nas konstrukcie zdefiniowanym przez twierdzenia użytej przez nas skali.

**Równoważność skalarna** (*scalar invariance*) lub pełna porównywalność wyniku (*full score comparability*). Ten poziom równoważności odpowiada nam na pytanie o to, czy danemu poziomowi badanej przez nas cechy odpowiada ta sama wartość odpowiedzi testowej badanych w porównywanych grupach. Innymi słowy, chcemy wiedzieć, czy wynik badania dwóch (lub więcej) osób o tym samym poziomie badanej cechy, np. niepokoju, jest ten sam, jeżeli tę cechę mierzymy naszą skalą. Najprościej mówiąc, potwierdzenie równoważności skalarnej oznacza, że badani w ten sam sposób używają skali.

Celem wykonania tej analizy testuje się równoważność stałych regresji (*intercepts*) pozycji testowych skali (np. van de Vijver, Leung, 1997), które określają jaka jest lokalizacja (punkt startowy) wartości odpowiedzi testowej danego twierdzenia w porównywanych grupach, kiedy poziom czynnika latentnego, którego twierdzenie jest wskaźnikiem, wynosi zero (np. jaka jest wartość stałej regresji twierdzenia X będącego wskaźnikiem skali niepokoju, jeżeli poziom niepokoju w danej grupie jest równy zero). Analogicznym parametrem w Teorii Odpowiedzi Testowych jest parametr trudności pytania (*item difficulty parameter*) (np. van de Vijver, Leung, 1997). Brak potwierdzenia tego poziomu równoważności sprowadza się do wniosku, że osoby w różnych grupach z tym samym poziomem cechy (np. niepokoju) w badaniu uzyskują różne średnie wyników, których porównywanie może prowadzić do błędnych wniosków o różnicy średnich wyników w obu grupach.

## Etapy i strategie analizy równoważności pomiarowej

Istnieją różne pakiety i metody analizy danych, które można wykorzystać do przetestowania równoważności pomiarowej. Jedną z nich jest eksploracyjna analiza czynnikowa (np. możliwa do przeprowadzenia w pakiecie SPSS), gdzie na macierzy korelacyjnej dokonuje się analizy czynnikowej manipulując przy tym przynależnością grupową badanych (Welkenhuysen-Gybels, van de Vijver, 2001) oraz wykorzystując do oszacowania uzyskania równoważności pomiarowej wartości wskaźnika $\phi$ (phi) Tuckera (np. Różycka-Tran, Boski, Wojciszke, 2014). Niemniej ta metoda jest dzisiaj już rzadko wykorzystywana z powodu skomplikowanej procedury obliczania oraz braku kontroli błędów pomiaru. Ponadto zasugerowałyśmy już wcześniej,

[2] Trafna translacja językowa nie musi wiązać się z trafnością ekologiczną tłumaczenia.

że chcąc przeprowadzić analizę równoważności pomiarowej, można wyjść poza Klasyczną Teorię Testów i skorzystać z Teorii Odpowiedzi Testowych (*Item Response Theory*). Takie analizy należą jednak w psychologii do rzadkości (z wyjątkiem badań edukacyjnych). Najczęściej stosowaną dzisiaj w badaniach psychologicznych strategią jest prawdopodobnie wykorzystanie konfirmacyjnej analizy czynnikowej na podstawie modelowania równań strukturalnych. Analiza ta bywa poszerzona o użycie estymatorów bayesowskich, co jest przydatne w warunkach, kiedy potwierdzenie równoważności przy założeniu pełnej równości szacowanych parametrów nie jest możliwe do osiągnięcia (np. Zercher i in., 2015).

Do analizy równoważności na podstawie konfirmacyjnej analizy czynnikowej wykorzystuje się dzisiaj różne pakiety statystyczne typu: Amos, Mplus (zainteresowanych odsyłamy do Cieciuch, Davidov, 2015), LISREL czy któryś z pakietów środowiska R. Z naszych doświadczeń wynika, że najszybciej i najskuteczniej[3] można dokonać tych analiz przy użyciu pakietu 'lavaan' (Rosseell, 2012) środowiska R. Ograniczeniem tego pakietu jest jednak póki co brak bezpośredniej możliwości wykorzystania analiz bayesowskich. Celem zrobienia tych analiz należy dodatkowo skorzystać z innych pakietów wykorzystujących grupę metod Markov chain Monte Carlo (MCMC), takich jak programy z grupy BUGS, JUGS czy Stan sprzężone z takimi pakietami środowiska R, jak: 'blavaan', 'rjags' czy 'Rstan'. Pakiet 'blavaan' (Merkle, Rosseel, 2015), oparty na programie JAGS, jest przy tym kompatybilny z pakietem 'lavaan'. Jeżeli zatem zależy nam na wykorzystaniu estymatorów bayesowskich, jednymi z najlepszych pakietów są: Mplus, który umożliwia przeprowadzenie wszystkich analiz przy użyciu jednego programu (jednak jego wadą jest to, że nie jest on darmowy), oraz darmowe pakiety 'lavaan' i 'blavaan' środowiska R.

Niemniej niezależnie od tego, który z pakietów wykorzystamy, strategia przeprowadzania analiz jest taka sama. Najpierw należy przetestować model bazowy, potem zaś model konfiguralny (równoważność konfiguralna) oraz kolejno modele zagnieżdżone z narzuconymi restrykcjami równości ładunków czynnikowych (równoważność metryczna) oraz stałych regresji (równoważność skalarna). Zanim opiszemy te poziomy analiz dokładniej wyjaśnimy strategie podejmowania decyzji dotyczące potwierdzenia (osiągnięcia) danego poziomu równoważności.

## Kryteria decyzyjne

Dwa etapy analiz związanych z testowaniem równoważności pomiarowej wymagają decyzji dotyczących dalszego postępowania. Po pierwsze, należy podjąć decyzję, czy model pomiarowy, który teoretycznie chcemy przetestować w naszym badaniu ma dobre parametry dopasowania do danych. Po wtóre, na etapie testowa-

[3] Przy porównywaniu parametrów modeli w 39 grupach i kilkunastu tysiącach danych, jedynie pakiet 'lavaan' nie zawiesił się przed ukończeniem analiz.

nia różnych poziomów równoważności pomiarowej naszego modelu należy podjąć decyzję dotyczącą dalszego postępowania w sytuacji braku potwierdzenia testowanego poziomu równoważności. Kryteria decyzyjne związane z tymi etapami opiszemy poniżej.

Przygotowując model pomiarowy, który chcemy przetestować najpierw należy zadbać o dobre parametry dopasowania modelu do danych w każdej z analizowanych grup oraz w całej próbie. Modele grupowe oraz model wielogrupowy powinny mieć przynajmniej akceptowalne parametry dopasowania do danych. Parametry, które poddaje się ocenie różnią się w zależności od preferencji badacza, niemniej warto kontrolować przynajmniej po jednym parametrze z danej grupy, z których każda dostarcza nam odmiennych informacji o poziomie dopasowania modelu do danych. Przykładowe parametry podajemy w tabeli 2. Jeżeli parametry dopasowania nie są zadowalające należy poszukać źródła problemu analizując parametry modelu oraz jego indeksy modyfikacyjne. W zależności od źródła problemu można dokonać w nim zmian, np. związanych z wprowadzeniem do modelu kowariancji błędów pomiaru zmiennych obserwowalnych lub z wyłączeniem z modelu zmiennych obserwowalnych (twierdzeń skali), których wartości ładunków czynnikowych lub stałych regresji nie są akceptowalne.

Następna grupa decyzji dotyczy tego, na czym opierać wniosek o potwierdzeniu danego poziomu równoważności pomiarowej. Przyjęte jest, że w tym zakresie kierujemy się kilkoma kryteriami opierającymi się na różnicy pomiędzy parametrami dopasowania analizowanego modelu a parametrami dopasowania modelu wcześniejszego, w którym testowaliśmy model na niższym poziomie równoważności aniżeli aktualnie testowany poziom (np. porównujemy parametry dopasowania modelu równoważności skalarnej z parametrami dopasowania modelu równoważności metrycznej lub parametry modelu konfiguralnego z modelem bazowym).

Kryteria decyzyjne o potwierdzeniu osiągnięcia danego poziomu równoważności oparte są na dwóch procedurach postępowania. Pierwsza z nich wymaga przeprowadzenia analizy istotności statystycznej różnicy pomiędzy chi kwadrat obu sąsiadujących modeli ($\Delta\chi^2$) za pomocą np. testu ANOVA. Jeżeli wynik wskazuje na brak istotnej statystycznie różnicy, oznacza to, że dany poziom równoważności może być potwierdzony jako osiągnięty.

Niemniej należy zaznaczyć, że w badaniach z wykorzystaniem bardzo dużych baz analizowanych danych oczekiwania, że przyrost $\chi^2$ pomiędzy modelami nie będzie istotny statystycznie, jest nierealistyczne ze względu na wrażliwość tego indeksu na wielkość próby (Cheung, Rensvold, 2002; Meade, Johnson, Braddy, 2008). W związku z tym zaproponowano nieco szerszą grupę kryteriów. Decyzja o potwierdzeniu każdego z poziomów równoważności opiera się tutaj na ustalonych, wskazanych poniżej kryteriach odcięcia:

(1) W dużych próbach badawczych ($N \geq 300$) zmiana CFI ($\Delta$CFI) do wartości $\leq$ -,01 pozwala na wnioskowanie o potwierdzeniu wszystkich poziomów równoważności pomiarowej (Cheung, Rensvold, 2002; Meade, Johnson, Braddy, 2008; Byrne, van de Vijver, 2010).

Tabela 2. Przykładowe indeksy dopasowania modelu do danych

| Grupa indeksów dopasowania | Parametr | | |
|---|---|---|---|
| | Nazwa | Opis | Kryteria oceny |
| Absolutne | $\chi^2$ | Wskaźnik dobroci dopasowania modelu do danych podatny na wielkość próby i odchylenia od normalności rozkładu danych. | Czym niższa wartość tym lepiej, a najlepiej jeżeli jest nieistotna statystycznie (mało realistyczne oczekiwanie w dużych próbach). |
| | $\chi^2/df$ | Znormalizowany parametr chi kwadrat wskazujący na dobroć dopasowania modelu do danych, w dużym stopniu odporny na wielkość próby. | 2,0-3,0 (≤ ,05) oznacza wystarczające dopasowanie modelu do danych. |
| | RMSEA (*Root Mean Square Error of Approximation*) | Indeks Steigera-Linda skorygowany ze względu na złożoność i oszczędność modelu. Wskazuje na wielkość błędu aproksymacji, czyli tego, jak źle model dopasowany jest do danych. | Ocena dopasowania modelu do danych:<br>• ≤ ,05 dobre<br>• ,05-,08 akceptowalne (rozsądny błąd aproksymacji)<br>• ≥ ,10 słabe dopasowanie. |
| Przyrostowe (relatywne) | CFI (*Comparative Fit Index*) | Porównawczy indeks dopasowania Bentlera. Wskazuje na różnicę dopasowania pomiędzy danymi i modelem hipotetycznym, korygując ze względu na wielkość próby. | Ocena dopasowania modelu do danych:<br>• ,90-,95 akceptowalne dopasowanie<br>• ,95 ≤ dobre dopasowanie. |
| Indeksy analizy reszt kowariancji | SRMR (*Standardized Root Mean Square Residual*) | Średnia wartość reszt dopasowania. | Ocena dopasowania modelu do danych:<br>• wartości bliskie zera oznaczają dobre dopasowanie modelu do danych.<br>• ≤ ,10 – wartości preferowane. |

Źródło: Browne, Cudeck, 1993; Hu, Bentler, 1999; Klein, 2005.

(2) W dużych próbach badawczych ($N \geq 300$) z równą liczebnością badanych w analizowanych grupach oraz mieszanym wzorcem odstępstw od równoważności, kryteria potwierdzania równoważności różnią się dla różnych jej poziomów (Chen, 2007): (a) przy testowaniu równoważności metrycznej, zmiana CFI ≤ -,010, uzupełniona przez zmianę RMSEA ≤ ,015 lub zmianę SRMR ≤ ,030; (b) przy testowaniu stałych regresji, zmiana CFI ≤ -,010, uzupełniona przez zmianę RMSEA ≤ ,015 lub zmianę SRMR ≤ ,010.
(3) W próbach małych ($N \leq 300$), z nierówną liczebnością badanych w grupach oraz jednolitym wzorcem odstępstw od równoważności, kryteria potwierdzania równoważności są następujące (Chen, 2007): (a) przy testowaniu równoważności metrycznej, zmiana CFI ≤ -,005, uzupełniona przez zmianę RMSEA ≤ ,010 lub zmianę SRMR ≤ ,025; (b) przy testowaniu stałych regresji, zmiana CFI ≤ -,005, uzupełniona przez zmianę RMSEA ≤ ,010 lub zmianę SRMR ≤ ,005.
(4) Zmiana parametrów *Akaike information criterion* (ΔAIC) oraz *Bayesian information criterion* (ΔBIC) o wartość większą niż 1 pozwala na wnioskowanie o równoważności (za Zercher i in., 2015).

## Kryteria decyzyjne w sytuacji braku potwierdzenia równoważności

W związku z tym, że wyniki analiz bardzo często nie są w stanie potwierdzić poziomu równoważności metrycznej czy skalarnej, następna grupa decyzji, przed którymi staje badacz, dotyczy tego, co zrobić, jeżeli dany poziom równoważności nie jest potwierdzony. Przynajmniej trzy ścieżki są tutaj możliwe. Po pierwsze, można zaprzestać dalszych analiz i ograniczyć poziom testowania hipotez głównych do potwierdzonego poziomu równoważności skali (do analizy związków pomiędzy zmiennymi przy potwierdzonej równoważności metrycznej oraz powstrzymanie się przed analizą różnic średnich przy braku potwierdzenia równoważności skalarnej).

Drugie wyjście polega na przystąpieniu do testowania **częściowej równoważności pomiarowej** (*partial measurement invariance*) (Byrne, Shavelson, Muthen, 1989). W tej sytuacji uwalniamy niektóre ładunki czynnikowe lub stałe regresji z restrykcji równości w analizowanych grupach. Decyzję o tym, które parametry uwolnić z restrykcji równości podejmuje się zwykle po analizie indeksów modyfikacyjnych oraz wielkości różnic analizowanych parametrów pomiędzy grupami (np. różnic w wielkości ładunków czynnikowych we wszystkich grupach). Podejście to jest coraz częściej krytykowane (np. Marsh i in., 2017). Jeden z problemów polega na tym, że indeksy modyfikacyjne, będące podstawą decyzji opierają się na danych obarczonych problemem kolinearności, w związku z czym wyniki wspierające częściową równoważność mogą być niereplikowalne. Pomimo że nie jest to podejście idealne, jest nadal często wykorzystywane w badaniach i zdecydowanie najlepsze (van de Schoot i in., 2013), jeżeli tylko niektóre parametry znacznie różnią się pomiędzy grupami

(np. kiedy ładunki czynnikowe jednego z kilku twierdzeń skali mają wyraźnie większą wartość w porównaniu z resztą twierdzeń).

Podejmując decyzję o ilości uwalnianych parametrów, przyjmuje się, że dopóki przynajmniej dwa ładunki czynnikowe lub stałe regresji mają narzucone parametry równości, można trafnie wnioskować o różnicach średnich latentnych w analizowanym modelu (za: van de Schoot, Lugtig, Hox, 2012). Przy tym jednak badania wskazują, że celem porównania sumy punktów lub średnich zmiennych obserwowalnych (np. średnich arytmetycznych twierdzeń), musimy wykazać pełną równoważność skalarną (Steinmetz, 2013). Częściowym rozwiązaniem tego problemu wydaje się uwolnienie z restrykcji równości mniej niż połowy parametrów puli zmiennych obserwowalnych (np. stałych regresji nie więcej niż 40% pozycji testowych skali) oraz wyłączenie z obliczania wartości średniej arytmetycznej (lub sumy) zmiennych obserwowalnych tych pytań skali, których ładunki lub stałe regresji zaburzały równoważność pomiarową skali (takie rozwiązanie zastosowano w Lubiewska, van de Vijver, 2015).

Trzecie rozwiązanie, przydatne szczególnie, jeżeli zależy nam na porównaniu średnich wyników skali w grupach, polega na przetestowaniu **przybliżonej równoważności pomiarowej** (*approximate measurement invariance*) opartej o modele bayesowskie (Muthen, Asparouhov, 2013; Verhagen, Fox, 2013). Po tę metodę sięga się tylko wtedy, kiedy nie udało się za pomocą wcześniej opisanych metod potwierdzić pełnej równoważności pomiarowej. Strategia ta, w przeciwieństwie do klasycznych estymatorów opartych na metodzie maksymalnego prawdopodobieństwa, zakłada, że stałe regresji czy wartości ładunków czynnikowych nie muszą być identyczne w analizowanych grupach. Przy tym założeniu pozostawia się pewien margines wariancji, na którą pozwala się w zakresie różnic pomiędzy porównywanymi parametrami. Margines wariancji określa się przed przetestowaniem modelu poprzez sformułowanie, na podstawie dostępnej badaczowi wiedzy, parametrów rozkładu *a priori*, który konfrontuje się potem z danymi uzyskując parametry rozkładu *a posteriori*, wskazujące czy nasz model zakładający pewien margines zmienności ma poparcie w danych w obliczu przyjętych założeń teoretycznych (*a priori*). Decyzje o przyjęciu modelu opiera się na wielkości parametru DIC (*deviance information criterion*) oraz wartości *posterior predictive p-value* (*ppp*), który powinien być większy lub zbliżony do ,050 (np. Verhagen, Fox, 2013). Jeżeli metoda przybliżonej równoważności pomiarowej wykaże brak równoważności pomiarowej na testowanym poziomie, można na podstawie analizy tego, które ładunki czynnikowe lub stałe regresji były różne w analizowanych grupach, uwolnić te parametry, które były różne z restrykcji równości, czyli zastosować strategię **częściowej przybliżonej równoważności pomiarowej** (np. van de Schoot i in., 2013; Zercher i in., 2015).

Metody symulacji Monte Carlo wykazały, że ta metoda oceny równoważności pomiarowej trafnie szacuje faktyczną wariancję danych zwiększając przy tym szanse na potwierdzenie danego poziomu równoważności pomiarowej (van de Schoot i in., 2013). Badania van de Schoota i współpracowników (2013) porównujące skuteczność

metody testowania częściowej równoważności oraz przybliżonej równoważności pomiarowej w zakresie równoważności stałych regresji wykazały, że jeżeli występują małe różnice w obrębie stałych regresji wielu wskaźników obserwowalnych zmiennej latentnej metoda przybliżonej równoważności pomiarowej sprawuje się lepiej aniżeli metoda częściowej równoważności. Ta ostatnia sprawuje się jednak lepiej w sytuacji, kiedy wartości stałych regresji nielicznych twierdzeń znacznie różnią się od wartości stałych regresji pozostałych twierdzeń.

W końcu, w sytuacji kiedy badacz chce testować w swoich badaniach hipotezy główne dotyczące różnic średnich wyników, zaś przeprowadzone przez niego analizy równoważności skalarnej wykazują brak wsparcia dla wniosku o chociażby częściowej równoważności skalarnej, może on zastosować metodę **wyrównywania** (*alignment*) zaproponowaną przez Muthén i Asparouhov (2013) przy użyciu estymatorów ML lub bayesowskich. Jest ona opracowana dla baz danych zawierających wiele grup. Wyrównywanie może być zastosowane do porównania średnich wartości czynników latentnych nawet jeżeli nie ma poparcia dla równoważności skalarnej. Metoda wyrównywania nie zakłada równoważności pomiarowej, a zamiast tego wykorzystując funkcję upraszczania poszukuje optymalnego wzorca równoważności pomiarowej. Funkcja ta jest podobna do kryteriów rotowania czynników w eksploracyjnej analizie czynnikowej, gdzie po rotacji generowane są wielkie lub małe wartości ładunków czynnikowych. W efekcie, możliwe staje się oszacowanie wszystkich parametrów modelu ograniczając zaburzenia równoważności do minimum oraz ocena tego, które parametry modelu są nierównoważne w analizowanych grupach. Jeżeli chcemy zastosować metodę wyrównywania nie tylko do porównywania średnich latentnych, a do innych analiz SEM, można zastosować metodę **wyrównywania-wewnątrz-CFA** (Muthén, Asparouhov, 2013), gdzie testuje się model, jeszcze raz używając wyników uzyskanego modelu jako wartości startowych modelu następnego (Marsh i in., 2017). Jako że opisanie tych bardziej zaawansowanych metod wykracza poza zakres niniejszego artykułu, którego celem jest przeprowadzenie czytelnika przez podstawowe analizy dotyczące równoważności, osoby zainteresowane odsyłamy do tekstów pokazujących przykłady analiz bayesowskich (np. Zercher i in., 2015), wyrównywania (np. van de Schoot i in., 2013; Asparouhov, Muthen, 2014) czy wyrównywania-wewnątrz-CFA (np. Marsh i in., 2017). Ponadto warto dodać, że możliwe jest również szukanie przyczyn braku równoważności skalarnej, która może wiązać się z moderującą rolą zmiennych drugiego stopnia (np. kultury). W tym celu można wykorzystać np. analizę wielopoziomowych modeli strukturalnych (*multilevel SEM*), wyjaśniając dlaczego mamy do czynienia w naszych danych z brakiem równoważności (Davidov i in., 2012).

## Przykładowa analiza równoważności skali

Poniżej opiszemy strategię analizy każdego z poziomów równoważności, dodatkowo podając komendy pakietu 'lavaan', który wykorzystamy do przeprowa-

dzenia analiz. Jako strategię radzenia sobie z brakiem potwierdzenia równoważności zastosujemy metodę częściowej równoważności, która pomimo ograniczeń jest dość przystępną i nadal często wykorzystywaną metodą radzenia sobie z problemem braku równoważności metrycznej czy skalarnej. W prezentowanym przykładzie przeanalizujemy skalę przywiązania złożoną z dwóch związanych ze sobą podskal Unikania i Niepokoju (skala *Adult Attachment Scale*, AAS; Collins, Read, 1990), którą przetestujemy w zakresie równoważności konfiguralnej, metrycznej i skalarnej. Analizy przeprowadzimy na danych zebranych od nastolatków w trzech krajach: w Niemczech, Turcji i Polsce (analizy te są przykładowe i nie były dotychczas publikowane).

**Krok 1. Ustanowienie modelu konfiguralnego i analiza równoważności konfiguralnej**

Analizę równoważności konfiguralnej w naszym przykładowym badaniu rozpoczęłyśmy od oddzielnej analizy modelu pomiarowego w każdej badanej grupie (Meade, Johnson, Braddy, 2008) w zakresie: ilości czynników latentnych skali; wzorca ładunków czynnikowych (wielkości i związku z czynnikiem głównym); oraz korelacji pomiędzy czynnikami latentnymi. Robiąc to przeanalizowałyśmy dane zapisane w osobnych dla każdej analizowanej grupy plikach, zaczynając od danych z Niemiec. Celem wprowadzenia tych danych do środowiska R wpisałyśmy następującą komendę[4]:

```
>mydataN<-read.table(„c:/daneR/Niemcy.txt",header=TRUE)
```

Potem zdefiniowałyśmy parametry modelu pomiarowego, który ma być przeanalizowany w pakiecie 'lavaan':

```
>myModel<-'
>Unikanie=~a03+a13+a14+a15+a17+a18
>Niepokój=~a04+a05+a06+a07+a08+a10+a11+a16
>Unikanie~~ Niepokój
>'
```

Celem oszacowania modelu przy wykorzystaniu konfirmacyjnej analizy czynnikowej (CFA) wpisałyśmy polecenie:

```
> fit<-cfa(myModel,data=mydataN)
```

oraz poprosiłyśmy o podsumowanie parametrów oszacowania modelu:

```
>summary(fit,fit.measures=T,standardized=T)
```

Po przeanalizowaniu struktury skali w danych niemieckich, zrobiłyśmy to samo na danych polskich oraz tureckich. Wyniki naszych analiz wykazały, że model ma zadowalające parametry dopasowania do danych w każdej grupie (kraju) oraz wykazuje tę samą strukturę i porównywalne wzorce związku ładunków czynnikowych twierdzeń z czynnikami latentnymi skali. Zakładając na podstawie wyników, że teo-

[4] Język środowiska R nie jest przedmiotem analizy niniejszego artykułu, stąd osoby zainteresowane odsyłam do strony https://www.r-project.org/ oraz opracowań, np. Kopczewska, Kopczewski, Wójcik (2009).

retycznie przez nas założony model pomiarowy trafnie opisuje strukturę konstruktu w każdej grupie, przeszłyśmy do sprawdzenia, jaka jest dobroć dopasowania tego modelu do danych wielogrupowych, który będzie stanowił na dalszych etapach analizy model konfiguralny (Byrne, 2008), od którego zaczniemy testowanie następnych poziomów równoważności pomiarowej naszej skali.

W tym celu połączyłyśmy wszystkie dane w jedną matrycę, w której dane z różnych krajów są wprowadzone wertykalnie oraz zawierają kolumnę ze zmienną grupującą (kraj). Polecenia wprowadzające analizowaną bazę danych do R oraz testujące równoważność konfiguralną wyglądały następująco[5]:

```
>mydata<-read.table(„c:/daneR/Total.txt",header=TRUE)
#plik „Total.txt" zawiera dane wielogrupowe ze wszystkich krajów, wraz z kolumną je identyfikującą
>fit.conf<-cfa(myModel,data=mydata,group="country")
#"country" wskazuje nazwę zmiennej określającej przynależność narodową badanego
>summary(fit.conf,fit.measures=T,standardized=T)
# ta komenda pozwala prześledzić wszystkie parametry modelu pomiarowego w każdej grupie
>fitMeasures(fit.conf)
#ta komenda jest węższa od komendy „summary" i zleca wygenerowanie tylko parametrów dopasowania modelu bazowego (konfiguralnego) do danych. Jest ona przydatna, jeżeli chcemy w wydruku mieć podane tylko te wyniki
```

Wyniki dopasowania tego modelu do danych, zaprezentowane w załączniku 1, wskazują na dobre parametry dopasowania. Na tej podstawie możemy wnioskować, że teoretycznie przyjęty model pomiarowy dobrze opisuje strukturę skali w trzech badanych przez nas grupach.

Analiza równoważności konfiguralnej jest najmniej restrykcyjna ponieważ w modelu nie narzuca się żadnych restrykcji równości parametrów w analizowanych grupach, zaś informacja o przynależności do grupy jest tutaj jedynie częścią modelu (wielogrupowej konfirmacyjnej analizy czynnikowej, *multigroup confirmational factor analysis*, MGCFA).

Parametry dopasowania do danych modelu konfiguralnego (nazwałyśmy go w naszym przykładzie modelem *conf*) stają się w dalszych analizach naszym punktem odniesienia (modelem bazowym) do porównań zmian w parametrach dopasowania modelu jeżeli narzucimy w nim restrykcje równości ładunków czynnikowych zmiennych obserwowalnych testując następny poziom równoważności metrycznej (Byrne, 2008). Często dopiero na etapie analizy dopasowania modelu konfiguralnego do danych, raportuje się indeksy dopasowania modelu w tabeli przedstawiającej

[5] Po znaku # podane są w niniejszym tekście informacje o treści komendy pomocne w jej rozumieniu, należy zatem pamiętać, że nie są to części komend, które należy wprowadzić do R celem dokonania analiz.

wyniki badania (przykład w załączniku 1 oraz w Lubiewska i in., 2016b). Jeżeli nasz model jest dopasowany do danych w sposób, co najmniej akceptowalny możemy przystąpić do przetestowania wyższego poziomu ekwiwalencji pomiarowej, czyli równoważności metrycznej.

**Krok 2: Równoważność metryczna**

W celu przetestowania równoważność jednostki pomiarowej narzuciłyśmy w testowanym przez nas modelu restrykcje równości na parametry ładunków czynnikowych w każdej grupie. W zależności od pakietu statystycznego, w którym przeprowadza się analizy robi się to różnie. Często należy wydać programowi polecenie oszacowania wielkości ładunków czynnikowych wskaźników obserwowalnych (twierdzeń) w jednej (np. największej) z grup przez nas badanych oraz narzucić w pozostałych grupach restrykcję równości ładunków czynnikowych wobec grupy, w której są one oceniane. W pakiecie 'lavaan' środowiska R, dla którego podajemy przykłady wystarczy dopisać komendę – group.equal="loadings" – w poleceniu oszacowania modelu i przeanalizować model ponownie:

```
>fit.metric<-cfa(myModel,data=mydata,group="country",group.equal="loadings")
>fitMeasures(fit.metric)
```

Model ten nazwałyśmy modelem *metric*[6], ponieważ nie jest on tożsamy z modelem wcześniejszym *conf*. Wyniki tej analizy, wskazane w załączniku 1, wskazują na pogorszenie parametrów dopasowania modelu *metric* do danych, w porównaniu z parametr mi modelu *conf*. Biorąc pod uwagę to, że nasza baza danych jest relatywnie duża ($N$ = 1100) powinniśmy przed przystąpieniem do analiz podjąć decyzję o zastosowaniu mniej restrykcyjnych kryteriów decyzji o potwierdzeniu poziomu równoważności, co doprowadziłoby nas do wniosku, że możemy potwierdzić w naszych badaniach poziom równoważności metrycznej ($\Delta$CFI < ,01 i $\Delta$RMSEA = 0). Niemniej celem instruktażowym sprawdziłyśmy także, czy wynik testu istotności różnic chi kwadrat wykaże problemy w zakresie potwierdzenia równoważności metrycznej. W tym celu wpisałyśmy komendę:

```
>anova(fit.conf,fit.metric)
```

Wynik tej analizy (załącznik 1) wykazał, że różnica chi kwadrat jest istotna statystycznie.

Warto tutaj zaznaczyć, że w raportach z badań na początku wskazuje się kryteria decyzyjne, na podstawie których będzie się podejmowało decyzję o potwierdzeniu (lub nie) każdego z poziomów równoważności. Jeżeli przyjmiemy kryteria odcięcia, takie jak $\Delta$CFI czy $\Delta$RMSEA zamiast $\Delta\chi^2$, nie raportujemy wtedy w rezultatach badań wyników testów $\Delta\chi^2$. Analiza przedstawiona przez nas w tym artykule

[6] Należy tutaj zaznaczyć, że nasz model pomiarowy jest jeden i został sprecyzowany w poleceniu „myModel" na początku analiz. Następne modele są tzw. modelami zagnieżdżonymi w tym modelu i nie różnią się od niego niczym, poza restrykcjami równości, które się narzuca w kolejnych analizach komendami typu „fit…<-cfa(…)".

oraz raportowanie wszystkich wyników ma tylko i wyłącznie cel szkoleniowy. Z tego też powodu potraktowałyśmy wynik $\Delta\chi^2$ jako wskazujący na brak równoważności metrycznej i przeszłyśmy do sprawdzenia, czy analiza częściowej równoważności będzie w stanie rozwiązać problem pogorszenia parametrów dopasowania modelu.

W tym celu prześledziłyśmy wielkość ładunków czynnikowych (modelu *conf*) we wszystkich analizowanych przez nas grupach poszukując twierdzeń, które w tym zakresie odstają od reszty. Zrobiłyśmy to korzystając z następującej komendy pakietu 'lavaan':

>fit.metric<-cfa(myModel,data=mydata,group="country",group.equal="loadings")

>summary(fit.metric,fit.measures=T,standardized=T)

#komenda umożliwia analizę ładunków czynnikowych we wszystkich grupach oddzielnie

Jest to metoda „na oko" i przy dużej baterii twierdzeń skali oraz analizowanych grup jest dość pracochłonna. Inną, nieco szybszą metodą jest metoda krokowa polegająca na narzucaniu restrykcji równości ładunków czynnikowych nie na wszystkie twierdzenia na raz (tak jak robiliśmy to, testując pełną równoważność metryczną komendą group.equal="loadings"), ale na każde twierdzenie z osobna (group.equal="loadings" oraz group.partial=…). Porównujemy przy tym jak bardzo przy uwolnieniu restrykcji równości każdego twierdzenia z osobna podwyższa się wartość $\chi^2$ (nie $\Delta\ \chi^2$) w stosunku do modelu wcześniejszego (*conf*). W tym celu wpisujemy następujące komendy 'lavaan':

>fit.Pmetric<-cfa(myModel,data=mydata,group="country",group.equal="loadings", group.partial=c(„Niepokój=~a17"))

# uwolniliśmy z restrykcji równości ładunek czynnikowy twierdzenia nr 17 w modelu z narzuconą restrykcją równości na ładunki czynnikowe wszystkich twierdzeń – ten model nazwaliśmy modelem *Pmetric*

>fitMeasures(fit.Pmetric)

Jeżeli wiemy już, w których grupach które wartości ładunków czynnikowych najbardziej przyczyniają się do obniżenia parametrów dopasowania modelu do danych[7], uwalniamy ładunki czynnikowe tych twierdzeń z restrykcji równości i sprawdzamy różnice pomiędzy modelami, do momentu kiedy $\Delta\chi^2$ nie wykaże istotnego statystycznie pogorszenia parametrów dopasowania modelu częściowej równoważności metrycznej (*Pmetric*) wobec modelu konfirmacyjnego (*conf*). W naszym przykładowym badaniu uwolnienie z restrykcji równości ładunku czynnikowego twierdzenia nr 10 wystarczyło do tego, żeby wynik testu ANOVA badający wielkość $\Delta\chi^2$ przestał być istotny statystycznie. Komendy z tą analizą związane przedstawiają się następująco:

[7] Wybieramy te twierdzenia, dla których wartość $\chi^2$ modelu oszacowanego oddzielnie dla każdego z nich jest najniższa w porównaniu w porównaniu z $\chi^2$ modelu konfiguralnego.

> fit.Pmetric<-cfa(myModel,data=mydata, group="country",group.equal="loadings",group.partial=c(„Niepokój=~a10"))

>fitMeasures(fit.Pmetric)

>anova(fit.conf,fit.Pmetric)

Jeżeli potwierdziliśmy równoważność metryczną (przynajmniej częściową) naszej skali, możemy przejść do następnego etapu testowania równoważności pomiarowej, jakim jest analiza równoważności skalarnej.

Jeżeli jednak, tak jak w naszym przykładzie, mamy model pomiarowy z dwoma czynnikami latentnymi (dwoma podskalami), które są połączone kowariancją, na tym etapie możemy przetestować również **równoważność kowariancji**, zwaną równoważnością strukturalną (*structural equivalence*) (Byrne, 2008). Testujemy ją dodając do dotychczas narzuconych restrykcji równości polecenie „lv.covariances" i sprawdzając istotność różnicy pomiędzy dopasowaniem modelu częściowej równoważności metrycznej (*Pmetric*) i tego samego modelu z narzuconą równością kowariancji (nazwijmy go modelem *cov*) do danych. W naszym przykładzie komenda pakietu 'lavaan' wyglądała następująco:

> fit.cov<-cfa(myModel,data=mydata, ,group="country", group.equal=c(„loadings","lv.covariances"),group.partial=c(„Niepokój=~a10"))

>anova(fit.Pmetric,fit.cov)

Wyniki w załączniku 1 wskazują na pełną równoważność kowariancji pomiędzy czynnikami latentnymi w analizowanych przez nas grupach. W związku z czym możemy przejść do testowania najwyższego poziomu równoważności pomiarowej.

**Krok 3: Równoważność skalarna**

Na tym etapie analiz do restrykcji równości ładunków czynnikowych i kowariancji dodajemy restrykcje równości stałych regresji (*intercepts*) wszystkich wskaźników obserwowalnych (twierdzeń) skali. W tym celu postępujemy podobnie jak wcześniej. W pakiecie 'lavaan' do komendy – group.equal=c(„loadings","lv.covariances") – dopisujemy „intercepts", co daje nam bardziej rozbudowaną komendę oszacowania naszego modelu pomiarowego (nazwanego modelem *scalar*):

> fit.scalar<-cfa(myModel,data=mydata,group="country", group.equal=c(„loadings","lv.covariances","intercepts"), group.partial=c(„Niepokój=~a10"))

>fitMeasures(fit.scalar)

Aby wnioskować o potwierdzeniu lub braku potwierdzenia równoważności skalarnej, ponownie oszacowałyśmy istotność statystyczną różnicy pomiędzy dopasowaniem do danych nowego (*scalar*) i wcześniej oszacowanego modelu (*cov*):

>anova(fit.cov,fit.scalar)

Wyniki naszych analiz, przedstawione w załączniku 1, wykazały znaczne pogorszenie parametrów dopasowania modelu *scalar* w porównaniu z modelem *cov*. Oznacza to, że nie możemy potwierdzić równoważności skalarnej analizowanej przez nas skali.

W tej sytuacji możemy spróbować sprawdzić jakie są warunki uzyskania częściowej równoważności skalarnej i dopiero na tej podstawie podjąć decyzję o: (1) teore-

tycznej i empirycznej sensowności traktowania wyniku jako podstawy do przystąpienia do porównywania średnich wyników w grupach; (2) zaprzestaniu dalszych analiz z wnioskiem o potwierdzeniu równoważności metrycznej (ale nie skalarnej); (3) przystąpieniu do analiz bayesowskich testujących przybliżoną równoważność pomiarową (np. przy użyciu pakietu 'blavaan') lub (4) porównania średnich po zastosowaniu metody wyrównywania (*alignment*).

Procedura testowania częściowej równoważności skalarnej jest niemal identyczna do testowania częściowej równoważności metrycznej, niemniej w tym przypadku musimy przeanalizować, które ze stałych regresji zmiennych obserwowalnych przejawiają wyraźnie odmienne wartości we wszystkich analizowanych grupach. W naszym badaniu zrobiłyśmy to poprzez uwalnianie pojedynczo stałych regresji każdego z pytań i testowanie zmian w zakresie wartości $\Delta\chi^2$. W pakiecie 'lavaan' wygląda to następująco:

```
> fit.Pscalar<-cfa(myModel,data=mydata,group="country", group.equal=c(„loadings","lv.covariances","intercepts"), group.partial=c(„Niepokój=~a10", „a03~1"))
#tutaj uwolniliśmy z restrykcji równości stałą regresji pytania nr 3
>fitMeasures(fit.Pscalar)
```

Niestety, analizy wykazały, że w naszych badaniach nie udało się uzyskać częściowej ekwiwalencji skalarnej. Zaprzestaliśmy dalszych prób w momencie, gdy uwolnienie 60% stałych regresji wskaźników obserwowalnych (pytań skali) w naszym modelu nadal nie pozwoliło na uzyskanie nieistotnej statystycznie różnicy pomiędzy modelem z restrykcją równości kowariancji (*cov*) i nowym modelem z częściowymi restrykcjami stałych regresji narzuconymi na 40% pytań skali.

Jeżeli podejmiemy decyzję o konieczności wnioskowania tylko o uzyskaniu w naszym badaniu równoważności metrycznej, tak jak w analizowanym przykładzie, możemy na tej podstawie w ramach hipotez głównych (modeli strukturalnych) testować związki pomiędzy zmiennymi, ale nie mamy uprawnień do porównywania średnich wyników w różnych grupach (Steinmetz, 2013). Przykłady konieczności podjęcia takiej decyzji wcale nie należą do rzadkości (np. Datta, Marcoen, Poortinga, 2005; Lubiewska i in., w recenzji). W jednym z badań chcąc przetestować tezę o różnicach międzykulturowych i międzypokoleniowych w poziomie przywiązania dorosłych przeanalizowaliśmy na wstępnym etapie analiz ekwiwalencję pomiarową skali przywiązania złożonej z 16 pozycji testowych w 39 grupach zróżnicowanych przynależnością do pokolenia i kultury (Lubiewska i in., w recenzji). Niestety, cała praca poszła na marne ponieważ po narzuceniu częściowych restrykcji celem uzyskania równoważności skalarnej udało nam się pozostać tylko z trzema pozycjami testowymi, których wyniki moglibyśmy porównać we wszystkich grupach. Tego typu rozwiązanie nie miałoby jednak sensu ani teoretycznego, ani empirycznego. Dodam przy tym, że równoważności skali nie udało nam się uzyskać również wewnątrz niektórych krajów, pomiędzy trzema grupami wiekowymi (pokoleniami).

## W pogoni za własnym ogonem

Analiza równoważności, szczególnie przy testowaniu złożonych modeli oraz dużych wielogrupowych baz danych jest zwykle bardzo czasochłonna. Należy zaznaczyć, że analiza równoważności pomiarowej jest tylko analizą wstępną, uprawniającą do testowania hipotez głównych, zaś potrafi zająć 95% czasu poświęconego na całą analizę danych. W dodatku, jak widzimy na naszym przykładzie, niejednokrotnie nie jest ona zakończona sukcesem a decyzją o zaprzestaniu kontynuacji analiz głównych. Wśród niektórych badaczy (np. Boehnke, 2012; Welzel, Inglehart, 2016) może pojawić się zatem wątpliwość dotycząca tego, czy czasami „goniąc za własnym ogonem", nie blokujemy rozwoju nauki, której potrzebne są tak badania, jak i ich replikacje w wielu zróżnicowanych próbach pochodzących z różnych populacji, w których często równoważności skali nie jesteśmy w stanie potwierdzić. Choć rozwianie tych wątpliwości nie jest proste, warto zwrócić uwagę na kilka kwestii.

Z jednej strony należy przyznać, że standardy wymagające analizy równoważności pomiarowej z pewnością komplikują badaczom życie (a dokładniej robią to recenzenci wymagający dowodu równoważności pomiarowej skal użytych w raportowanym badaniu). Przeprowadzenie analizy równoważności wymaga wiedzy i zręczności w analizie danych ilościowych, co podwyższa standardy wobec tak raportów z badań, jak i szkolenia doktorantów. Z drugiej jednak strony, jeżeli popatrzymy na rozwój nauki, to postęp wiąże się właśnie z podwyższaniem standardów, np. w zakresie precyzji analiz czy stosowanych narzędzi pomiarowych.

Odpowiadając na pytanie, czy nie tracimy czasu na „gonienie za własnym ogonem", należy przede wszystkim przeanalizować, co dla psychologii oznacza wymóg przeprowadzania analizy ekwiwalencji pomiarowej w badaniach. Po pierwsze, analiza równoważności pomiarowej skal samoopisowych wydaje się być częściową odpowiedzią na zarzuty formułowane wobec pomiaru kwestionariuszowego (tam analiza równoważności najczęściej jest stosowana). Badacze przyzwyczaili się już do masowego przepraszania w swoich raportach z badań opartych na pomiarze kwestionariuszowym za to, że nie zastosowali w swoim badaniu pomiaru obserwacyjnego lub eksperymentalnego (sekcja „ograniczenia badań"). Postęp jednak nie wiąże się z przepraszaniem, a z eliminacją napotkanych problemów oraz ograniczeń.

Narzędzia samoopisowe dostarczają wiedzy na temat tego, co ludzie myślą i czują, jak postrzegają siebie i świat. Jest to wiedza, której nie uzyskamy podczas obserwacji czy manipulacji eksperymentalnej. Mają one rozliczne wady, do których m.in. zaliczyć należy narzucanie formatu odpowiedzi (poprzez skalę Likerta i sformułowane już twierdzenie). Niemniej są dobrym papierkiem lakmusowym do testowania trendów populacyjnych w badaniach psychologicznych, które mogą być dalej poddawane mikroanalizie obserwacyjnej, eksperymentalnej czy tej związanej z wywiadem pogłębionym. Stąd warto je rozwijać, w czym pomocna jest m.in. analiza równoważności pomiarowej czy też analiza wspólnej wariancji metody pomiaru. Dzisiaj do tego nie wystarczy już tylko analiza rzetelności skali w nowej próbie, w której skalę stosujemy.

Ponadto analiza równoważności odpowiada na ważne pytania dotyczące tego, czy analizowany przez nas konstrukt wskazywany przez naszą pulę twierdzeń testowych ma tę samą strukturę, czy metoda pomiaru jest ta sama oraz czy badani traktują skalę oraz dane jej twierdzenie podobnie, odpowiadając na nie. Przez to wyniki naszych badań dostarczają znacznie pewniejszych i mocniejszych wniosków. Odkrywając brak równoważności pomiarowej w stosowanym przez nas narzędziu, mamy pewność, że nie wygenerujemy wniosków, którym nie można zaufać. Brak uzyskania równoważności pomiarowej skali mówi nam ważną rzecz „w tym zakresie ludzie w badanych grupach się różnią". Wykazując to, przyczyniamy się do powszechnego zrozumienia ważnego i długo ignorowanego faktu – konstrukty psychologiczne nie są całkowicie uniwersalne i różnią się w zależności od kontekstu i próby, w której są mierzone, zaś narzędzia pomiarowe nie mają stałych parametrów. Ta sama pozycja testowa może być świetnym wskaźnikiem konstruktu w jednej grupie i nieprzydatnym w innej.

Między innymi dzięki analizie równoważności pomiarowej potrafimy dokładniej zdiagnozować, czym i dlaczego analizowane konstrukty się różnią w odmiennych grupach i czy tak samo powinny być mierzone. Inną analizą, która jest w tym zakresie przydatna niezależnie od przynależności grupowej badanych, jest analiza wspólnej wariancji metody (np. Lubiewska i in., 2016a). Warto dodać, że wspólna wariancja metody, wykryta w skali mierzącej dany konstrukt, może wpływać na wyniki analizy równoważności pomiarowej (Butts, Vandenberg, Williams, 2006). Obie analizy dostarczają nam możliwość precyzyjnej kalibracji narzędzia w każdym badaniu. Czym większą ilością analiz diagnostycznych dotyczących narzędzi pomiarowych w psychologii dysponujemy, tym bardziej godne zaufania są nasze wyniki, przez co i większy staje się wkład naszych badań do głównego nurtu wiedzy psychologicznej.

Problem, który pozostaje na tym etapie naszej wiedzy dotyczy raczej nie tego, czy dbać o precyzję naszych badań poprzez stosowanie analizy równoważności narzędzi pomiarowych, lecz raczej, co zrobić, jeżeli nie udaje nam się wykazać równoważności pomiarowej w badaniu. Pytanie to dotyczy właściwie nie tyle konkretnej skali, co w ogóle wskaźników badanych przez nas konstruktów, np. twierdzeń kwestionariusza. W obrębie badań psychologii międzykulturowej Boehnke (2012) zaproponował porzucenie w tym zakresie analiz statystycznych dokonywanych w różnych kulturach z użyciem tych samych skal typu *etic* (powszechnie znanych, których większość powstała na Zachodzie i jest tłumaczona na inne języki), na rzecz badań typu *emic*, w których wskaźniki, np. twierdzenia skali, mierzące dany konstrukt są rozwijane oddzielnie w każdej analizowanej kulturze, rozpoczynając od analizy natury konstruktu (np. Boski, 2009). Do tego typu analiz potrzebne są wywiady, a często też badania psycholeksykalne lub obserwacyjne w każdej analizowanej kulturze. Są to niezwykle cenne analizy, ponieważ dzięki takim właśnie badaniom dowiedzieliśmy się, że należy dopełnić dobrze znany pięcioczynnikowy model osobowości o nowe wymiary osobowości w Chinach (Cheung i in., 2001) czy

w Południowej Afryce (np. Nel i in., 2012). Niemniej podejście *emic*, czy też połączenie badań typu *emic* i *etic*, jest niezwykle kosztowne oraz czasochłonne. Badania tego typu dostarczają cennej wiedzy, niemniej są poza zasięgiem większości badaczy.

Ponadto analizy typu *emic* nie są w stanie powiedzieć nam jak równoważność semantyczna (pozycji testowej skali znanej, powszechnie stosowanej, typu *etic*) przekłada się na równoważność matematyczną, pozwalającą na zrozumienie, czym się różni i w czym jest podobna rzeczywistość psychologiczna badanych w różnych grupach (van de Vijver, 2012). Myśląc perspektywicznie, warto dysponować pulą wskaźników danego konstruktu, które dobrze sprawdzają się w większości analizowanych grup oraz pulą wskaźników tegoż konstruktu, które są kontekstualnie-zależne. Pomimo tego, że daleko nam jeszcze do tego miejsca, od czegoś trzeba zacząć. Wymóg analizy równoważności pomiarowej wydaje się być właśnie tym miejscem i standardem

**Literatura cytowana**


Alessandri, G., Vecchione, M., Eisenberg, N., Łaguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychological Assessment*, 27, 621-635, doi: 10.1037/pas0000073

Boehnke, K. (2012). On Comparing Apples and Oranges: Towards a Quantitative Emic Cross-Cultural Psychology. *Baltic Journal of Psychology*, 13 (1), 8-15.

Boski, P. (2009). *Kulturowe ramy zachowań społecznych*. Warszawa: Wydawnictwo Naukowe PWN i Academica.

Browne, M.W., Cudeck, R. (1993). Alternative ways of assessing model fit. W: K.A. Bollen, J.S. Long (red.), *Testing structural equation models* (s. 136-162). Newbury Park, CA: Sage.

Butts, M.M., Vandenberg, R.J., Williams, L.J. (2006). Investigating the susceptibility of measurement invariance test: The effects of common method variance. *Academy of Management Proceedings*, 1, D1-D6, doi: 10.5465/AMBPP.2006.27182126

Byrne, B.M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, 20, 872-882.

Byrne, B.M., Shavelson, R.J., Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105 (3), 456-466, doi: 10.1037/0033-2909.105.3.456

Byrne, M.B., van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of non-equivalence. *International Journal of Testing*, 10, 107-132.

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, doi: 10.1080/10705510701301834

Cheung, F.M., Leung, K., Zhang, J.X., Sun, H.F., Gan, Y.Q., Song, W.Z., Xie, D. (2001). Indigenous Chinese Personality Constructs: Is the Five-Factor Model Complete? *Journal of Cross-Cultural Psychology*, 32 (4), 407-433, doi: 10.1177/0022022101032004003

Cheung, G.W. (2007). Testing Equivalence in the Structure, Means, and Variances of Higher-Order Constructs with Structural Equation Modeling. *Organizational Research Methods*, 11, 593-613, doi: 10.1177/1094428106298973

Cheung, G.W., Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 13, 531-542.

Cieciuch, J., Davidov, E. (2015). Establishing measurement invariance across online and offline samples. A tutorial with the software packages Amos and MPlus. *Studia Psychologica*, 15, 83-99.

Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C., Schwartz, S.H. (2014). The Cross-National Invariance Properties of a New Scale to Measure 19 Basic Human Values: A Test Across Eight Countries. *Journal of Cross-Cultural Psychology*, 45, 764-776, doi: 10.1177/0022022114527348

Collins, N.L., Read, J.R. (1990). Adult attachment, working models, and relationship quality in dating couples. *Journal of Personality and Social Psychology*, 58 (4), 644--663, doi: 10.1037/0022-3514.58.4.644

Datta, P., Marcoen, A., Poortinga, Y.H. (2005). Recalled early maternal bonding and mother- and self-related attitudes in young adult daughters: A cross-cultural study in India and Belgium. *International Journal of Psychology*, 40, 324-338, doi: 10.1080/00207590444000366

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, 43, 558-575, doi: 10.1177/0022022112438397

De Beuckelaer, A., Lievens, F. (2009). Measurement equivalence of paper-and-pencil and internet organisational surveys: A large scale examination in 16 countries. *Applied Psychology: An International Review*, 58, 336-361, doi: 10.1111/j.1464-0597.2008.00350.x

Haltigan, J.D., Leerkes, E.M., Wong, M.S., Fortuna, K., Roisman, G.I., Supple, A.J., …, Plamondon, A. (2014). Adult Attachment States of Mind: Measurement Invariance Across Ethnicity and Associations With Maternal Sensitivity. *Child Development*, 85, 1019-1035, doi: 10.1111/cdev.12180

Hu, L.T., Bentler, P.M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55, doi: 10.1080/10705519909540118

Hui, C.H., Triandis, H.C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, 16, 131-152.

Kline, R.B. (2005). *Principles and practice of structural equation modelling* (wyd. 2). New York: The Guilford Press.

Knight, G.P., Zerr, A.A. (2010a). Introduction to the special section: Measurement equivalence in child development research. *Child Development Perspectives*, 4, 1-4, doi: 10.1111/j.1750-8606.2009.00112.x

Knight, G.P., Zerr, A.A. (2010b). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, 4, 25-30, doi: 10.1111/j.1750-8606.2009.00112.x

Kopczewska, K., Kopczewski, T., Wójcik, P. (2009). *Metody ilościowe w R. aplikacje ekonomiczne i finansowe*. Warszawa: CeDeWu.
Lubiewska, K., Głogowska, K., Mickiewicz, K., Wojtynkiewicz, E., Wiśniewski, C., Izdebski, P. (2016a). Skala Experience in Close Relationships-Revised: Struktura, Rzetelność oraz Skrócona Wersja Skali w Polskiej Próbie. *Psychologia Rozwojowa*, 21, 49-63, doi: 10.4467/20843879PR.16.004.4793
Lubiewska, K., Mayer, B., Albert, I., Trommsdorff, G. (w recenzji). Relations between parenting and adolescents' attachment in diverse cultures.
Lubiewska, K., van de Vijver, F.J.R. (2015). *Attachment types or dimensions: Evidence from the Adult Attachment Scale across three generations*. Nieopublikowany manuskrypt.
Lubiewska, K., Wojtynkiewicz, E., Głogowska, K., Mickiewicz, K., Wiśneiwski, C., Izdebski, P. (2016b). Ekwiwalencja pomiarowa skali *Experience in Close Relationships-Revised* w grupach zróżnicowanych pod względem wieku oraz płci badanych. *Przegląd Psychologiczny*, 59, 245-262.
Marsh, H.W., Kingdom, U., Guo, J., Parker, P., Nagengast, B., Asparouhov, T., …, Dicke, T. (2017). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparisons of latent means across many groups. *Psychological Methods*, doi: 10.1037/met0000113.
Meade, A.W., Johnson, E.C., Braddy, P.W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, 93, 568-592, doi: 10.1037/0021-9010.93.3.568
Merkle, E.C., Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *arXiv*. Retrieved from http://arxiv.org/abs/1511.05604
Muthén, B., Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313-335, doi: 10.1037/a0026802
Nel, J.A., Valchev, V.H., Rothmann, S., van de Vijver, F.J.R., Meiring, D., de Bruin, G.P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality*, 80 (4), 915-948, doi: 10.1111/j.1467-6494.2011.00751.x
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716, doi: 10.1126/science.aac4716
Pentz, M.A., Chou, C.P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, 62, 450-462, doi: 10.1037/0022-006X.62.3.450
Pokropek, A. (2018). Wybrane statystyczne metody radzenia sobie z brakami danych. *Polskie Forum Psychologiczne*, 23 (2), 291-310, doi: 10.14656/PFP20180205
Razmus, W., Mielniczuk, E. (2018). Błąd wspólnej metody w badaniach kwestionariuszowych. *Polskie Forum Psychologiczne*, 23 (2), 277-290, doi: 10.14656/PFP20180204
Reynolds, C.R., Suzuki, L. (2013). Bias in psychological assessment: An empirical review and recommendations. W J. R. Graham, J. A. Naglieri, I. B. Weiner
Rohner, E.C., Rohner, R.P., Roll, S. (1980). Perceived parental acceptance-rejection and children's reported behavioral dispositions. A comparative and intacultural study of American and Mexican children. *Journal of Cross Cultural Psychology*, 11 (2), 213-231, doi: 10.1177/0022022180112006.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48, 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Różycka-Tran, J., Boski, P., Wojciszke, B. (2014). Wiara w grę o sumie zerowej jako aksjomat społeczny: Badanie w 37 krajach. *Psychologia Społeczna*, 9, 92-109, doi: 10.7366/189618002014280106

Schmitt, N., Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18, 210-222, doi: 10.1016/j.hrmr.2008.03.003

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, 9 (1), 1-12, doi: 10.1027/1614-2241/a000049

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4, 770, doi: 10.3389/fpsyg.2013.00770

Van de Schoot, R., Lugtig, P., Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9 (4), 486-492, doi: 10.1080/17405629.2012.686740

Van de Vijver, F.J.R. (2012). Should We Develop a Quantitative Emic Cross-Cultural Psychology? *Baltic Journal of Psychology*, 13, 16-18.

Van de Vijver, F.J.R., Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology*, 79, 852-859.

Van de Vijver, F.J.R., Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.

Vandenberg, R., Lance, C.E. (2000) A review and synthesis of the measurement invariance literature: Suggestion, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-69.

Verhagen, A.J., Fox, J.P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, 66 (3), 383-401, doi: 10.1111/j.2044-8317.2012.02059.x

Welkenhuysen-Gybels, J.G.J., van de Vijver, F.J.R. (2001). A comparison of methods for the evaluation of construct equivalence in a multigroup setting. *Proceedings of the Annual Meeting of the American Statistical Association*, 9, 357-371.

Welzel, C., Inglehart, R.F. (2016). Misconceptions of Measurement Equivalence: Time for a Paradigm Shift. *Comparative Political Studies*, 1-27, doi: 10.1177/0010414016628275

Zawadzki, B. (2006). *Kwestionariusze osobowości. Strategie i procedury konstruowania.* Warszawa: Wydawnictwo Naukowe Scholar.

Zercher, F., Schmidt, P., Cieciuch, J., Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology*, 6 (733), 207-216, doi: 10.3389/fpsycg.2015.00733

Załącznik 1. Przykład tabeli raportującej wyniki analizy równoważności pomiarowej skali unikania i niepokoju przywiązaniowego (AAS) w Turcji, Niemczech i Polsce

| Poziom równoważności skali AAS | *df* | $\chi 2$ | RMSEA | 90%CI | CFI | SRMR | $\Delta\chi^2$ (*df*) |
|---|---|---|---|---|---|---|---|
| Model bazowy (konfiguralny) | 129 | 267,582 | ,052 | ,044; ,060 | ,952 | ,046 | - |
| Równoważność metryczna | 147 | 382,119 | ,052 | ,045; ,059 | ,946 | ,055 | 36,380** (18) |
| Częściowa równoważność metryczna (uwolniony a10) | 145 | 292,939 | ,051 | ,043; ,058 | ,949 | ,052 | 23,096 (16) |
| Równoważność kowariancji czynników latentnych | 147 | 294,922 | ,050 | ,043; ,058 | ,949 | ,055 | 1,4832 (2) |
| Równoważność skalarna | 165 | 562,558 | ,078 | ,071; ,084 | ,863 | ,074 | 613,4*** (18) |

Wszystkie wartości $\chi 2$ są istotne statystycznie na poziomie $p < ,001$.
* $p < ,05$; ** $p < ,01$; *** $p < ,001$.

**Streszczenie**. Analiza równoważności pomiarowej staje się powoli lecz konsekwentnie standardem analiz ilościowych w badaniach psychologicznych. Jej celem jest sprawdzenie czy stosowane przez badacza narzędzie pomiarowe (np. skala) ma takie same właściwości pomiarowe w analizowanych przez niego grupach (np. zróżnicowanych w zakresie wieku, kultury czy formy badania). Wykazanie na wstępnym etapie analiz określonego poziomu równoważności pomiarowej uprawnia do dalszego testowania głównych hipotez badawczych dotyczących związków pomiędzy zmiennymi lub różnic średnich badanych konstruktów. W artykule dokonujemy przeglądu strategii i problemów związanych z analizą równoważności pomiarowej celem szerszego przybliżenia metody polskim badaczom. Najpierw opisujemy poziomy równoważności pomiarowej oraz kryteria decyzyjne związane z jej potwierdzeniem. Przedstawiamy również strategie radzenia sobie z brakiem równoważności wskazując na metody bayesowskie. W końcu, prezentujemy przykład przeprowadzenia analiz krok po kroku przy użyciu pakietu 'lavaan' środowiska R. Podsumowując, poddajemy refleksji znaczenie tego rodzaju analiz dla badań psychologicznych. Postulujemy przy tym, że analiza równoważności pomiarowej nie ogranicza badaczy, a raczej zwiększa precyzję wnioskowania w badaniach psychologicznych.
**Słowa kluczowe**: równoważność pomiarowa, psychometria, badania międzykulturowe, psychologia porównawcza

Data wpłynięcia: 5.10.2017
Data wpłynięcia po poprawkach: 15.02.2018
Data zatwierdzenia tekstu do druku: 31.03.2018