

Adam Sagan

Uniwersytet Ekonomiczny w Krakowie

RZETELNOŚĆ I FUNKCJONOWANIE POZYCJI SKALI A UKRYTA NIEJEDNORODNOŚĆ POPULACJI

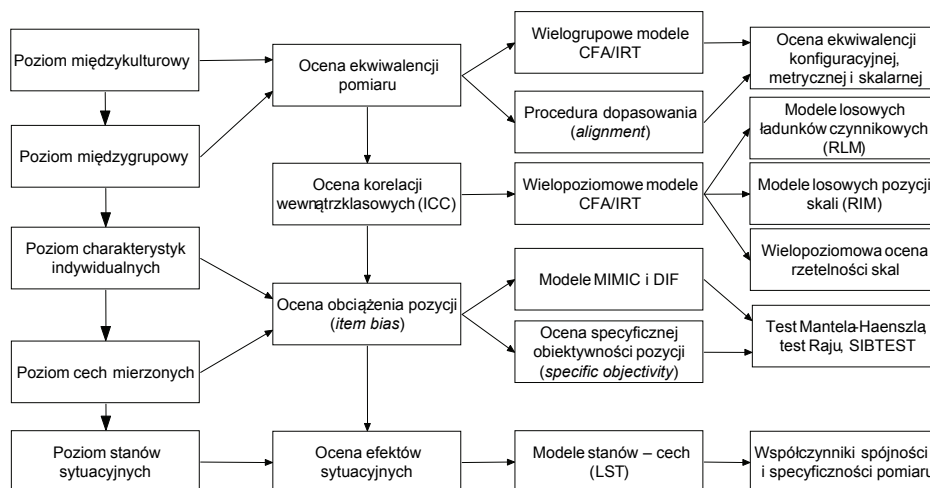
Wprowadzenie

Do podstawowych źródeł obciążenia rzetelności pomiaru (i tym samym wszystkich innych parametrów modeli statystycznych wynikających z występowania błędów w zmiennych) należą: a) hierarchiczne czynniki instytucjonalne wywołujące współzależność obserwacji i występowanie silnych korelacji wewnątrzklasowych (zagnieżdżenie obserwacji), b) przynależność respondentów do grup narodowościowych, kręgów kulturowych i klas społecznych (wpływ systematycznych czynników kulturowych), c) cechy społeczno-demograficzne badanych, d) czynniki sytuacyjne i kontekstowe pomiaru oraz e) styl odpowiadania na pozycje skali przez respondentów (kształtowanie obrazu siebie przez respondenta).

W celu oceny siły oddziaływania powyższych czynników na rzetelność pomiaru stosowanych jest wiele podejść badawczych. Analiza różnic międzykulturowych jest najczęściej dokonywana na podstawie wielogrupowych modeli confirmacyjnej analizy czynnikowej. Wpływ czynników instytucjonalnych, wynikających z mikrospołecznego lub ekonomicznego zróżnicowania odpowiedzi (przynależność respondentów do lokalnych jednostek terytorialnych, gospodarstw domowych, ośrodków handlowych, szkół, szpitali, pierwotnych jednostek losowania w doborze zespołowym itp.) jest uwzględniana w wielopoziomowych modelach czynnikowych z efektami stałymi i losowymi (losowymi ładunkami i wyrazami wolnymi). Na poziomie cech społeczno-demograficznych respondentów kontrola rzetelności pomiaru jest dokonywana na podstawie modeli zróżnicowanego funkcjonowania pozycji skali (DIF) oraz modeli wielorakich wskaźników – wielorakich skutków (MIMIC). Czynniki kontekstowe i sytuacyjne oraz wynikająca z nich dekompozycja rzetelności skal na składową stałą

i sytuacyjną jest podstawą oceny wskaźników spójności i specyficzności pomiaru w modelach stanów – cech (LS-T). Zróznicowanie parametrów modelu pomiarowego może wynikać z pomiaru samej cechy mierzonej. Diagnoza wpływu poziomu cechy mierzonej na charakterystyki modelu pomiaru jest możliwa w analizie specyficznej obiektywności skal Rascha (*objective specificity*).

Rys. 1 przedstawia strukturę błędów pomiarowych i metody ich diagnozy.



Rys. 1. Struktura źródeł rzetelności pomiaru i metody jej oceny

Celem artykułu jest analiza źródeł obciążenia jakości skal pomiarowych wynikających z hierarchicznych źródeł heterogeniczności populacji i mających wpływ na obciążenie parametrów ścieżkowych w modelu strukturalnym. Obszar ten jest bardzo często pomijany w analizach źródeł tzw. błędów w zmiennych, które najczęściej uwzględniają nierzetelność wskaźników, bez korekty dotyczącej błędu systematycznego w oszacowaniach tychże współczynników rzetelności.

Ocena siły obciążenia pozycji i stopnia niejednorodności współczynników rzetelności zostanie dokonana na ogólnopolskiej zagnieżdżonej próbie obszarowo-kwotowej 1100 respondentów zlokalizowanych w 440 gospodarstwach domowych (badaniami objęto ojca, matkę i najstarsze dziecko w rodzinie) na podstawie:

- 1) wielopoziomowych współczynników rzetelności skali stylu podejmowania decyzji w gospodarstwach domowych (skala DMS),
- 2) oceny stopnia obciążenia współczynników rzetelności w układach wielogrupowych (procedura wyrównywania – *alignment*),
- 3) kontroli wpływu cech społeczno-ekonomicznych respondentów w modelach *MIMIC*,

4) parametrycznych i nieparametrycznych metod oceny zróżnicowanego funkcjonowania pozycji skal (*DIF*).

Analizowaną skalą jest skala altruistycznego stylu podejmowania decyzji w rodzinie (ASPD)¹.

1. Analiza inwariancji pomiaru w wielogrupowych modelach confirmacyjnej analizy czynnikowej

W ocenie międzykulturowej ekwiwalencji pomiaru można wyróżnić dwa podstawowe podejścia. W pierwszym stosowana jest analiza porównawcza wyników wielogrupowej analizy czynnikowej, w której modele z ustalonymi ładunkami czynnikowymi (ocena inwariancji metrycznej), wyrazami wolnymi (ocena inwariancji skalarnej) lub z ustalonymi wariancjami błędu są porównywane ze sobą na podstawie istotności różnic w statystykach χ^2 i stopniach swobody. Brak odrzucenia bardziej restrykcyjnego modelu jest wskazówką do przyjęcia hipotezy o występowaniu ekwiwalencji metrycznej (słabej) lub skalarnej (silnej) dla układów wielogrupowych.

Ograniczeniem tego podejścia jest możliwość testowania inwariancji w nielicznych przekrojach (3-4 grupy). W alternatywnym podejściu wykorzystywane są modele confirmacyjnej analizy czynnikowej z oszacowaniem wartości średnich dla czynników (*means and covariance structure*). Podstawą oceny inwariancji jest procedura wyrównywania (*alignment*). Identyfikacja modelu z średnimi i wariancjami dla czynników jest dokonywana poprzez restrykcje optymalizujące tzw. funkcję straty (*F*) (*loss function*), która prowadzi do uzyskania niewielu parametrów (ładunków czynnikowych i wyrazów wolnych) o bardzo niskiej inwariancji pomiaru oraz dużej liczby parametrów inwariantnych w przekroju grup. Rozwiązanie to jest w istocie podobne do roli rotacji ortogonalnej w analizie czynnikowej.

Optymalizacja inwariancji jest dokonywana na podstawie znalezienia minimum funkcji straty. W pierwszym kroku dokonuje się estymacji modelu z uwolnionymi parametrami (jednorodnego – *configural*), a w drugim następuje

¹ Pozycje skali: P23: Uważam, że rodzina powinna ograniczać wydatki na indywidualne potrzeby poszczególnych osób na rzecz zaspokojenia wspólnych potrzeb; P24: Dobro wspólne całej rodziny jest ważniejsze niż zaspokajanie zachcianek, dążeń i przyjemności każdego z osobna; P25: Lepiej realizować się dzięki dobrom i usługom służącym całemu gospodarstwu domowemu niż indywidualnym potrzebom każdego z osobna; P26: Radość życia czerpię w większym stopniu z dóbr i usług, które służą wszystkim członkom rodziny, niż tych kupowanych na moje osobiste potrzeby.

oszacowanie średnich i wariancji dla zmiennych ukrytych na poziomie minimum funkcji straty (dla każdej pary ładunków i wyrazów wolnych w przekrojach uporządkowanych grup względem rosnącej wartości średniej dla czynników). Funkcja upraszczająca ma postać²:

$$F = \sum_p \sum_{g1 < g2} w_{g1,g2} f(\lambda_{pg1,1} - \lambda_{pg2,1}) + \sum_p \sum_{g1 < g2} w_{g1,g2} f(c_{pg1,1} - c_{pg2,1}), \quad (1)$$

gdzie wagi w_{g1} w_{g2} stanowią średnie geometryczne liczebności grup:

$$w_{g1,g2} = \sqrt{N_{g1} N_{g2}}. \quad (2)$$

W efekcie zastosowania tej procedury występuje niewiele parametrów silnie nieinwariantnych i wiele parametrów inwariantnych. Tabela 1 przedstawia wyniki porównań wielogrupowej confirmacyjnej analizy czynnikowej skali ASPD w przekrojach ról społecznych w rodzinie (ojca, matki, dziecka).

Tabela 1

Międzygrupowa inwariancja pomiaru

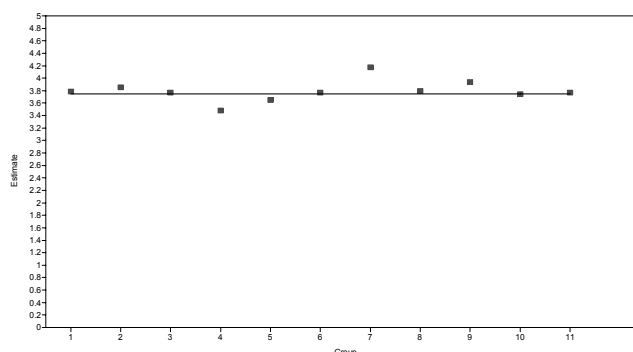
Model	Statystyka χ^2	Stopnie swobody	Poziom p
Jednorodny	13,6	6	0,03
Metryczny	18,1	12	0,11
Skalarny	32,5	18	0,02
Jednorodny a metryczny	4,6	6	0,60
Jednorodny a skalarny	18,9	12	0,09
Metryczny a skalarny	14,3	6	0,03

Źródło: Opracowano na podstawie programu Mplus 7.12.

Przyrostowe testy istotności wskazują, że model z inwariancją metryczną jest nieistotnie gorzej dopasowany niż model jednorodny ($p = 0,60$), lecz istotnie gorszy niż model skalarny ($p = 0,03$). Należy przyjąć, że pomiar stylu podejmowania decyzji jest inwariantny w przekroju członków gospodarstwa domowego i może stanowić podstawę porównań strategii podejmowania decyzji wśród członków rodzin.

W sytuacji, gdy występuje duża liczba porównywanych grup, podejście to jest niepraktyczne i zwielokrotnia liczbę testów sprawdzających. Procedura wyrównywania pozwala na wykorzystanie funkcji straty pozwalającej na uszeregowanie grup względem średnich czynnikowych dla optymalnej struktury nieinwariancji. Rys. 2 zawiera wyniki analizy porównawczej dla nieinwariantnej pozycji skali (P24) w przekroju 11 analizowanych województw.

² B. Muthen, T. Asparouhov: New Methods for the Study of Measurement Invariance with Many Groups, s. 9, Mplus.statmodel.com [12.04.2014].



Rys. 2. Wykres parametrów pozycji P24

Źródło: Ibid.

W województwie 7 (podlaskim) wartość wyrazu wolnego świadczy o braku inwariancji skalarnej skali ASPD (jest to efekt stylu odpowiedzi wynikający z ujawniania raczej pozytywnych ocen na skali – *ARS effect*).

2. Efekty losowe w modelach pomiarowych

Zaprezentowana powyżej wielogrupowa analiza czynnikowa (MACS) wykorzystuje model efektów stałych, w których parametry modelu nie są traktowane jako zmienna losowa, lecz są niezależnie estymowane w przekrojach grup. Podejście to jest zasadne w sytuacji, gdy występuje mała liczba wskaźników (pozycji skali) i grup oraz badacz zna potencjalne źródła braku inwariancji. W przypadku dużej liczby porównywanych grup (np. 30-100), małej ich liczebności (3-10 badanych) i braku znajomości struktury inwariancji stosowane jest podejście wielopoziomowe (wielopoziomowa confirmacyjna analiza czynnikowa) w ocenie modeli pomiarowych. W ramach tego podejścia wyróżnić można modele: 1) losowych wyrazów wolnych i nielosowych (inwariantnych) ładunków, 2) nielosowych wyrazów wolnych i losowych ładunków (model inwariancji pomiaru), 3) losowych wyrazów wolnych i losowych ładunków czynnikowych (*random intercepts and slopes*)³.

Ograniczeniem wykorzystania modeli efektów losowych w ocenie skal pomiarowych w przekroju grup jest założenie losowości klas, które jest spełnione jedynie w przypadku złożonych schematów losowania (np. wykorzystania doborów zespołowych, obszarowych lub wielostopniowych próby). Ich zastosowanie

³ Ibid., s. 13.

wiąże się z koniecznością uwzględnienia hierarchicznych czynników instytucjonalnych wywołujących współzależność obserwacji i występowanie silnych korelacji wewnątrzklasowych wynikających z zagnieżdżenia obserwacji (np. członkowie rodzin, pracownicy w przedsiębiorstwach, studenci w grupach dziekańskich, pacjenci w oddziałach szpitalnych itp.).

Skala ASPD została wykorzystana w badaniach stylu podejmowania decyzji w 440 gospodarstwach domowych (wśród członków gospodarstw). Ocena inwariancji pomiaru w przekroju rodzin została przeprowadzona na podstawie dwupoziomowego modelu confirmacyjnej analizy czynnikowej. Istota wielopoziomowego modelu CFA wynika z dekompozycji ocen wyrażanych na skalach pomiarowych na zmienność wewnątrzklasową wynikającą z odchyień obserwacji od średniej dla danych klasy i zmienność międzyklasową:

$$y_{ij} = (y_{ij} - y_{.j}) + (y_{.j}), \quad (3)$$

gdzie $(y_{ij} - y_{.j})$ oznacza odchylenie indywidualnych (i) odpowiedzi od średniej dla klasy (j), a $(y_{.j})$ określa zmienność w przekroju klas. Wielkość wpływu zmienności międzyklasowej na ogólną zmienność odpowiedzi określa współczynnik korelacji wewnątrzklasowej (*intraclass correlation coefficient – ICC*).

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}. \quad (4)$$

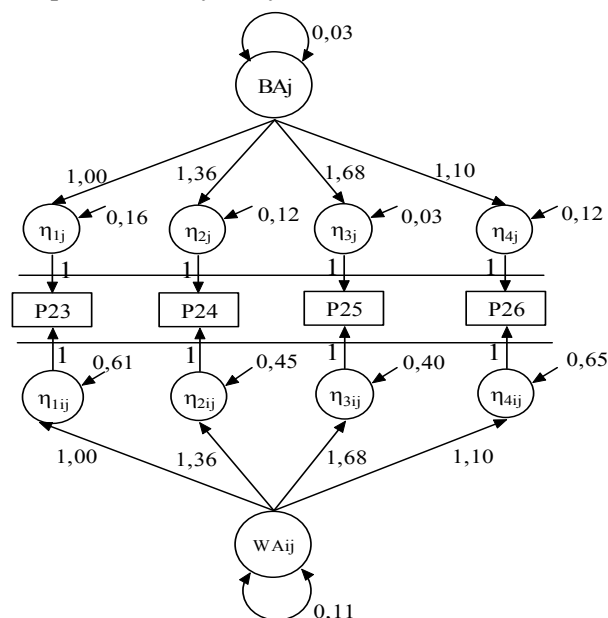
Biorąc pod uwagę hierarchiczną strukturę pomiaru, ocena rzetelności skal w modelach wielopoziomowych musi uwzględniać oba poziomy analizy i wpływ zmienności międzygrupowej na strukturę rzetelności⁴:

$$R_1 = R_B ICC + R_W (1 - ICC). \quad (5)$$

Ogólna rzetelność skali (R1) jest to ważona (współczynnikiem ICC dla skali) suma rzetelności wewnątrzklasowej i międzyklasowej. Dekompozycja rzetelności pozwala na ocenę wiarygodności pomiaru na obu poziomach analizy i diagnozę źródła braku rzetelności (akceptowalna rzetelność skali może wynikać z bardzo niskiej rzetelności na poziomie respondentów i bardzo wysokiej na poziomie klas). W przypadku analizowanej skali wartości ICC wskazują, że wystarczająca część zmienności wskaźników (poza P25) jest tłumaczona przez zmienność międzygrupową. Wartości ICC dla pozycji wynoszą odpowiednio: P23 = 0,22, P24 = 0,20, P25 = 0,14, P26 = 0,21.

⁴ G.J. Geldhof, K.J. Preacher, M.J. Zyphur: Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework. „Psychological Methods” 2013. DOI: 10.1037/a0032138.

Model dwupoziomowy skali ASPD z inwariancją metryczną (*random intercept*) został zaprezentowany na rys. 3.



Rys. 3. Dwugrupowy model wielopoziomowy z inwariancją metryczną
Źródło: Ibid.

Rys. 3 obejmuje model CFA skali altruizmu dla poziomu wewnętrznego (WA_{ij}) oraz gospodarstw domowych (BA_j). Wyrazy wolne modelu są zmiennymi losowymi w przekroju grup. Dopasowanie modelu z ograniczonymi ładunkami jest akceptowalne. Wartość statystyki χ^2 wynosi 18,48, (7) i poziom $p = 0,01$. Średniokwadratowy pierwiastek błędów aproksymacji (RMSEA) jest równy 0,039, a przyrostowe wskaźniki CFI i TLI przyjęły wartość odpowiednio 0,978 i 0,962.

Przy uwzględnieniu dwupoziomowego charakteru modelu zostały oszacowane wartości współczynników rzetelności dla modelu wewnętrznego i zewnętrznego. Zostały one podane w tabeli 2.

Tabela 2

Wartości współczynników rzetelności

Poziom analizy	Współczynniki rzetelności			
	α – Cronbacha	Ω – McDonalda	H – Bentlera	AVE Larckera-Fornella
Ogółem	0,60	0,64	0,68	0,34
Wewnątrzklasowa	0,61	0,66	0,68	0,33
Międzyklasowa	0,69	0,72	0,68	0,33
Wartość ICC dla skali	0,006			

Źródło: Ibid.

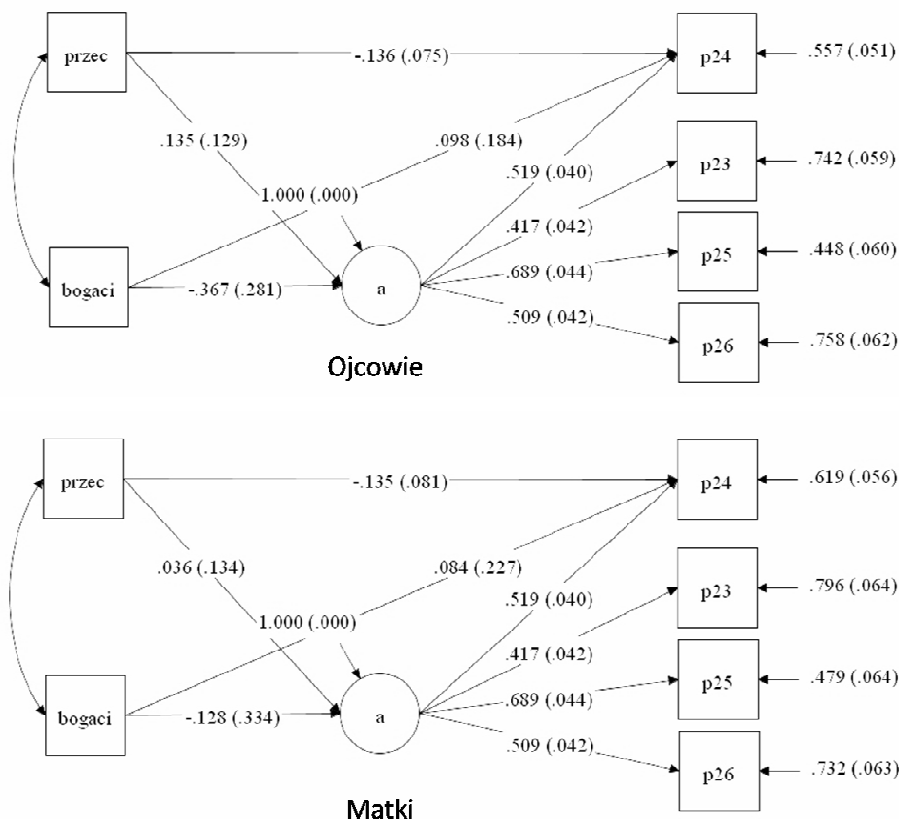
Tabela 2 prezentuje cztery rodzaje wskaźników rzetelności skal. Z powodu ujemnego obciążenia współczynnika Cronbacha wykorzystane zostały również współczynniki zbudowane na podstawie modelu (McDonalda, Bentlera i Larckera-Fornella). Wskazują one na akceptowalną, chociaż niską, rzetelność skali ASPD. Na obu poziomach analizy zaobserwowano podobne oceny rzetelności skali. Świadczy o tym również współczynnik korelacji wewnątrzklasowej obliczony dla całej skali.

3. Funkcjonowanie skali na poziomie cech demograficznych

Obciążenie pozycji skali może również wynikać ze zróżnicowanego wpływu cech społeczno-demograficznych respondentów (płci, wieku, wykształcenia itp.) na reakcje na pozycje. Ocena tego wpływu jest dokonywana na podstawie modeli wielorakich wskaźników – wielorakich przyczyn (MIMIC) lub w obszarze IRT za pomocą metod diagnozy zróżnicowanego funkcjonowania pozycji skal (DIF) i porównawczych testów ilorazu wiarygodności, testu Mantela-Haenszla, testu Raju i Breslowa-Daya⁵. Tabela 3 zawiera modele MIMIC estymowane w przekroju członków gospodarstw domowych (ojca i matki) z wykorzystaniem kowariantów jakościowych (subiektywnej oceny sytuacji materialnej) oraz płci respondenta w przypadku dziecka. Rys. 4 przedstawia strukturę modelu MIMIC dla wskaźnika P24 wśród rodziców.

Parametry ścieżkowe dla zmiennych sztucznych określających subiektywną ocenę sytuacji materialnej wskazują na stopień obciążenia rzetelności pozycji P24 ze względu na poziomy tej zmiennej. Istotny parametr dla zmiennej dotyczącej przeciętnej sytuacji materialnej (przec) wskazuje na ujemny i podobny w przekroju ról społecznych wpływ tego czynnika na ocenę pozycji P24. Jest ona nieinwariantna z punktu widzenia tej cechy społeczno-demograficznej (w porównaniu do osób oceniających sytuację materialną jako złą i bardzo dobrą). Tabela 3 przedstawia wpływ poszczególnych zmiennych społeczno-demograficznych na funkcjonowanie pozycji skali ASPD.

⁵ MacIntosh, S. Hashim: Variance Estimation for Converting MIMIC Model Parameters into IRT Parameters in DIF Analysis, Applied. „Psychological Measurement” 2003, No. 27(5), s. 372-379.



Rys. 4. Model MIMIC w przekroju ról społecznych

Źródło: Ibid.

Tabela 3

Parametry ścieżkowe modelu MIMIC

Zmienna demograficzna	P23	P24	P25	P26
Przeciętna sytuacja materialna – ojciec	0,14 (0,08)	0,14 (0,07)*	-0,01 (0,08)	0,06 (0,08)
Dobra sytuacja materialna – ojciec	0,18 (0,19)	0,09 (0,18)	-0,26 (0,21)	0,07 (0,20)
Przeciętna sytuacja materialna – matka	0,19 (0,08)*	-0,13 (0,08)*	-0,08 (0,09)	0,09 (0,08)
Dobra sytuacja materialna – matka	0,24 (0,24)	0,08 (0,23)	0,10 (0,25)	-4,00 (0,25)
Dziecko – chłopiec	-0,09 (0,11)	1,17 (0,09)*	-0,09 (0,11)	-0,12 (0,11)

Źródło: Ibid.

Struktura parametrów ścieżkowych określających wpływ tych zmiennych na wskaźnik (przy ustalonej wartości zmiennej ukrytej) wskazuje na obciążenie odpowiedzi na pozycję P24 ze względu na płeć dziecka i cechy społeczno-

-demograficzne rodziców. Dla pozycji P23 jedynie ocena przez matki sytuacji materialnej jako przeciętnej wpływa na sposób odpowiedzi na pozycję skali. W przypadku pozostałych pozycji analizowane cechy społeczno-demograficzne nie wpływają na zróżnicowanie odpowiedzi na pozycje.

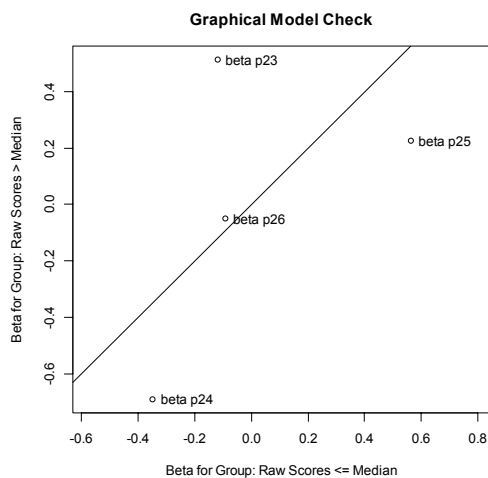
4. Specyficzna obiektywność skali

Pozycje skali mogą różnicować wypowiedzi badanych nie tylko ze względu na brak porównywalności międzykulturowej, czynniki hierarchiczne, cechy społeczno-demograficzne respondentów i kontekstowe. Zróżnicowanie to może wynikać z położenia respondenta na *continuum* mierzonej cechy ukrytej. Respondenci uzyskujący bardzo niskie lub wysokie wartości na *continuum* cechy ukrytej mogą inaczej reagować na pozycje skali.

Analiza specyficznej obiektywności skali jest dokonywana w obrębie modeli Rascha, zakładających jednakową moc dyskryminacyjną pozycji oraz skumulowany ich charakter (pozycje różnią się tylko poziomem trudności). Ocena specyficznej obiektywności jest dokonywana na podstawie testu ilorazu wiarygodności (*likelihood ratio*) w grupach wyodrębnionych na podstawie mediany ocen wartości zmiennych ukrytych (grupę „dolną” stanowią osoby poniżej mediany a „górną” – osoby powyżej mediany). W sytuacji obiektywności pomiaru różnice w parametrach trudności pozycji między pozycjami są takie same dla wszystkich badanych oraz różnice w parametrach zdolności (w poziomie cechy ukrytej) między parami respondentów są takie same dla wszystkich pozycji.

Na rys. 5 znajduje się wykres rozrzutu parametrów trudności modelu Rascha dla pozycji P23-P26. Ułożenie parametrów trudności względem linii referencyjnej wskazuje na równość parametrów pozycji w dolnej i górnej grupie badanych. Z wykresu wynika, że jedynie pozycja P26 spełnia założenie specyficznej obiektywności.

Wartość testu ilorazu wiarygodności Andersena (LR-test) pozwala na odrzucenie hipotezy o równości parametrów w grupach. Wartość testu LR wynosi 29 844 dla 3 stopni swobody i poziomu $p = 0,00$. Pozycje P23-P25 nie spełniają założeń specyficznej obiektywności.



Rys. 5. Test specyficznej obiektywności skali

Źródło: Opracowano na podstawie pakietu eRm programu R.

Podsumowanie

Uwzględnienie ukrytej niejednorodności populacji pozwala na uzyskanie nieobciążonych współczynników rzetelności i tym samym nieobciążonych parametrów modeli strukturalnych. Brak ich uwzględnienia prowadzi do występowania błędów w zmiennych, tłumienia współczynników korelacji/regresji i wielkości błędu standardowego przy maskującym efekcie współczynnika determinacji R^2 . W badaniach wpływu zanieczyszczenia środowiska na stopę śmiertelności⁶ analiza regresji wskazywała, że zanieczyszczenie ma istotny wpływ na śmiertelność. Ponowna analiza zjawiska dokonana przez Bollen wskazywała, że mierniki zanieczyszczenia zawierają do 50% błędu pomiaru, co powoduje obciążenie współczynników regresji o 30%-40%. Po uwzględnieniu błędów pomiaru żaden z parametrów nie był istotny (przy praktycznie niezmiennym współczynniku R^2). Tabela 4 przedstawia wyniki analizy regresji między altruistycznym stylem podejmowania decyzji a oceną sytuacji materialnej z uwzględnieniem obciążenia rzetelności pomiaru.

⁶ L.B. Lave, E.P. Seskin: Air Pollution an Human Health. Johns Hopkins Press, Baltimore 1977.

Tabela 4

Parametry i ocena dopasowania modeli regresji z błędami w zmiennych

Współczynnik rzetelności	Współczynnik B	Przedział ufności	Współczynnik determinacji
1	-0,039*	-0,078 ; -0,001	0,04
0,72	-0,055	-0,109 ; -0,001	0,04
0,68	-0,059	-0,116 ; -0,001	0,04
0,60	-0,066	-0,131 ; -0,002	0,05
0,68 (P24)	-0,072	-0,141 ; -0,03	

Źródło: Opracowano na podstawie programu Stata 13.

Struktura parametrów modelu potwierdza obserwacje Bollen. Po pierwsze, występuje tłumienie współczynników regresji (bez uwzględnienia nierzetelności skali są one zbyt niskie). Po drugie, przedziały ufności dla współczynników B są zbyt wąskie. Po uwzględnieniu błędów pomiarowych wzrasta współczynnik kierunkowy regresji i jednocześnie rośnie prawdopodobieństwo otrzymania jego nieistotnych wartości.

Podsumowując: uwzględnienie ukrytej heterogeniczności populacji pozwala na uzyskanie nieobciążonych współczynników rzetelności i tym samym nieobciążonych parametrów modeli strukturalnych. Heterogeniczność ma charakter hierarchiczny, występujący na poziomie międzykulturowym (diagnozowana na podstawie przyrostowych testów χ^2), międzyobszarowym (modele wielopoziomowe), międzygrupowym (modele MIMIC i DIF), kontekstowo-sytuacyjnym (modele LS-T) i wewnątrzkonstrukcyjnym (testy specyficznej obiektywności pozycji). Ich rola jest szczególnie ważna w naukach społecznych, które najczęściej wykorzystują skale pomiarowe nie tylko o względnie niskiej rzetelności, lecz przede wszystkim nieinwariantne metrycznie o dodatkowo obciążonej ich rzetelności. Powoduje to, że mierzone są nie tylko nie te same zmienne w przekroju międzypopulacyjnym, lecz i nie tak samo w procesie pomiaru.

Literatura

- Geldhof G.J., Preacher K.J., Zyphur M.J.: Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework. „Psychological Methods” 2013. DOI: 10.1037/a0032138.
- Lave L.B., Seskin E.P.: Air Pollution and Human Health. Johns Hopkins Press, Baltimore 1977.
- Muthen B., Asparouhov T.: New Methods for the Study of Measurement Invariance with Many Groups. Mplus. statmodel.com [12.04.2014].
- MacIntosh R., Hashim S.: Variance Estimation for Converting MIMIC Model Parameters into IRT Parameters in DIF Analysis, Applied. „Psychological Measurement” 2003, No. 27(5).

**RELIABILITY AND ITEM FUNCTIONING
IN UNOBSERVED HETEROGENEITY OF POPULATION****Summary**

The aim of paper is to evaluate the reliability and item bias in the measurement in cross-cultural comparative research and heterogenous populations. Well known reliability coefficients (Cronbach's, Armor's and McDonald's) are used in homogenous population with common frame of reference. We present the holistic approach to scale reliability assessment, when homogeneity assumption is violated in research practice. This approach enables full control of measurement bias in heterogenous population.