dr Artur Pokropek
Instytut Badań Edukacyjnych
Zespół EWD

Wielogrupowy Model IRT
Analizy Symulacyjne

Wstęp

Każdy model statystyczny zawiera szereg założeń, niekiedy spełnianych dokładnie, czasami do pewnego stopnia, a czasami wyraźnie pogwałconych w toku przeprowadzania analiz. Pogwałcenia założeń modelu prowadzą do błędów estymacji, te zaś do błędnych wniosków. Nie inaczej jest w modelowaniu IRT (*Item Response Theory*). W tym tekście zajmiemy się jednym z założeń prostego modelowania IRT założeniem, iż poziom umiejętności uczniów szacowany jest dla jednej grupy (populacji). Jeżeli grup jest więcej, prosty model IRT nie spełnia wyłożonego założenia i należy się odwołać do bardziej złożonych modeli IRT - wielogrupowych modeli IRT (*Multiple Group IRT*). W tekście model wielogrupowy zostanie formalnie wyspecyfikowany. Następnie przedstawionych zostanie szereg symulacji, w których sprawdzona zostanie implementacja tego modelu w różnych programach statystycznych. Ponadto wyniki modelu wielogrupowego zostaną porównane z łamiącym założenia modelowania jednogrupowym modelem IRT.

Prosty a wielogrupowy model IRT

Punktem wyjścia dla modelowania IRT jest funkcja $P_i()$ określająca prawdopodobieństwo zaobserwowania konkretnej wartości odpowiedzi h dla zadania i w zależności od poziomu umiejętności θ oraz parametrówβ_i charakteryzujących zadanie. Dla modelu jednoparametrycznego $\beta_i=(b_i)$, dla modelu dwuparametrycznego $\beta_i=(a_i,b_i)$, a dla trzyparametrycznego $\beta_i=(a_i,b_i,c_i)$. Funkcja ta określana jest mianem *krzywej charakterystycznej* zadania lub *funkcji charakterystycznej* zadania:

$$P_{ih}(\theta)=P_i(U_i=h|\ \theta,\ \beta_i$$

Wymaga się od niej, by była podwójnie różniczkowalna oraz wspólna dla grup i jednostek (Bock i Zimowski, 1997).

Zakładając warunkową niezależność odpowiedzi jednostek, IRT umożliwia opisanie prawdopodobieństwa uzyskania wektora odpowiedzi $Uj=[U_1j, U_2j,..., U_nj]$, warunkowo ze względu na poziomu umiejętności θ i parametry zadań, za pomocą iloczynu *funkcji charakterystycznych* dla n liczby zadań (Beker i Kim, 2004):

$$P(\mathbf{U_j}\,|\,\theta,\boldsymbol{\beta}_i)=\prod_{i=1}^{n}P_i(U_{ij}=h\,|\,\theta,\boldsymbol{\beta}_i)$$

Poprzestając na tej formule, łatwo można zapomnieć, iż w prostym modelu IRT zakładamy, że jednostki pochodzą z jednej grupy (populacji) charakte-

ryzowanej przez funkcję gęstości prawdopodobieństwa określającą rozkład θ dla danej grupy. Widać to dokładnie w zapisie modelu, który wyraża brzegowe prawdopodobieństwo wektora odpowiedzi *Uj* (warunkowanego ze względu na $\beta_i$ale nie na θ):

$$P_k(\mathbf{U_j} \mid \boldsymbol{\beta}_i) = \int P(\mathbf{U_j} \mid \theta, \boldsymbol{\beta}_i) g_k(\theta) d\theta$$

gdzie $g_k(\theta)$ jest funkcją gęstości prawdopodobieństwa określającą rozkład zmiennej w grupie *k*. Aby rozszerzyć prosty model IRT o możliwość szacowania parametrów dla kilku grup jednocześnie, należy wprowadzić zestaw parametrów η_k charakteryzujących rozkład θ w różnych grupach (najczęściej chodzi o średnią i odchylenie standardowe):

$$P_k(\mathbf{U_j} \mid \boldsymbol{\beta}_i) = \int P(\mathbf{U_j} \mid \theta, \boldsymbol{\beta}_i) g(\theta \mid \boldsymbol{\eta_k}) d\theta$$

Przedstawiony powyżej model jest rozszerzeniem klasycznego modelu IRT i opisywany jest w literaturze (Bock i Zimowski, 1997) jako wielogrupowy model IRT (*Multiple Group IRT*). Oprócz parametrów zadań, tak jak w prostych modelach IRT, model umożliwia szacowanie parametrów rozkładów θ w różnych grupach, dając jednocześnie nieobarczone estymacje θ poszczególnych jednostek.

Z modelem wielogrupowym wiążą się jednak trudne kwestie natury technicznej. Do jego oszacowania (tak jak i innych modeli IRT) niezbędna jest procedura estymacji największej wiarygodności, z tym że w przeciwieństwie do prostego modelu IRT, w modelu wielogrupowym funkcja wiarygodności (a precyzyjniej całka stanowiąca jej część) nie ma analitycznego rozwiązania (szczegóły można znaleźć w Bock i Zimowski, 1997 oraz Beker i Kim, 2004). Estymacja jest zatem trudna i wymaga wprowadzenia procedur całkowania numerycznego, które nie jest implementowane we wszystkich programach służących do modelowania IRT.

Programy szacujące model wielogrupowy

Tylko kilka programów przeznaczonych do analiz psychometrycznych jest w stanie szacować model wielogrupowy. Wśród nich znajduje się program napisany przez Ceesa Glasa (2010), profesora Uniwersytetu Twente. Program oprócz estymacji modeli wielogrupowych ma wiele innych możliwości estymacji złożonych modeli IRT (za pomocą niego można modelować między innymi modele wielowymiarowe), program ten ma też ogromną zaletę w obecnej wersji jest programem bezpłatnym. Modele wielogrupowe można szacować również za pomocą płatnego pakietu Mplus (Muthen i Muthen, 2010). Pakiet oferuje estymacje różnorodnych modeli ze zmiennymi ukrytymi (*Generalized Latent Variable Modeling*), w tym modelowania IRT. Modele wielogrupowe można zaimplementować również dzięki pakietowi Winbugs (Ntzoufras, 2009). Winbugs jest programem, a w zasadzie silnikiem estymacyjnym, służącym do szacowania dowolnych modeli w ramach podejścia bayesowskiego. Jakkolwiek podejście do modelowania IRT stosowane w Winbugs różni się od pozostałych programów, to uzyskiwane wyniki (w większości sytuacji) są bardzo podobne do tych uzyskiwanych w klasycznej estymacji IRT (szczegóły dotyczące estymacji modeli IRT w podejściu bayesowskim można znaleźć w Fox, 2010).

Punktem odniesienia dla wyników estymowanych przez programy szacujące modele wielogrupowe będą wyniki uzyskane za pomocą programu Parscale 4.1 (Muraki i Bok, 1997). Parscale jest uznanym, sprawdzonym w różnych sytuacjach i wielokrotnie programem mogącym szacować jednogrupowe modele IRT. W przedstawionych analizach został on wykorzystany w dwojaki sposób. Po pierwsze, za jego pomocą szacowane były parametry jednogrupowego modelu IRT, na danych o charakterze wielogrupowym, aby zobaczyć wielkość błędów mogących powstać poprzez zastosowanie modelu niedostosowanego w pełni do posiadanych danych. Po drugie, sprawdzone zostało, na ile wprowadzenie zmiennej charakteryzującej przynależność do grupy, traktowanej jako jedno z zadań, może poprawić oszacowania. Takie podejście jest prostą próbą estymacji różnicy między grupami analogiczną do wprowadzenia zmiennej zero-jedynkowej jako zmiennej wyjaśniającej w modelu regresji. Podejście to ma oczywiste wady formalne (chociażby zakłada się równą wariancję międzygrupową), lecz prostota tego rozwiązania może wydawać się kusząca. Warto zatem sprawdzić, czy rzeczywiście taki wybieg może poprawić jakość szacowania. W obydwu metodach z wykorzystaniem programu Parscale najpierw szacowane były θ dla poszczególnych uczniów, dalej za ich pomocą szacowano parametry rozkładów.

Wszystkie wyniki symulacyjne, przedstawiane w dalszej części tego tekstu, uzyskane zostały bez zmian kryteriów wykorzystywanych w procedurach iteracyjnych, czy bez liczby punktów całkowania. Wszystkie ustawienia domyślne zostały zachowane w każdym z programów.

Symulacje

Schemat badania symulacyjnego został przedstawiony w tabeli 1. Symulacje zostały zaprojektowane, aby przetestować dwa scenariusze. W obydwu scenariuszach mamy dwie grupy (1 i 2). W pierwszym scenariuszu wszyscy uczniowie odpowiadają na ten sam zestaw zadań (zadania 1-50). Przy czym grupa 2. charakteryzuje się wyższym poziomem umiejętności i zarazem większym ich zróżnicowaniem. Średnia wartość θ w grupie 1. wynosi 0, a odchylenie standardowe 1. Dla drugiej grupy przewidziane zostały dwa warianty:

- 2a średni poziom θ wynosi 0,2, a odchylenie standardowe 1,2 oraz
- 2b średni poziom θ wynosi 0,5, a odchylenie standardowe 1,5.

Scenariusz I odzwierciedla kalibracje jednego testu na dwóch różnych populacjach charakteryzujących się różnymi rozkładami θ. Scenariusz taki może odnosić się do sytuacji, w której jeden test rozwiązywany jest przez różne grupy wiekowe, istnieje różnica między dziewczętami a chłopcami, różnymi szkołami lub dowolnymi innymi grupami.

Scenariusz II, i jego warianty, jest analogiczny do pierwszego z jedną różnicą - grupy „rozwiązują" różne zestawy zadań. Pierwsza grupa zadania od 1 do 30, druga od 21 do 50. Jest to scenariusz odzwierciedlający schemat zrównywania nieekwiwalentnych grup z testem kotwiczącym. Schematy takie wykorzystuje się do zrównywania tak horyzontalnego, jak i wertykalnego (Kolen i Brennan, 2004).

Tabela 1. Schemat badania symulacyjnego

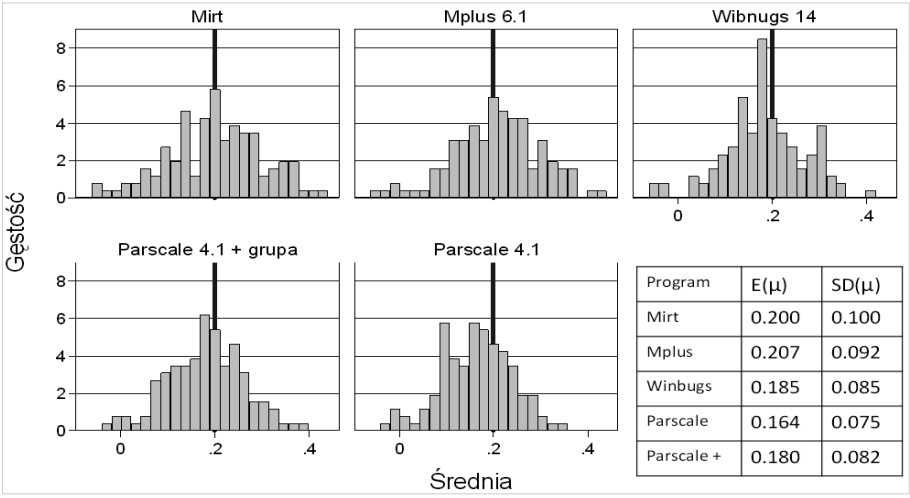
Scenariusz	Grupa/Wariant	Zadania		
I	1: ($\mu = 0$; $\sigma = 1$)	Zadania 1-50		
	2a: ($\mu = 0.2$; $\sigma = 1.2$) 2b: ($\mu = 0.5$; $\sigma = 1.5$)	Zadania 1-50		
II	1: ($\mu = 0$; $\sigma = 1$)	Zadania 1-20	Zadania 21-30	
	22a: ($\mu = 0.2$; $\sigma = 1.2$) 2b: ($\mu = 0.5$; $\sigma = 1.5$)		Zadania 21-30	Zadania 31-50

Dla każdego wariantu symulacji generowano 400 jednostek dla każdej z grup (razem 800), tak aby rozkład θ w grupach pokrywał się z zadanymi w różnych wariantach charakterystykami grup. Następnie generowano zbiory odpowiedzi do zadań o losowo wyznaczonych parametrach. Parametr trudności (b) losowano z rozkładu o średniej 0 i odchyleniu standardowym 1, parametr dyskryminacji (a) z rozkładu o średniej 1 i odchyleniu standardowym 0,2, parametr zgadywania dla każdego zadania (c) wynosił 0. Dla każdego z wariantów symulacji wygenerowano 100 zbiorów danych. Na każdym zbiorze wypróbowano trzy pogramy mogące estymować wielogrupowe modele IRT: Mirt, Mplus 6.1, Winbugs 14 oraz program do estymacji modeli jednogrupowych Pparscale 4.1 z indykatorem grupy oraz estymowany w sposób klasyczny. Każdy z programów estymował model dwuparametryczny.

Wyniki symulacji

Na rysunku 1. przedstawiono rozkłady estymowanych średnich dla grupy drugiej przy zastosowaniu różnych programów i podejść na 100 symulowanych zbiorach danych. Na rysunku grubą pionową linią zaznaczono wartość prawdziwą (0,2). W prawym dolnym rogu rysunku znajduje się tabela przedstawiająca średnią i odchylenie standardowe estymatorów z poszczególnych symulacji.

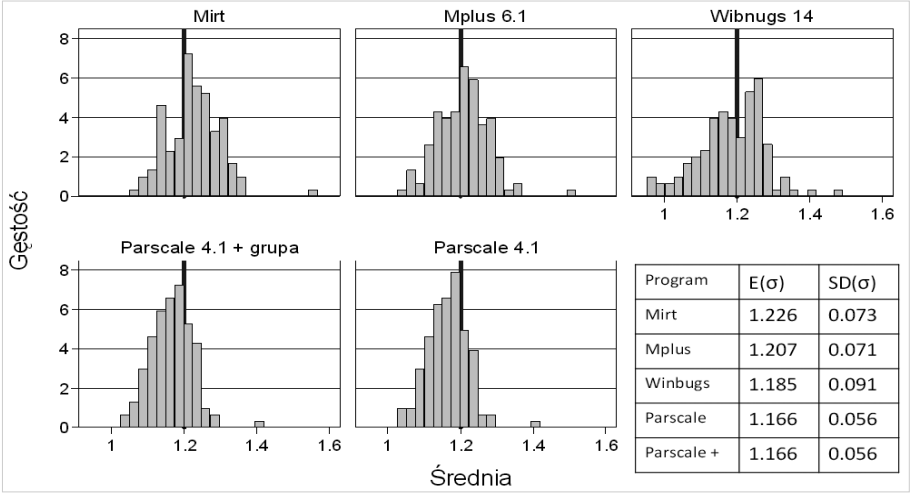
W tym scenariuszu przy estymacji średniej grupowej najlepiej radziły sobie programy Mirt i Mplus. Lekkie obciążenie ku zeru zauważalne jest dla pakietu Winbugs (rozkład *prior* dla θ w obydwu grupach wynosił 0 przy odchyleniu standardowym 10, średnia grupowa została zatem ściągnięta ku zeru). Średnia estymatorów w Parscale ze zmienną oznaczającą grupę jest bliska wynikom uzyskanym przez Winsbugs. Jak można było oczekiwać, największe obarczenia estymatorów obserwowane są dla programu Parscale w klasycznej estymacji.



Program	E(μ)	SD(μ)
Mirt	0.200	0.100
Mplus	0.207	0.092
Winbugs	0.185	0.085
Parscale	0.164	0.075
Parscale +	0.180	0.082

Rysunek 1. Rozkłady oszacowań średniej dla grupy drugiej. ***Schemat I wariant a***

Rysunek 2. jest analogiczny do rysunku 1. z tym, że zostały na nim przedstawione oszacowania odchylenia standardowego dla grupy drugiej na symulowanych zbiorach danych obliczone przez różne programy.

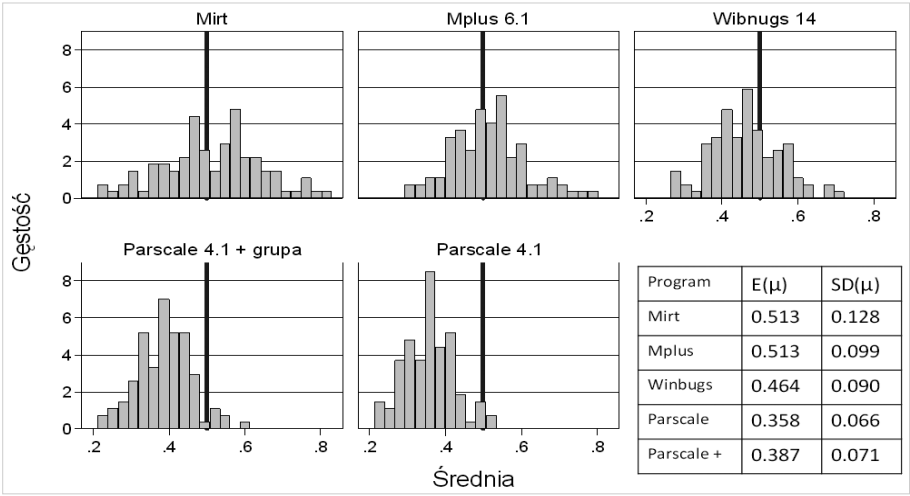


Program	E(σ)	SD(σ)
Mirt	1.226	0.073
Mplus	1.207	0.071
Winbugs	1.185	0.091
Parscale	1.166	0.056
Parscale +	1.166	0.056

Rysunek 2. Rozkłady oszacowań odchylenia standardowego dla grupy drugiej. ***Schemat I wariant a***

Tak jak w przypadku oszacowania średniej, najmniej obarczone rezultaty daje Mplus i Mirt. Winbugs, tak jak w przypadku oszacowań średniej, daje wyniki nieznacznie ściągnięte do rozkładu *prior* (średnia 1 odchylenie standardowe 10). Jak poprzednio najsłabiej wypada Parscale. W obydwu parametryzacjach daje identyczne wyniki, zaniżające odchylenie standardowe.

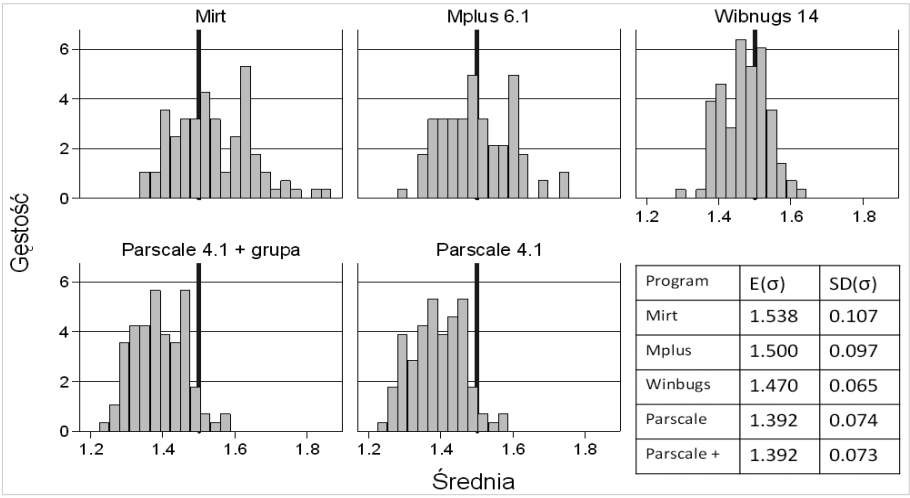
Na rysunku 3. przedstawiono wyniki symulacji analogiczne do wyników przedstawionych na rysunku 1. Z tym, że odnoszące się do wariantu, w którym różnice międzygrupowe były większe (wynosiły 0,5 odchylenia standardowego). Powtarza się wzór znany z rysunku 1. Mirt wraz z Mplus nie charakteryzują się dostrzegalnym obciążeniem (przy czym warto dodać, że estymatory szacowane przez Mirt mają największą wariancję spośród wszystkich testowanych tu programów). Słabiej od Mirt i Mplus wypada Winbugs, obciążenie nie jest jednak bardzo duże. Parscale z dodatkową zmienną daje nieznacznie lepsze wyniki niż klasyczna estymacja, jednak w obydwu estymacjach z wykorzystaniem Parscale obciążenie estymatorów jest znaczne.



Program	E(μ)	SD(μ)
Mirt	0.513	0.128
Mplus	0.513	0.099
Winbugs	0.464	0.090
Parscale	0.358	0.066
Parscale +	0.387	0.071

Rysunek 3. Rozkłady oszacowań średniej dla grupy drugiej. ***Schemat I wariant b***

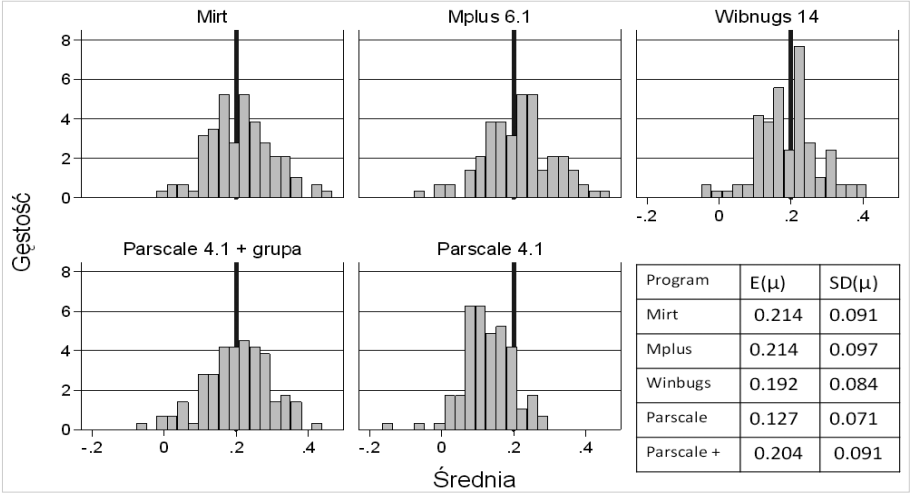
Rysunek 4. jest analogiczny do rysunku 2. i przedstawia estymowane odchylenie standardowe dla grupy drugiej. Wartość prawdziwa odchylenia standardowego wynosi w tym wariancie 1,5. Tak jak dla mniejszych różnic w średniej i w odchyleniu standardowym, i w tym wariancie bardzo słabo wypada Parscale (w obydwu wariantach). Średnio rzecz biorąc, Mplus szacuje najmniej obciążone estymatory. Mirt i Winbugs dają zadowalające wyniki, przy czym Mirt lekko przeszacował odchylenie standardowe, a Winbugs go zaniżył.



Program	E(σ)	SD(σ)
Mirt	1.538	0.107
Mplus	1.500	0.097
Winbugs	1.470	0.065
Parscale	1.392	0.074
Parscale +	1.392	0.073

Rysunek 4. Rozkłady oszacowań odchylenia standardowego dla grupy drugiej. ***Schemat I wariant b***

Kolejne rysunki odnoszą się do scenariusza II, w którym testowana jest sytuacja odzwierciedlająca zrównywanie nieekwiwalentnych grup z testem kotwiczącym. Rysunek 5. przedstawia estymatory średniej oszacowane dla grupy drugiej na 100 wygenerowanych zbiorach danych, przy założeniu, że średnia dla tej grupy wynosi 0,2.

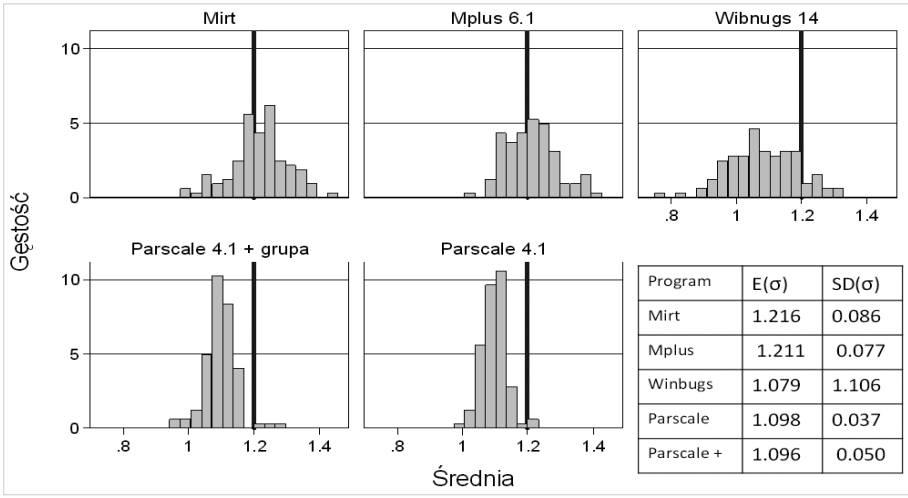


Program	E(μ)	SD(μ)
Mirt	0.214	0.091
Mplus	0.214	0.097
Winbugs	0.192	0.084
Parscale	0.127	0.071
Parscale +	0.204	0.091

Rysunek 5. Rozkłady oszacowań średniej dla grupy drugiej. ***Schemat II wariant a***

Wszystkie trzy programy zdolne do szacowania modelu wielogrupowego dają zadowalające, nieobarczone oszacowania średniej. Klasyczny model IRT obciążony jest ku średniej. Model ze zmienną identyfikującą przynależność grupową wypada bardzo dobrze, porównywalnie do wyników modelu wielogrupowego (a nawet odrobinę lepiej od nich).

Jeśli chodzi o oszacowanie odchylenia standardowego grupowej dla scenariusza II wariantu a, to wyniki symulacji przedstawione zostały na rysunku 6.

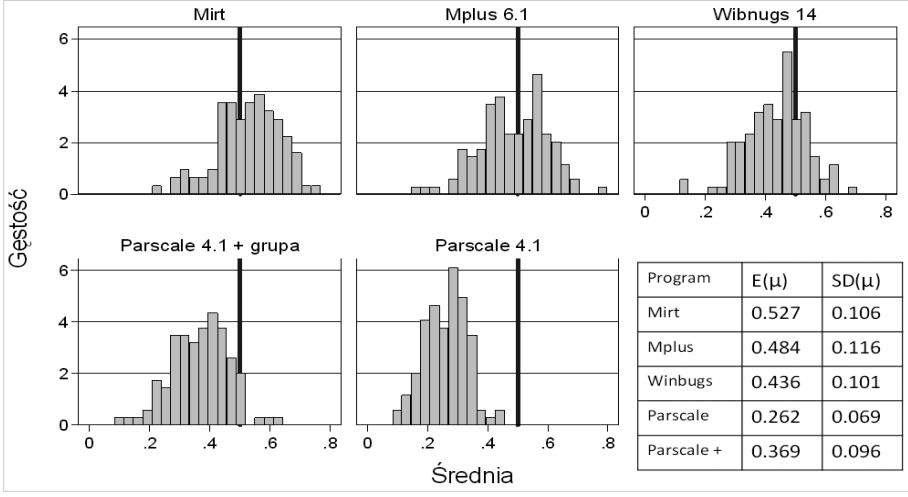


Program	E(σ)	SD(σ)
Mirt	1.216	0.086
Mplus	1.211	0.077
Winbugs	1.079	1.106
Parscale	1.098	0.037
Parscale +	1.096	0.050

Rysunek 6. Rozkłady oszacowań odchylenia standardowego dla grupy drugiej. ***Schemat II wariant a***

Tylko Mirt i Mplus okazały się produkować zadowalające, nieobarczone estymatory odchylenia standardowego. Winbugs i Parscale (w dwóch parametryzacjach) zaniżają odchylenia standardowe w grupie 2.

Na rysunku 7. i 8. przedstawione zostały wyniki symulacji, w których szacowana była średnia i odchylenie standardowe dla grupy drugiej zgodnie ze scenariuszem II w wariancie b, czyli przy założeniu, że średnia grupowa (grupy drugiej) wynosi 0,5, a odchylenie standardowe 1,5.

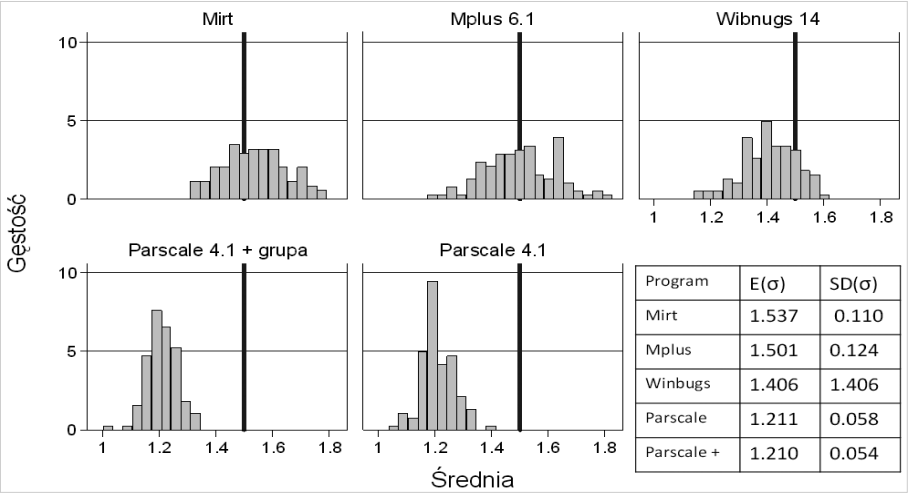


Program	E(μ)	SD(μ)
Mirt	0.527	0.106
Mplus	0.484	0.116
Winbugs	0.436	0.101
Parscale	0.262	0.069
Parscale +	0.369	0.096

Rysunek 7. Rozkłady oszacowań średniej dla grupy drugiej. ***Schemat II wariant b***

Podobnie jak w poprzednim wariancie, estymatory średniej grupowej dla Mirt i Mplus są praktycznie nieobciążone. Lekkie ściągnięcie ku średniej obserwowane jest dla Winbugs. Duże obciążenie zauważalne jest dla modelu jednogrupowego estymowanego w programie Parscale. Model ze zmienną identyfikującą grupę, średnio rzecz biorąc, charakteryzuje się znacznie mniejszym obciążeniem w porównaniu z modelem jednogrupowym, lecz obciążenie nadal pozostaje duże w stosunku do wartości wyznaczonej w tym scenariuszu.

Jeżeli chodzi o estymacje odchylenia standardowego (rysunek 8.), Mplus i Mirt dają zadowalające wyniki. Przy czym Mirt lekko przeszacowuje wariancje. Podobnie jak we wcześniejszych symulacjach, Winbugs nieznacznie zaniża różnice między grupami, ale wyniki nie charakteryzują się znaczącym obciążeniem. Parscale w obydwu parametryzacjach nie radzi sobie z poprawnym oszacowaniem odchylenia standardowego.



Program	E(σ)	SD(σ)
Mirt	1.537	0.110
Mplus	1.501	0.124
Winbugs	1.406	1.406
Parscale	1.211	0.058
Parscale +	1.210	0.054

Rysunek 8. Rozkłady oszacowań odchylenia standardowego dla grupy drugiej. Schemat II wariant b

Podsumowanie

Stosowanie jednogrupowego modelu IRT do danych, gdzie jednostki pochodzą z różnych grup o różnych rozkładach umiejętności, może prowadzić do znacznych błędów w szacowaniu parametrów grupowych. Modele wielogrupowe powinny być stosowane do zrównywania tam, gdzie założona jest nieekwiwalentość grup, zewnętrzny test kotwiczący i kalibracja łączna. Dodanie zmiennej identyfikującej grupę do zwykłego modelu IRT zmniejsza obciążenie estymacji średniej, lecz nie likwiduje go, przynajmniej w sytuacjach, gdy różnice między grupami są duże. Nie poprawia natomiast oszacowań zróżnicowania rozkładu. Darmowe programy służące do estymacji modeli wielogrupowych dobrze spełniają wyznaczone im cele estymacyjne (Mirt lepiej, Winbugs trochę gorzej), podobnie jak komercyjne pakiety (Mplus).

Bibliografia:

1. Baker, F. B. i Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
2. Bock, R. D. i Zimowski, M. F. (1997). „Multiple group IRT" [w:] W. J. van der Linden i R. K. Hambleton (red.), *Handbook of modern item response theory*. New York: Springer- Verlag.
3. Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications.* New York: Springer.
4. Glas, C. A. W.(2010). Preliminary Manual of the software program Multidimensional Item Response Theory (MIRT). (http://www.utwente.nl/gw/omd/afdeling/temp_test/mirt-manual.pdf)
5. Kolen, M. J. i Brennan R. L. (2004). *Test equating, scaling, and linking: Method and practice* (2nd ed.). New York, NY: Springer-Verlag.
6. Muraki, E. i Bock, R.D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data. *Scientific Software International*.
7. Muthen, L. K. i Muthen, B.O. (2010). Mplus. Statistical Analysis With Latent Variables. User's Guide. http://www.statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v6.pdf
8. Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS. Hoboken*. New Jersey: Wiley and Sons.