



EDUKACYJNA WARTOŚĆ DODANA

Zróżnicowanie funkcjonowania zadań między różnymi wersjami testu z wykorzystaniem Testlet Response Theory

Przemysław Majkut

Maciej Koniewski

Paulina Skórska



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE



EDUKACYJNA
WARTOŚĆ
DODANA

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY





Problem

- W systemie egzaminów zewnętrznych funkcjonują dwie wersje testu, które różnią się kolejnością odpowiedzi w ramach zadań
- Czy istnieje związek między rozwiązywaną przez uczniów wersją testu a funkcjonowaniem poszczególnych pytań testowych?
- Problem ten zostanie przeanalizowany dla wyników egzaminu gimnazjalnego z historii i wiedzy o społeczeństwie rozwiązywanych w 2013 r.



Schemat analizy danych

Czy istnieją różnice w poziomach wykonania zadań między dwiema wersjami testu?



Czy różnice w poziomach wykonania zadań między dwiema wersjami testu są istotne?



Jak silny jest efekt zróżnicowanego funkcjonowania zadań między dwiema wersjami testu?



Jakie są konsekwencje dla wyników uzyskanych na całym teście?



Czym charakteryzują się zadania o największym zróżnicowaniu między wersjami arkusza?



Schemat analizy danych – używane narzędzia analityczne

Analiza zróżnicowania funkcjonowania pozycji testowych (DIF)

test Mantel-Haenszel

W jakim stopniu odpowiedź na pozycję testową zależy od rozwiązywania danej wersji arkusza egzaminacyjnego przy kontroli poziomu umiejętności

Pozwala na mierzenie wielkości efektu w postaci tzw. stałego ilorazu szans (*constant odds ratio*) (cOR),

Ocena skali konsekwencji różnic między formami testu

Model IRT - dwuparametryczny model logistyczny (2PLM)

Stopień w jakim różnice między zadaniami w dwóch wersjach testu wpływają na wyniki uzyskane na całym teście

Oszacowania poziomu wiedzy uczniów są niezależne od testu; właściwości psychometryczne pozycji testowych są niezależne od próby



Czy istnieją różnice w poziomach wykonania zadań między dwiema wersjami testu?

	Poziom wykonania zadania (w nawiasie prawidłowa odpowiedź)		Różnica w poziomach wykonania zadania między wersjami A i B arkusza
	wersja A	wersja B	
z11_1	0,55 (C)	0,43 (C)	0,12
z11_2	0,91 (B)	0,91 (C)	0,00
z11_3	0,66 (C)	0,54 (C)	0,12
z18_1	0,98 (A)	0,98 (A)	0,00
z18_2	0,43 (A)	0,57 (B)	0,13
z18_3	0,64 (A)	0,70 (C)	0,06

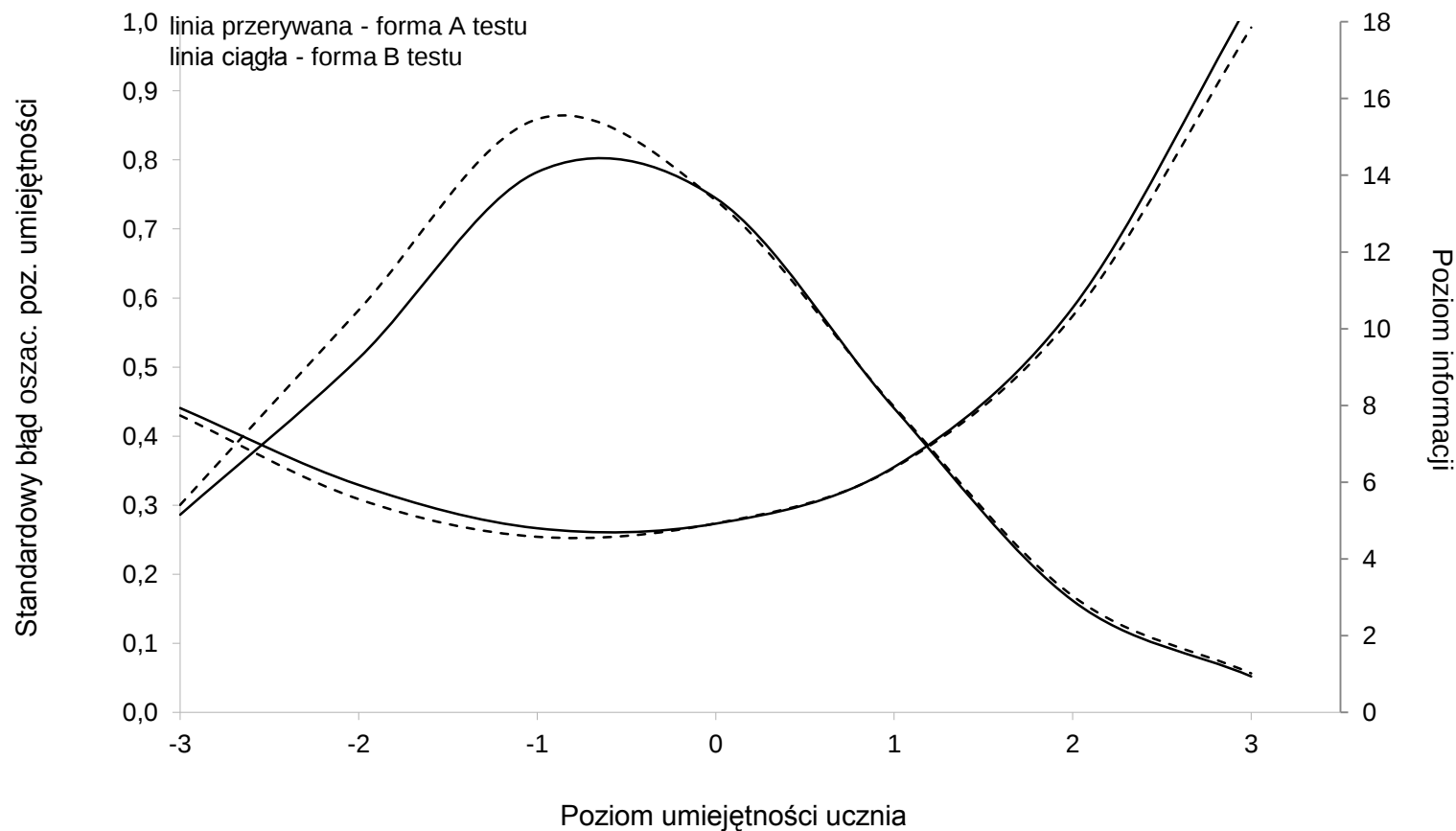


Czy różnice w poziomach wykonania zadań między dwiema wersjami testu są istotne oraz jak silny jest to efekt?

	chi-kwadrat MH	istotność	cOR	ln(cOR)	MH P-DIF	STD P-DIF
z11_1	1390,44	0,000	0,55	-0,59	0,15	0,12
z11_3	1529,21	0,000	0,55	-0,60	0,14	0,13
z18_2	1622,67	0,000	1,83	0,60	-0,15	-0,13



Jakie są konsekwencje dla wyników uzyskanych na całym teście?

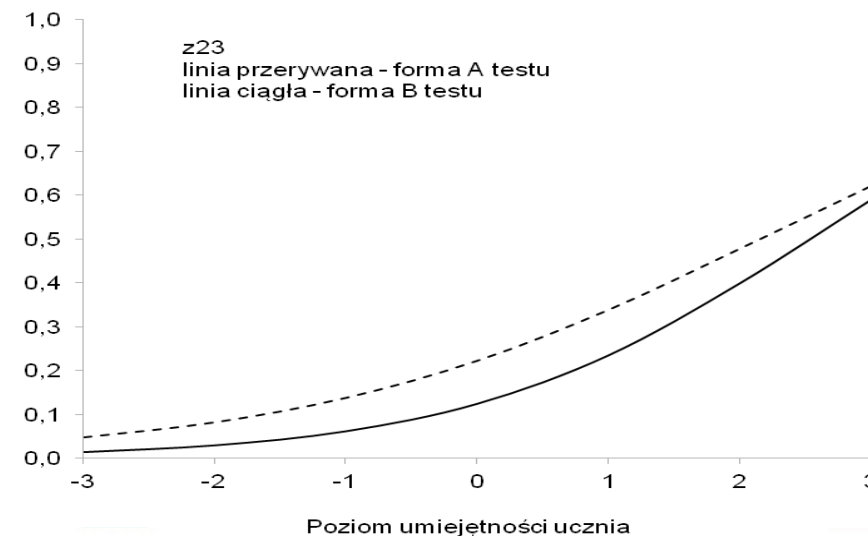
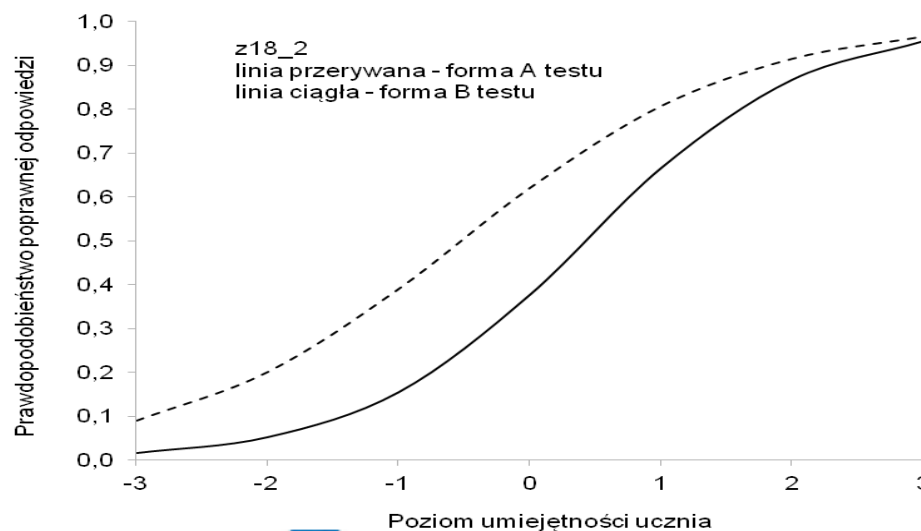
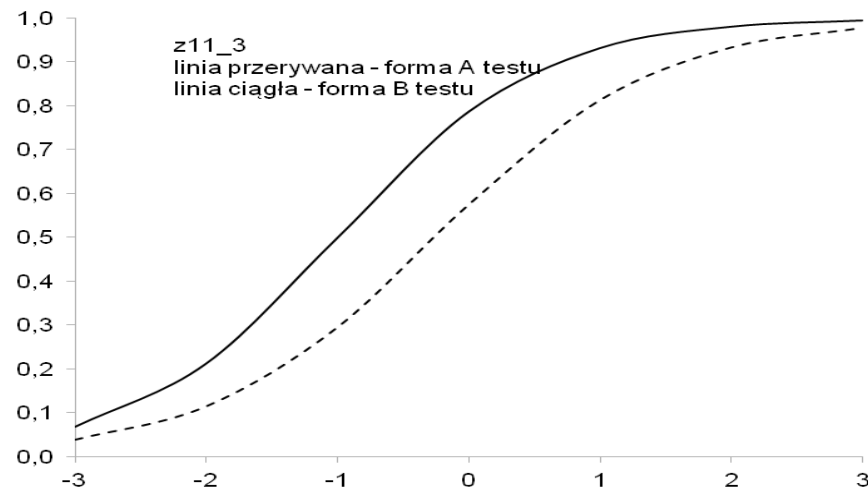
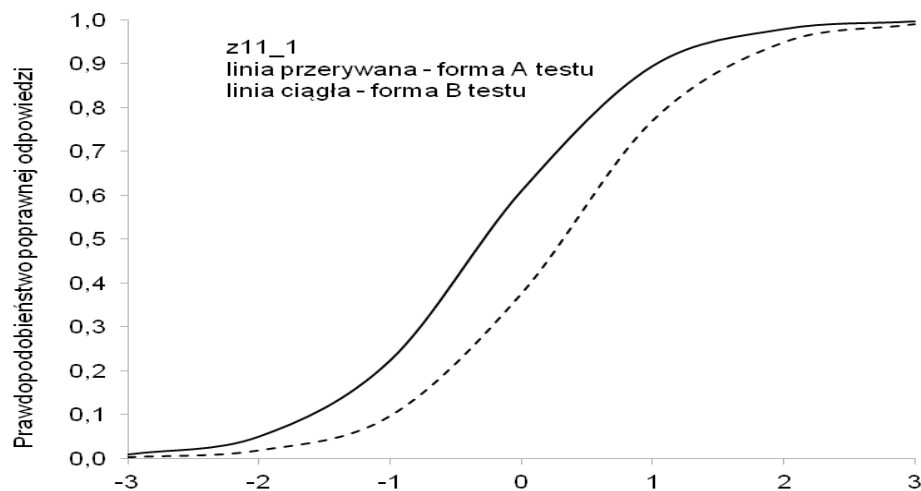


Istnieją różnice między krzywymi informacyjnymi oraz poziomami błędów dla wersji A i B



Jakie są konsekwencje dla wyników uzyskanych na całym teście?

Krzywe charakterystyczne wybranych pytań o wysokich DIF w wersji A i B arkusza z historii i WOS za 2013 r.





Czy zadania, których funkcjonowanie różni w zależności od wersji arkusza, charakteryzują się czymś szczególnym?

Hipoteza 1 – odpowiada za to forma zadawania pytań

- Wśród obciążonych pytań są pytania zamknięte różnego rodzaju, tj. zadanie na dobieranie, wielokrotnego wyboru, zadanie z luką, itp.
- Nie zdiagnozowano, że to forma pytania odpowiada za różne funkcjonowanie pytań między wersjami arkusza. W arkuszach egzaminacyjnych w 2012 i 2013 r. są inne zadania, o takiej samej konstrukcji, które nie powodują obciążenia.

Hipoteza 2 – jest to związane z układem odpowiedzi

- Zadania o największym DIF w analizowanym egzaminie są umieszczone w wiązkach zadań
- Tego rodzaju konstrukcja zadań zwiększa znacznie możliwość złamania założenia o lokalnej niezależności odpowiedzi na poszczególne zadania w teście



Skutki złamania założenia o lokalnej niezależności

- zadania zamknięte w wiązkę są ze sobą związane -> wariancja wyników testu inną niż związaną z różnicami w poziomie badanej umiejętności (*construct irrelevant variance*, CIV);
- błędne oszacowanie rzetelności i błędów standardowych poziomu umiejętności uczniów (Bradlow, Wainer i Wang, 1999);
- niepoprawne zrównywanie wyników testów (Lee, Kolen, Frisbie i Ankenmann, 2002);
- błędna interpretacja wielkości parametrów dyskryminacji zadań (Wainer i Wang, 2000);
- błędna diagnoza zadań jako niedopasowanych do struktury danych (Marais i Andrich, 2008).



Hipoteza 2 – DIF jest związany z układem odpowiedzi

	Symbol poprawnej odpowiedzi	
	wersja A	wersja B
z11_1	C	C
z11_2	B	C
z11_3	C	C
z18_1	A	A
z18_2	A	B
z18_3	A	C

- W zadaniu 11, prawidłowy wzór odpowiedzi w wersji B to trzy odpowiedzi C pod rząd.
- W zadaniu z18, w wersji A wzór poprawnych odpowiedzi to trzy A pod rząd



Hipoteza 2 – DIF jest związany z układem odpowiedzi

Czy wzór odpowiedzi poprawnych w wiązce może wpłynąć na udzielanie poprawnych odpowiedzi na zadania w ramach tej wiązki?

Sprawdzono to, wykonując serię regresji logistycznych, gdzie zmiennymi zależnymi były zadania o najmocniejszym DIF w wiązках 11 i 18.

Jako zmienne wyjaśniające w modelach umieszczano sumę poprawnych odpowiedzi na wszystkie pozycje w teście (dzięki temu kontrolowano poziom umiejętności uczniów) oraz interakcję odpowiedzi na pozostałe zadania w danej wiązce (informacja o tym czy uczeń rozwiązał poprawnie **dwa** pozostałe pytania w wiązce, czy nie)



Hipoteza 2 – DIF jest związany z układem odpowiedzi

Wyniki regresji logistycznej dla zadania z11.1

	B	SE	Exp(B)
Wynik testu	0,216**	0,003	1,241
z11.2*z11.3	-0,595**	0,025	0,551
Stała	-4,247**	0,049	0,014

Prawidłowe odpowiedzi na pozostałe pytania w wiązce obniżają szansę poprawnej odpowiedzi na pozycję z11.1 o 45% (przy kontroli wpływu poziomu umiejętności uczniów)

Wyniki regresji logistycznej dla zadania z11.3

	B	SE	Exp(B)
Wynik testu	0,170**	0,002	1,186
z11.1*z11.2	-0,575**	0,026	0,563
Stała	-2,918**	0,044	0,054

Prawidłowe odpowiedzi na pozostałe pytania w wiązce obniża szansę poprawnej odpowiedzi na pozycję z11.3 o 44% (przy kontroli wpływu poziomu umiejętności uczniów)

Wyniki regresji logistycznej dla zadania z18.2

	B	SE	Exp(B)
Wynik testu	0,175**	0,002	1,191
z18.1*z18.3	-0,905**	0,025	0,404
Stała	-3,159**	0,043	0,042

Prawidłowe odpowiedzi na pozostałe pytania w wiązce obniża szansę poprawnej odpowiedzi na pozycję z18.2 o 60% (przy kontroli wpływu poziomu umiejętności uczniów)



Hipoteza 2 – DIF jest związany z układem odpowiedzi

- Przyczyną występowania DIF w tych zadaniach może być mechanizm powiązany ze złudzeniem gracza (*Gambler's fallacy*)
- Złudzenie gracza jest błędem poznawczym opierającym się na błędnym rozumieniu zjawiska prawdopodobieństwa (Kahneman, Tversky, 1971, 1972).
- Można je rozumieć jako błędne przekonanie o tym, że właściwości dużych prób losowych stosuje się również do małych prób.
- Drugie wyjaśnienie wskazuje na wpływ heurystyki reprezentatywności w szacowaniu prawdopodobieństwa wystąpienia danego zjawiska - dla osób mniej prawdopodobne jest występowanie zjawiska, które w ich świadomości jest rzadkie, mimo iż nie ma podstaw by w jego przypadku oczekiwać jakiegś regularności.



Problem lokalnej niezależności czy złudzenie gracza

- Obserwowane rezultaty DIF można wiązać z występowaniem złudzenia gracza
- Jednak zadania w wiązce mają złamane założenie o lokalnej niezależności: odpowiedzi na poszczególne zadania w wiązce są ze sobą związane.
- Czy zastosowanie metod IRT dedykowanych do analizy zadań w wiązkach pozwoli na zmniejszenie zróżnicowania funkcjonowania tych zadań?

Metody zorientowane na wynik

sumowanie punktów za zadania w ramach wiązki i tworzenie z nich jednego zadania, tzw. „super zadania

najczęściej stosuje się jednowymiarowe modele IRT do zadań wielopunktowych (najczęściej generalized partial credit model, GPCM)

pomniejsza pulę zadań, w oparciu o które szacowany jest poziom konstruktu głównego, przez co obniża precyzję jego pomiaru

nie jest brany pod uwagę wzór odpowiedzi ucznia na zadania składające się na wiązkę

Metody zorientowane na zadanie

wprowadzanie dodatkowych wymiarów (*testlet factors*): pozwala modelować wariancję specyficzną dla wiązki zadań równocześnie z wariancją czynnika głównego

model podwójnego czynnika (*bifactor model*): każde zadanie jest wskaźnikiem konstruktów głównego oraz ładuje jeden z wymiarów reprezentujących wiązkę zadań, do której zadanie należy

model z czynnikiem wyższego rzędu (*higher order factor*): zadania ładują tylko wymiary specyficzne. Korelacje między wymiarami specyficznymi są modelowane przez czynnik wyższego rzędu.

model wiązki zadań (*testlet model*) model podwójnego czynnika mieszanych efektów, w którym ładunki zadań na wymiary specyficzne są zdefiniowane jako proporcjonalne do ładunków tych zadań na konstrukt główny w ramach każdej wiązki zadań



Testlet Response Theory

- Analizy mające na celu detekcję DIF powtórzono ale jako miarę poziomu umiejętności zastosowano oszacowania z modelu 2PL podwójnego czynnika (Reise, 2012)
- **model podwójnego czynnika** (*bifactor model*; Gibbons i Hedeker, 1992), w którym każde zadanie testowe jest wskaźnikiem konstruktów głównego oraz dodatkowo ładuje jeden ze specyficznych wymiarów reprezentujących wiązkę zadań, do której dane zadanie należy



Decyzja czy modelować wiązki?

- Modelowanie efektów wiązek zadań, które są nieistotne sprawia, że model jest niepotrzebnie skomplikowany i zwiększa błędy oszacowań parametrów zadań
- Efekt wiązki zadań jest istotny, gdy złamane jest założenie o lokalnej niezależności w parach lub grupach zadań ujętych w wiązki
- Dlatego, analizę należy rozpocząć od identyfikacji zadań, dla których złamane jest założenie o lokalnej niezależności
- Chen i Thissen (1997) zaproponowali statystykę $LD Xsq$ do detekcji złamania założenia o lokalnej niezależności
- Wartości statystyki powyżej 10 wskazują na złamanie założenia o lokalnej niezależności (Cai, Thissen i du Toit, 2011)



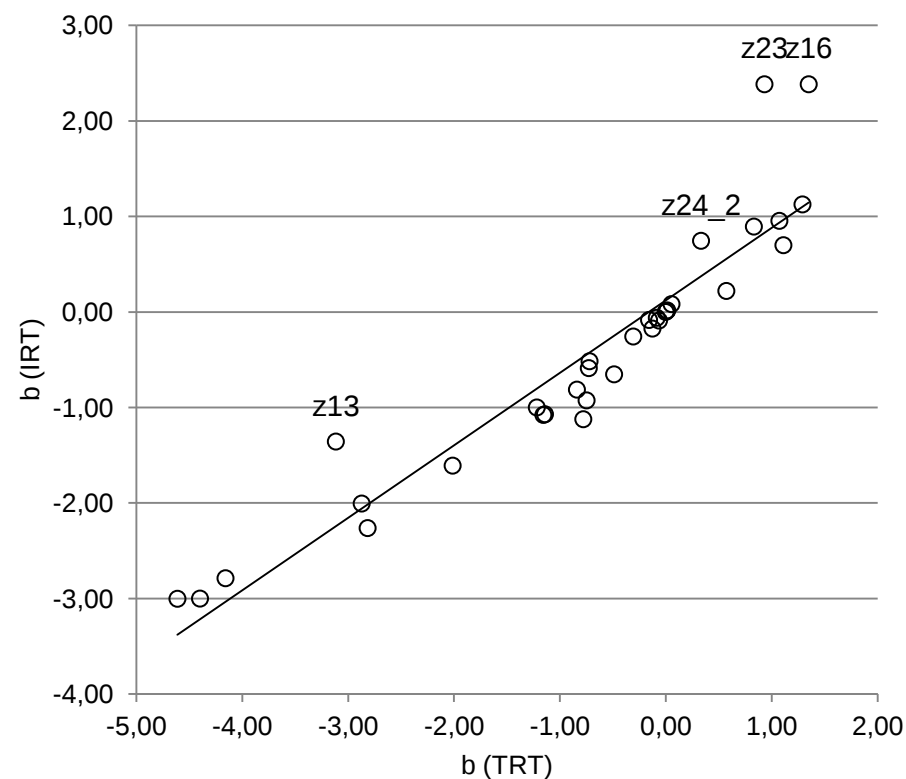
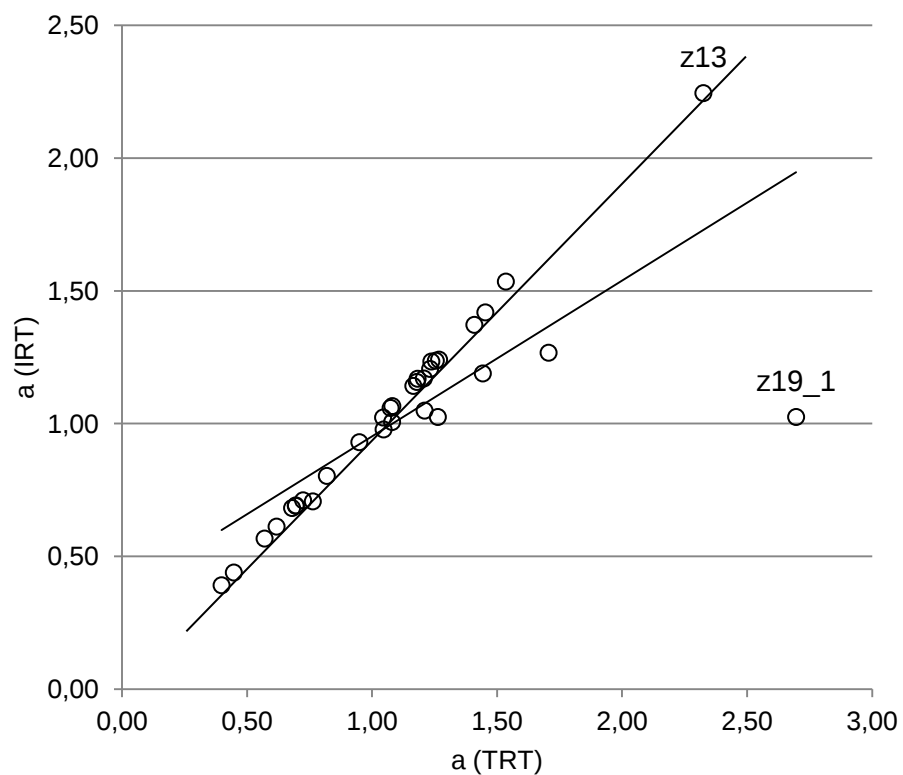
Decyzja czy modelować wiązki?

- Test niezależności Chi-sq na populacji >80K jest niekonkluzywny, dlatego do oceny złamania zał. o lok. niezal. wykorzystano losową próbkę $n=1100$ (analizę wykonano raz, można ją powtarzać n-razy dla zwiększenia pewności)
- Istotne (>10) statystyki testowe X-sq zaobserwowano w wiązках z11, z19, z20, z24 ale nie w z18
- Spośród 5 wiązek w teście, zidentyfikowano 4, w których złamane zostało założenie o lokalnej niezależności



Konsekwencje nieuwzględnienia wiązek

N=81 545

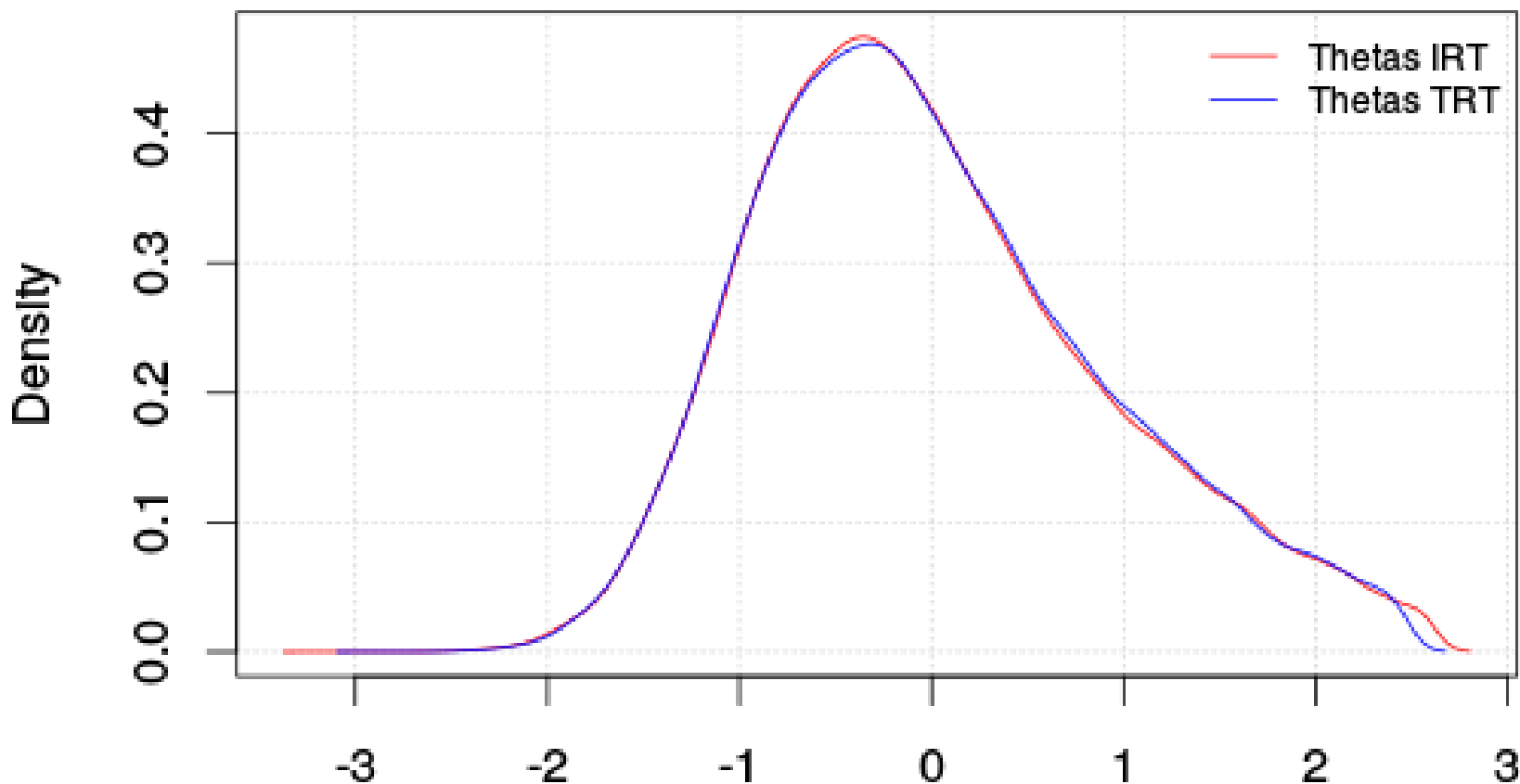




Konsekwencje nieuwzględnienia wiązek

n=81 545

- Znikomy wpływ na oszacowania umiejętności, co powinno skutkować brakiem różnic w detekcji DIF ze względu na wersję arkusza z wykorzystaniem sumy punktów, tradycyjnego 2PL (tu nie przedstawiono), oraz 2PL TRT





Iloraz wiarygodności			Test D ilorazu wiarygodności		Interpretacja	
Model a	Model b	Model c	Niejedno-rodny DIF	Jedno-rodny DIF	Niejedno-rodny DIF	Jedno-rodny DIF

Suma punktów

z11.1	-48017.2	-47320.9	-47320.0	1.8	1392.6***	Nie	Tak
z11.3	-49461.3	-48687.4	-48681.2	12.4***	1547.8***	Tak	Nie
z18.2	-51757.5	-50950.1	-50922.7	54.8***	1614.8***	Tak	Nie

Theta EAP 2PL TRT

z11.1	-47309.3	-46628.6	-46623.4	10.5***	1361.4***	Tak	Nie
z11.3	-49846.0	-49102.6	-49099.6	6.0***	1486.8***	Tak	Nie
z18.2	-52447.4	-51633.4	-51631.3	4.3***	1627.9***	Tak	Nie

(a) $\ln \left(\frac{P(u=1)}{P(u=0)} \right) = \beta_0 + \beta_1 X$, gdzie X to wynik końcowy ucznia na teście.

(b) $\ln \left(\frac{P(u=1)}{P(u=0)} \right) = \beta_0 + \beta_1 X + \beta_2 G$, gdzie G to wersja testu (A lub B).

(c) $\ln \left(\frac{P(u=1)}{P(u=0)} \right) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$, gdzie XG to interakcja między wynikiem końcowym a wersją arkusza.



EDUKACYJNA WARTOŚĆ DODANA

Dziękuję za uwagę.

p.majkut@uj.edu.pl

Instytut Badań Edukacyjnych

ul. Górczewska 8, 01-180 Warszawa

tel.: (22) 241 71 00, e-mail: ewd@ewd.edu.pl



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI



EDUKACYJNA
WARTOŚĆ
DODANA

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

