



# INSTYTUT BADAŃ EDUKACYJNYCH

## Wykorzystanie teorii odpowiedzi na wiązki zadań (*Testlet Response Theory*) w analizie wyników testów egzaminacyjnych

Przemysław Majkut  
Maciej Koniewski  
Paulina Skórska



KAPITAŁ LUDZKI  
NARODOWA STRATEGIA SPÓJNOŚCI

IBE



INSTYTUT  
BADAŃ  
EDUKACYJNYCH

Unia Europejska  
Europejski Fundusz Społeczny



**Wiązka zadań** - rdzeń (np. fragment tekstu, wykres, mapa, rysunek lub inny element graficzny, itp.), do którego zadawane są pytania (zadania testowe)

- efektywny sposób na przygotowywanie nowych zadań
- łatwe w administracji testu
- nie są oderwane od jakiegokolwiek kontekstu, jak zadania wielokrotnego wyboru (*multiple choice questions*, MCQ)
- dobrze sprawdzają się w pomiarze procesów umysłowych, wiedzy i umiejętności wyższego rzędu i tych bardziej złożonych
- krótszy czas ucznia potrzebny na zapoznanie się z treścią zadań

# Skutki złamania założenia lokalnej niezależności

- zadania zamknięte w wiązkę są ze sobą powiązane – generują wariację 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)

# Wiązki zadań w arkuszach egzaminu gimnazjalnego w latach 2004 - 2015

IBE



INSTYTUT  
BADAŃ  
EDUKACYJNYCH

| Część egzaminu            | Rok  | L. wiązek | L. pytań | Część egzaminu          | Rok  | L. wiązek | L. pytań |
|---------------------------|------|-----------|----------|-------------------------|------|-----------|----------|
| humanistyczna             | 2004 | 4         | 32       | matematyka              | 2012 | 2         | 23       |
| matematyczno-przyrodnicza | 2004 | 2         | 35       | j. polski               | 2012 | 4         | 23       |
| humanistyczna             | 2005 | 4         | 34       | przedmioty przyrodnicze | 2012 | 7         | 26       |
| matematyczno-przyrodnicza | 2005 | 7         | 36       | historia i WOS          | 2012 | 8         | 34       |
| humanistyczna             | 2006 | 5         | 27       | matematyka              | 2013 | 2         | 23       |
| matematyczno-przyrodnicza | 2006 | 6         | 35       | j. polski               | 2013 | 4         | 23       |
| humanistyczna             | 2007 | 4         | 31       | przedmioty przyrodnicze | 2013 | 4         | 28       |
| matematyczno-przyrodnicza | 2007 | 6         | 35       | historia i WOS          | 2013 | 7         | 33       |
| humanistyczna             | 2008 | 5         | 33       | matematyka              | 2014 | 2         | 23       |
| matematyczno-przyrodnicza | 2008 | 7         | 34       | j. polski               | 2014 | 4         | 23       |
| humanistyczna             | 2009 | 5         | 30       | przedmioty przyrodnicze | 2014 | 4         | 28       |
| matematyczno-przyrodnicza | 2009 | 9         | 37       | historia i WOS          | 2014 | 5         | 34       |
| humanistyczna             | 2010 | 6         | 30       | matematyka              | 2015 | 1         | 23       |
| matematyczno-przyrodnicza | 2010 | 9         | 37       | j. polski               | 2015 | 3         | 22       |
| humanistyczna             | 2011 | 4         | 30       | przedmioty przyrodnicze | 2015 | 4         | 28       |
| matematyczno-przyrodnicza | 2011 | 8         | 37       | historia i WOS          | 2015 | 2         | 32       |

## Metody zorientowane na wynik:

- sumowanie punktów za zadania w ramach wiązki i tworzenie z nich jednego zadania, tzw. „super zadania” (Bishop i Omar, 2002)
- najczęściej stosuje się jednowymiarowe modele IRT do zadań wielopunktowych (najczęściej *generalized partial credit model*, GPCM; Muraki, 1992)
- 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 (Yen, 1993)

## Metody zorientowane na zadanie:

- wprowadzają dodatkowe wymiary (czynniki specyficzne wiązki, testlet factors), które pozwalają modelować wariancję specyficzną dla wiązki zadań równocześnie z wariancją właściwą dla czynnika głównego (Bradlow, Wainer i Wang, 1999; Wainer, Bradlow i Du, 2000)

# Metody zorientowane na zadanie: modele *Testlet Response Theory*

- **model podwójnego czynnika** (*bifactor model*; Gibbons i Hedeker, 1992), w którym każde zadanie testowe jest wskaźnikiem konstruktu głównego oraz dodatkowo ładuje jeden ze specyficznych wymiarów reprezentujących wiązkę zadań, do której dane zadanie należy
- **model wiązki zadań** (*testlet model*; Bradlow, Wainer i Wang, 1999) to model podwójnego czynnika mieszanych efektów, w którym ładunki (parametry dyskryminacji) zadań na wymiary specyficzne są zdefiniowane jako proporcjonalne do ładunków tych zadań na konstrukt główny w ramach każdej wiązki zadań (Li, Bolt i Fu, 2006)
- **model z czynnikiem wyższego rzędu** (*higher order factor*, HOF), zadania testowe ładują tylko wymiary specyficzne; korelacje między wymiarami specyficznymi są modelowane przez czynnik wyższego rzędu

- Dla zilustrowania zastosowania modelowania wyników testów za pomocą modeli uwzględniających wiązki zadań, wybrano wyniki części egzaminu gimnazjalnego z historii i wiedzy o społeczeństwie z 2014 r. dla losowej próbki 1000 uczniów , pobranej z ogólnopolskiej populacji 362 752 zdających
- Wybór testu z historii i wiedzy o społeczeństwie był podyktowany faktem, że od 2012 w polskich egzaminach gimnazjalnych w tej części egzaminu występowało najwięcej wiązek zadań
- Test składał się z 33 zadań, z czego 14 tworzyło pięć wiązek:
  - k1: 'z6.1', 'z6.2'; k2: 'z8.1', 'z8.2', 'z8.3';
  - k3: 'z10.1', 'z10.2', 'z10.3';
  - k4: 'z15.1', 'z15.2', 'z15.3';
  - k5: 'z22.1', 'z22.2', 'z22.3'

# 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)
- Statystyki powyżej 10 odnotowano w grupach zadań 'z6.1', 'z6.2'; 'z8.1', 'z8.2', 'z8.3'; oraz 'z22.1', 'z22.2', 'z22.3'
- Spośród pięciu wiązek w teście, zidentyfikowano trzy, w których złamane zostało założenie o lokalnej niezależności

- model Rascha (jednoparametryczny) wiązki zadań (mod1; Wang i Wilson, 2005)
  - model dwuparametryczny podwójnego czynnika (mod2; Reise, 2012)
- 
- porównanie między modelem z wszystkimi potencjalnymi wiązkami a modelem z wszystkimi potencjalnymi wiązkami za wyjątkiem jednej wiązki (porównanie „all-but-one”) – odpowie na pytanie czy model uwzględniający wiązki jest lepiej dopasowany do danych?

# Wyniki – podsumowanie porównania modeli „all-but-one”

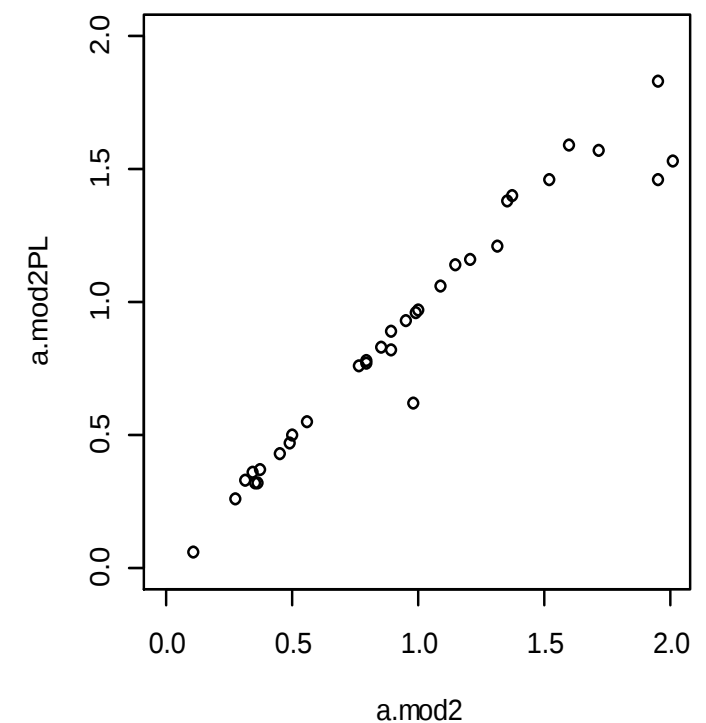
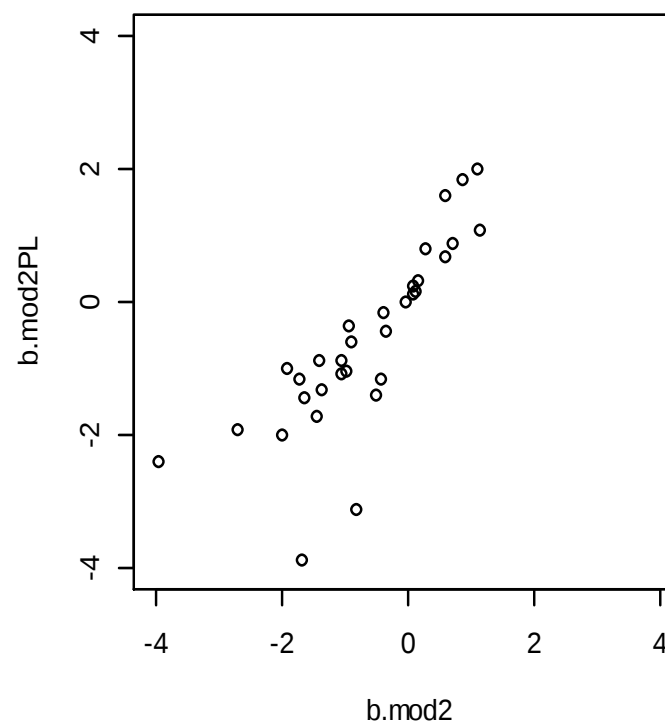
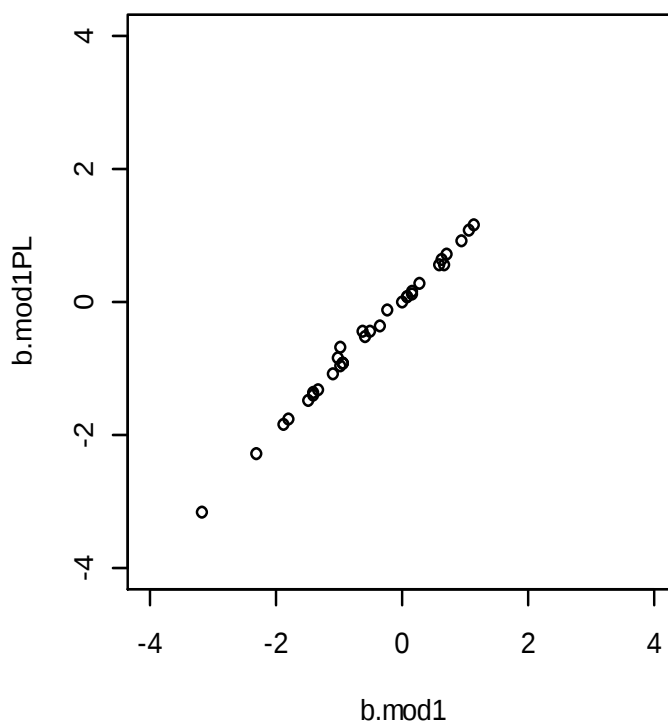
| Krótką nazwa modelu | L. parametrów | LL              | AIC            | SA-BIC         | Rzetelność EAP dim1 | k1           | k2           | k5           |
|---------------------|---------------|-----------------|----------------|----------------|---------------------|--------------|--------------|--------------|
| <b>mod1</b>         | <b>37</b>     | <b>-18777,6</b> | <b>37629,1</b> | <b>37693,0</b> | <b>0,758</b>        | <b>0,283</b> | <b>0,500</b> | <b>0,395</b> |
| mod1-k1             | 36            | -18791,9        | 37655,9        | 37718,1        | 0,763               | nd           | 0,503        | 0,406        |
| mod1-k2             | 36            | -18864,2        | 37800,4        | 37862,6        | 0,779               | 0,292        | nd           | 0,420        |
| mod1-k5             | 36            | -18838,6        | 37749,2        | 37811,4        | 0,755               | 0,287        | 0,507        | nd           |
| <b>mod2</b>         | <b>74</b>     | <b>-18477,8</b> | <b>37103,6</b> | <b>37231,5</b> | <b>0,806</b>        | <b>0,381</b> | <b>0,473</b> | <b>0,633</b> |
| mod2-k1             | 72            | -18497,1        | 37138,3        | 37491,6        | 0,807               | nd           | 0,473        | 0,631        |
| mod2-k2             | 71            | -18573,8        | 37289,6        | 37412,3        | 0,819               | 0,318        | nd           | 0,628        |
| mod2-k5             | 71            | -18542,5        | 37227,0        | 37349,7        | 0,806               | 0,361        | 0,471        | nd           |

- Miary AIC i SA-BIC sugerują, że najlepsze modele to te, które uwzględniają wszystkie trzy wiązki zadań, zidentyfikowane jako istotne za pomocą statystyki LD Xsq
- Żadna z tych wiązek nie powinna być ignorowana

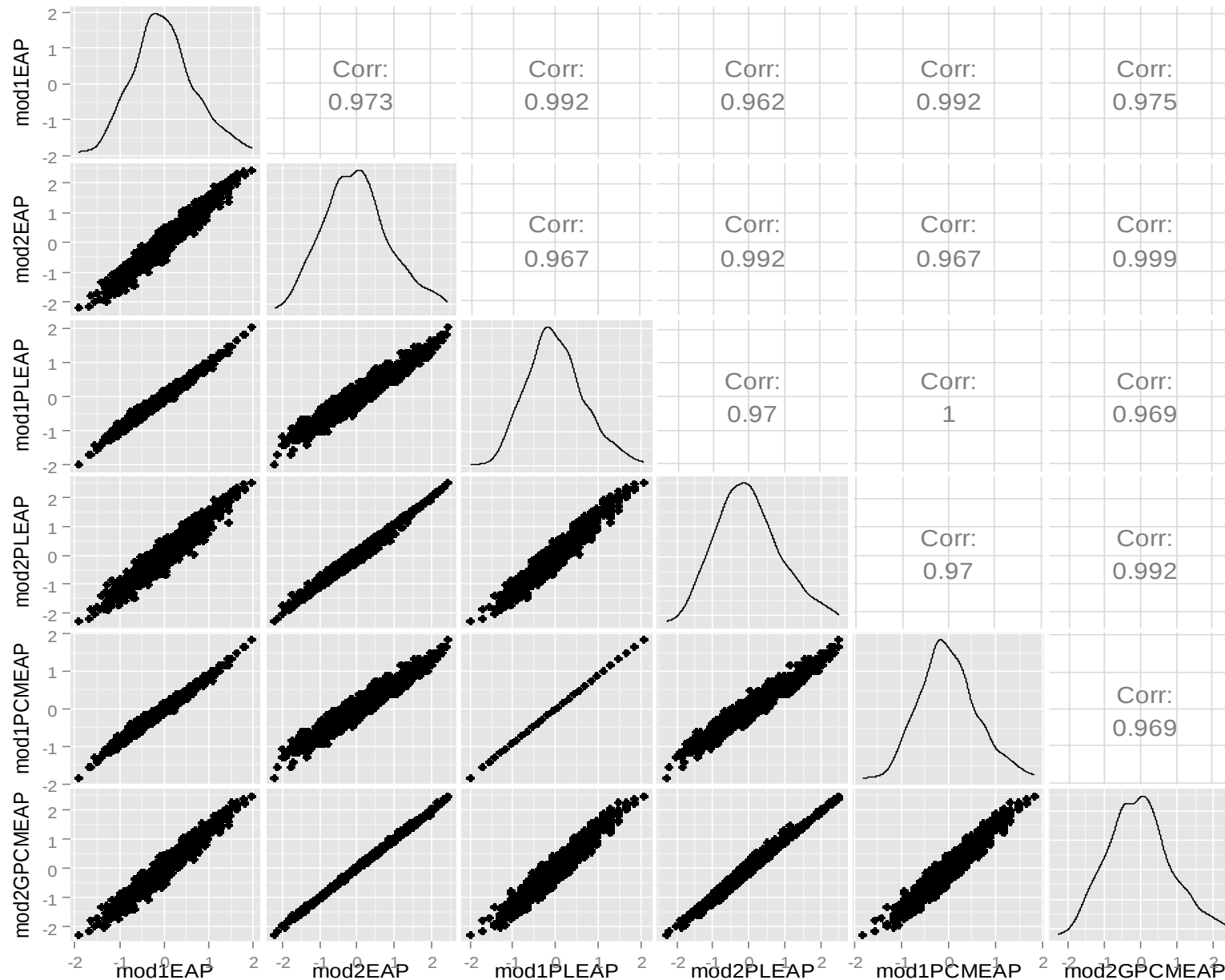
# Wyniki – statystyki podsumowania analizowanych modeli

| Krótką nazwa modelu | Opis modelu                                  | L. par. | LL       | AIC     | SA-BIC  | Rzeteln<br>ość<br>EAP<br>dim1 | k1    | k2    | k5    |
|---------------------|----------------------------------------------|---------|----------|---------|---------|-------------------------------|-------|-------|-------|
| mod1                | model Rascha wiązki zadań                    | 37      | -18777,6 | 37629,1 | 37693,0 | 0,758                         | 0,283 | 0,500 | 0,395 |
| mod1PL              | klasyczny model Rascha                       | 33      | -18948,2 | 37964,4 | 38023,3 | 0,778                         | nd    |       |       |
| mod1PCM             | klasyczny <i>Partial Credit Model</i>        | 30      | -17451,9 | 34971,7 | 35030,6 | 0,759                         |       |       |       |
| mod2                | model dwupar. podwójnego czynnika            | 74      | -18477,8 | 37103,6 | 37231,5 | 0,806                         | 0,381 | 0,473 | 0,633 |
| mod2PL              | klasyczny model dwuparametryczny             | 66      | -18655,3 | 37442,7 | 37557,0 | 0,820                         | nd    |       |       |
| mod2GPCM            | klasyczny <i>Graded Partial Credit Model</i> | 60      | -17183,8 | 34489,5 | 34595,1 | 0,759                         |       |       |       |

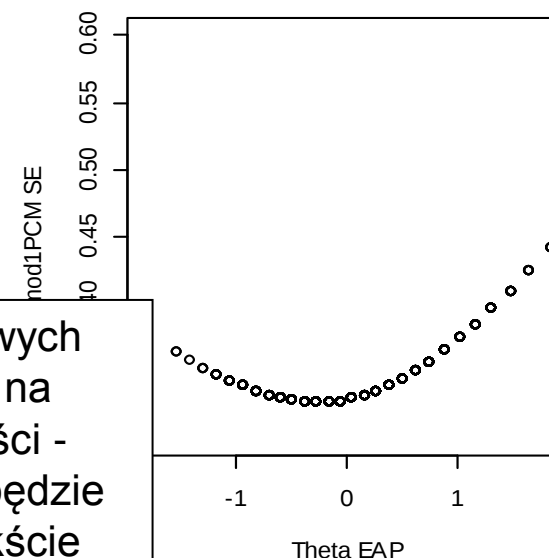
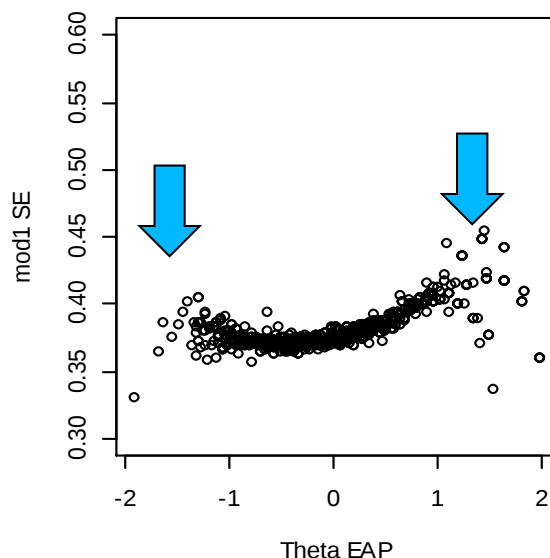
- Porównanie parametrów trudności:
  - w modelu Rascha wiązki zadań vs. modelu 1PL (Wykres 1);
  - w modelu podwójnego czynnika vs. modelu 2PL (Wykres 2);
- oraz porównanie parametrów dyskryminacji w modelu podwójnego czynnika vs. modelu 2PL (Wykres 3)
- Na Wykresie 2 nie umieszczono parametru b dla zadania z22.3, który w modelu 2PL wyniósł 8,912, a w modelu podwójnego czynnika 4,467
- Użycie klasycznego modelu IRT prowadziło do wniosku, że zadanie to jest dwukrotnie trudniejsze niż pozwala stwierdzić model wiązki zadania



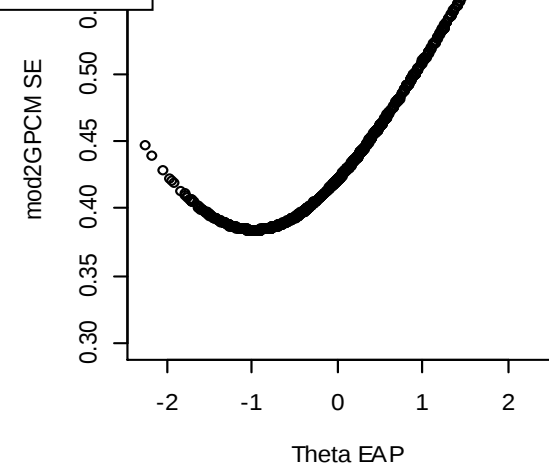
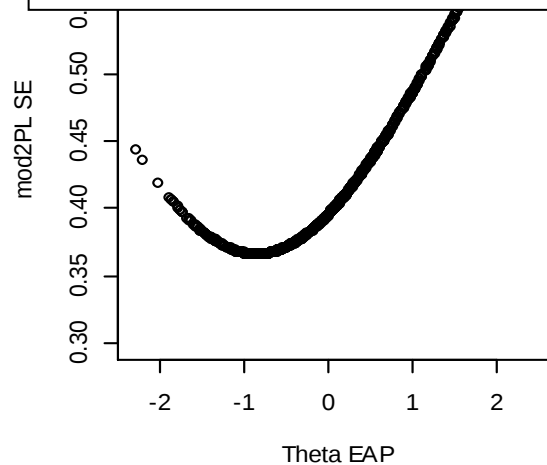
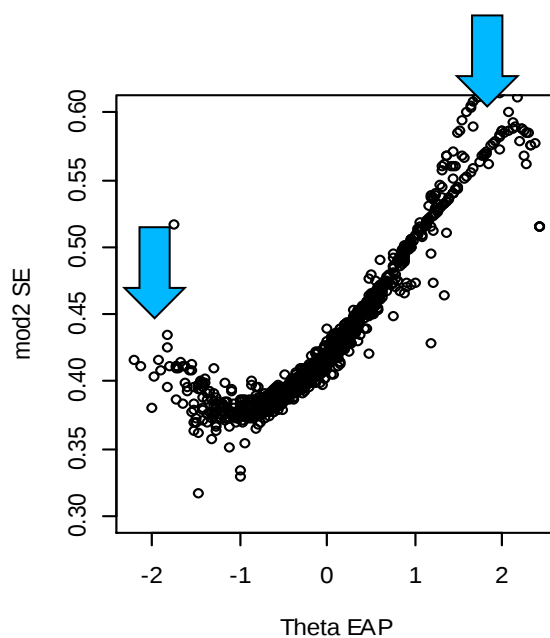
# Rozkłady umiejętności uczniów i ich korelacje w zastosowanych modelach



# Oszacowania błędów standardowych i parametrów uczniów w analizowanych modelach



Obniżenie błędów standardowych dla uczniów znajdujących się na krańcach rozkładu umiejętności - wybór modelu wiązki zadań będzie miał więc znaczenie w kontekście zwiększenia precyzji oszacowań umiejętności uczniów o najwyższych i najniższych umiejętnościach





# INSTYTUT BADAŃ EDUKACYJNYCH

Dziękuję za uwagę!

[p.majkut@ibe.edu.pl](mailto:p.majkut@ibe.edu.pl)

[m.koniewski@ibe.edu.pl](mailto:m.koniewski@ibe.edu.pl)

[p.skorska@ibe.edu.pl](mailto:p.skorska@ibe.edu.pl)



**KAPITAŁ LUDZKI**  
NARODOWA STRATEGIA SPÓJNOŚCI

**IBE**



INSTYTUT  
BADAŃ  
EDUKACYJNYCH

**Unia Europejska**  
Europejski Fundusz Społeczny

