INSTYTUT BADAŃ EDUKACYJNYCH

O zróżnicowanym funkcjonowaniu zadania (DIF) – w teorii i w praktyce

Tomasz Żółtak
Karolina Świst

Kraków, 23 września 2016

IBE INSTYTUT BADAŃ EDUKACYJNYCH

Czym jest DIF? (ze statystycznego punktu widzenia)

Ze zróżnicowanym funkcjonowaniem zadania mamy do czynienia, gdy osoby o **tym samym poziomie** mierzonej przez test umiejętności, ale należące do rożnych grup – typowo dwóch, tzw. referencyjnej i ogniskowej – charakteryzują się **różnym rozkładem odpowiedzi** na to zadanie.

Co oznacza, że udzielenie prawidłowej odpowiedzi na zadanie zależy również od innych czynników niż tylko mierzonej umiejętności

W zapisie matematycznym:

$$P(U_i = 1 | \theta, G = f) \neq P(U_i = 1 | \theta, G = r),$$

Na dzisiejszych warsztatach koncentrujemy się na detekcji oraz interpretacji występowania DIF dla:

- Jednowymiarowego testu.
- Dwóch grup będących przedmiotem zainteresowania – referencyjnej (*reference*, *r*) oraz ogniskowej (*f*, *focal*).

Przykłady DIF (1)

- Grudniewska i Kondratek (2013) – analiza egzaminu matematyczno-przyrodniczego z roku 2002:

Miłośnicy pływania

Zadanie 29. (0–3)

Marcin przebywa autobusem $\frac{3}{4}$ drogi do jeziora, a pozostałą część piechotą. Oblicz odległość między domem Marcina a jeziorem, jeżeli trasa, którą przebywa pieszo, jest o 8 km krótsza niż trasa, którą przebywa autobusem. Zapisz obliczenia.

3 kryteria oceny:

- Ustalenie zależności między poszczególnymi odcinkami szukanej drogi - 1 pkt
- Ułożenie równania - 1 pkt
- Rozwiązanie równania (zapisanie poprawnego wyniku) - 1 pkt.

Przy tym samym poziomie umiejętności, dziewczęta wypadają lepiej niż chłopcy na pierwszych dwóch kryteriach.

Przykłady DIF (2)

- Koniewski, Majkut i Skórska (2014) – zróżnicowane funkcjonowanie zadań **ze względu na wersję arkusza**.

Przedstawione w tekście porozumienie zawarto w 11.1. ____. W cytowanym fragmencie dokumentu szlachta gwarantowała sobie 11.2. ____. Wspomniana w dokumencie forma wyboru króla została wprowadzona w dobie 11.3. ____.

Układ odpowiedzi na zadanie 11 w wersji A arkusza:

11.1. A. XIV w.	B. XV w.	C. XVI w.
11.2. A. przywileje ekonomiczne	B. tolerancję religijną	C. wzajemną pomoc
11.3. A. monarchii patrymonialnej	B. rządów absolutnych	C. demokracji szlacheckiej

Układ odpowiedzi na zadanie 11 w wersji B arkusza:

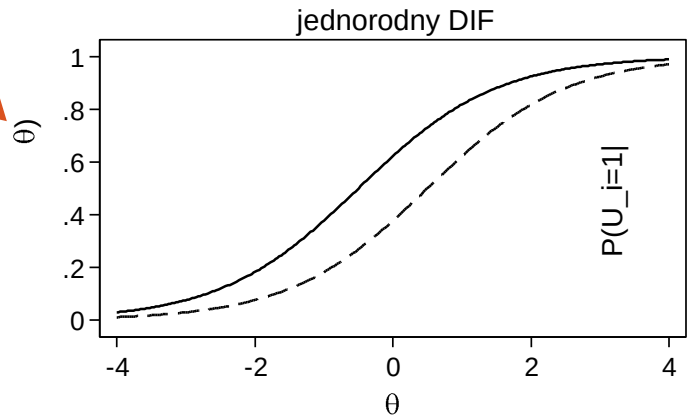
11.1. A. XIV w.	B. XV w.	C. XVI w.
11.2. A. przywileje ekonomiczne	B. wzajemną pomoc	C. tolerancję religijną
11.3. A. monarchii patrymonialnej	B. rządów absolutnych	C. demokracji szlacheckiej

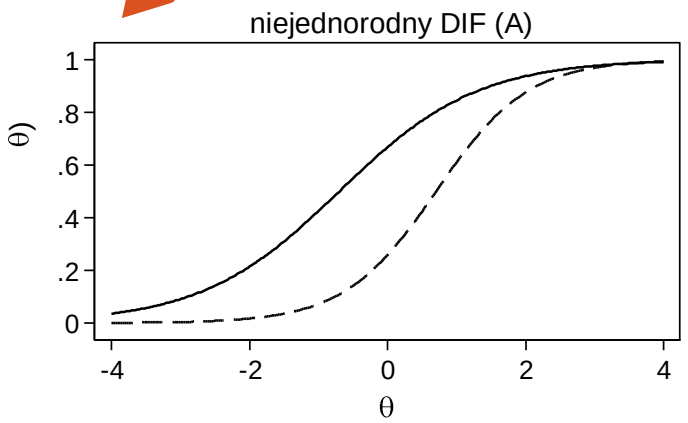
Czym jest DIF (ze statystycznego punktu widzenia) (2)

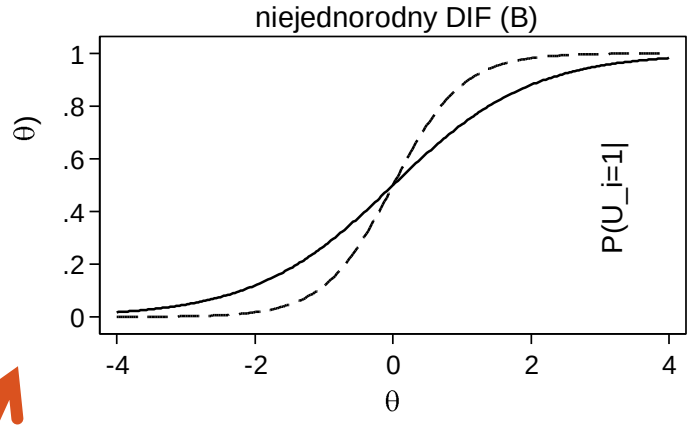
- **DIF jednorodne vs DIF niejednorodne**

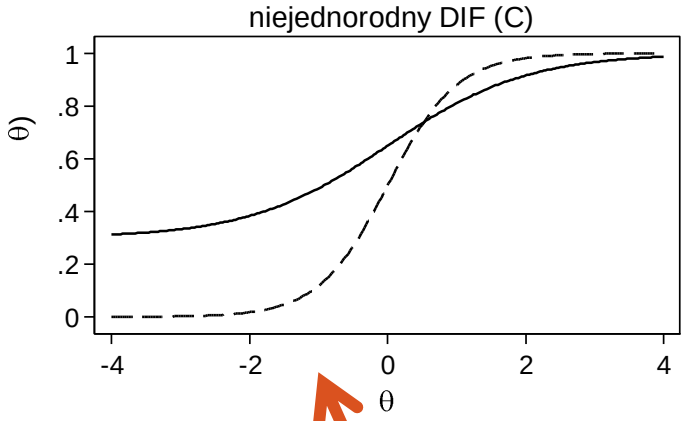
Stała (i różna od 1) wartość ilorazu szans między grupami

Różna trudność, różna dyskryminacja









Taka sama trudność, różna dyskryminacja

Różna dyskryminacja + różnica w parametrze pseudozgadywania

DIF a niezmienniczość/równoważność pomiaru (1)

W podejściach opartych na analizie czynnikowej bardzo często sprawdza się niezmienniczość pomiaru - czy ta sama cecha ukryta (zaufanie, religijność, wartości, etc) jest mierzona w analizowanych grupach (np. Cieciuch i in. 2014, Davidov, 2009 i wielu innych)

Jest to nic innego jak analog DIF:

- testowanie niezmienniczości **metrycznej** = testowanie **niejednorodnego** DIF

- testowanie niezmienniczości **skalarnej** (w sytuacji, gdy stwierdzimy, że niejednorodne DIF **nie występuje**) **=** testowanie **jednorodnego** DIF

DIF a stronniczość testu

Wykrycie zróżnicowanego funkcjonowania zadania jest warunkiem **koniecznym, ale niewystarczającym** do stwierdzenia, że jest ono stronnicze!

Stronniczość zadania = faworyzowanie jednej z grup wskutek występowania wpływu na wyniki danej pozycji testowej czynników wykraczających poza badaną przez dany test umiejętność (cechę).

Stronniczość to zaburzenie trafności testu, DIF to pojęcie czysto statystyczne.

Stwierdzenie stronniczości wymaga analizy eksperckiej.

Metody detekcji DIF (najbardziej popularne)

- **Test Mantela-Haenszela:**
  - porównanie liczby poprawnych i niepoprawnych odpowiedzi na dane zadanie w grupie ogniskowej i grupie odniesienia, przy kontroli poziomu umiejętności (najczęściej wyniku sumarycznego).
  - prosty, łatwy w interpretacji, dla jednorodnego DIF najmocniejszy test dla weryfikacji hipotezy zerowej o braku DIF …ALE jedynie dla zadań dychotomicznych + nie wykrywa niejednorodnego DIF.
- **Regresja logistyczna:**
  - szacuje się prawdopodobieństwo udzielenia poprawnej odpowiedzi przez ucznia w zadaniu ocenianym dychotomicznie z wykorzystaniem takich zmiennych niezależnych jak: całkowity wynik ucznia w teście, przynależność do grupy oraz interakcja wyniku w teście i przynależności do grupy.
  - Możemy modelować dwa rodzaje DIF w jednym równaniu, jednak nie da się oszacować w sposób bezpośredni rozkładów umiejętności.
- **Test oparty na ilorazie wiarygodności i podejściu IRT -> nim zajmiemy się dzisiaj.**

Test IRT-LR (1)

Ogólnie rzecz biorąc: przeprowadzamy test ilorazu wiarygodności (*likelihood ratio test*, LR) porównujący:

a) model zakładający, że między grupami parametry zadania potencjalnie posiadającego DIF mogą się różnić:

$$P(U=u\,|\,G)=\int\left[\prod_{n\in\{1,\ldots,N\}\setminus\{i\}} f(u_n,\theta,\beta_n)\right] f(u_{i,}\theta,\beta_{i,G})\psi_G(\theta)d\theta.$$

b) model zakładający, że zadanie funkcjonuje dla obu grup zgodnie z tym samym zestawem parametrów:

$$P(U=u\,|\,G)=\int\left[\prod_{n\in\{1,\ldots,N\}} f(u_{n,}\theta,\beta_n)\right]\psi_G(\theta)d\theta,$$

Są to dwa modele zagnieżdżone, które testujemy przy pomocy statystyki LR:

$$LR=-2\ln(\frac{L_0}{L_1}),$$

Test IRT-LR (2)

Co zyskujemy przy pomocy testu IRT-LR?

- Kontrola błędu pomiaru dla poziomu umiejętności.
- Możliwość uwzględnienie parametru (pseudo)zgadywania.
- Radzimy sobie z brakami danych najróżniejszych typów: zarówno losowymi, jak i wynikającymi ze schematu badania.
- Jesteśmy w stanie wykryć jednorodne i niejednorodne DIF.
- Umiemy wykryć DIF dla zadań wielokategorialnych (o wielu poziomach wykonania).

Ale: ten test wymaga odpowiedniego oprogramowania umożliwiającego szacowanie modeli wielogrupowych IRT (na przykład pakiet *mirt* w R).

Test IRT-LR (3) – IRT P-DIF

W praktyce, gdy analizujemy duże zbiory danych (jakimi są dane egzaminacyjne), testy statystyczne świadczące o występowaniu DIF stają się mniej przydatne -> ważna staje się wielkość efektu DIF

Wymienione wcześniej testy (MH, regresja logistyczna) mają swoje miary wielkości efektów, podejście IRT-LR posługuje się miarą IRT P-DIF.

$$\text{IRT P} - \text{DIF} = \int \sum_{x} x \left[f\left(x, \theta, \boldsymbol{\beta}_{i,f}\right) - f\left(x, \theta, \boldsymbol{\beta}_{i,r}\right) \right] \psi_f\left(\theta\right) d\theta$$

- informuje o ile różniłby się średni wynik w zadaniu *i*, gdyby funkcjonowało ono w grupie ogniskowej (f) zgodnie z właściwościami, jakie ma ono w grupie odniesienia (r).
- dla zadań dychotomicznych: o ile różniłaby się łatwość analizowanego zadania w grupie f, gdyby funkcjonowało ono w tej grupie tak, jak funkcjonuje w grupie r

Test IRT-LR (4) – kategoryzacja wielkości efektów

Miary wielkości efektu DIF wyrażone są na skali łatwości zadania (Monahan i in., 2007):

• Kategoria A – gdy test weryfikujący statystyczną istotność DIF dał wynik negatywny, lub gdy wynik testu jest pozytywny, ale absolutna wartość P-DIF jest mniejsza od 0,05;

• Kategoria B – gdy DIF jest statystycznie istotne oraz absolutna wartość P-DIF znajduje się w przedziale od 0,05 do 0,1;

• Kategoria C – gdy DIF jest statystycznie istotne oraz absolutna wartość P-DIF wykracza poza przedział (0,1).

DZIĘKUJEMY ZA UWAGĘ!

t.zoltak@ibe.edu.pl
k.swist@ibe.edu.pl