

# Přednáška 8–9: Teorie odpovědi na položku

---

24. a 31. 10. 2022 | PSYn4790 | Psychometrika: Měření v psychologii  
Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler | [hynek.cigler@mail.muni.cz](mailto:hynek.cigler@mail.muni.cz)

# Fundamentální („základní“) měření

---

**1. Přímé měření:** není odvozené z jiného měření, měří se přímo objekt...

- Délka (metr), váha (rovnoramenné váhy)...

... nebo **2. nepřímé měření:** je odvozené pomocí aditivních operací z naměřených hodnot.

- Nepřímé měření: Objem, čas, teplota, barva či síla zemětřesení (Richterova stupnice).

Podobné staršímu dělení na intenzivní vs. extenzivní veličiny, avšak vlastnost měření.

- Změna preferovaného principu měření u některých veličin (skládací metr vs. laserový dálkoměr).

Výsledkem je intervalová (příp. poměrová) škála s aditivní strukturou.

- Aditivita: možnost převést funkci „+“ do „×“ a základní aritmetické operace. Např.  $f(a + b) = f(a) + f(b)$ .
- Hodnoty tak lze „sčítat“ a „odčítat“.

**Důsledky:**

- Měření je „nezávislé“ na měřicím nástroji.
- Měřicí škála stále stejná pro všechny úrovně naměřených hodnot.

# Připomenutí měření v rámci CTT

---

Měření v rámci CTT je založeno na Stevensově definici.

- Výsledná (položková) data jsou proto nominální nebo ordinální, málokdy intervalová.
- Ze Stevensova pohledu je „měřením“ již odpověď na položku.
- Numerická data ale neznamenaají, že jde o „čísla“ v pravém slova smyslu.

Další CTT analýza ordinální není (součet položek...).

- CTT **pouze předpokládá**, že standardizované skóry odvozené z hrubých skóru jsou intervalová data. Dodržení aditivity neřeší.
- Pro výpočty používá míry centrální tendence a rozptylu (regrese, FA).
- Zachází tedy se škálami, jako kdyby fundamentální byly.

**Kdy zejména to vadí?**

# Jde o měření? | Likertova škála

Rosenber Self-Esteem Scale (první 4 položky)	souhlasím	spíše souhlasím	spíše nesouhlasím	nesouhlasím
Jsem se sebou vcelku spokojený/spokojená.	<b>3</b>	<b>2</b>	<b>1</b>	<b>0</b>
Občas si myslím, že jsem k ničemu.	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>
Cítím, že mám řadu dobrých vlastností.	<b>3</b>	<b>2</b>	<b>1</b>	<b>0</b>
Cítím, že toho není mnoho, na co bych u sebe mohl/mohla být hrdý/hradá.	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>

Celkový skór: **suma počtu bodů z dílčích položek.**

# Jde o měření? | Měření pozornosti

---

p	d	p	p	d	d	d	d	p	d
d	d	d	d	p	p	d	p	d	p

## Test pozornosti d2

Postupujte po řádcích a zaškrtněte všechna „d“ s 2 značkami nad nebo pod písmenem.

Celkový skór 1: **Počet prvků/řádků za jednotku času.**

Alternativní skór 1: **Čas průchodu testem.**

Celkový skór 2: **Počet chyb.**

# Měření v rámci CTT

Dotazník pro pacienty s anorexií  
(př. Bond & Fox, 2009):

- 1. Pravidelně zvracím, abych si udržel/a svou váhu.
- 2. Počítám gramy tuku na jídle, které jím.
- 3. Tvrdě cvičím, abych spálil/a kalorie.

**Odpovědi:** nesouhlasím (1), spíše nesouhlasím (2),  
tak napůl (3), spíše souhlasím (4), souhlasím (5)

- $r_{xx'} = 0,75$ ;  $M = 3$ ;  $SD = 3$ ;
- $SE = 1,5$ ,  $CI_{95\%} = 2,94$

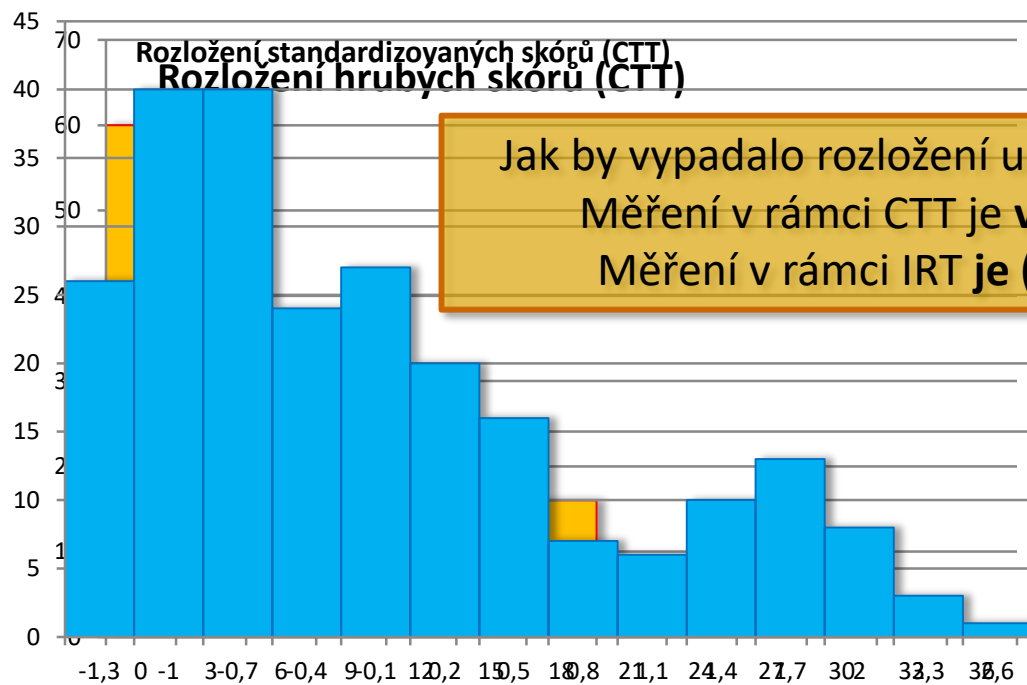
otázka	respondent 1	respondent 2
1	spíše nesouhlasím (2)	souhlasím (5)
2	spíše souhlasím (4)	souhlasím (5)
3	souhlasím (5)	nesouhlasím (1)
hrubý skór:	11 (6,06–11,94)	11 (6,06–11,94)

- CTT: oba lidé mají z hlediska CTTstejný hrubý skór, a tedy i míru anorexie i intervaly spolehlivosti.
- IRT: výsledky nejsou rovnocenné – jiný „person-fit“ (1PL), případně i chyby měření a skóry (2PL).

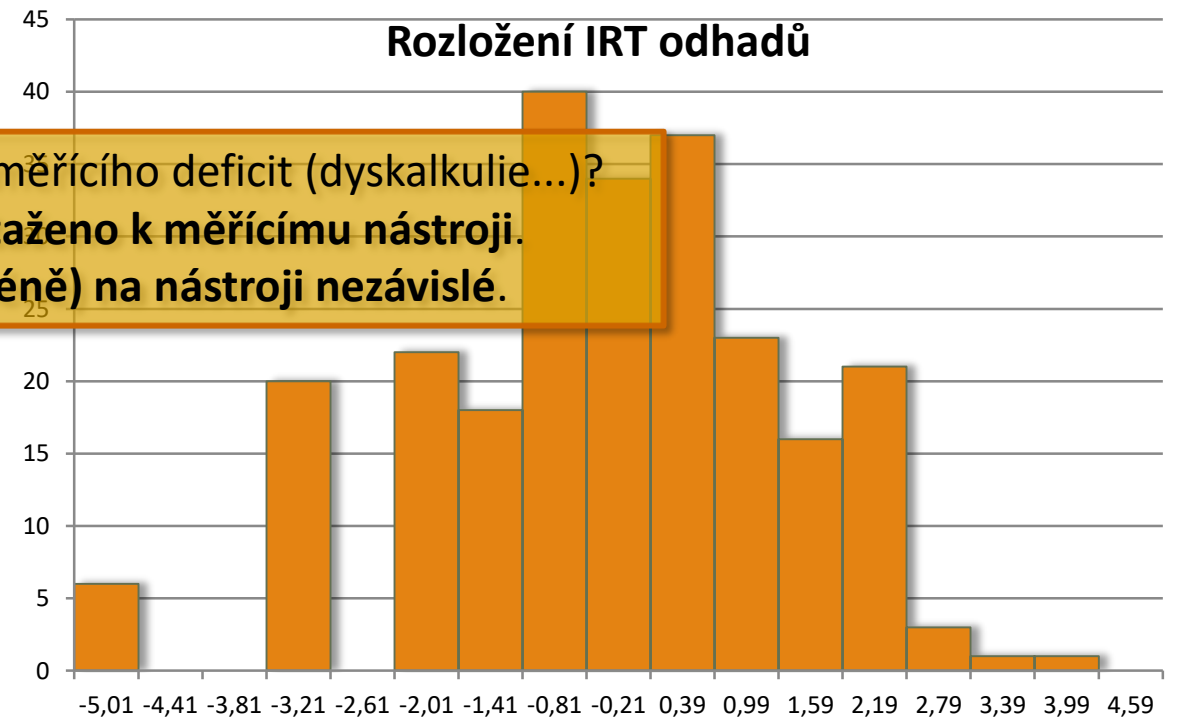
# Příklad: Nezávislost měření na nástroji

TIM<sup>3-5</sup>: Test pro identifikaci matematicky nadaných dětí

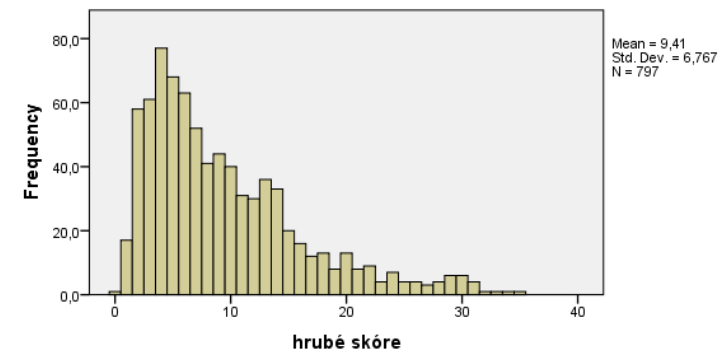
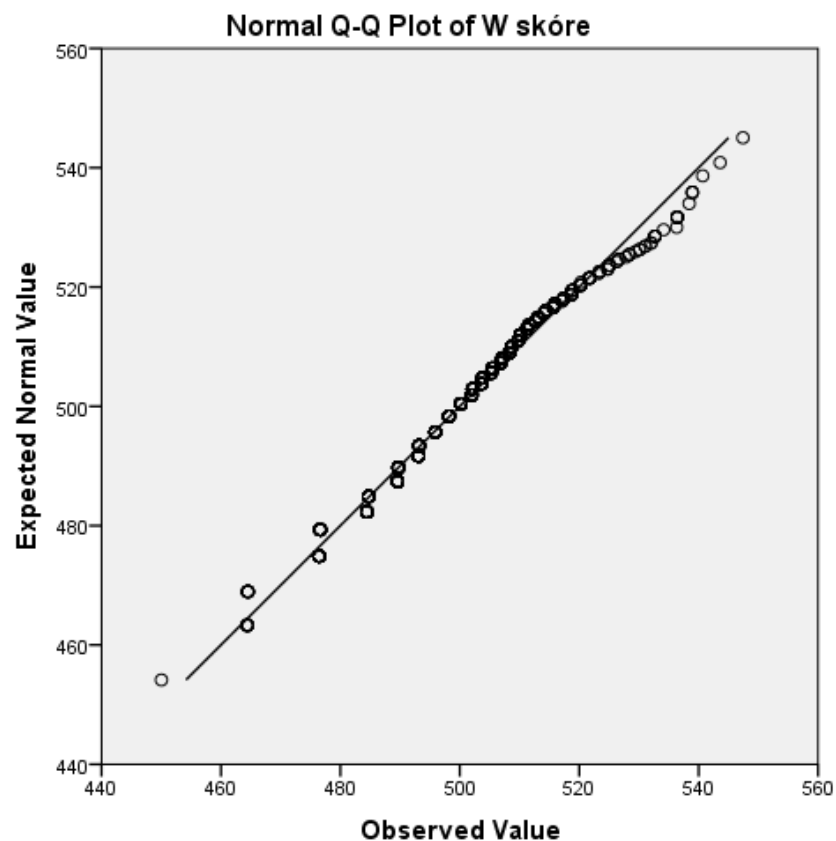
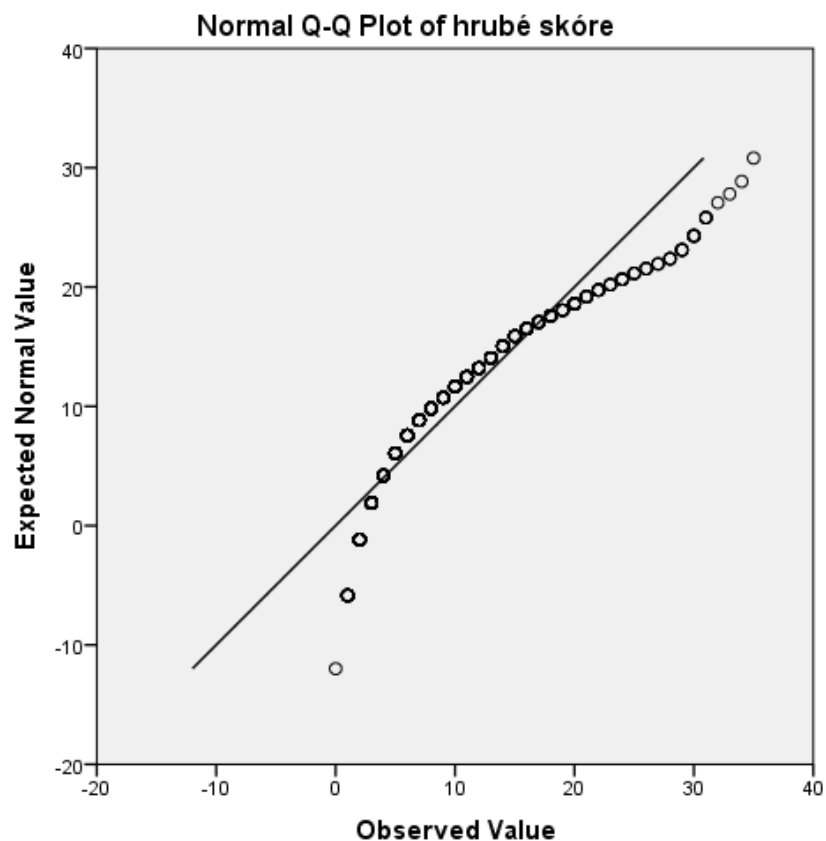
- Test je **velmi obtížný**, aby dobře měřil nadprůměr.
- $r_{xx'} = 0,82$ ;  $M = 8,51$ ;  $SD = 6,72$ ;  $\min = 0$ ;  $\max = 33$
- **Předpoklad:** Rozložení matematických schopností je v populaci normálně rozložené.
- **Závěr:** Jaké budou naměřené skóry?



Jak by vypadalo rozložení u testu, měřícího deficit (dyskalkulie...)?  
Měření v rámci CTT je **vždy vztaženo k měřicímu nástroji**.  
Měření v rámci IRT je **(více méně) na nástroji nezávislé**.



# Příklad: Nezávislost měření na nástroji



## Kolmogorův-Smirnovův test (MC, p-value)

ročník	3 (n=243)	4 (n=276)	5 (n=278)
hrubé skóre	<0,001	0,001	0,001
W- skóre	<0,001	,065	,061



# Vývoj teorií odpovědi na položku

50. a 60. léta, další rozvoj v 80. letech (počítače).

Nezávisle na sobě G. Rasch (matematik), F. M. Lord (psycholog, psychometrik) a P. F. Lazarsfeld (sociolog).

Jde o stochastickou úpravu původně deterministického Guttmanova modelu.

Tři hlavní stádia vývoje:

- Předchůdci, do 50. let (Binet, Guttman, Thurstone...)
- Raný vývoj, 50.–60. léta (Rasch, Novick, Lord...)
- Rozvoj, 70.–80./90. léta (Bock, Samejima...)
- Sjednocování a zobecňování (od 90. let)



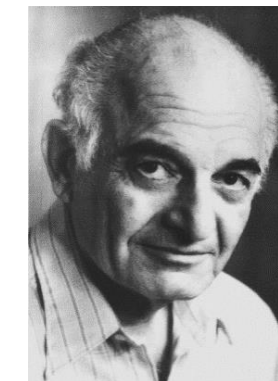
Paul Felix Lazarsfeld  
(1901–1976)



Georg Rasch (1901–1980)



Frederic M. Lord  
(1912–2000)



Louis Guttman  
(1916–1987)

# Extrémní příklad

Máme položku  
ve faktorové analýze

- Skórovaná ne=0,  
tak napůl=1, ano=2.
- Průsečík (intercept):  $b = 1$ .
- Faktorový náboj:  $\lambda = 0,5$ .

Faktor má průměr 0 (SD=1).

$$E(x_{ip}) = \lambda_i \theta_p + b_i$$

Jaká je očekávaná odpověď,  $E(x_i)$ ,  
respondenta s hodnotou faktoru...

...  $\theta = 0$  ?

- $E(x_i) = 1$

...  $\theta = 1$  ?

- $E(x_i) = 1,5$

...  $\theta = -1$  ?

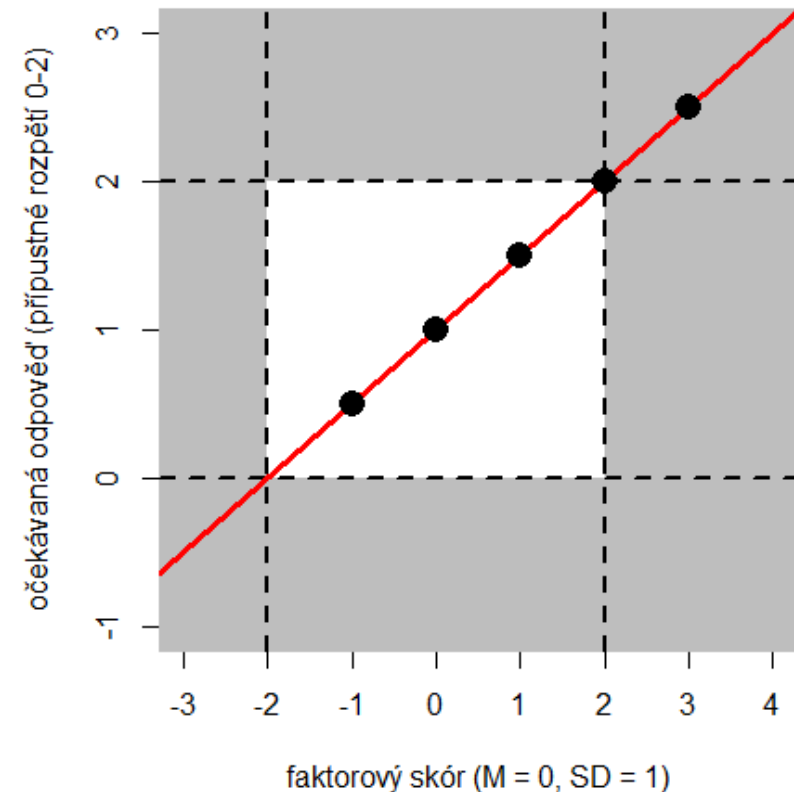
- $E(x_i) = 0,5$

...  $\theta = 2$  ?

- $E(x_i) = 2$

... a konečně  $\theta = 3$  ?

- $E(x_i) = 2,5$



Jaký je vztah měřeného rysu  
a odpovědi na binární položku  
(správně/špatně)?

Například vztah „fluidní inteligence“ a správné/špatné odpovědi  
na jednu úlohu v Ravenových progresivních maticích.

# Základy IRT:

## Charakteristická funkce položky (ICC)

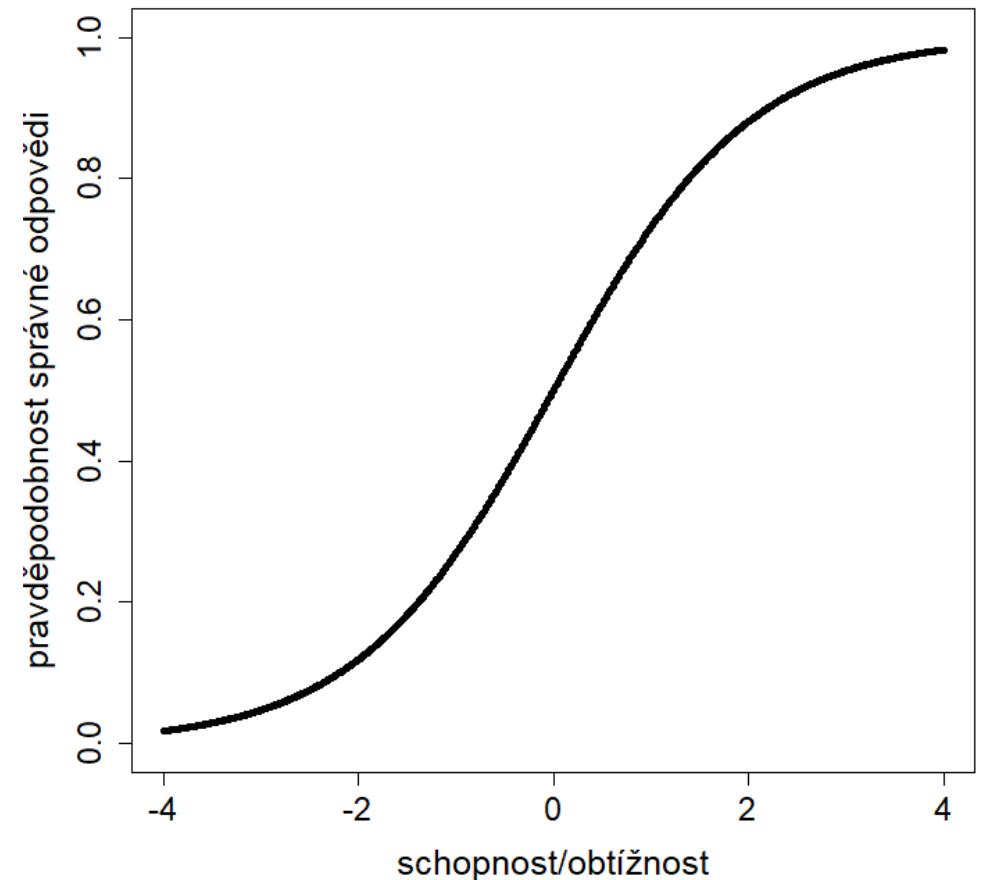
---

Výkon probanda v položce lze odhadnout pomocí množiny latentních rysů.

- Schopnosti respondenta.
- Parametry položek.

### Item Characteristic Curve (ICC):

- Má (zpravidla) přibližně tvar kumulativního normálního rozdělení.
- Popisuje vztah mezi schopnostmi probandů a očekávaným výkonem v dané položce.
- Pravděpodobnost správné odpovědi podle parametrů položky a probanda.
- Tvar ale může být prakticky libovolný (různé modely).



# Srovnání modelů měření (Borsboom, 2005)

---

## KLASICKÁ TESTOVÁ TEORIE

Měřený atribut: **Pravý skór daného člověka v daném testu.**

**Lineární vztah** pravého a pozorovaného skóre.

### **Homoskedasticita**

- Stejný chybový rozptyl pro všechny respondenty a všechny úrovně pravého skóre

## MODEL Y S LATENTNÍMI PROMĚNNÝMI

Měřený atribut: **Předpokládaný latentní rys.**

### **Faktorová analýza**

- **Lineární vztah** pozorované odpovědi a latentního rysu.
- Homoskedasticita reziduí.

### **Teorie odpovědi na položku**

- **Nelineární** (zpravidla logistický) **vztah** pozorované odpovědi a latentního rysu.

# FA jako specifický příklad IRT

---

FA lze chápat jako specifický případ IRT.

- Charakteristická funkce (vztah odpovědi a rysu) je lineární.
  - Mellenbergh, G.J. (2016). Models for Continuous Responses. In W.J. van der Linden (ed.), *Handbook of Item Response Theory* (vol. 1), 181-192. Chapman and Hall/CRC Press.

FA „váží“ odpovědi.

- V předchozím příkladu s anorexií by obě dívky měly odlišný odhad faktorového skóru.

Někdy totiž lze lineární vztah předpokládat.

- Např. hierarchická struktura v CHC, kdy „položkou“ je celý „subtest“.
- Např. reakční časy (jsou-li dostatečně dlouhé a normálně rozložené).
- Jiné dostatečně „jemné“ položky (jsou-li normálně rozložené).

Nedodržení předpokladu linearitě ale působí řadu obtíží.

- Vícedimenzionalita, zejm. tzv. „difficulty factor“ v inteligenčních testech (McDonald, [1965](#); ten Berge, [1972](#)).

# Jednoparametrový Raschův model (1PL)

---

Logistický vztah rysu a odpovědi:

$$P(x_i = 1|\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Analogicky po úpravě:

$$\ln \frac{P_{ip}}{1 - P_{ip}} = \theta_p - b_i$$

- e = Eulerova konstanta
- ln = přirozený logaritmus (se základem e)
- Pro zjednodušení zápisu  $P(x_i = 1|\theta_p) = P_{ip}$

$P(x_i = 1|\theta)$  je pravděpodobnost správné odpovědi na položku  $i$  při schopnosti  $\theta$ .

- Tato pravděpodobnost se někdy nazývá také „odhad pravého skóre“ respondenta v dané položce (u binárních položek), analogie k  $E(\tau_{pi})$ .

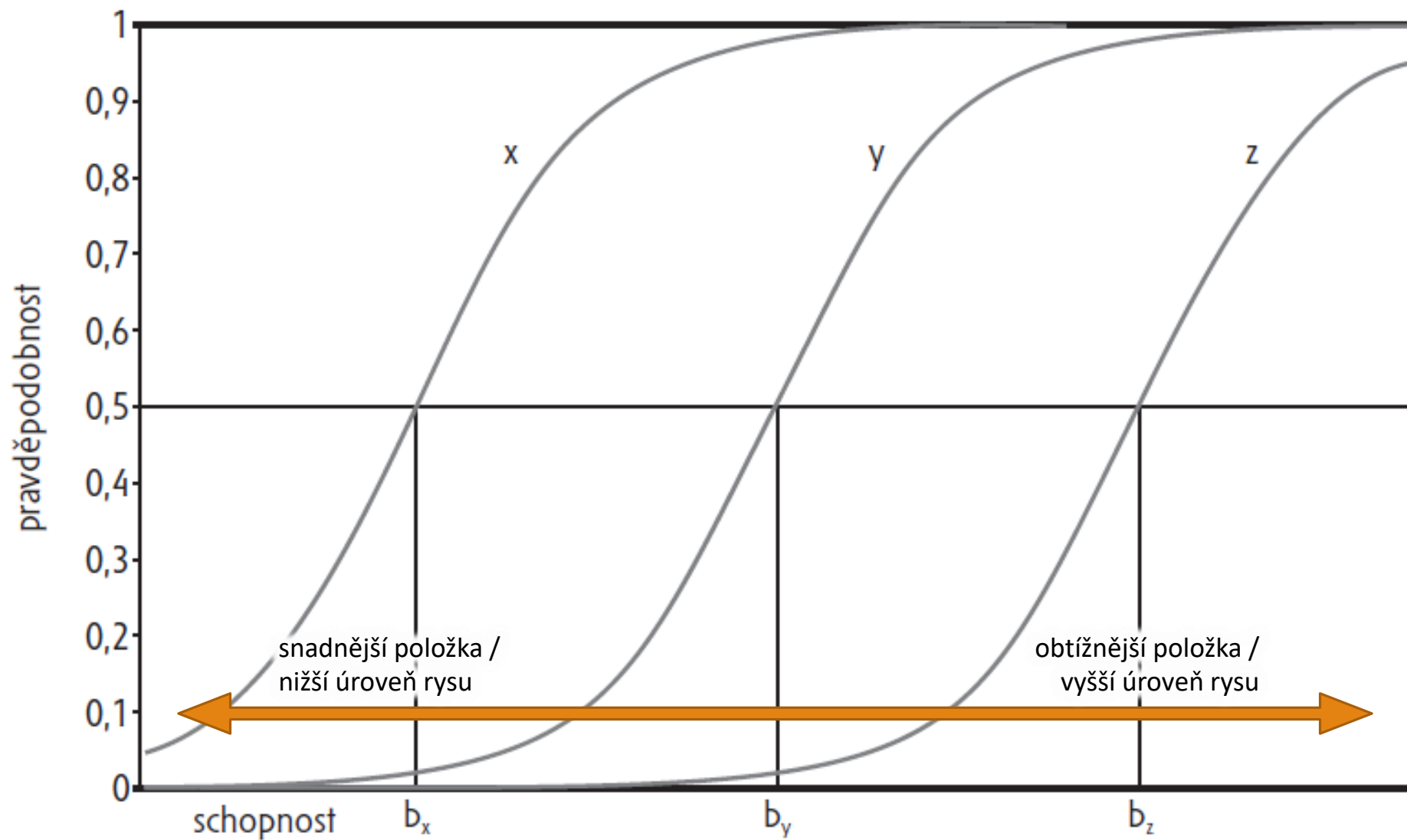
Theta ( $\theta_p$ ) je úroveň schopnosti respondenta  $p$ .

- Subskript  $p$  se zpravidla vynechává.

$b_i$  je parametr obtížnosti položky  $i$ .

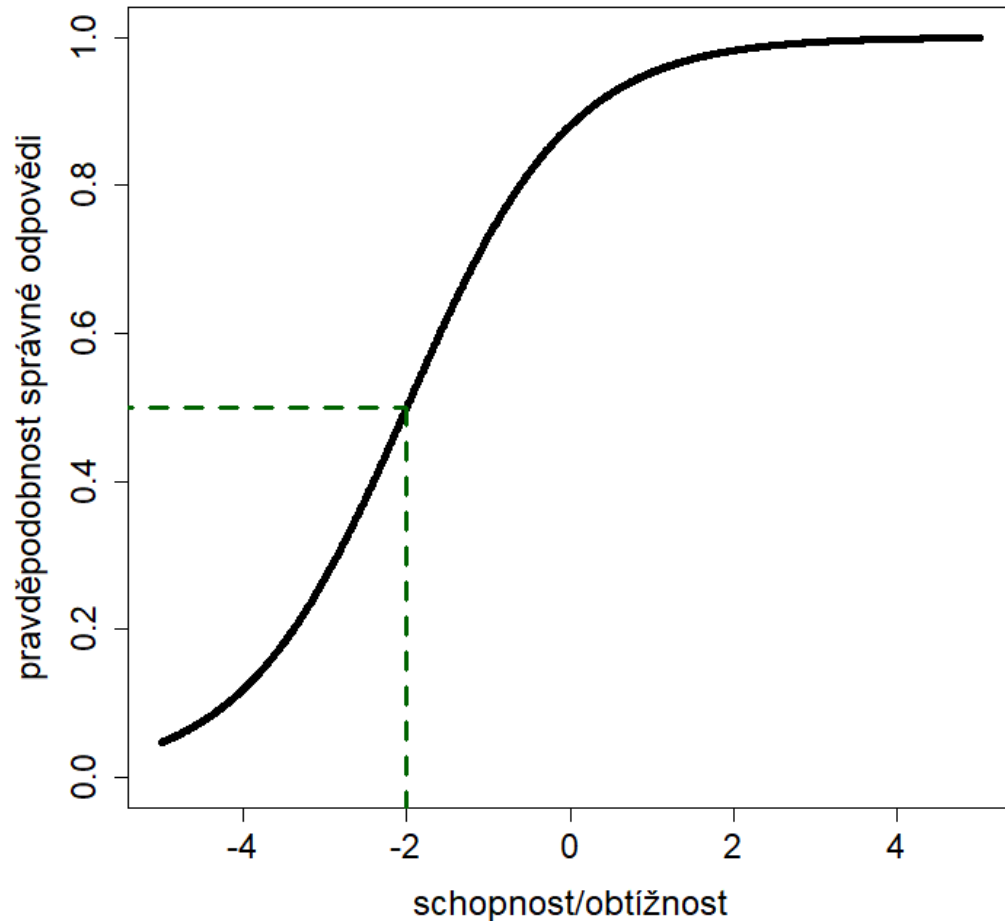
- Parametr obtížnosti  $b_i$  položky  $i$  je bod na škále schopnosti, v němž je pravděpodobnost správné odpovědi respondenta  $j$  se stejnou mírou schopnosti ( $\theta_p = b_i$ ) na danou položku  $P(x_i = 1|\theta) = 0,5$ .

<http://fssvm6.fss.muni.cz/ICC/>





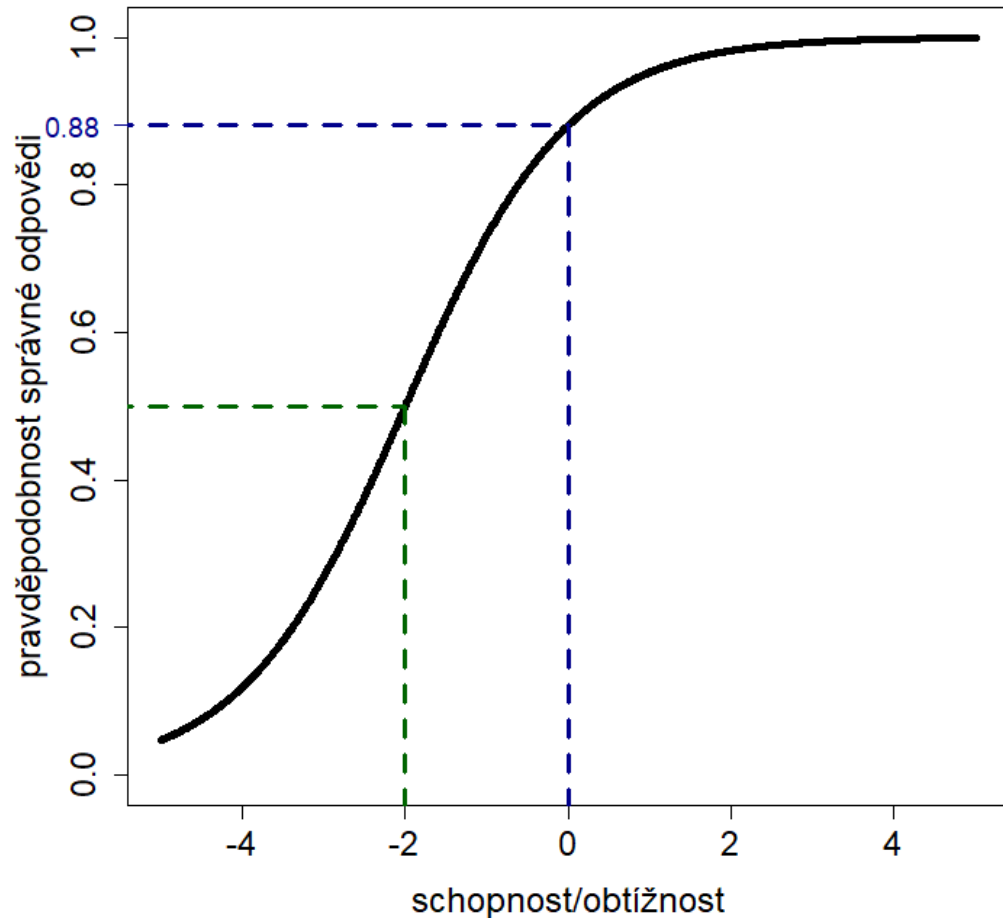
# Raschův model (jednparametrový)



Položka s obtížností  $b_i = -2$ .

Respondent se schopností  $\theta = b_i = -2$  má 50 % pravděpodobnost správné odpovědi.

# Raschův model (jednparametrový)



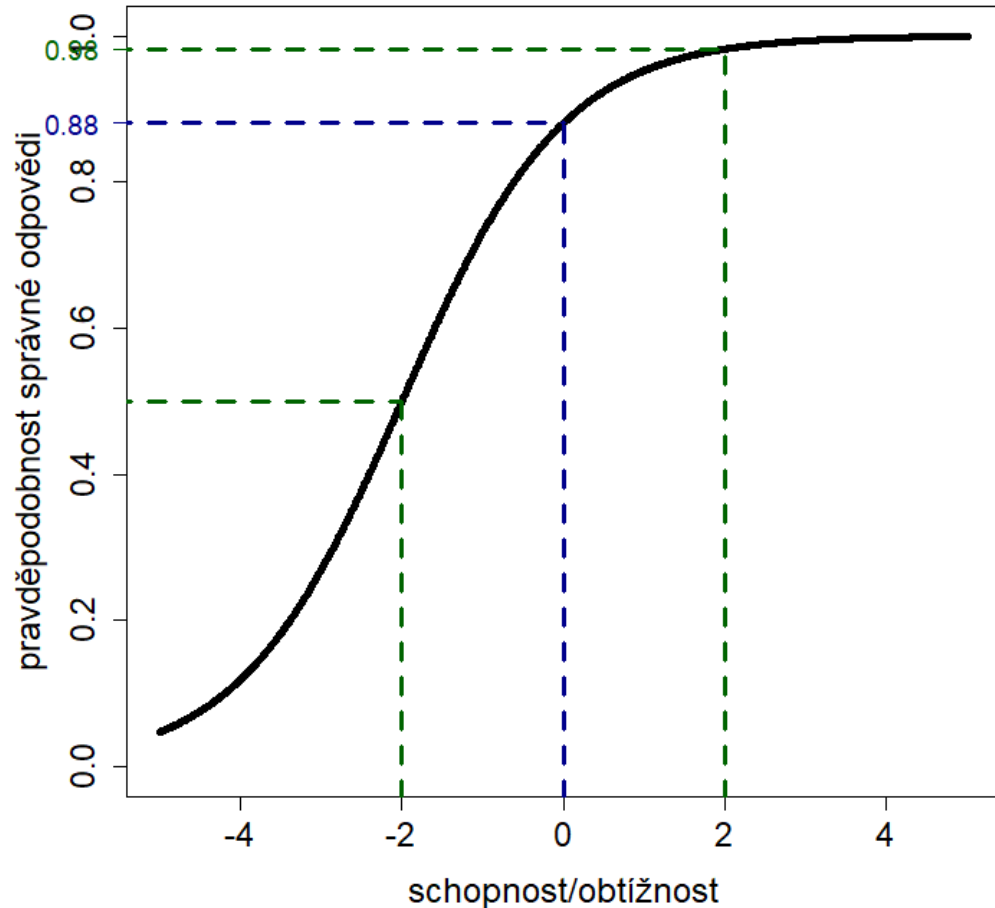
Položka s obtížností  $b_i = -2$ .

Respondent se schopností  $\theta = b_i = -2$  má 50 % pravděpodobnost správné odpovědi.

- Analogicky respondent s  $\theta = 0$  odpoví správně s 88% pravděpodobností:

- $P_i(\theta) = \frac{e^{(0+2)}}{1+e^{(0+2)}} = 0,88$ .

# Raschův model (jednparametrový)



Položka s obtížností  $b_i = -2$ .

Respondent se schopností  $\theta = b_i = -2$  má 50 % pravděpodobnost správné odpovědi.

- Analogicky respondent s  $\theta = 0$  odpoví správně s 88% pravděpodobností:

- $P_i(\theta) = \frac{e^{(0+2)}}{1+e^{(0+2)}} = 0,88.$

- A respondent s  $\theta = 2 \rightarrow 95 \%$ .

- $P_i(\theta) = \frac{e^{(0+4)}}{1+e^{(0+4)}} = 0,98.$

# Dvouparametrový model (2PL)

---

**Diskriminační parametr** je rozlišovací schopnost položky: ukazuje, jak moc se liší „dobří“ a „špatní“ respondenti v očekávané pravděpodobnosti správné odpovědi.

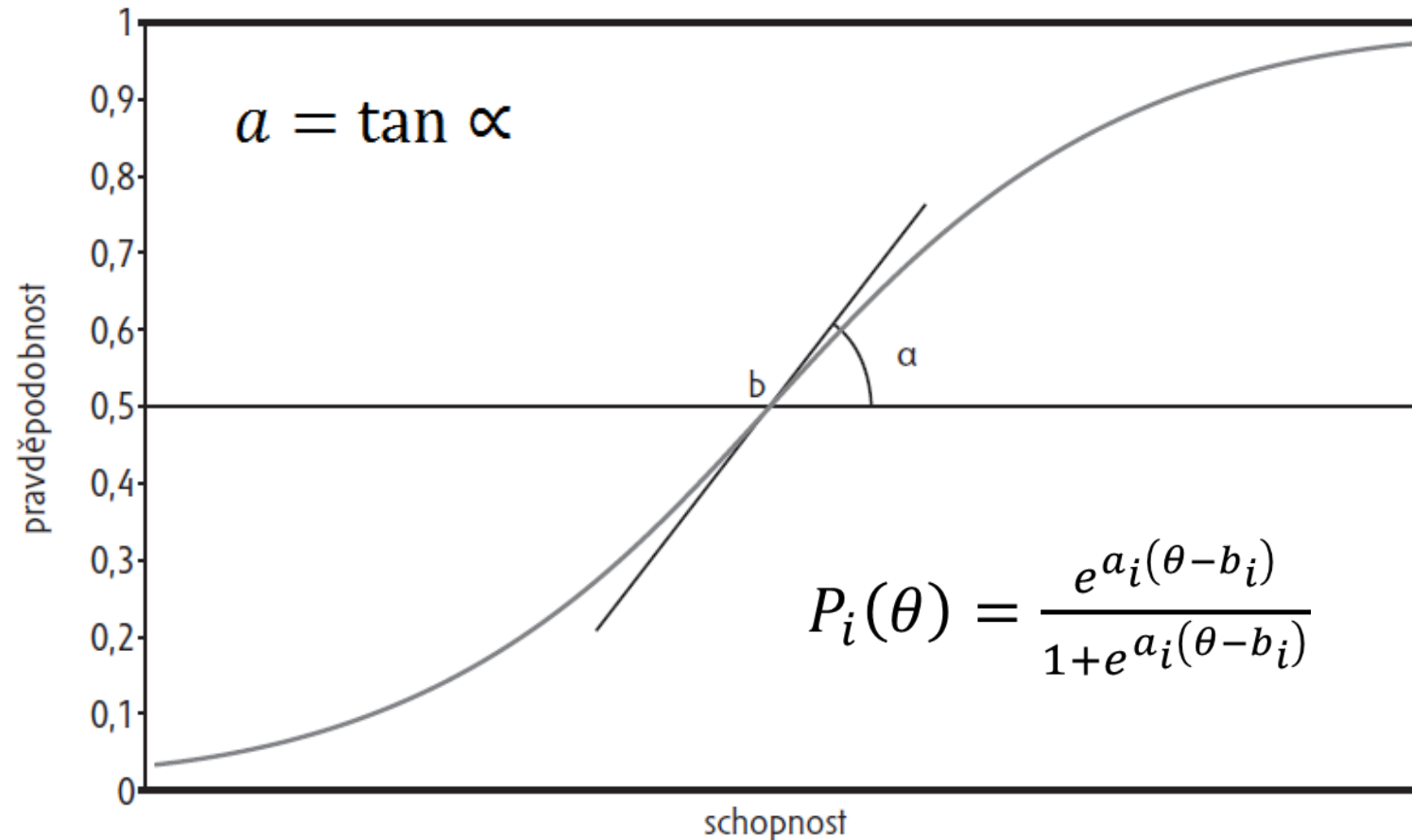
$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

$a_i$  je diskriminační parametr pol.  $i$   
– naklonění ICC v bodě  $b$ .

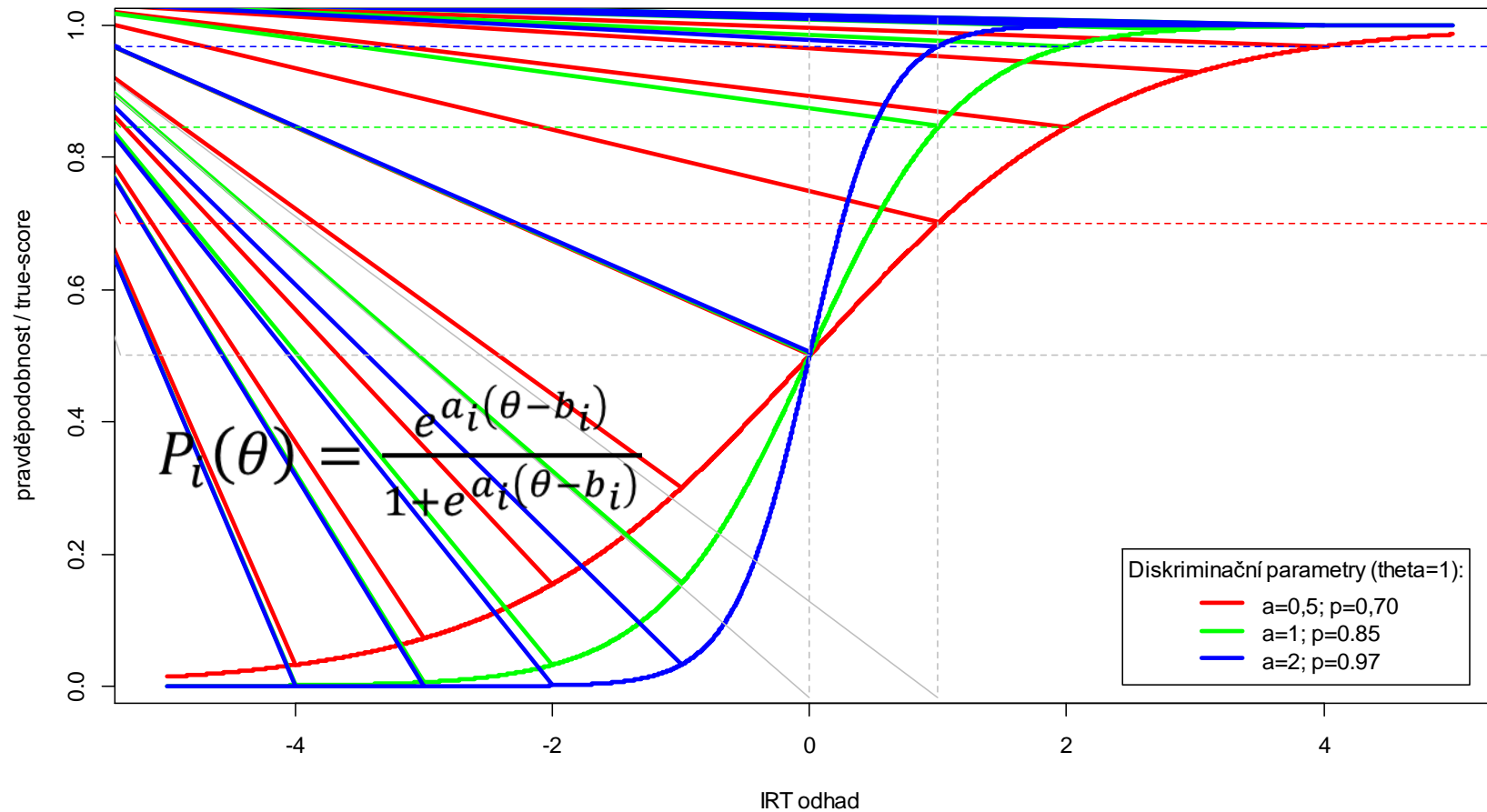
- čím je křivka „plošší“, tím méně rozlišuje

Analogií  $a_i$  je ve faktorové analýze faktorový náboj.

# Charakteristická křivka položky 2PL



# Charakteristická křivka položky 2PL



# Tříparametrový model (3PL)

---

Zavádí parametr pseudouhádnutelnosti  $c_i$  pro položky vícenásobné volby (multiple-choice):

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

- $c_i$  je parametr (pseudo)uhádnutelnosti pro položku  $i$ .

V multiple-choice testech lze nahradit Bockovým NRM modelem.

- Modeluje přímo jednotlivé odpověďové možnosti (distraktory).

Při prostém tipování je pravděpodobnost „náhodně správné“ odpovědi teoreticky  $1/n$ , kde  $n$  je počet možných odpovědí.

- Tedy  $n-1$  distraktorů a právě 1 správné odpovědi.

Tento předpoklad je příliš silný, proto je lepší pro každou položku tuto pravděpodobnost odhadnout zvlášť.

- Některé distraktory mohou být evidentně chybné a respondent je vyloučí.
- Ideálně by se takové distraktory samozřejmě neměly vyskytovat... chytáky nefungují.

# Charakteristické křivky položek 3PL

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

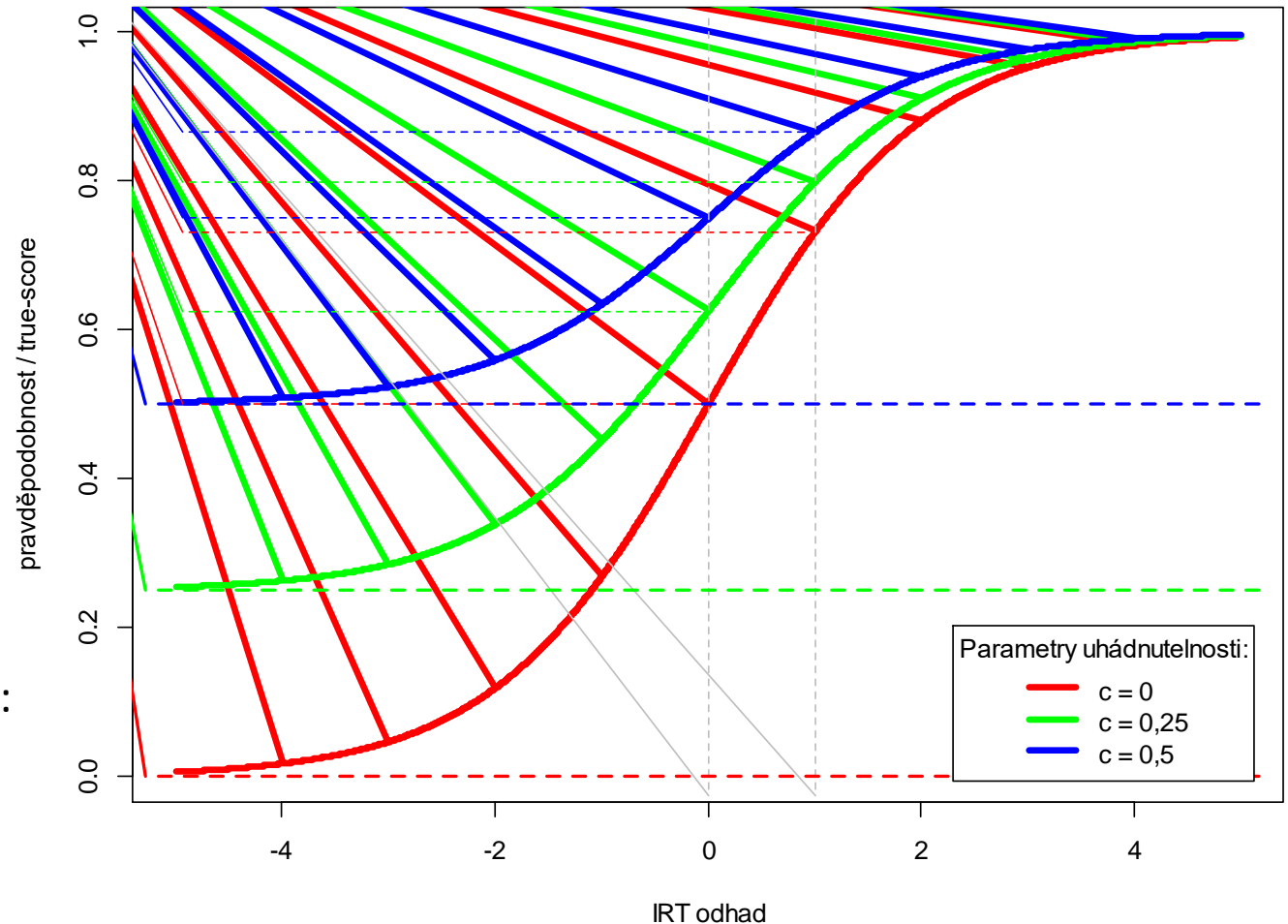
c	P(θ=0)	P(θ=1)
0	0,5	0,73
0,25	0,63	0,80
0,5	0,75	0,87

$b_i = 0$  pro všechny položky

Pozor – přestává platit poučka ze 2PL modelu:

$(\theta_p = b_i) \Rightarrow (P_{ij} = 0,5)$ !

V bodě  $b_i$  je ale ICC nejstrmější.





# Čtyřparametrový model (4PL)

---

Použití spíše výjimečně pro specifické účely.

Zpravidla malé výhody, zahrnutím dalších parametrů se naopak významně zhoršují vlastnosti modelu.

- Někdy je ale výhodné pracovat s horní namísto spodní asymptotou.

4PL: **parametr „ledabylosti“** – ani nejlepší respondent nemá pravděpodobnost správné odpovědi rovnu 100 %.

$$P_i(\theta) = c_i + (d_i - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

- $d_i$  je parametr ledabylosti; zpravidla bývá blízký 1.

# Charakteristická křivka 4PL modelu

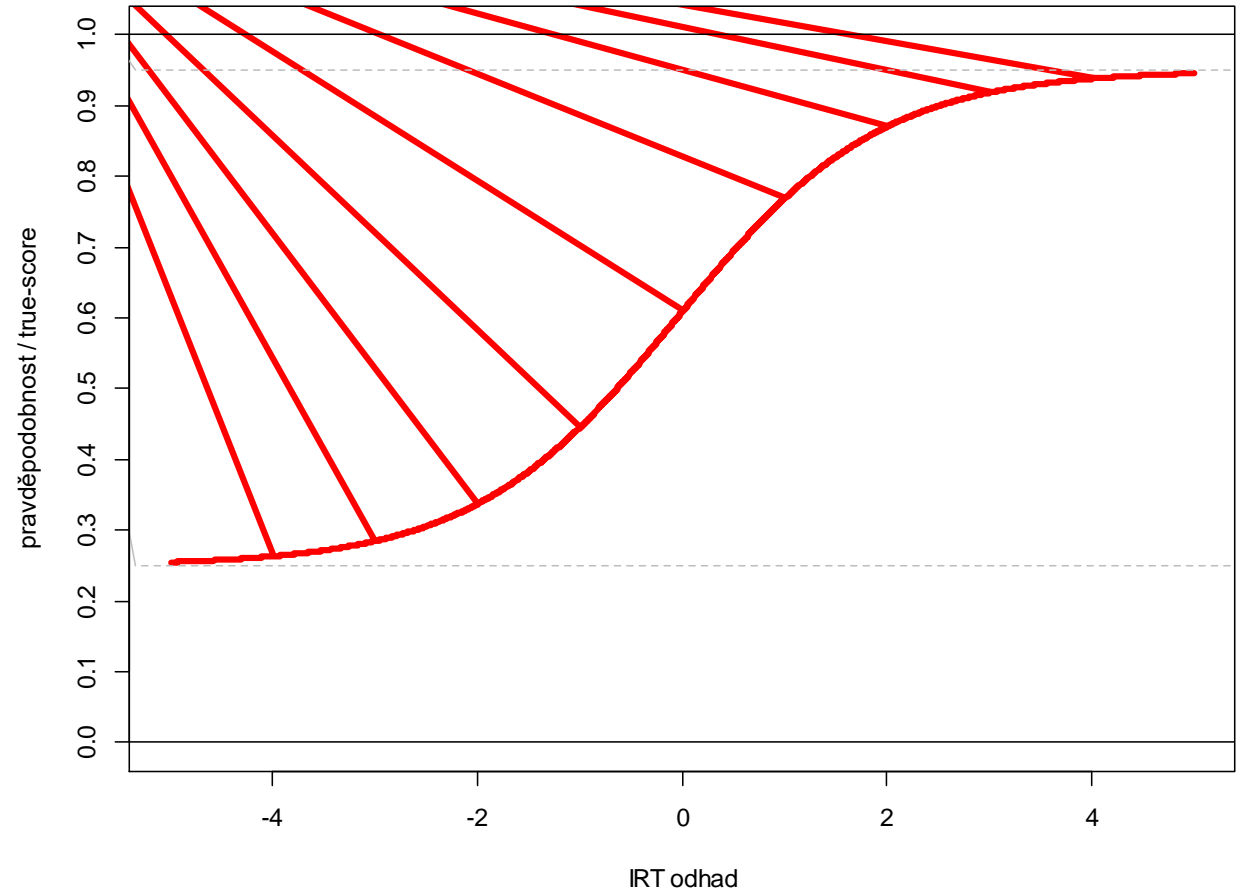
## ► Parametry:

- $a = 1$
- $b = 0$
- $c = 0,25$
- $d = 0,95$

## ► Pravěpodobnost:

- $P_i(\theta=0)=0,61$
- $P_i(\theta=1)=0,77$

$$P_i(\theta) = c_i + (d_i - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$



# Srovnání 1PL–3PL modelů

---

## jednparametrový model

- pouze parametr obtížnosti položky  $b_i$

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}$$

## dvouparametrový model

- přidává diskriminační parametr  $a_i$

$$P_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}$$

## tříparametrový model

- přidává parametr pseudo-uhádnutelnosti  $c_i$

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}$$

- Ostatní symboly:

- schopnost respondenta:  $\theta$
- pravděpodobnost správné odp.:  $P_i$
- $i$  – číslo položky

- 4PL:  $d_i = 1$  → 3PL

- 3PL:  $c_i = 0$  → 2PL

- 2PL:  $a_i = 1$  (nebo  $a_i = a$ ) → 1PL

# On-line ilustrace

---

<http://fssvm6.fss.muni.cz/ICC/>

<https://shiny.cs.cas.cz/ShinyItemAnalysis/>

# Raschův model

---

1PL model bývá označován jako Raschův.

To ale není tak docela přesné.

Raschovy modely jsou specifická kategorie v rámci IRT modelů.

- Odlišná epistemologická východiska.
- Zpravidla odlišný účel.
- Zpravidla odlišná identifikace modelu.
  - IRT modely – zpravidla fixován rozptyl faktoru ( $SD = 1$ ).
  - Raschovy modely – zpravidla fixován diskriminační parametr ( $a = 1$ ).

# Srovnání Raschova a 1PL–3PL přístupu

---

## RASCHŮV MODEL (1PL)

Spíše konfirmační princip  
(data musí odpovídat modelu).

Pouze 1. parametr,  $a=1$ , zbytek je „šum“.

- Všechny pol. diskriminují (teoreticky) stejně.

Cílem je fundamentalita škály, invariance odhadu.

Menší závislost odhadů na položkách/respondentech.

Nižší počet parametrů → nižší počet respondentů.

Vhodnější pro konstrukci diagnostických testů (SB-V, Leiter-3, v ČR pak WJ-IV, KIT a další)

Možnost žádných předpokladů o rozložení latentního rysu (JML estimátor).

## IRT (1PL, 2PL, 3PL...)

Spíše explorační princip  
(přizpůsobuje model datům).

Počet parametrů, který nejlépe popíše data.

- Diskriminace položek se může lišit.

Důraz je kladen na výběr „nejlepšího“ modelu.

Vyšší závislost odhadů na položkách/respondentech.

Vyšší počet parametrů → vyšší počet respondentů.

Vhodnější pro test-equating v high-stakes testech (SAT, GRE, SCIO, SK maturita) a adaptivní testování.

Zpravidla předpoklad normálního rozdělení (MML, CML aj. estimátory).

# Různé formáty parametrizace a zápisu

---

Rozdílné zápisy modelované pravděpodobnosti:

$$\begin{aligned} P(x_{ip} = 1 | \theta_p) &= P_i(\theta) = P_{ip} \\ &= P(x_{ip} = 1 | \theta_p, b_i, a_i, c_i) \end{aligned}$$

Rozdílné možnosti zápisu (zde 1PL) modelu:

$$\begin{aligned} P_{ip} &= \frac{e^{(\theta_p - b_i)}}{1 + e^{(\theta_p - b_i)}} = \frac{1}{1 + e^{-(\theta_p - b_i)}} \\ &= \frac{\exp(\theta_p - b_i)}{1 - \exp(\theta_p - b_i)} = \frac{1}{1 + \exp(b_i - \theta_p)} \end{aligned}$$

Exponenciální vs. logistický zápis:

$$P_{ip} = \frac{e^{(\theta_p - b_i)}}{1 + e^{(\theta_p - b_i)}} \sim \ln \frac{P_{ip}}{1 - P_{ip}} = \theta_p - b_i$$

Tradiční IRT parametrizace (2PL modelu):

$$P_{ip} = \frac{e^{a_i(\theta_p - b_i)}}{1 + e^{a_i(\theta_p - b_i)}}$$

Intercept-slope parametrizace:

$$P_{ip} = \frac{e^{a_i\theta_p + b_i}}{1 + e^{a_i\theta_p + b_i}}$$

# Výhody intercept-slope parametrizace

---

Výhoda 1: multidimenzionální (Reckaseho, kompenzatorní) model

$$P_{ip} = \frac{e^{a_{i1}\theta_{p1} + a_{i2}\theta_{p2} + \dots + a_{in}\theta_{pn} + b_i}}{1 + e^{a_{i1}\theta_{p1} + a_{i2}\theta_{p2} + \dots + a_{in}\theta_{pn} + b_i}}$$

Výhoda 2: srovnání s faktorovou analýzou

Faktorová analýza:  $E(x_{ip}) = a_{i1}\theta_{p1} + a_{i2}\theta_{p2} + \dots + a_{in}\theta_{pn} + b_i$

- S reziduálním rozptylem  $\sigma_i^2$  shodným pro všechny odpovědi na danou položku.
- faktorový náboj  $a_i$  se zpravidla značí jako  $\lambda_i$

IRT:  $\ln \frac{P_{ip}}{1 - P_{ip}} = a_{i1}\theta_{p1} + a_{i2}\theta_{p2} + \dots + a_{in}\theta_{pn} + b_i$

$$E(x_{ip}) = P_{ip}$$

- S reziduálním rozptylem  $P_{ip}(1 - P_{ip})$  (rozptyl binární proměnné) různým napříč respondenty.



# Přesnost měření v IRT

## IRT skóry

## Další modely IRT

---

2. ČÁST PŘEDNÁŠKY

# Předpoklady IRT

---

Latentní rys existuje a jde o spojitou intervalovou proměnnou.

- Zpravidla navíc i normálně rozloženou.
- Ale... diskrétní IRT modely, LCA, estimátory s nenormálně rozloženým latentním rysem.

Lokální nezávislost položek.

- Veškeré vzájemné vztahy položek lze vysvětlit působením modelovaných latentních rysů.
  - Tzn. parciální vztah položek po kontrole úrovně latentního rysu je nulový.
- V případě jediného rysu: jednodimenzionalita.

Odpovědi lidí na položku lze modelovat prostřednictvím ICF.

- Charakteristická funkce položky (ICF = Item Characteristic Function)
- Někdy též Item Response Function (IRF), Item Characteristic Curve (ICC) atd.
- Ale... Mokkenovo škálování.

# Obsah druhé části přednášky

---

Charakteristická funkce testu.

Odhad míry latentního rysu, IRT škálování, IRT skóry.

Práce s chybou: Informační funkce položky, testu, chyba měření.

Shoda modelu s daty.

IRT modely pro polytomní data.

Ordinální faktorová analýza (item-factor analysis).

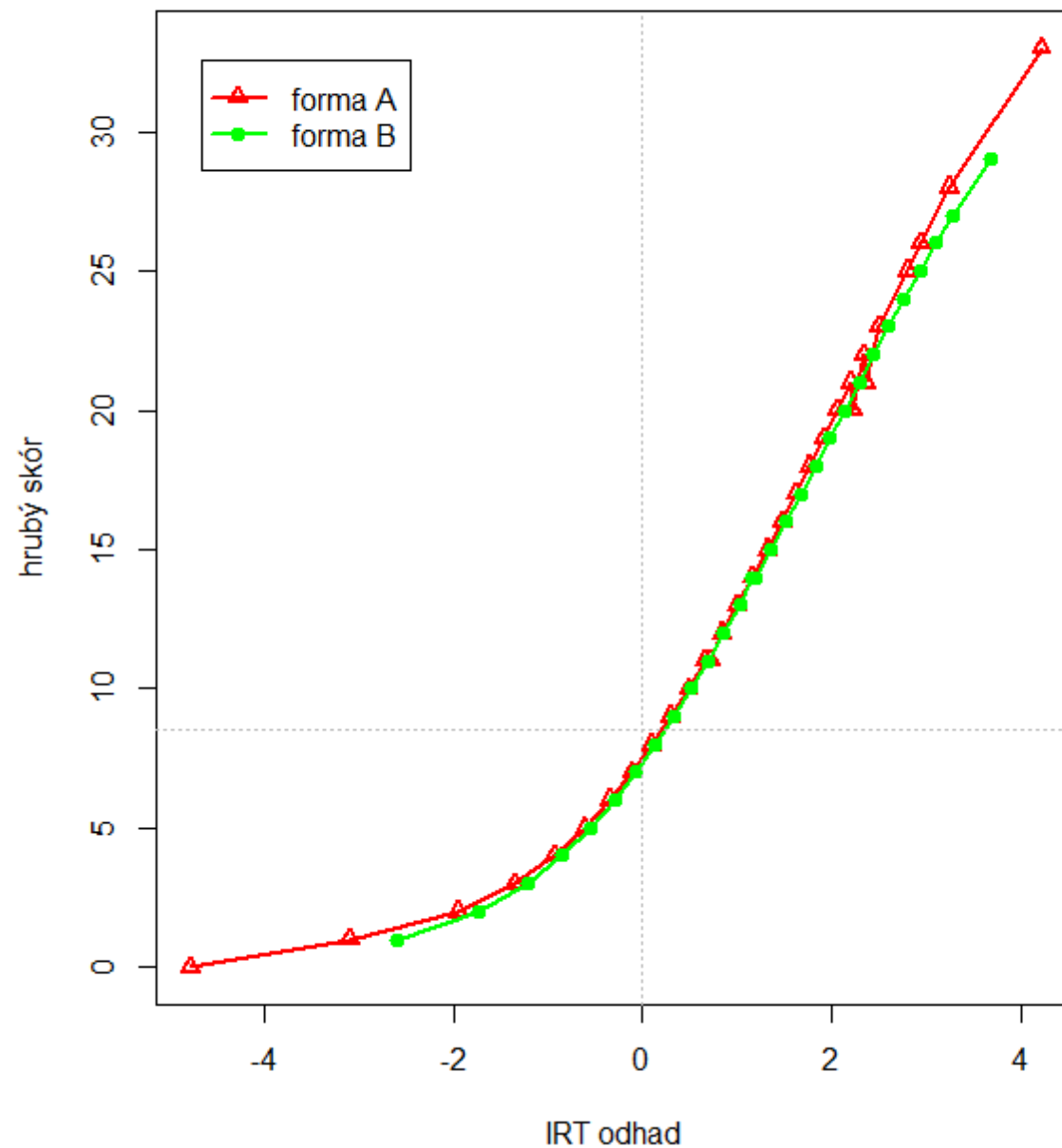
Klíčové oblasti využití IRT.

- Počítačově adaptivní testování.
- Vyvažování paralelních forem testu.

# Charakteristická funkce testu

Cígler, H., Jabůrek, M., Straka, O., & Portešová, Š. (2017). *Psychometrická analýza TIM<sup>3-5</sup> – Testu pro identifikaci nadaných žáků v matematice pro 3.–5. třídu*. Brno: Masarykova univerzita. Retrieved from <https://munispace.muni.cz/index.php/munispace/catalog/book/968>

Srovnání hrubého skóru a IRT odhadu



# Charakteristická funkce testu (TCF)

---

Test Characteristic Function/Curve (TCF/TCC).

Jde o prostý součet jednotlivých ICC:

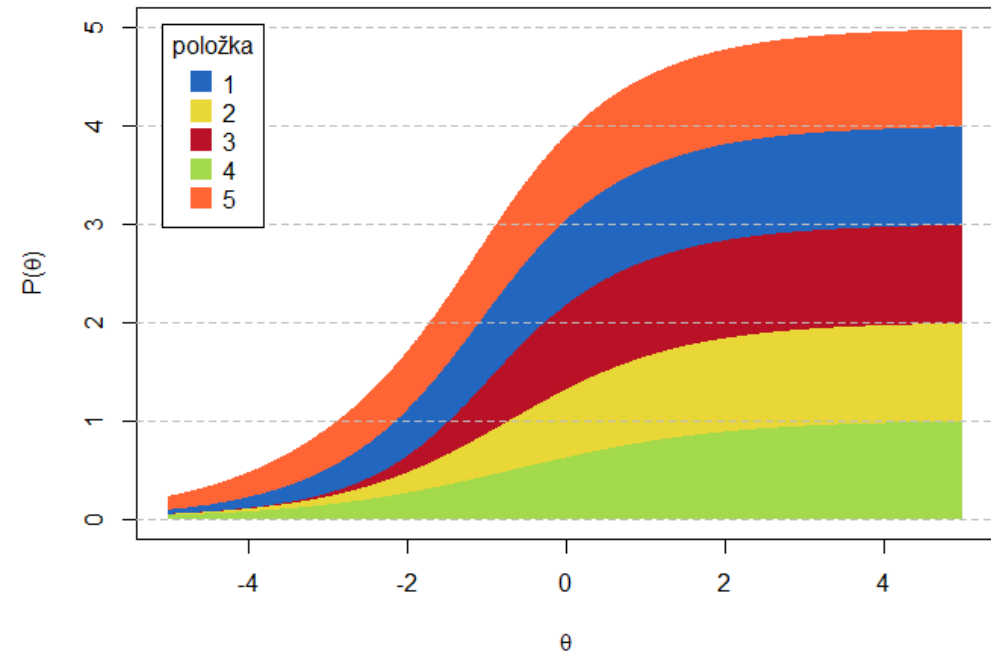
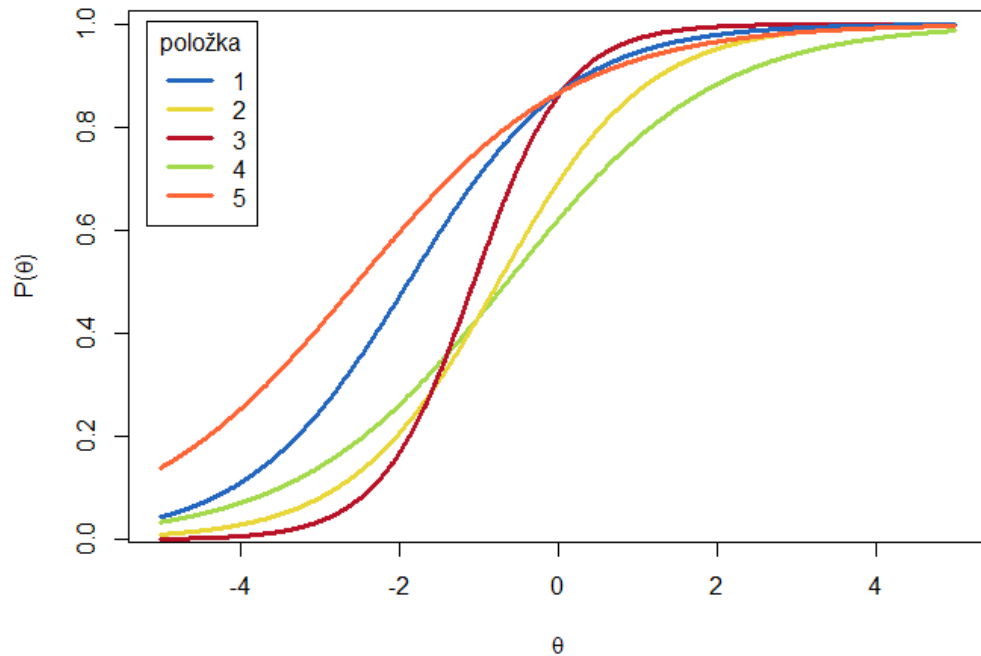
$$TCC(\theta) = \sum_{i=1}^n ICC_i(\theta) = \sum_{i=1}^n P_i(\theta) = E(T|\theta)$$

- kde  $n$  je počet položek.

Hodnota očekávaného pravého ( $E(T|\theta)$ ), případně pozorovaného ( $E(X|\theta)$ )<sup>1</sup> skóre u respondentů s určitou mírou latentního rysu  $\theta$ .

<sup>1</sup> Záleží na nastavení modelu, estimátoru atd., není to podstatné.

# Charakteristická funkce testu (TCF)



# Charakteristická funkce testu (TCF)

---

TCF lze využít při skórování testu.

**1PL:** TCC izomorfní, každému  $X$  odpovídá právě jedno  $\theta$ . Toho se využívá v psychologických testech.  $TCC(\theta) \leftrightarrow X$

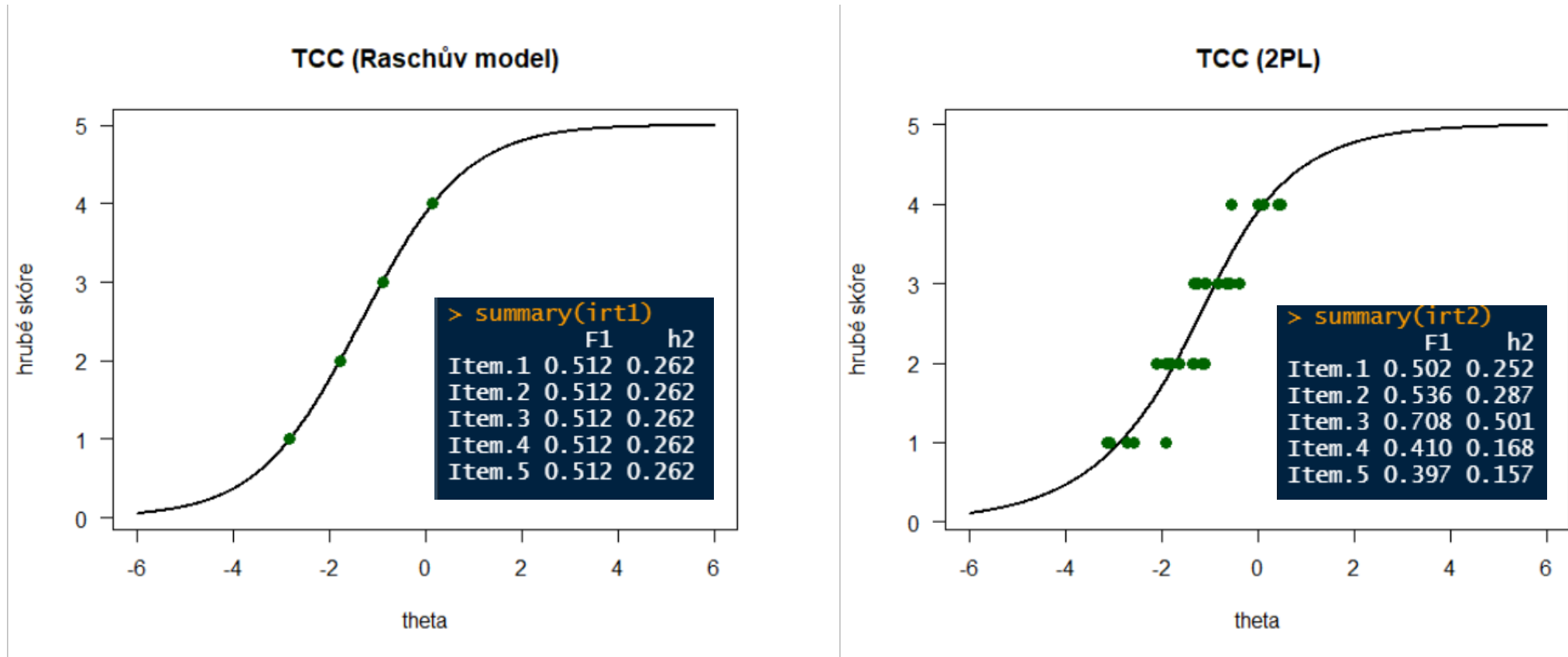
**2PL:** vztah není jednoznačný; diskriminační parametr dává rozdílné váhy položkám. Záleží, které byly zodpovězeny správně.  $TCC(\theta) \rightarrow X$

- Každému HS odpovídá konečný počet odhadů latentních rysů podle konkrétních odpovědí.
- Z hrubého skóre lze na úroveň latentního rysu usuzovat jen se ztrátou reliability.
- Zpravidla se pro skórování používají přímo odpovědi na jednotlivé položky.

Řada dalších využití, např.:

- Observed score IRT equating.
- Differential test functioning (DTF).

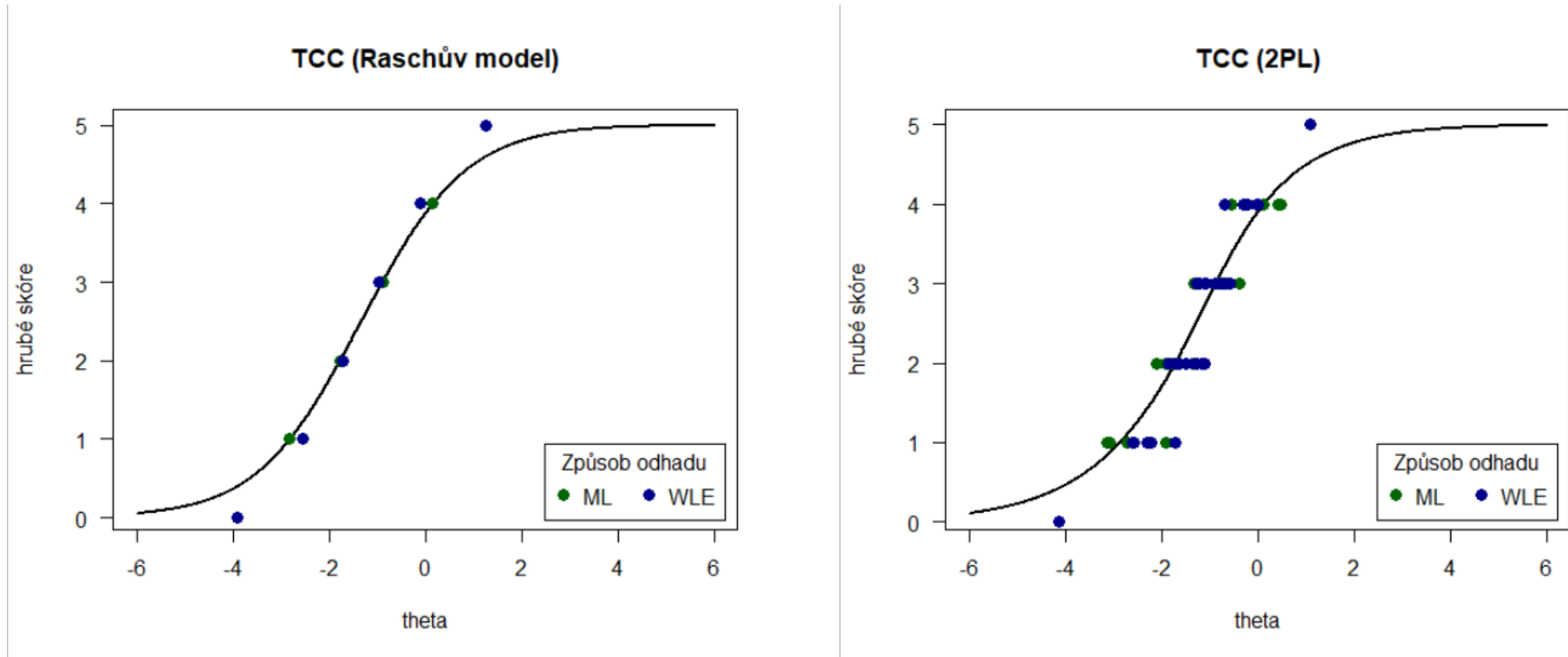
# Srovnání TCC Raschova a 2PL modelu



LSAT7 data v mirt balíčku (5 binárních položek)



# Srovnání TCC Raschova a 2PL modelu

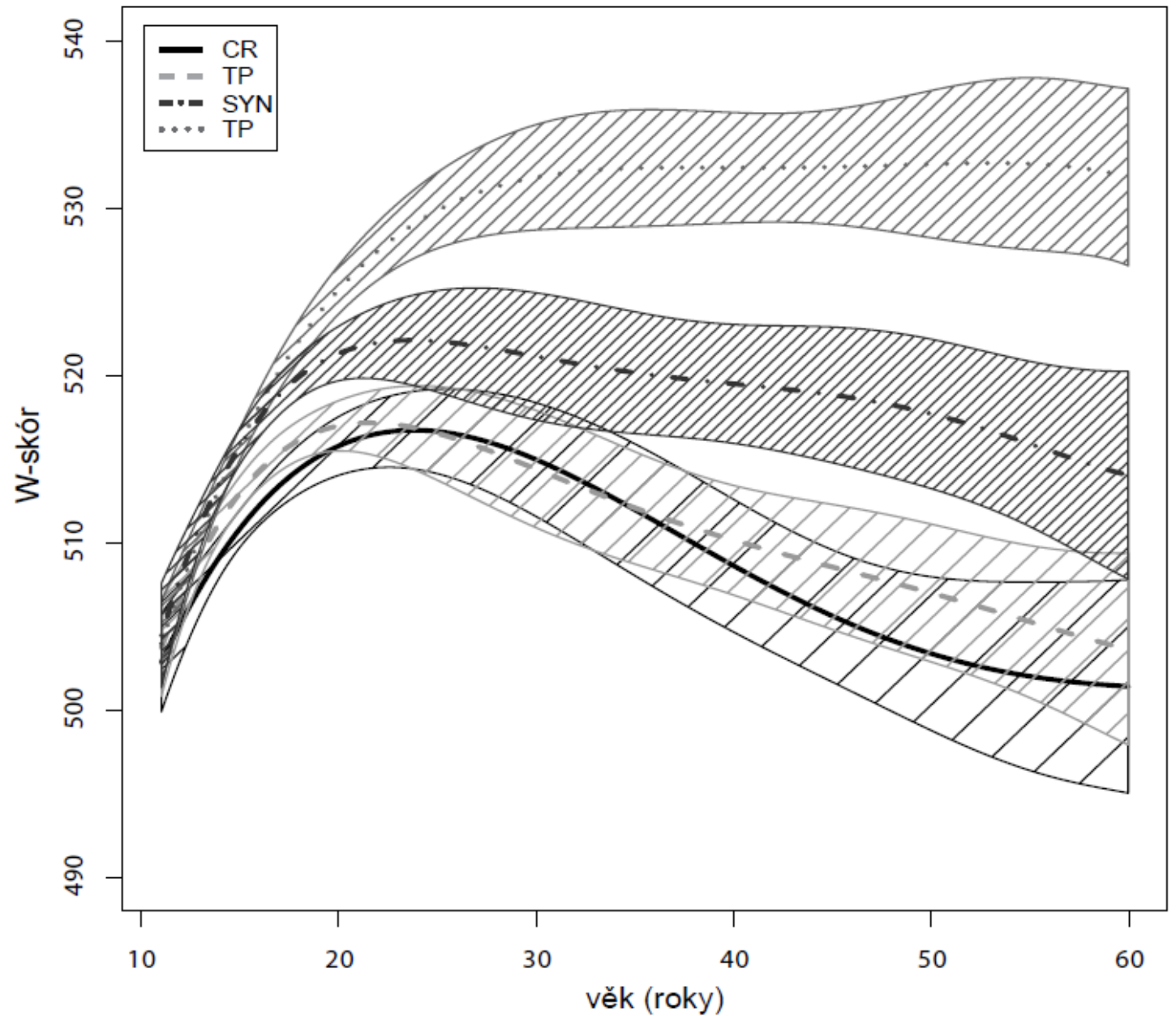


LSAT7 data v mirt balíčku (5 binárních položek)

# IRT škálování

IRT skóry

IRT škály



# Kde je (sakra) to celkové skóre?

---

Problém zpětné inference (epistemologie).

- **Model:** Latentní rys způsobuje odpovědi na položky.
- **Praxe:** Z odpovědí na položky usuzujeme na míru rysu.
- Známe-li parametry (obtížnost...) položek, můžeme odhadnout nejpravděpodobnější úroveň latentního rysu, pro kterou bychom právě takové odpovědi pozorovali.

Při výzkumu (např. standardizace metody):

- Odhadujeme parametry položek i osob naráz.
- Parametry položek uschováme pro budoucí použití, parametry osob se použijí pro tvorbu norem (IQ, T-skóry, percentily...)

Při praktickém použití již standardizované metody:

- Z dopředu „nakalibrovaných“ položek usuzujeme na míru rysu, kterou pak převedeme na standardní skóry.

# Logitový skór

Výstupem IRT (Raschova modelu, 2PL+ to může být komplikovanější) je skór v logitech.

- Analogie hrubého skóre v CTT.

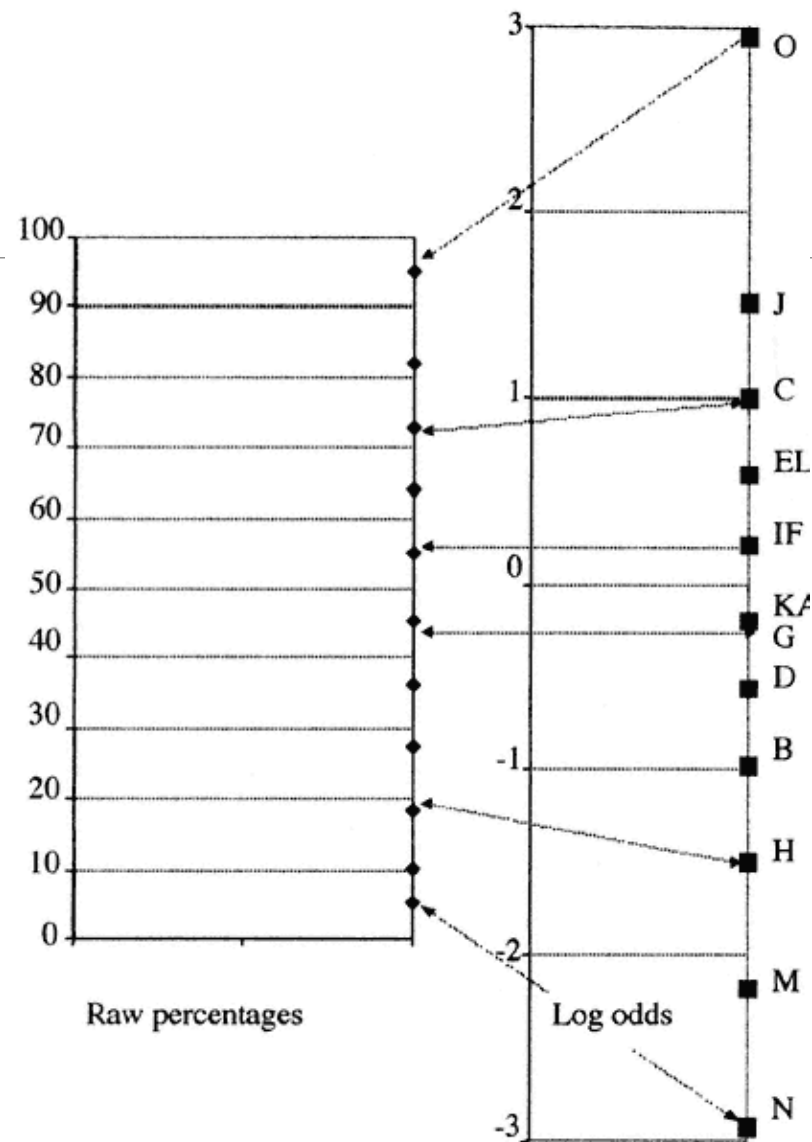
Interpretace:

$$\text{logit} = \ln \frac{P(\theta)}{1 - P(\theta)}$$

Kde  $P(\theta)$  je typicky podíl položek, které respondent zvládne splnit správně.

- Platí jen přibližně!

Logity převádějí pravděpodobnost (resp. percentil) na intervalovou proměnnou.



$\theta - b_i$	P
-5	0,7%
-4,5	1,1%
-4	1,8%
-3,5	2,9%
-3	4,7%
-2,5	7,6%
-2	11,9%
-1,5	18,2%
-1	26,9%
-0,5	37,8%
0	50,0%
0,5	62,2%
1	73,1%
1,5	81,8%
2	88,1%
2,5	92,4%
3	95,3%
3,5	97,1%
4	98,2%
4,5	98,9%
5	99,3%

# IRT škálování

---

Samotný skór v logitech se pro praktické použití dále standardizuje.

- Intervalová škála rysu napříč všemi skupinami respondentů.
- Z ní IQ, T-skóry apod. pro daný ročník/věk/pohlaví atd.

Kromě toho specifické (typicky Raschovské) skóry:

- **W-skóry:** Vhodné pro sledování růstu či vývoje, nezávisí na vzorku.
  - W 500 ve věku 10;0 (příp. na začátku 5. ročníku)
  - Vzdálenost  $b - \theta = 10W$  odpovídá změně pravděpodobnosti správné odpovědi z 50 % na 75 % (resp. 25 %).
  - Lze predikovat úspěch v položkách/subtestech.
- **RPI (Relative Proficiency Index):**  $X/_{90}$ , závisí na vzorku.
  - **Index relativní výkonnosti.** Jaká je pravděpodobnost X správné odpovědi na položky, které lidé ze stejné normalizační skupiny odpovídají s 90% pravděpodobností správně? (Pro jiné základy zlomku [kalkulačka zde.](#))

W DIFF	RPI	W DIFF	RPI	W DIFF	RPI
29 and above	100 <sup>1</sup> /90	-1	89/90	-36	15/90
28	99/90	-2	88/90	-37	13/90
27	99/90	-3	87/90	-38	12/90
26	99/90	-4	85/90	-39	11/90
25	99/90	-5	84/90	-40	10/90
24	99/90	-6	82/90	-41	9/90
23	99/90	-7	81/90	-42	8/90
22	99/90	-8	79/90	-43	7/90
21	99/90	-9	77/90	-44	7/90
20	99/90	-10	75/90	-45	6/90
19	98/90	-11	73/90	-46	5/90
18	98/90	-12	71/90	-47	5/90
17	98/90	-13	68/90	-48	4/90
16	98/90	-14	66/90	-49	4/90
15	98/90	-15	63/90	-50	4/90
14	98/90	-16	61/90	-51	3/90
13	97/90	-17	58/90	-52	3/90
12	97/90	-18	55/90	-53	3/90
11	97/90	-19	53/90	-54	2/90
10	96/90	-20	50/90	-55	2/90
9	96/90	-21	47/90	-56	2/90
8	96/90	-22	45/90	-57	2/90
7	95/90	-23	42/90	-58	2/90
6	95/90	-24	39/90	-59	1/90
5	94/90	-25	37/90	-60	1/90
4	93/90	-26	34/90	-61	1/90

Ability Minus Difficulty ( $W_{A-D}$ )	Probability of Success ( $P$ )
+50	.996
+45	.993
+40	.988
+35	.979
+30	.964
+25	.940
+20	.900
+15	.839
+10	.750
+5	.634
0	.500

$$W = \frac{10}{\ln 3} (\theta - \bar{\theta}_{10}) + 500$$

$$W = 9,1(\theta - \bar{\theta}_{10}) + 500$$

- kde  $\bar{\theta}_{10}$  = průměrný skór 10letých
- W-skóre má 9,1krát užší měřítko než logit.

# IRT škálování

Klíčová výhoda IRT škálování:  
 Odhad latentního rysu není závislý na  
 použitých položkách.

- V CTT je naopak pravý skór „operacionalizován“ položkami.
- Chybějící data nejsou problém

Toho využívají IRT metody, např.:

- Subtesty dělené podle věku, ale stále srovnatelné pomocí W-skóru.
- Různé „startovací položky“.
- Pravidla ukončení.

## Subtest M11) Procedurální znalosti

### Pomůcky

- Psací potřeby pro testovanou osobu
- Pracovní list „Procedurální znalosti“
- Záznamový sešit „Matematika“

### Výchozí bod

- Do 5. třídy: začínáme blokem A (položka 1)
- Od 5. třídy: začínáme blokem B (položka 18)
- Od 8. třídy: začínáme blokem C (položka 31)

### Časový limit

Časový limit na položku v případě tohoto subtestu není stanoven. Pokud však testovaná osoba nad některým příkladem přemýšlí delší dobu (přibližně 30 sekund), aniž by příklad viditelně řešila (počítala), povzbudíme ji, např.: „*Pokud si nejsi jistý/á, zkus si tipnout.*“ Pokud ani po tom nezačne s počítáním, vyzveme ji, aby začala řešit další příklad.

### Bazální úroveň

Pro dosažení bazální úrovně musí testovaná osoba získat alespoň 4 body v rámci prvních 5 administrovaných položek daného vstupního bodu. Pokud testovaná osoba nedosáhne bazální úrovně, pokračujte v administraci položek, dokud nebude dosaženo pravidla ukončení. Teprve poté administrujeme celý blok položek pro předchozí vstupní bod. V případě, že jste začínali blokem C a testovaná osoba nedosáhla bazální úrovně ani po návratu v rámci bloku B, administrujte všechny položky bloku B a následně zadejte zbývající blok A od položky 1.

### Pravidlo ukončení

Subtest ukončete po 7 chybně zodpovězených či nezodpovězených položkách jdoucích bezprostředně za sebou. Pokud má položka více částí (např. část a a b), pracujte pro tyto účely s každou z nich jako se samostatnou položkou. Více viz způsob administrace.

9	D	senioři
x	x	
x	x	
x	x	x
x	x	
x	x	x
x	x	x

CJ7/I	Diktát I				x	x	x	x	x	x	+	+		x <sup>a</sup>
CJ7/II	Diktát II										x	x	x	
CJ8/I	Opravy chyb I				x	x	x	x	+					x
CJ8/II	Opravy chyb II								+	x	x	x	x	

Bednářová, J., Cígler, H., & Jabůrek, M. (2019). *Standardizace BACH: Testy školních dovedností: Obecné pokyny*. Verze dokumentu 1.02. Masarykova univerzita a Propsyco.

Bednářová, J., Cígler, H., & Jabůrek, M. (2019). *Testy školních dovedností (BACH): Matematika*. Masarykova univerzita a Propsyco.

# IRT škálování

---

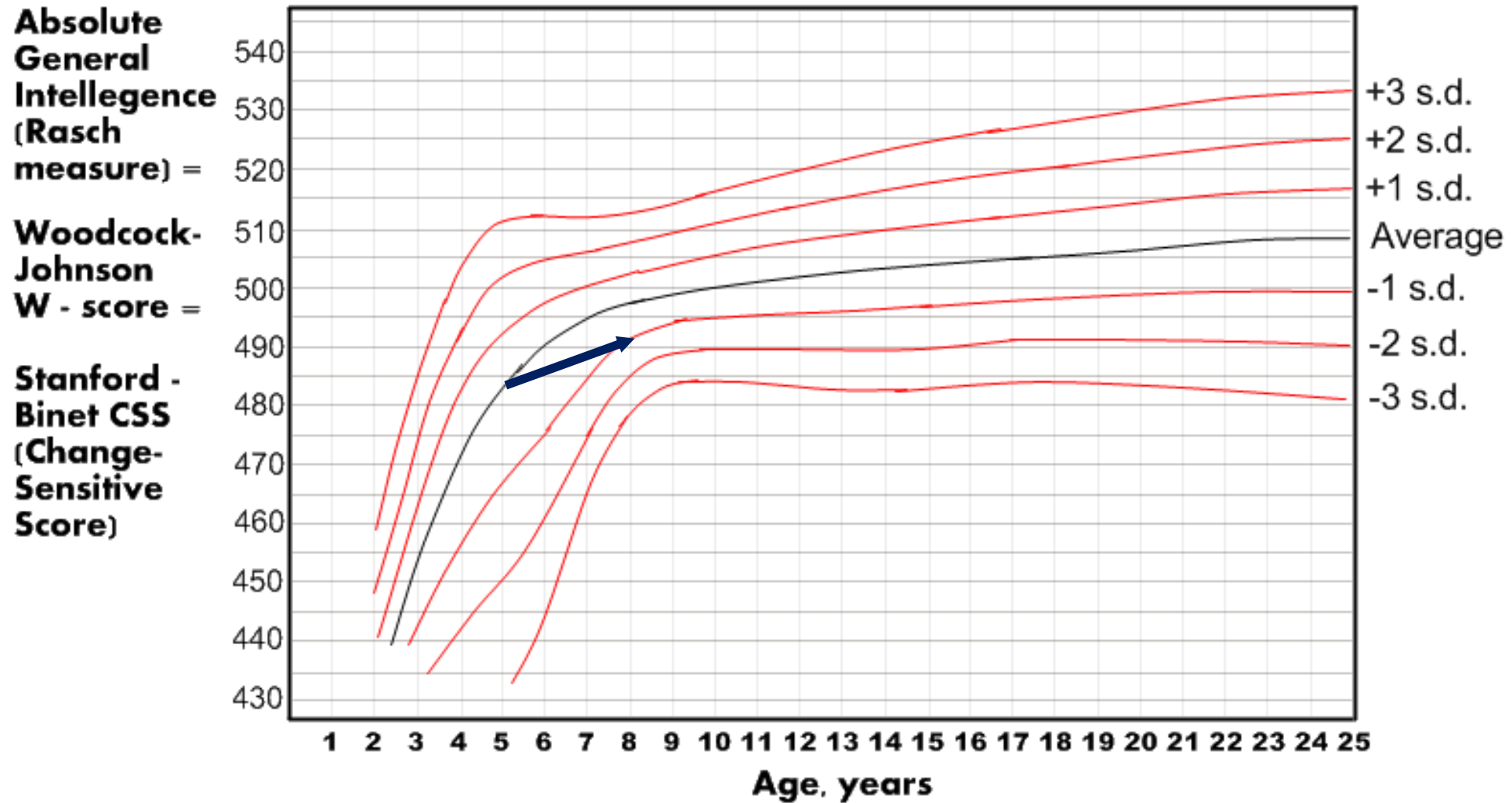
## Příklad z měření fluidní inteligence:

- Dítěti v 5 letech jsme naměřili IQ 100.
- Při retestu v 8 letech má IQ 85.

## Intelligence dítěte se: ... ?

- a) zvýšila
- b) nezměnila
- c) snížila
- d) nelze říci
- e) nechci odpovídat





Remake of Woodcock-Johnson block rotation subtest graph from "Applied Psych Test Design Part C - Use of Rasch scaling technology - Slide 19 "- full test should be similar but not identical.

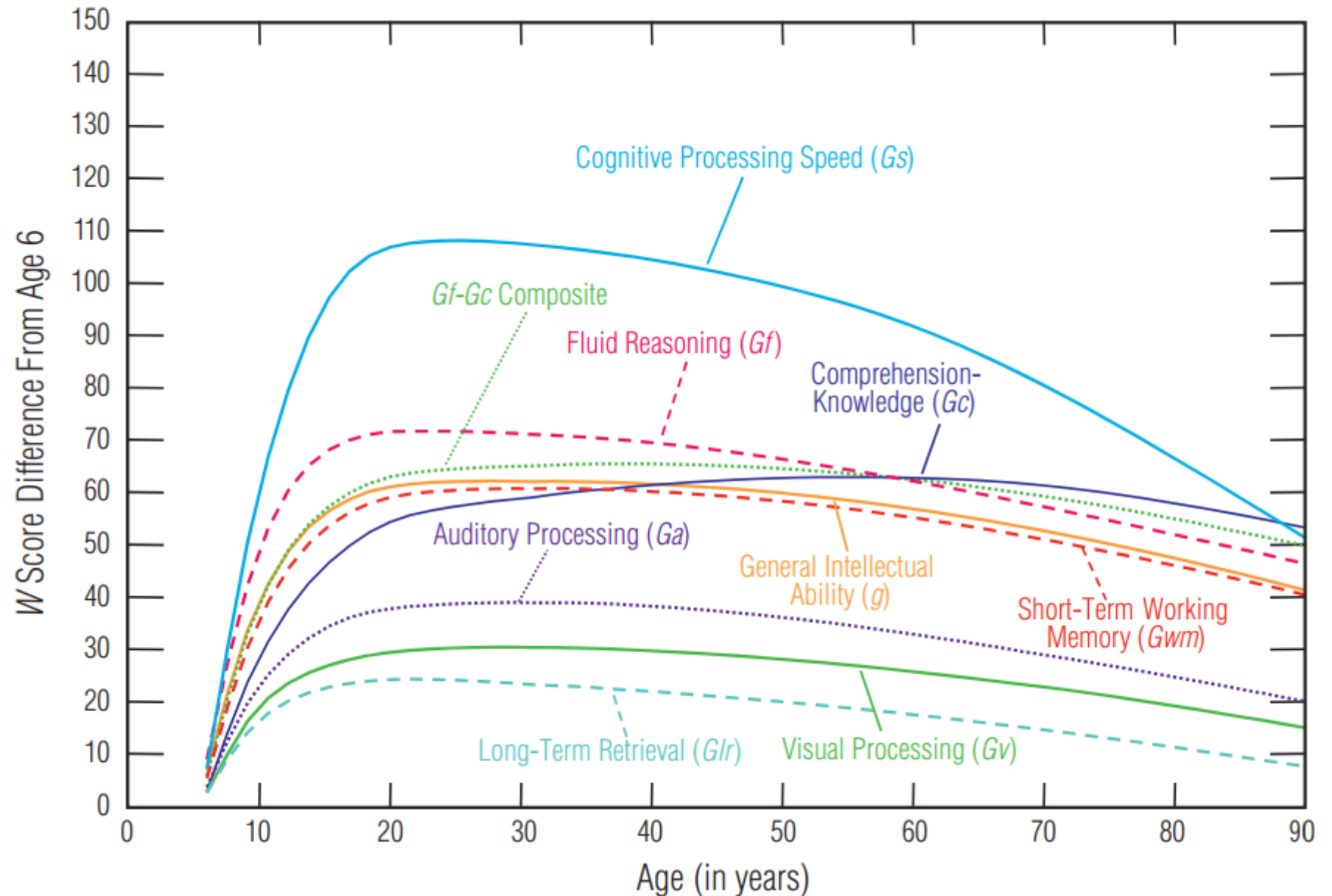
### Figure 5-3.

Plot of WJ IV COG GIA, seven CHC factor clusters, and the Gf-Gc Composite W score difference curves by age.

### Vývoj indexů ve WJ-IV v závislosti na věku.

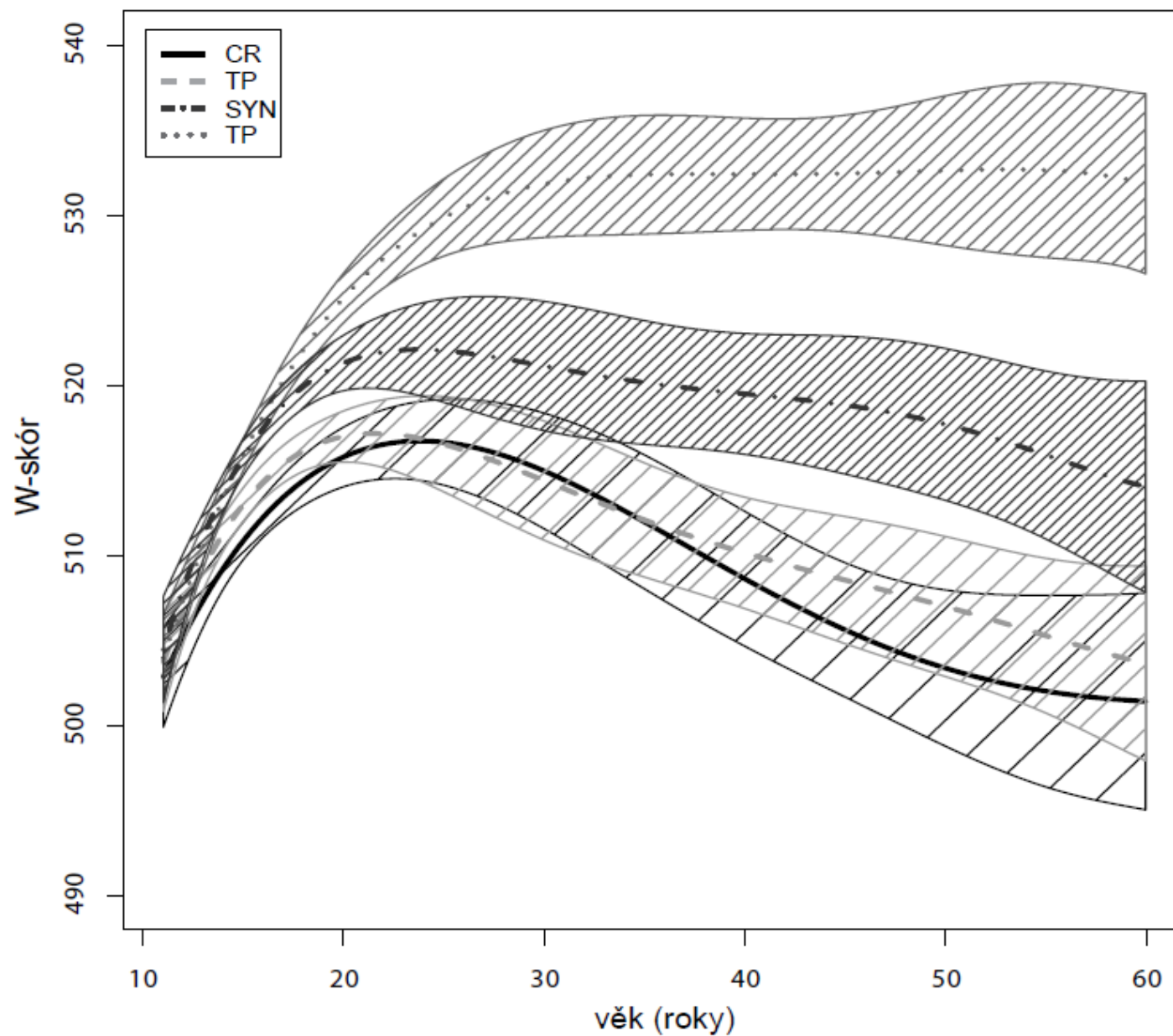
Raschův model umožňuje srovnávání vývoje průměrné úrovně rysů v čase.

Ve vícePL IRT modelech problematické (nestejná „škála“).



## Krátký inteligenční test (KIT)

Srovnání vývojových křivek  
použito jako důkaz  
konstruktové validity.

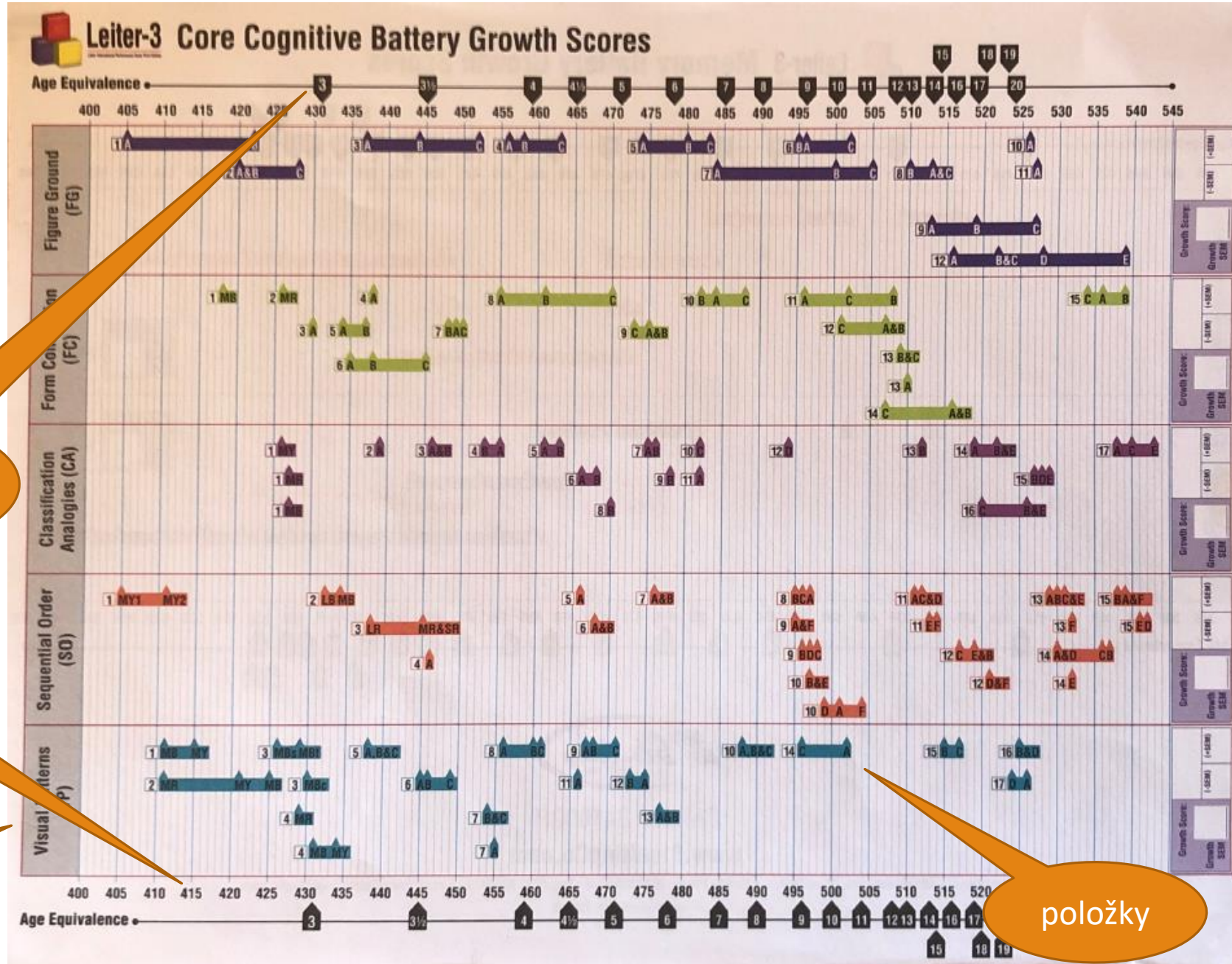


LEITER-3  
(Leiter International  
Performance Scale)

věkové  
ekvivalenty

W-škála

jednotlivé  
subtesty



položky

# Přehled různých typů skóru

---

**Hrubé skóry** (CTT součtové skóry, IRT odhady) – nelze samy o sobě interpretovat.

**Odvozené skóry** (percentily, IQ a další standardní skóry) poskytují normativní srovnání s referenční skupinou. Jsou závislé na vlastnostech škály a vzorku (M, SD).

**Ipsativní skóry** poskytují intraindividuální srovnání odvozených skóru (diagnostika profilu atp.).

- Statisticky, klinicky významný rozdíl...

**W-skóry** zasazují výkon člověk na škálu nezávislou na věku a populaci společnou pro typ testů.

- Do jisté míry nezávislou na počtu a konkrétním znění položek.

**RPI index** poskytuje měřítko pro srovnání rozdílu výkonu probanda a referenční skupiny na snadno představitelné škále. Závislý na průměru (M), ale nikoli na variabilitě (SD).

- Rozdíl 30 IQ v pěti a dvaceti letech znamená velmi odlišný rozdíl v reálném výkonu, protože  $SD_5 > SD_{20}$ .

**Věkové a ročníkové ekvivalenty** zasazují respondenta na vývojovou škálu. Zóna nejbližšího vývoje.

# Chyba měření v IRT

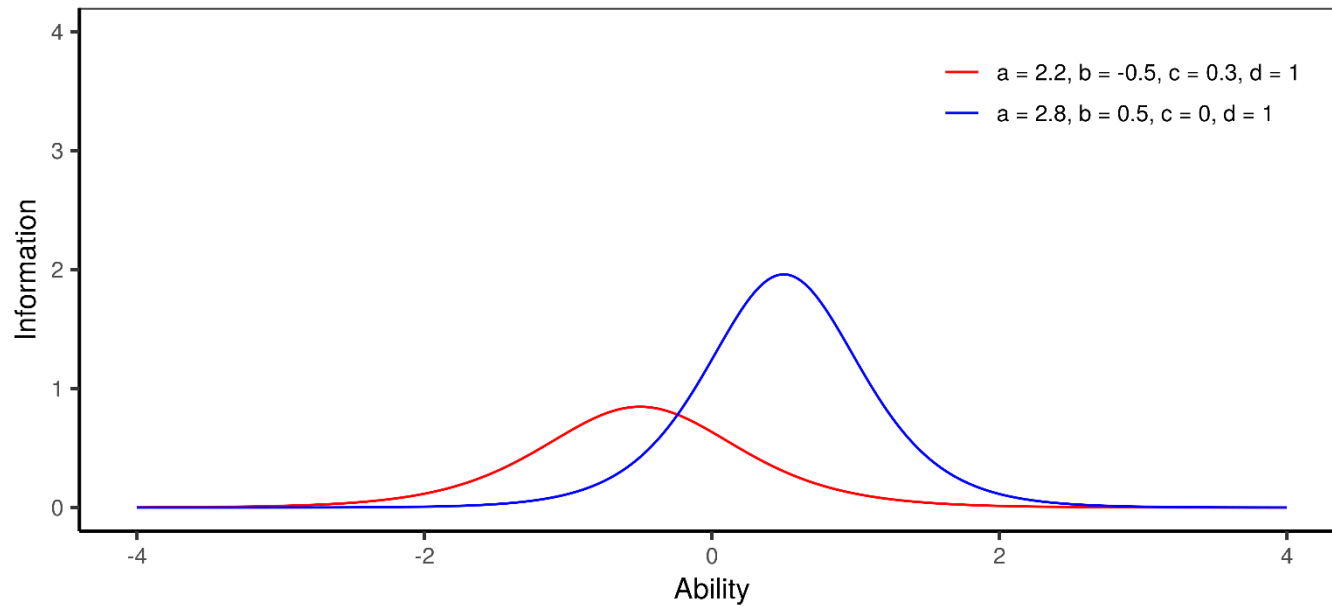
Informační funkce položky

Informační funkce testu

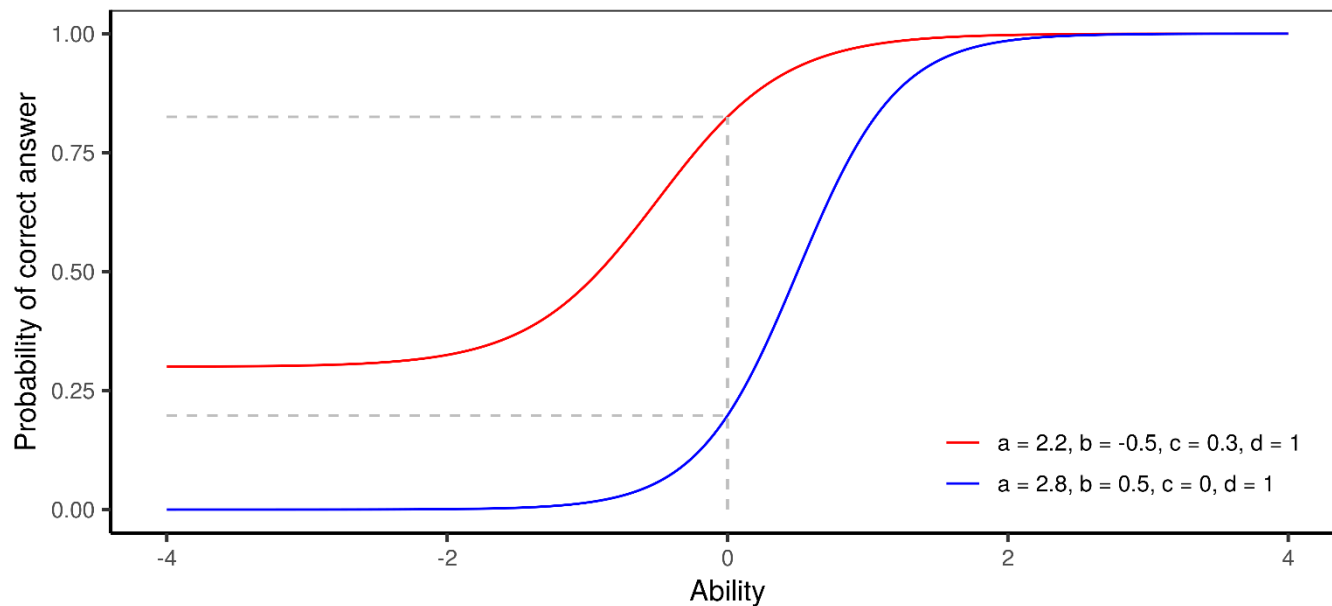
Chyba měření

Martinkova P., & Drabinova A. (2018). *ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests*. The R Journal, 10(2), 503-515.  
doi: [10.32614/RJ-2018-074](https://doi.org/10.32614/RJ-2018-074)

### Item information function



### Item characteristic curve



# Pojetí reliability a přesnosti měření v IRT

---

IRT odděluje úvahu o:

- Chybě měření (a intervalech spolehlivosti odhadu).
  - Tzv. **informační funkce položky/testu**.
  - Teoreticky nezávislá na výzkumném souboru.
- Reliabilitě, celkové spolehlivosti testu.
  - Odhadnuté na základě parametrů vzorku a chyb měření.

V IRT je tedy odhad SE používán pro odhad reliability.

- V CTT spíše naopak (ale srov. GT).

# Informační funkce položky (IIF)

---

Item Information Function/Curve (IIF/IIC)

Informační funkce položky  $I_i(\theta)$  je funkcí jednotlivých parametrů modelu.

- Pro každou úroveň schopnosti  $\theta$  jiná.

Binární položky:

$$I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta)(1 - P_i(\theta))}$$

- $P_i(\theta)$  = Charakteristická funkce položky
- $P_i'(\theta)$  = první derivace této funkce.
- $1 - P_i(\theta)$  = pravděpodobnost jiné než správné odpovědi.



# Informační funkce položky (IIF)

---

## 1PL MODEL (RASCHŮV)

Pro **1PL** model platí

$$P'_i(\theta) = P_i(\theta)[1 - P_i(\theta)]$$

- a lze tedy zjednodušit:

$$I_i(\theta) = P_i(\theta)[1 - P_i(\theta)]$$

- V Raschově binárním modelu mají všechny položky stejný průběh funkce (diskriminační parametr), liší se jen umístěním maxima.
  - Maximum je v bodě obtížnosti pol. ( $b_i$ ).
  - Maximum funkce je vždy  $0,5 \cdot 0,5 = 0,25$ .

## 2PL, 3PL MODELY

Pro **2PL** model platí

$$P'_i(\theta) = a_i^2 P_i(\theta)[1 - P_i(\theta)]$$

- a lze tedy zjednodušit:

$$I_i(\theta) = a_i^2 P_i(\theta)[1 - P_i(\theta)]$$

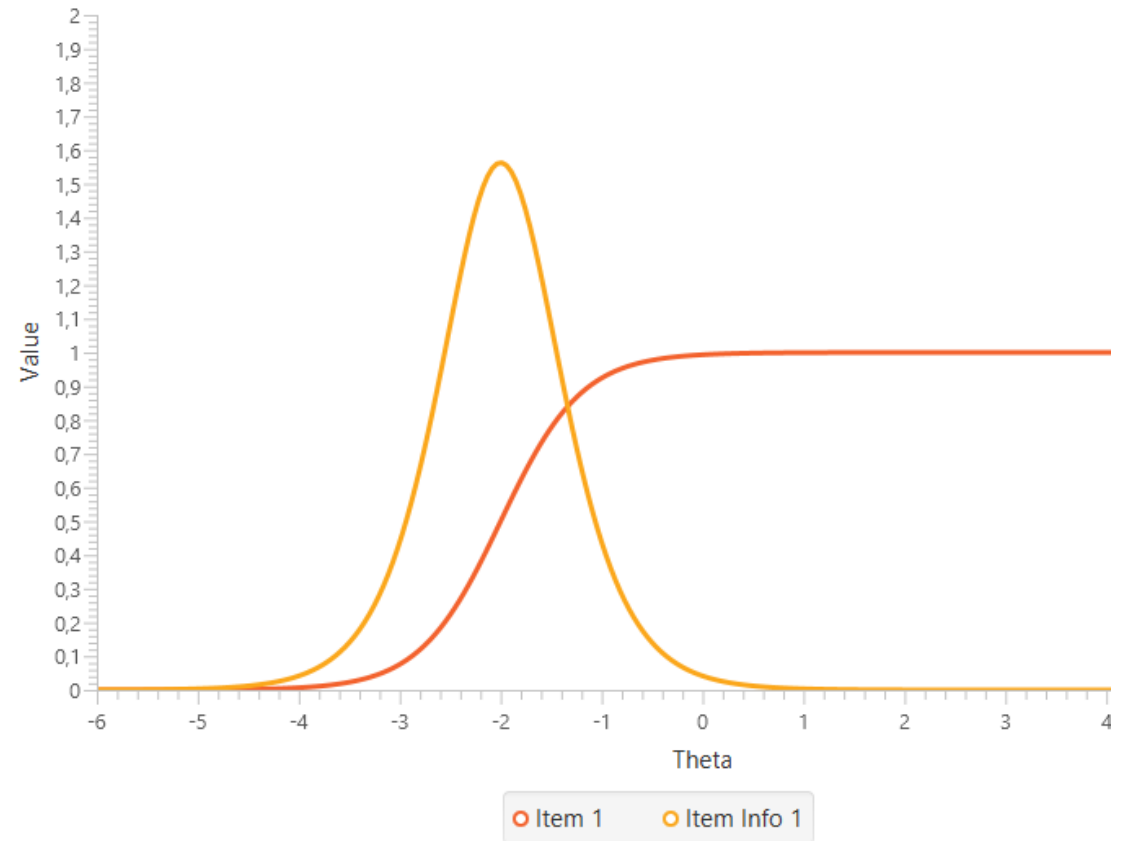
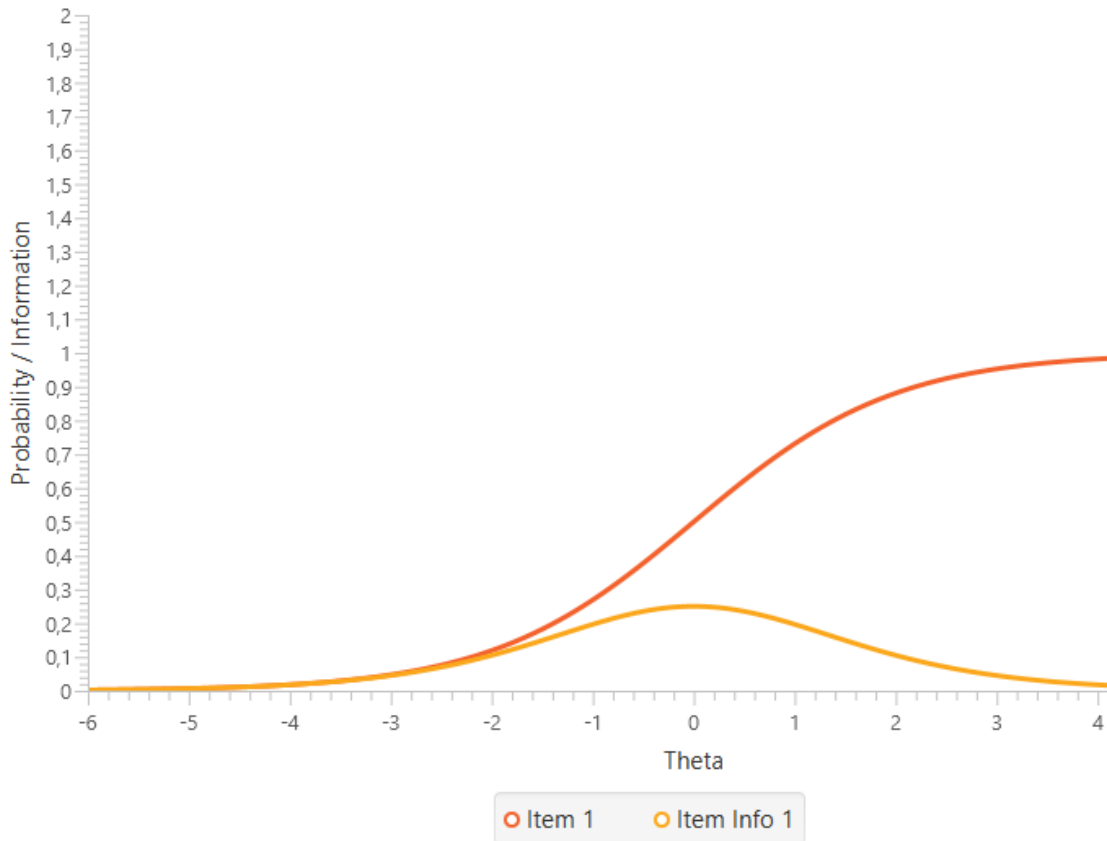
Informační funkce **3PL** modelu je:

$$I_i(\theta) = a_i^2 \frac{(P_i(\theta) - c_i)^2}{(1 - c_i)^2} \frac{1 - P_i(\theta)}{P_i(\theta)}$$

- fixováním  $c_i = 0$ , resp.  $a_i = 1$  lze dosáhnout 2PL, resp. 1PL IIF.
- U 3PL není maximum v bodě obtížnosti.

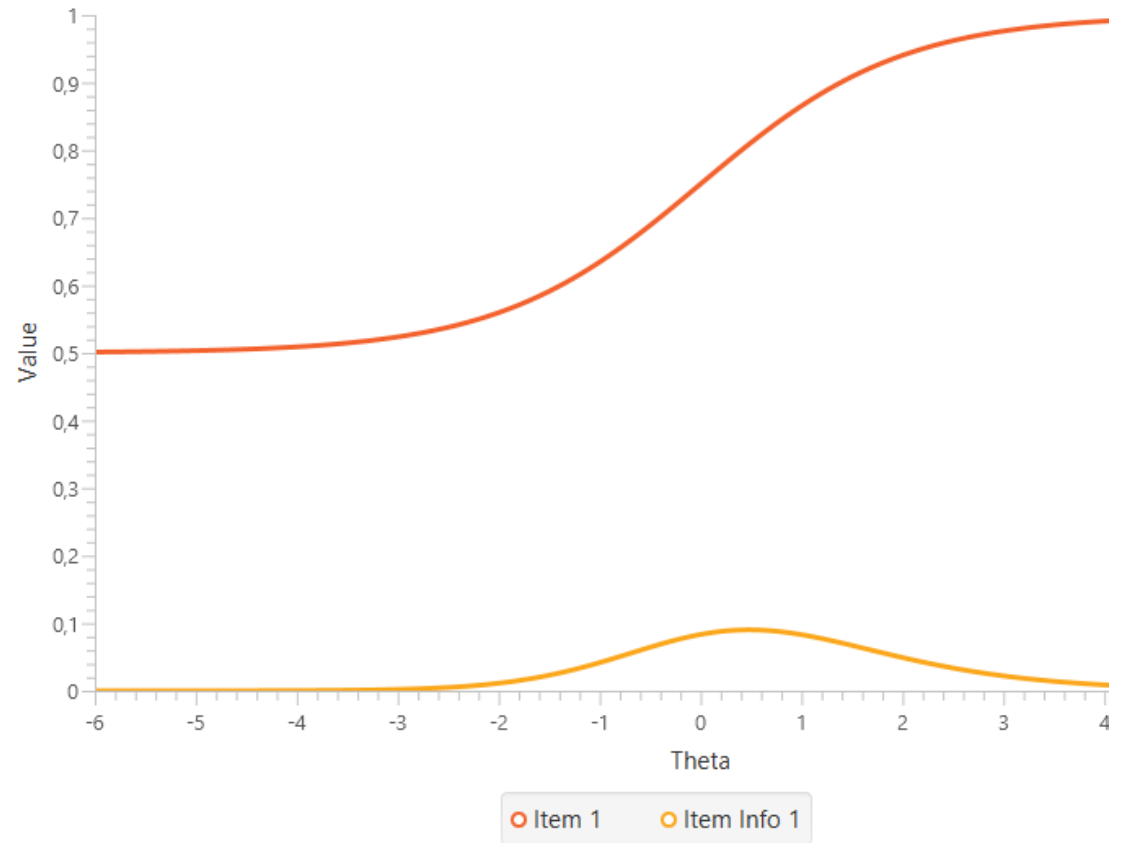
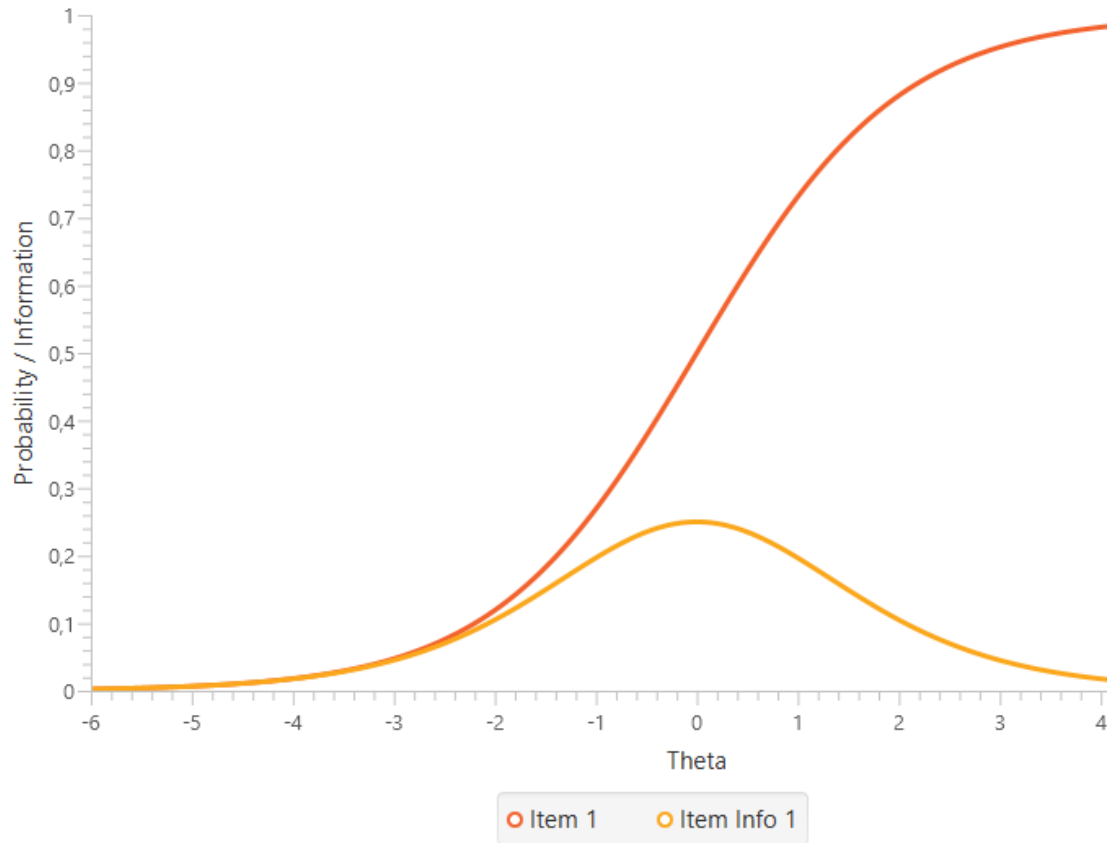
# Informační funkce položky

Vlevo:  $a=1$ ;  $b=0$ ;  $c=0$ ;  $d=1$  | Vpravo:  $a=2,5$ ;  $b=-2$ ;  $c=0$ ;  $d=1$



# Informační funkce položky

Vlevo:  $a=1$ ;  $b=0$ ;  $c=0$ ;  $d=1$  | Vpravo:  $a=1$ ;  $b=0$ ;  $c=0,5$ ;  $d=1$



# Informační funkce položky

---

Celková informační funkce položky (plocha pod křivkou) závisí na:

- Diskriminačním parametru (+).
- Parametru pseudouhádnutelnosti (-).

Velikost informace položky se liší pro jednotlivé respondenty podle jejich schopnosti  $\theta$  a závisí dále na:

- Blízkosti parametru obtížnosti a latentního rysu respondenta.
- Položka přináší nejvíce informace, když je ICC nejstrmější, a tedy pravděpodobnost správné odpovědi  $\theta = b_i$  (1PL, 2PL).
- Toho se využívá při počítačově adaptivním testování (CAT).

# Informační funkce testu (TIF) a chyba měření

---

Informační funkce testu  $I(\theta)$  je součtem informačních funkcí jednotlivých položek:

$$I(\theta) = \sum_{i=1}^n I_i(\theta)$$

- (Analogie k CTF.)

Lze ji chápat jako relativní nepřítomnost chybového rozptylu, a proto se **chyba měření**  $SE$  liší podle odhadu úrovně lat. rysu  $\hat{\theta}$ :

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

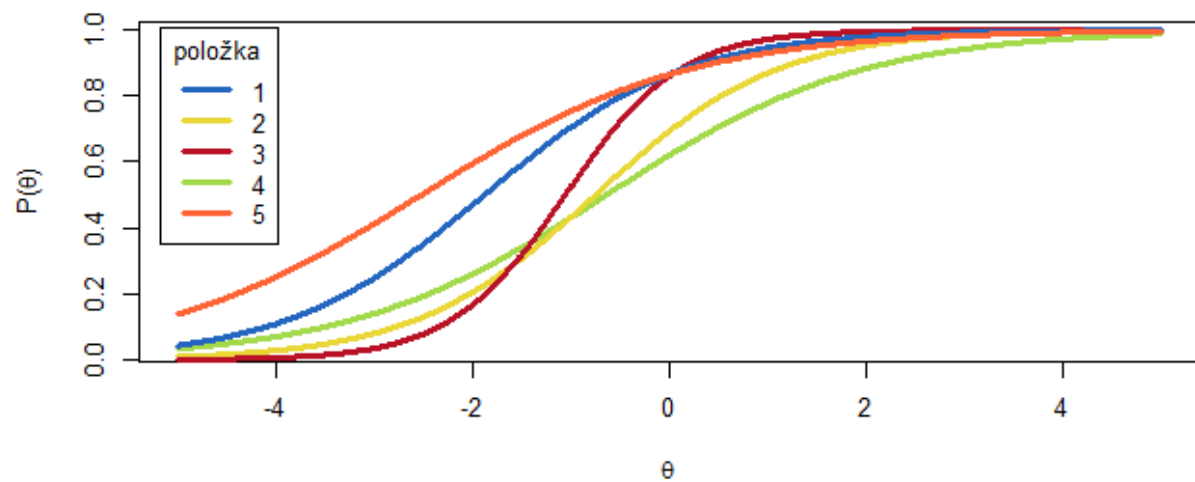
- (tedy čím vyšší informační funkce, tím přesnější měření/menší chyba měření)

Interval spolehlivosti potom získáme jednoduše jako:

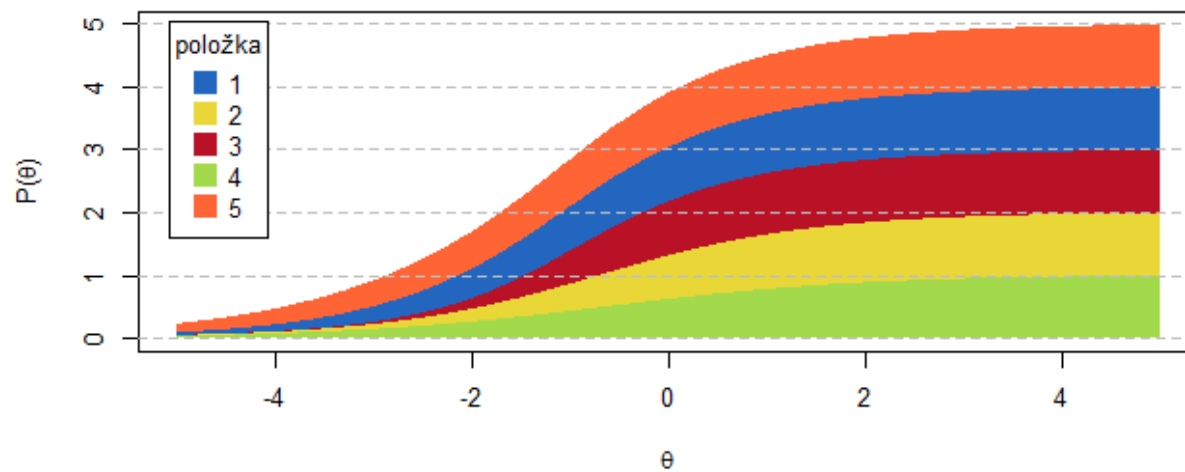
$$CI_{95\%}(\hat{\theta}) = \theta \pm z_{97,5\%} \cdot SE_{\hat{\theta}}$$

- (Reálně se ale používají různé přesnější bootstrapové techniky).

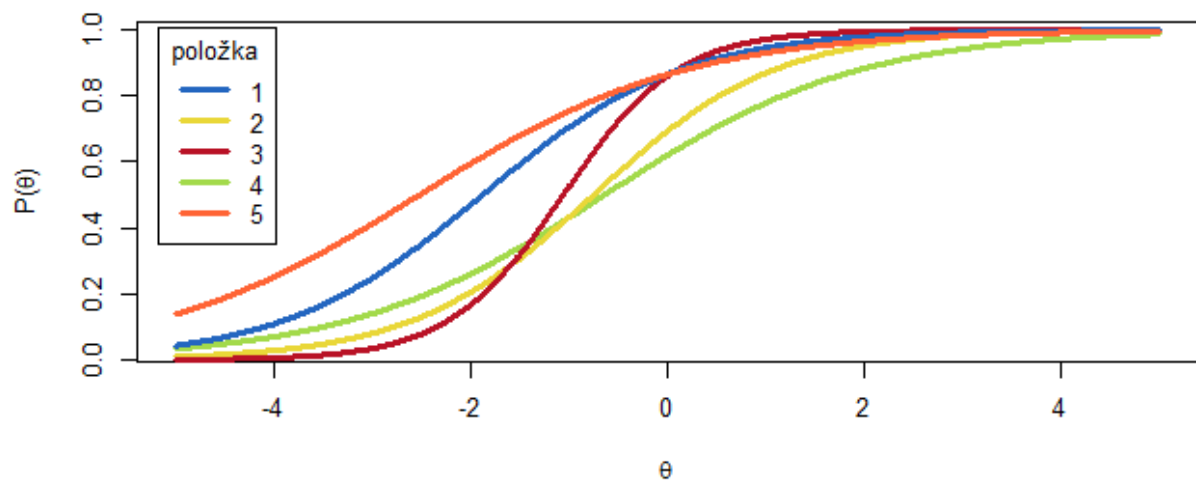
## Charakteristická funkce položek



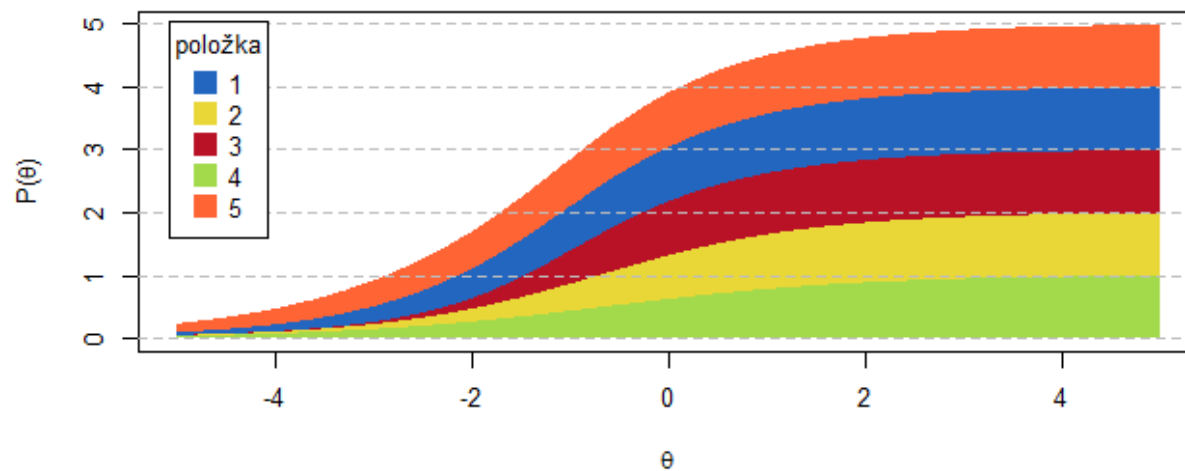
## Charakteristická funkce testu



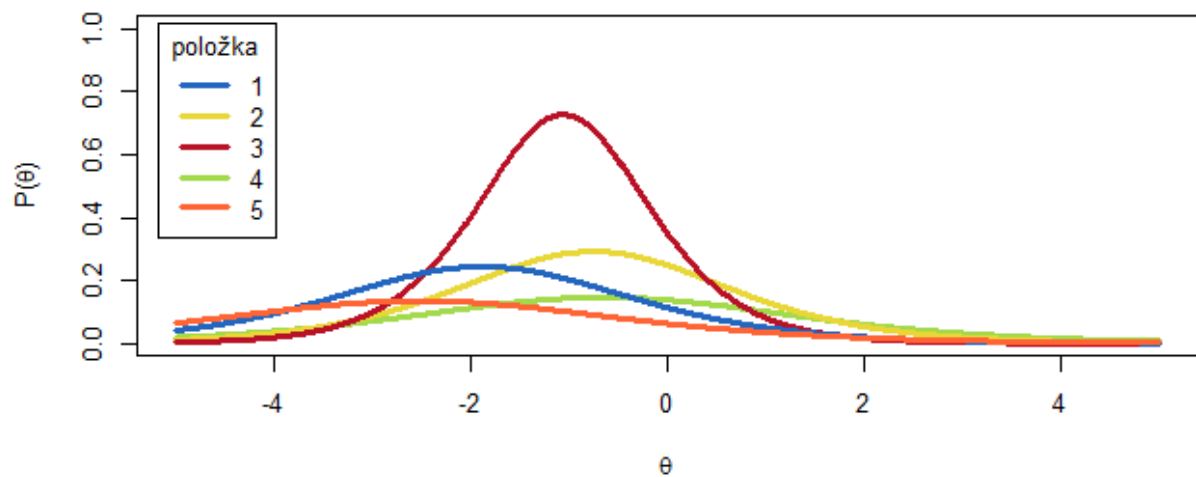
### Charakteristická funkce položek



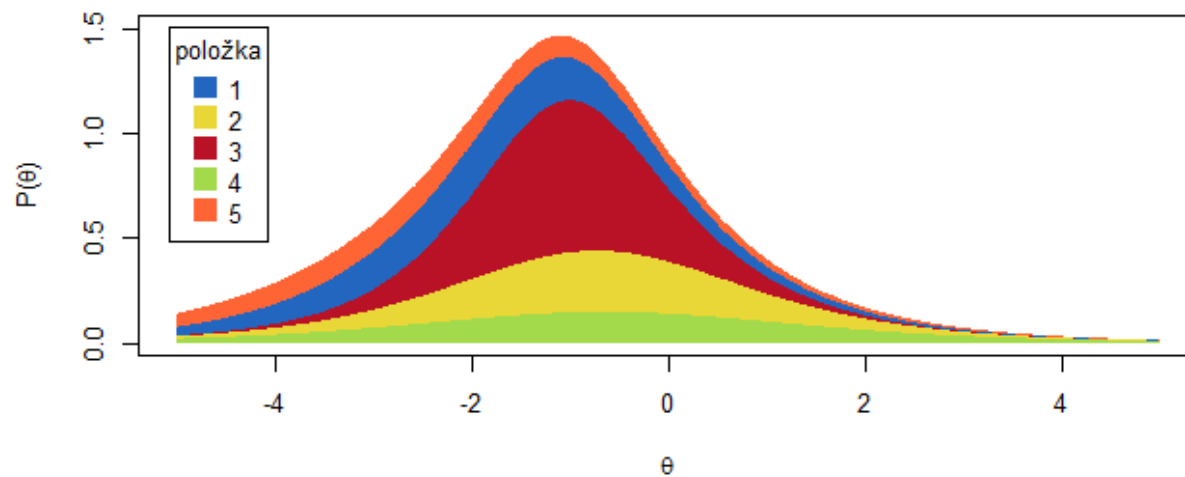
### Charakteristická funkce testu



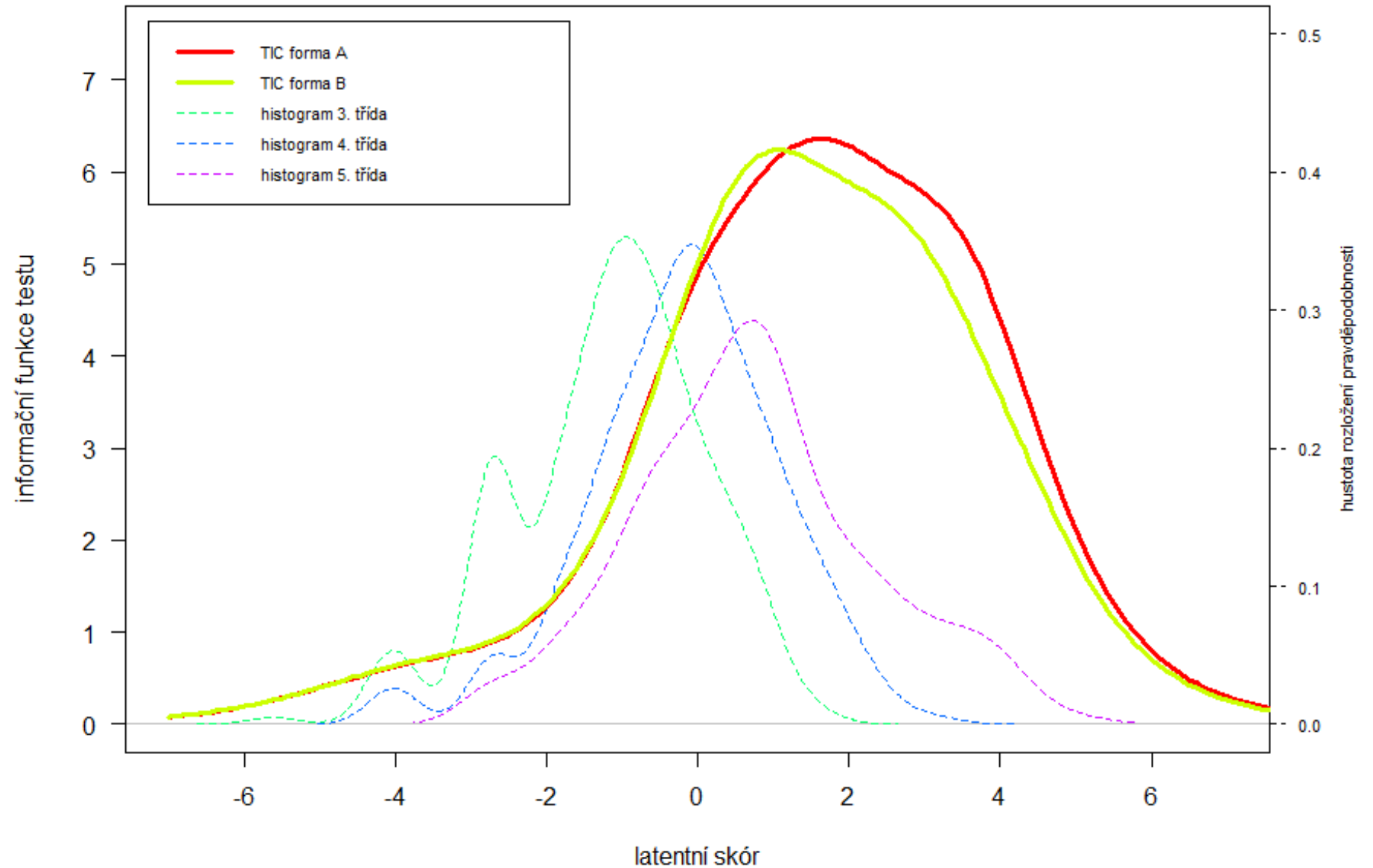
### Informační funkce položek



### Informační funkce testu



# Informační funkce testu a chyba měření





# Reliabilita v IRT

Stejná definice reliability jako v CTT:  $r_{xx'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} = \frac{\sigma_X^2 - \sigma_e^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$

- Interpretace je stejná, jako v CTT.

Odhad reliability:

- Do vzorce výše dosadíme za  $\sigma_X$  pozorovanou SD odhadů latentních rysů.
- A  $\sigma_e = RMSE = \sqrt{\frac{\sum_{p=1}^N SE_p^2}{N}}$ , kde  $SE_p$  je standardní chyba každého z N respondentů, a RMSE je tzv. root mean-square error (odmocnina průměrného chybového rozptylu). Takže:

$$r_{xx'} = 1 - \frac{RMSE^2}{\sigma_X^2} = 1 - \frac{\sum_{p=1}^N SE_p^2}{N\sigma_X^2}$$

Komplikace: Záleží na estimátoru.

- CML, MML a resp. EAP, MAP odhady pracují s odhadem latentního rysu (regrese k průměru) a tedy je odhadován nikoliv  $\sigma_X^2$ , ale přímo  $\sigma_T^2$ .

A tedy:  $r_{xx'} = \frac{\sigma_T^2}{\sigma_T^2 + RMSE^2}$

# Reliabilita v IRT

---

Interpretace: poněkud komplikovanější než v CTT.

V zásadě: reliabilita jako vysvětlený rozptyl.

- Podíl rozptylu odhadů faktorových skóre, který lze vysvětlit latentním rysem.

Interpretace jako korelace problematická.

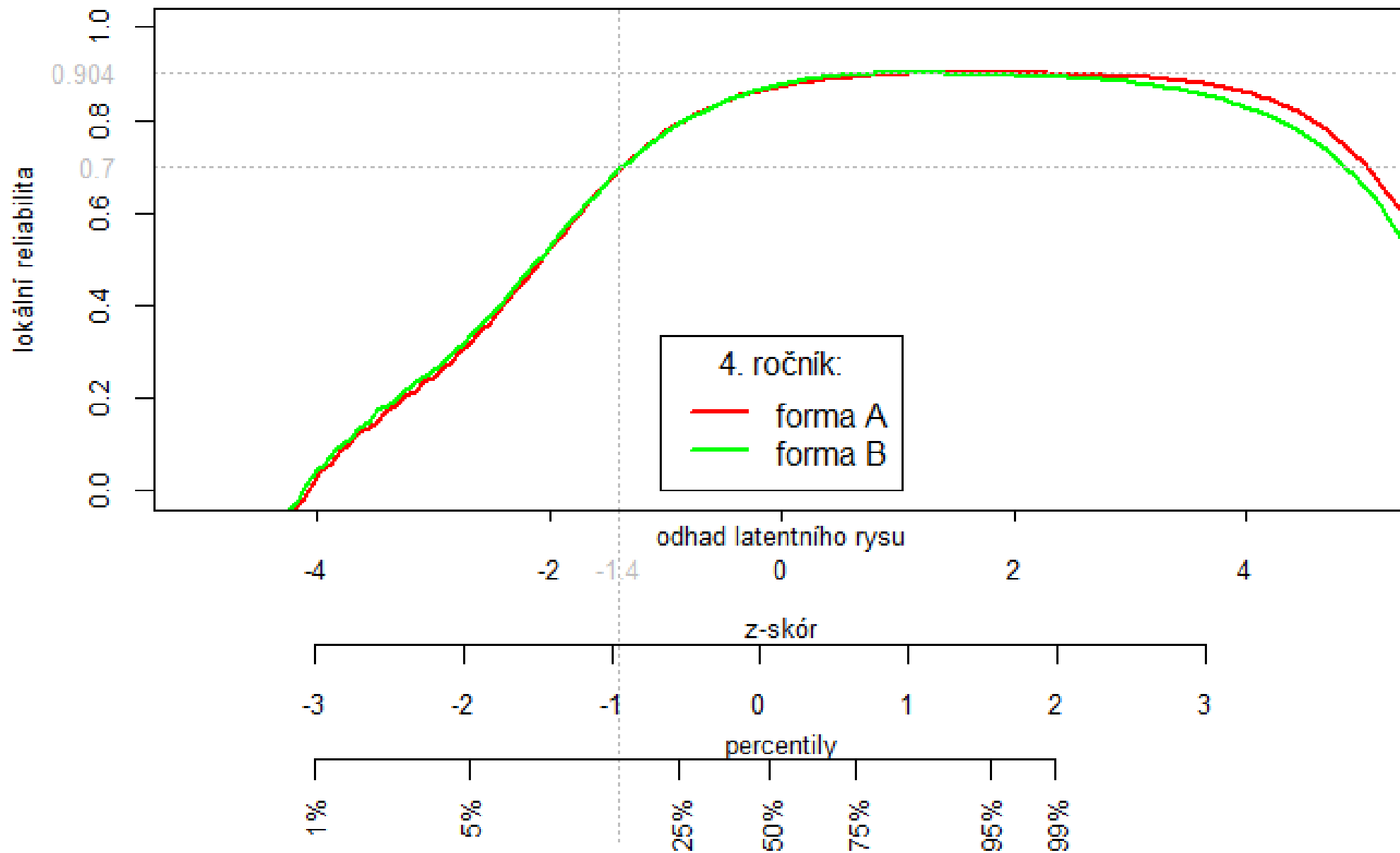
- Jen přibližně.
- Heteroskedascidita chyb odhadu.

# Lokální reliabilita

---

Pro reliabilitu měření konkrétního respondenta nebo konkrétní skupiny dosadíme za  $\sigma_e$  přímo SE daného odhadu či RMSE spočítaného pro konkrétní skupinu (Daniel, 1999): tzv. „**lokální reliabilita**“.

- Reliabilita testu, „pokud by fungoval všude stejně, jako pro dané respondenty“.
- Umožňuje zacílit výběr položek pro určitý testový záměr.
- Není reliabilitou v pravém slova smyslu (tj. „statisticky“), ale pro praktické použití je velmi užitečná.



# Shoda modelu s daty

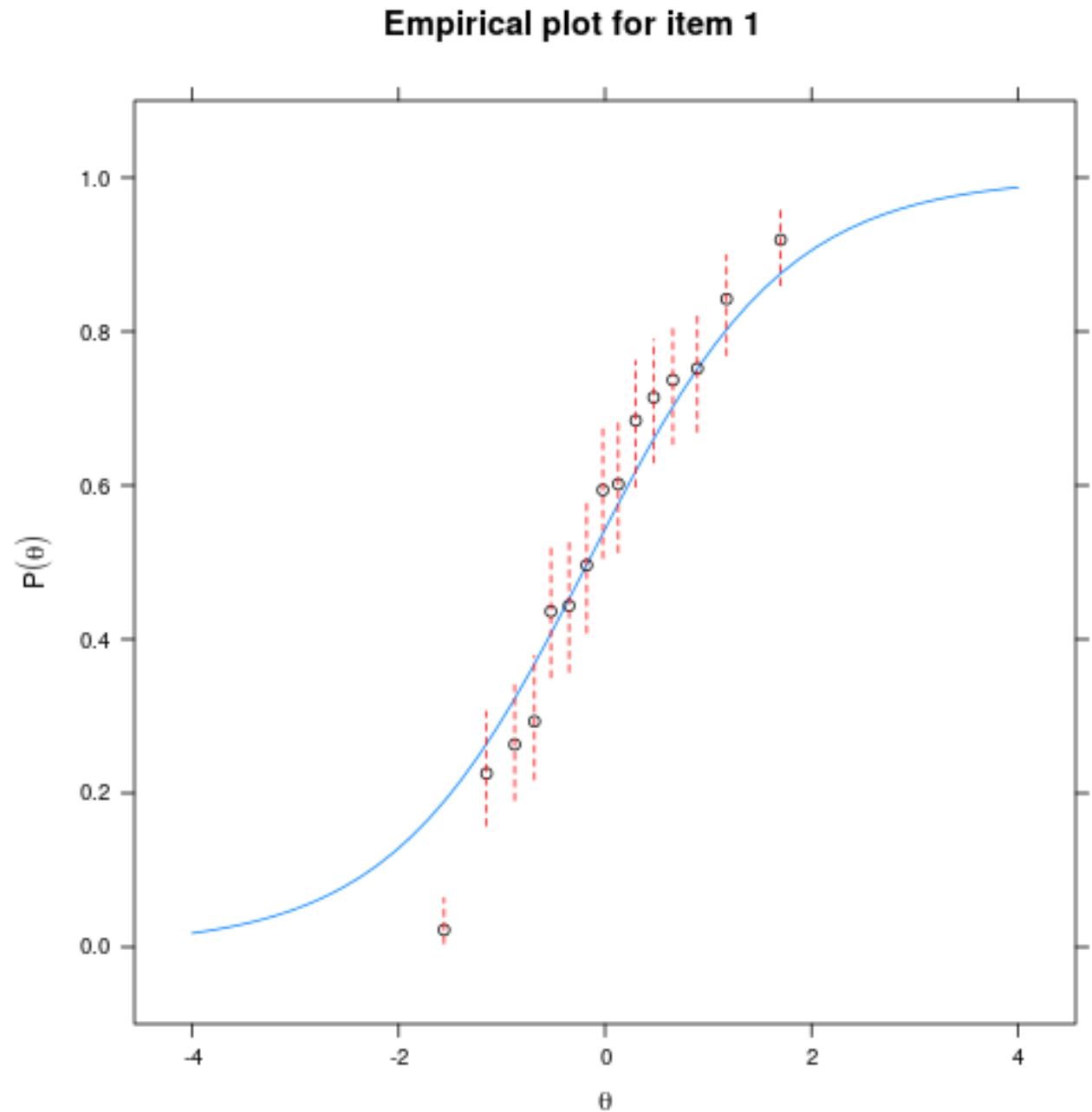
Na úrovni položky.

Na úrovni respondenta.

Pravděpodobnost konkrétní odpovědi.

Lokální závislost položek.

Na úrovni modelu.



# Shoda modelu s daty

---

## NA ÚROVNI CELÉHO MODELU

Odpovídají pozorovaná data IRT modelu?

Obdobný přístup jako v konfirmační faktorové analýze

- $\chi^2$ , TLI, CFI, RMSEA...
- Na hrubých datech zkreslené velkým počtem d.f., proto reprodukované kovarianční matice ([Maydeu-Olivares a Joe, 2006](#); [Cai a Hansen, 2013](#))

Umožňuje srovnání modelů navzájem

- 1PL vs. 2PL vs. 3PL... (nejen pomocí LRT).

IRT lze v tomto ohledu použít namísto běžné EFA/CFA

## NA ÚROVNI POLOŽKY/RESPONDENTA

Na kolik dobře odpovídají pozorované odpovědi 1 respondenta nebo odpovědi na 1 položku zvolenému IRT modelu?

Celá řada indexů.

- **Person fit:** identifikace aberantních odpovědí.
  - Např. pro účely purifikace dat při standardizaci.
- **Item fit:** doplňková informace o kvalitě položky (vedle parametrů modelu)
- Testy lokální nezávislosti (analogie reziduálních korelací a modifikačních indexů v FA).

# Shoda na úrovni respondenta/položky

---

Na rozdíl od CFA lze uvažovat o shodě modelu s daty na úrovni položky/respondenta.

- „Odpovídá univariační frekvenční tabulka pozorovaných odpovědí predikovaným odpovědím?“

Využití shody položky s daty:

- Vyřazování nefungujících položek, kontrola položek při equatingu, MG IRT a podobně.
- Úprava IRT modelu (ICC) pro konkrétní položku.

Využití shody respondenta s daty

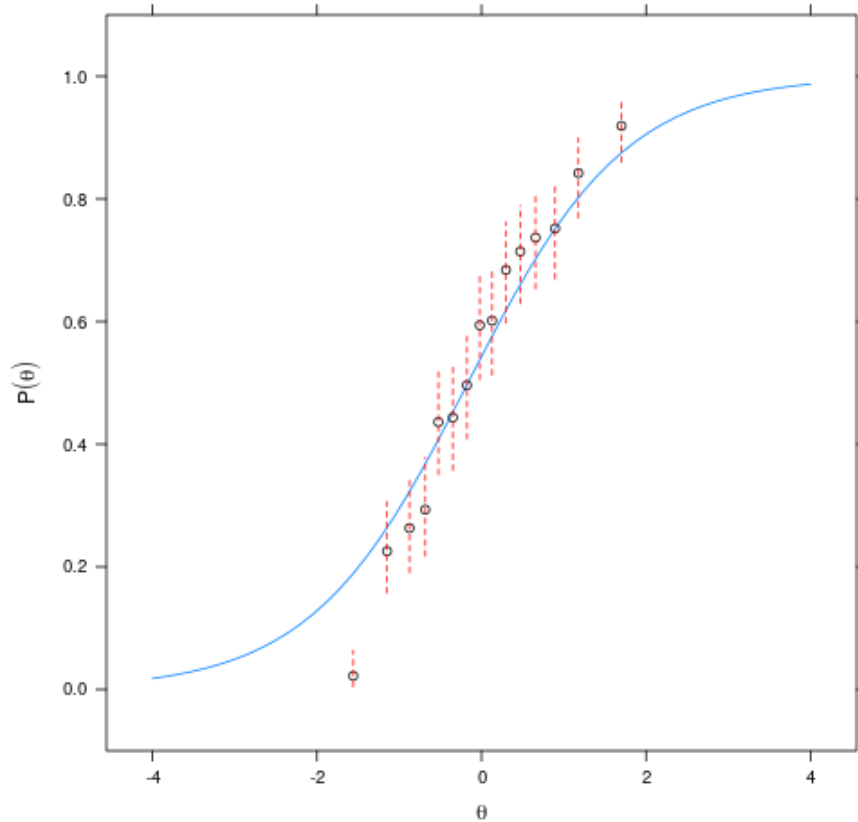
- Identifikace aberantního odpovídání.
- Vyřazení respondentů odpovídajících nahodile při standardizačních studiích.

Občas se využívá i identifikace konkrétní nepravděpodobné odpovědi.

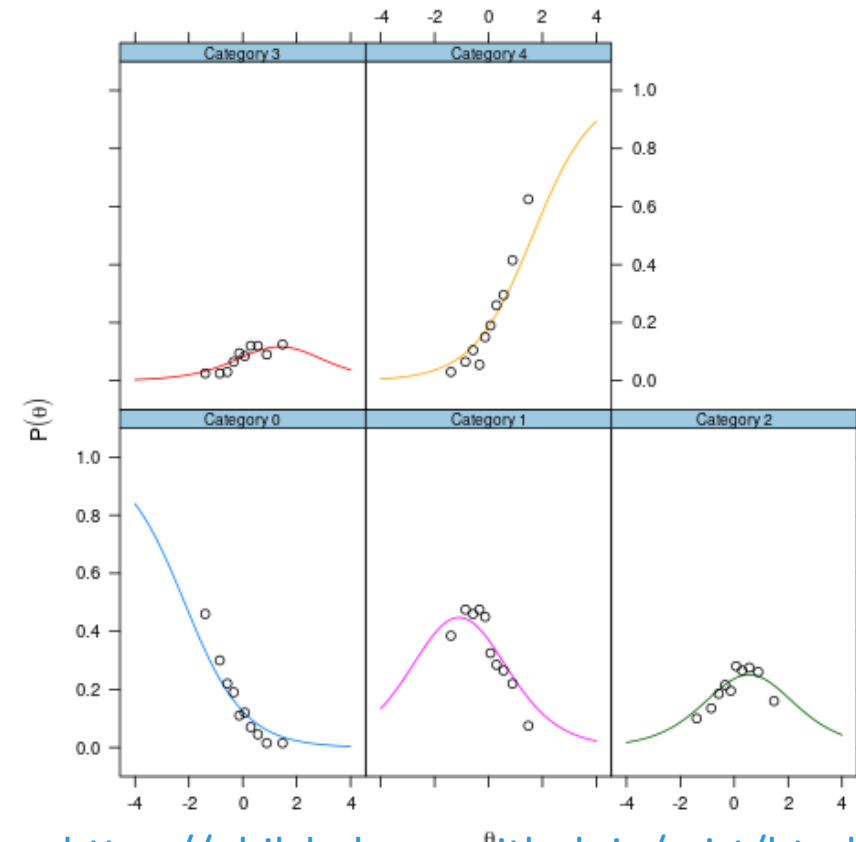
- WJ-IV COG: jsou vyřazeny odpovědi podle tzv. pravidla  $5\sigma$  ( $p = 0,00000057$ ).
- Například respondent odpoví chybně z důvodů nesouvisejících s měřeným rysem.

# Shoda položky s daty (item fit)

Empirical plot for item 1



Empirical plot for item 1



<https://philchalmers.github.io/mirt/html/itemfit.html>



# Lokální závislost položek

---

Explorace, zda dvě položky nesouvisí silněji či slaběji, než by odpovídalo modelu.

- „Odpovídá bivariační frekvenční tabulka dvou položek tomu, co predikuje model?“

Lze identifikovat prostřednictvím chí-kvadrát testu a odvozených metod.

Analogie k reziduální kovarianční matici, případně modifikačním indexům (M.I.) v CFA, nicméně výrazně výpočetně náročnější.

- Reziduální kovariance jsou přímo spočítané v rámci modelu.
- M.I. lze získat jednoduchými maticovými operacemi, zde je potřeba počítat pro každý pár zvlášť.

Velikost efektu (např. Cramerovo V) vs. signifikance...

# Shoda celého modelu s daty

---

Založen na chí-kvadrát testu stejně jako v CFA.

- CFI, TLI, RMSEA, SRMSR, AIC, BIC, saBIC a další.

Full-information statistiky:  $\chi^2$ ,  $G^2$ .

- Založené na diskrepanční likelihood funkci ( $G^2$ ), resp. diskrepanci pozorované a modelem predikované matice odpovědí ( $\chi^2$ ).
- Jinými slovy: diskrepance multivariační frekvenční tabulky všech položek.
- Jaké jsou předpoklady  $\chi^2$ ? Jsou dodrženy?

Proto limited-information statistiky:  $M_2$ ,  $M_2^*$ ,  $C_2$ .

- $M_2$ ,  $M_2^*$  – univariační a bivariační frekvence, binární ( $M_2$ ) a polytomické ( $M_2^*$ ) položky.
- $C_2$  – varianta pro kratší testy s delší odpověďovou škálou, pouze bivariační frekvenční tabulky.

Interpretace indexů CFI, TLI, RMSEA a dalších založených na  $M_2$ ,  $M_2^*$ ,  $C_2$  analogická indexům v CFA.

# Polytomní IRT modely

Graded Response Model

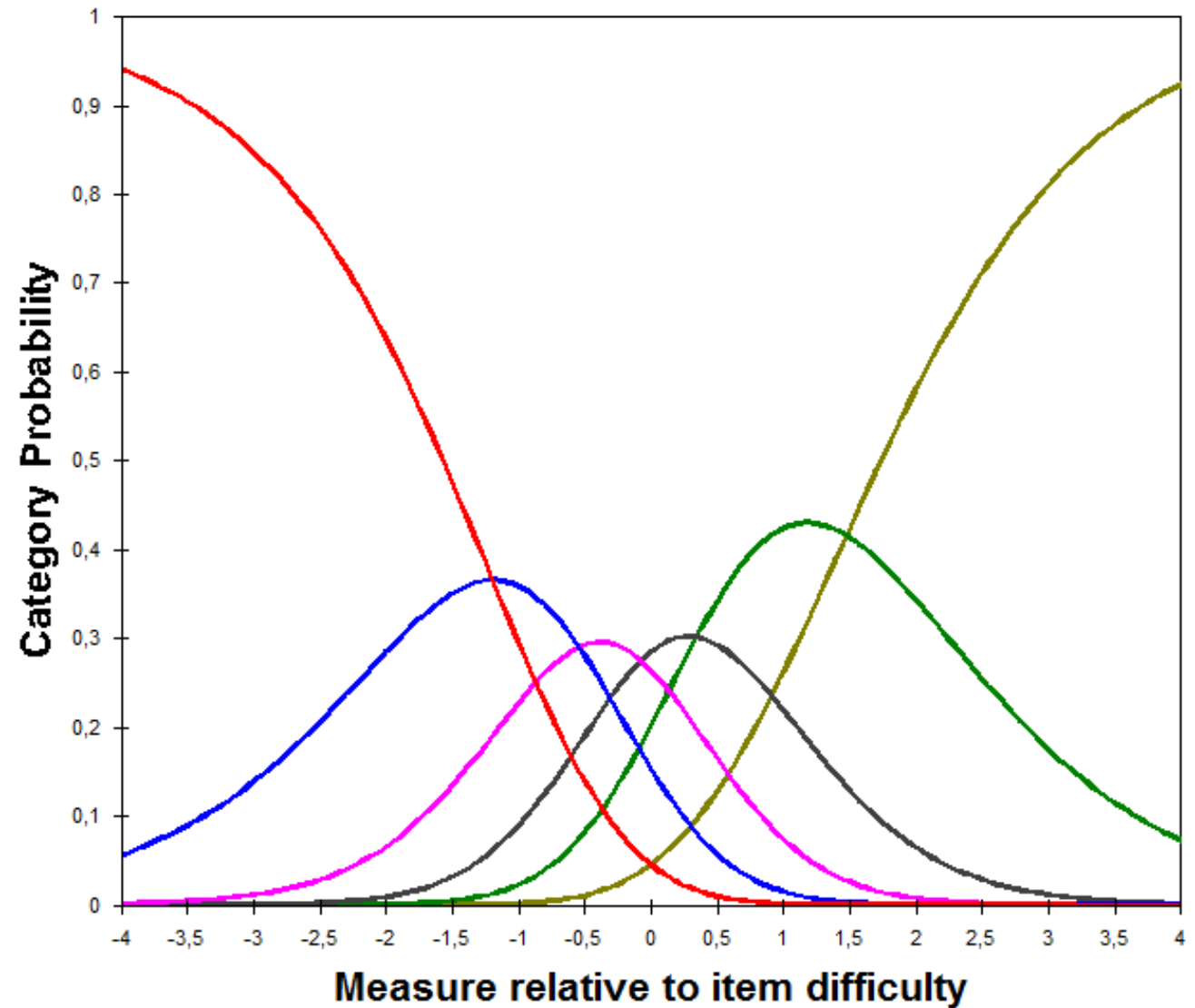
Generalized Partial Credit Models

Tutzův sekvenční model

Bockův Nominal Response Model

Ordinální faktorová analýza

## 1. Jsem spíše vyšší než muži mého věku



# Polytomní IRT modely

---

Určeny pro práci s položkami s více odpověďmi.

- Např. Likertova škála 1-7, parciálně správné odpovědi ve výkonovém testu nebo multiple-choice položky.
- Na rozdíl od CTT mohou vést k doporučení zvýšit či snížit počet kategorií položek.
- Zpravidla 1PL či 2PL.

Modely pro nominální či nominální kategorie.

3 hlavní kategorie polytomních modelů<sup>1</sup>:

- difference models (GRM, MGRM) – výhradně ordinální kategorie
- divide-by-totals (PCM, GPCM, NRM)
- sekvenční modely (Tutzův sekvenční model)

<sup>1</sup> Sijtsma, K., & Hemker, B. (2000). A Taxonomy of IRT Models for Ordering Persons and Items Using Simple Sum Scores. *Journal of Educational and Behavioral Statistics*, 25(4), 391-415. <http://www.doi.org/10.2307/1165222>

# Polytomní modely (z rychlíku)

---

## Ordinální data

- (Generalized) Partial Credit Model (GPCM, PCM) – původně určený pro výkonová data, kde se skóre položky sestává z dílčích samostatně skórovaných kategorií.
- Graded Response Model (GRM) – původně určený pro dotazníky, kde respondent zaznamenává spojitou, kontinuální míru „souhlasu“ na ordinální škále.

## Nominální data

- Nominal Response Model (NRM) – každá odpověďová kategorie je modelovaná zvlášť.
- Multiple-choice Model (MCM) – dílčí úprava NRM vhodné pro MC data.

# Graded Response Model (GRM)

---

Zobecnění 2PL modelu (Samejima, 1969): série 2PL modelů:

$$P_{ix}^*(\theta) = \frac{e^{a_i(\theta - b_{ix})}}{1 + e^{a_i(\theta - b_{ix})}}$$

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta)$$

Dvoukrokový odhad pravděpodobnosti:

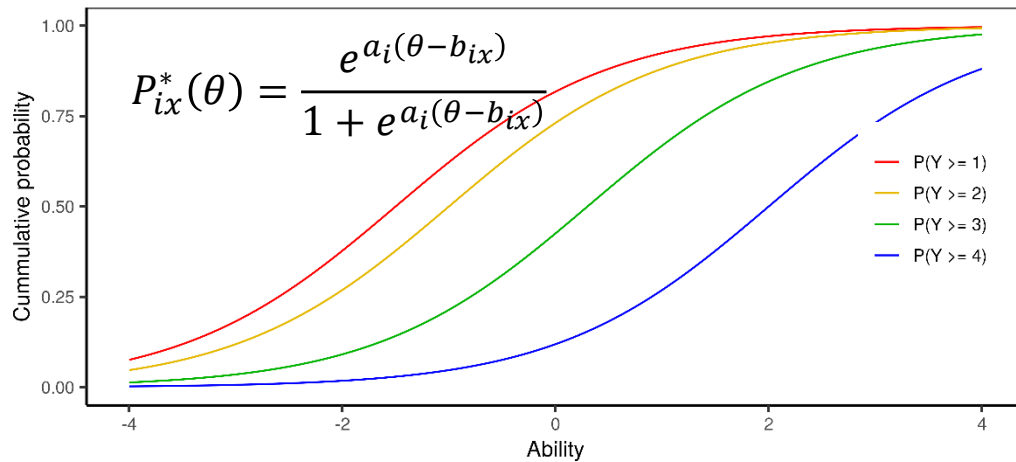
- Pro každou odpověď  $x$  je odhadnuta pravděpodobnost  $P_{ix}^*(\theta)$ , že respondent odpoví touto nebo vyšší odpovědí (vs. nižší).  $b_{ix}$  - obtížnost kategorie  $x$  na položce  $i$ . Pro účely výpočtu je nejnižší kategorie  $P_{i(x=0)}^*(\theta) = 1$
- Výsledná pravděpodobnost konkrétní odpovědi  $P_{ix}(\theta)$  je rozdílem odhadnuté pravděpodobnosti a pravděpodobnosti o jedna „vyšší/těžší“ odpovědi.

Modified Graded Response Model (MGRM, Muraki, 1990); někdy též GRSM.

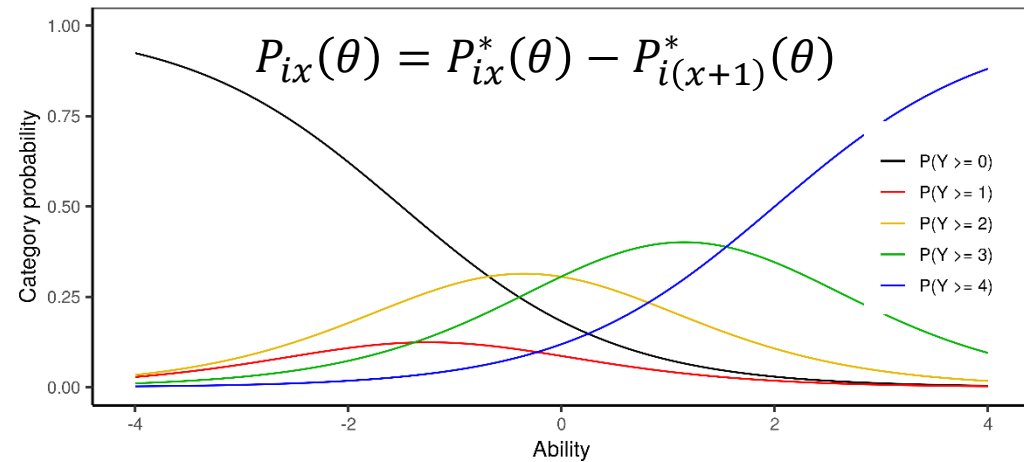
- $P_{ix}^*(\theta) = \frac{e^{a_i[\theta - (b_i - c_j)]}}{1 + e^{a_i[\theta - (b_i - c_j)]}}$ , kde  $c_j$  jsou parametry jednotlivých prahů  $j$  a  $b_i$  obtížnost položky  $i$ .

# Graded Response Model (GRM)

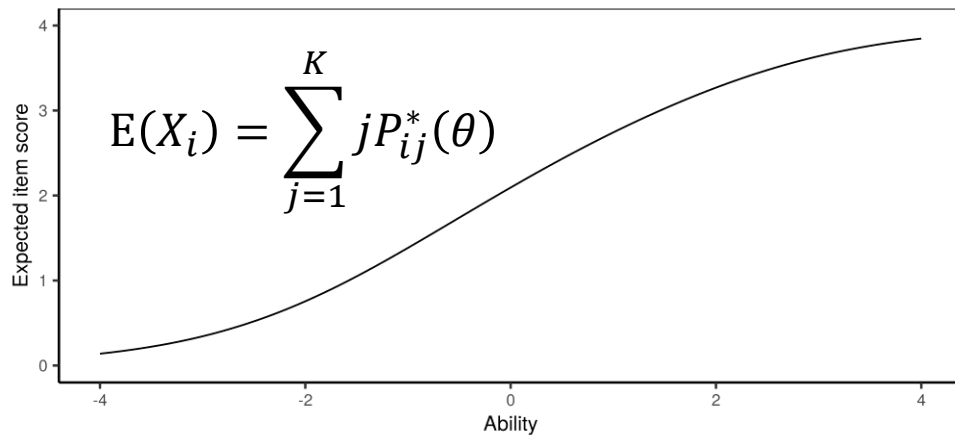
Cummulative probabilities



Category probabilities



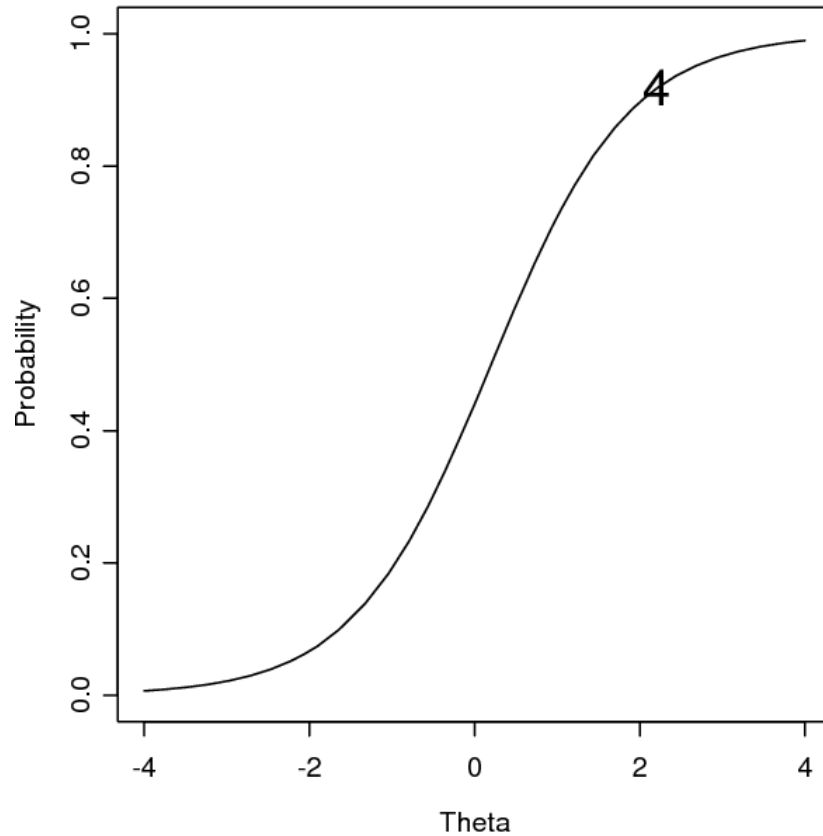
Expected item score



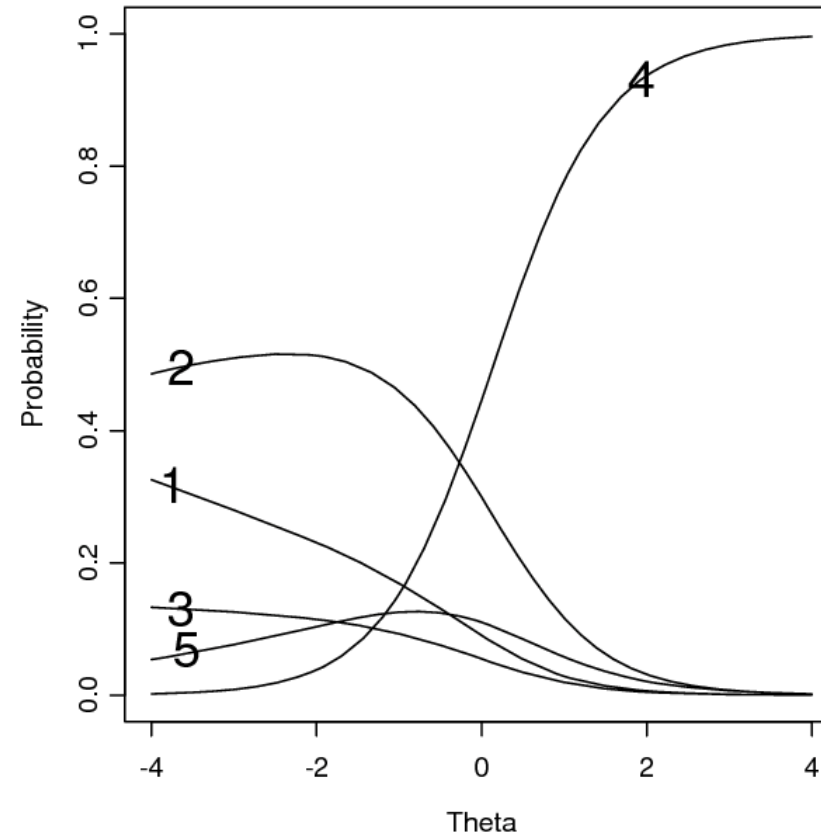
Martinkova P., & Drabinova A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, 10(2), 503-515. doi: 10.32614/RJ-2018-074

# Nominal Response Model

Two Parameter Logistic Model

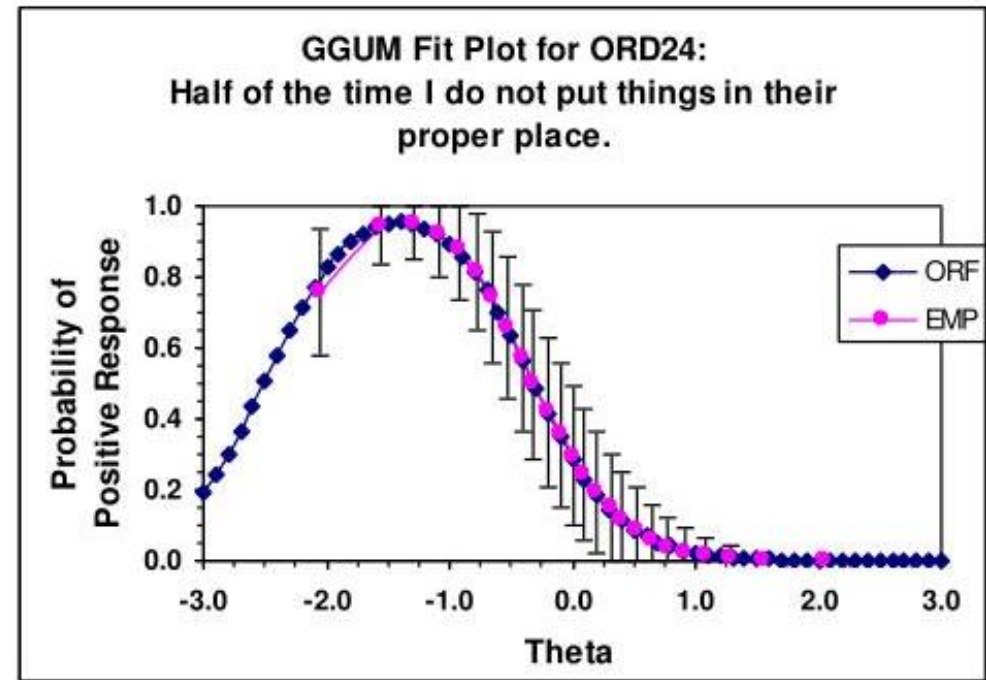
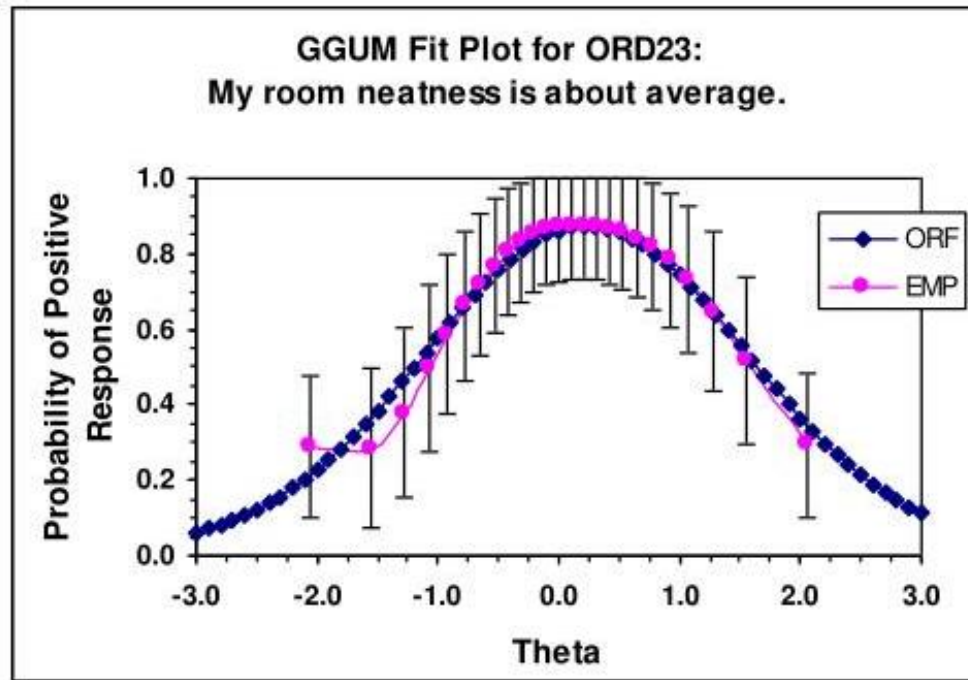


Nominal Response Model





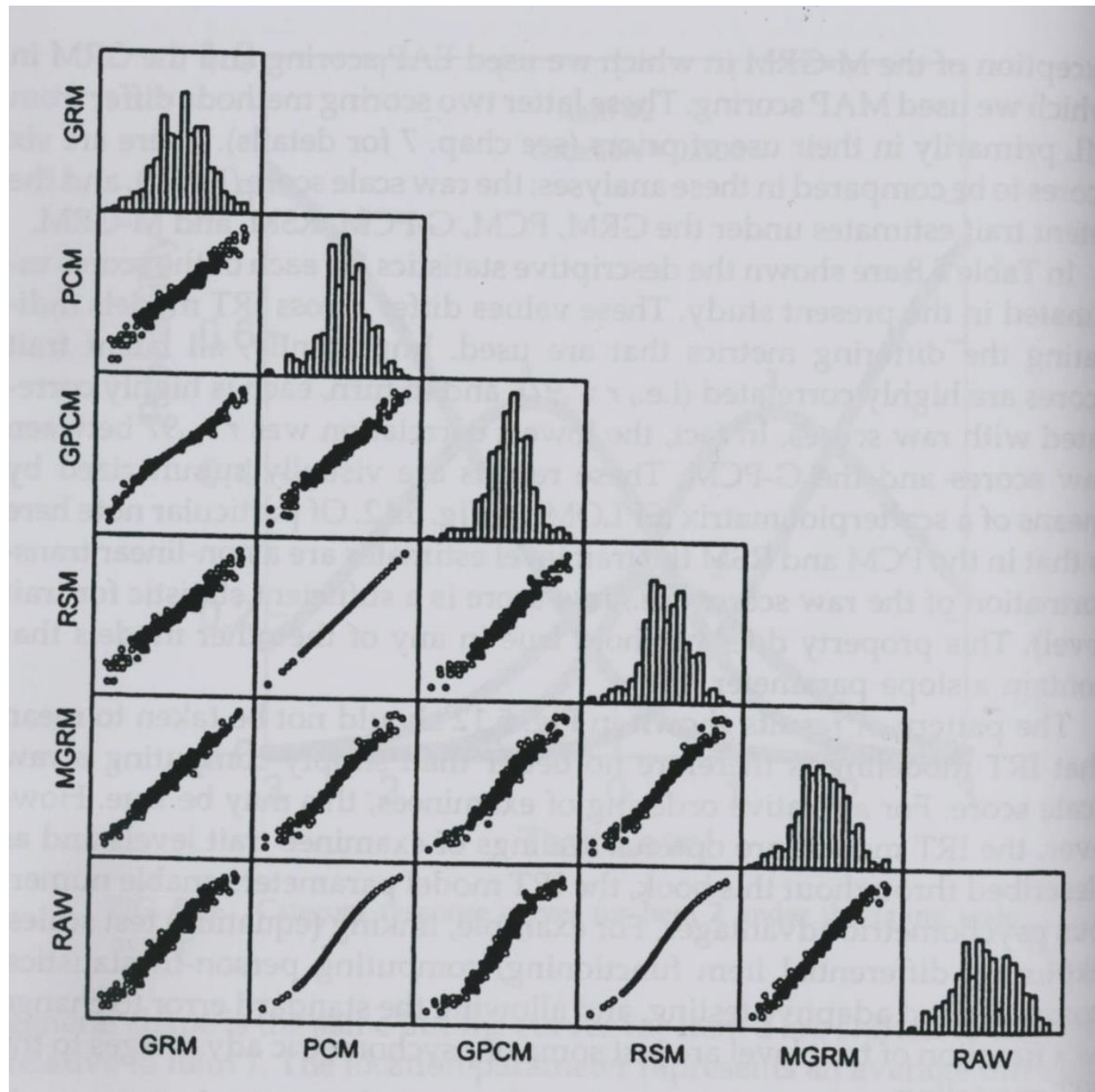
# Ukázka ideal-point modelu



# Srovnání modelů

Běžné modely:  
divided-by-total a graded modely.

Embretson a Reise (2009)



# Ordinální faktorová analýza

---

Ordinální faktorová analýza je založená na tetrachorických (binární položky), respektive polychorických korelacích (ordinální položky).

Tetrachorická/polychorická korelace:

- Existuje spojitá, intervalová, normálně rozložená latentní odpověď (LR, Latent Response).
- Ta není přímo pozorovaná (je latentní).
- Manifestuje se pouze jako ordinální kategorie.
- Pokud LR překročí příslušný *práh* položky, pozorujeme vyšší kategorii.

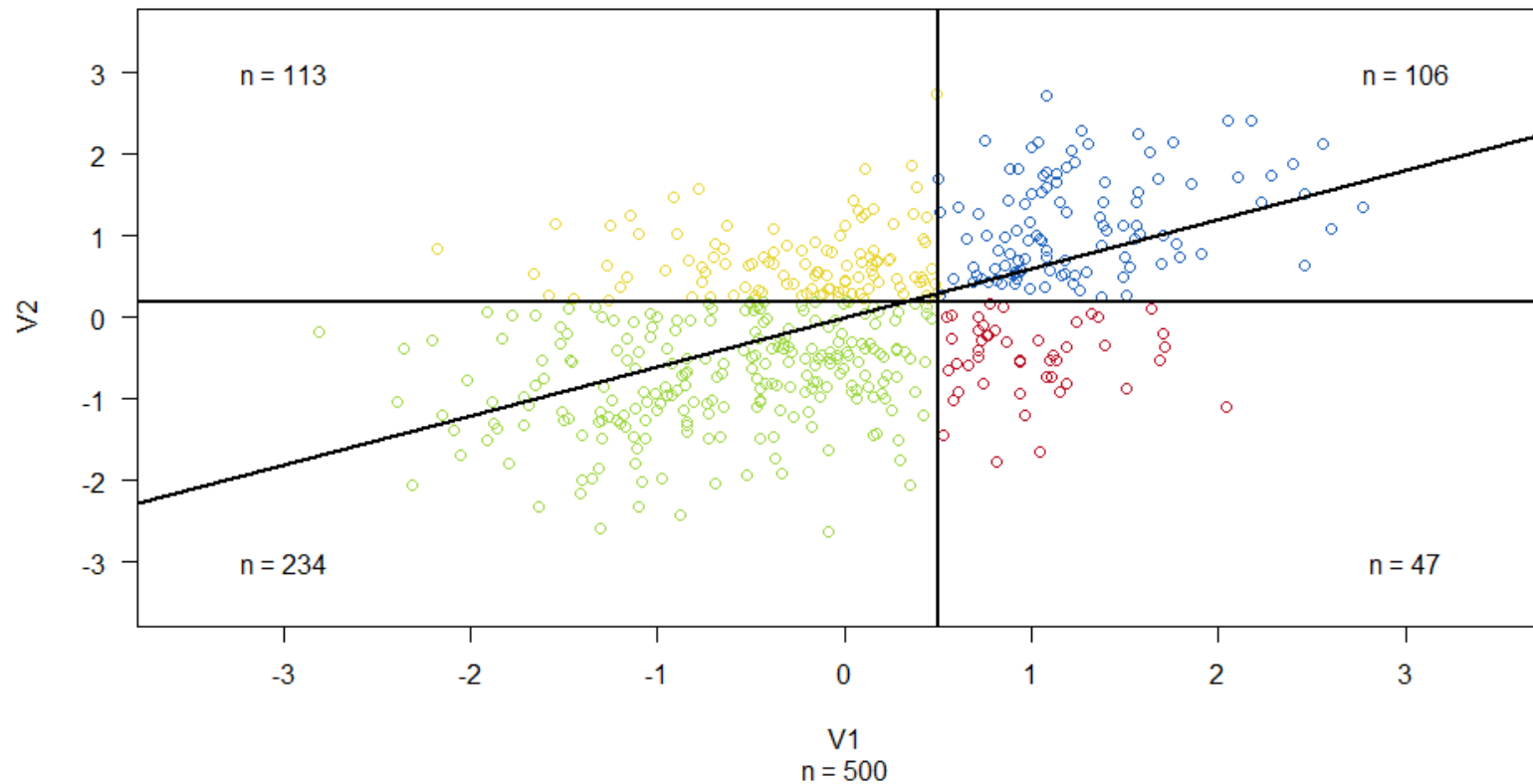
Tetra/poly korelace jsou odhadovány na základě bivariačních frekvenčních tabulek.

Ordinální FA tedy faktoruje matici polychorických korelací.

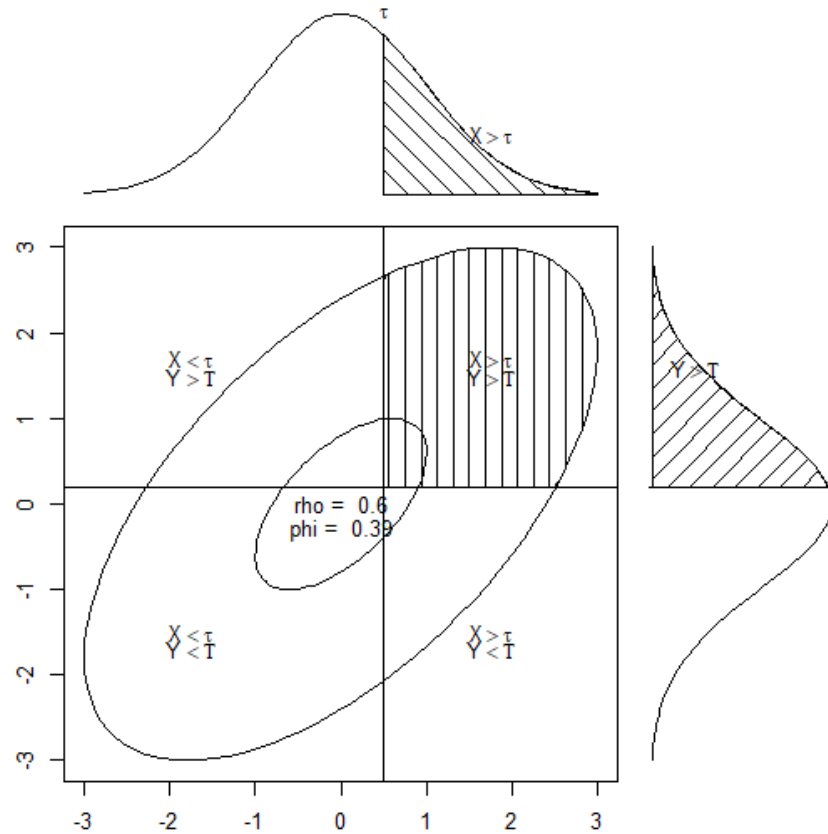
- Tradiční postup: Odhadne se polychorická matice a ta vložena do EFA.
- Modernější postup: polychorická matice a parametry FA jsou odhadovány naráz pomocí DWLS/WLSMV estimátoru.

# Tetrachorická korelace ( $\rho = 0,6$ )

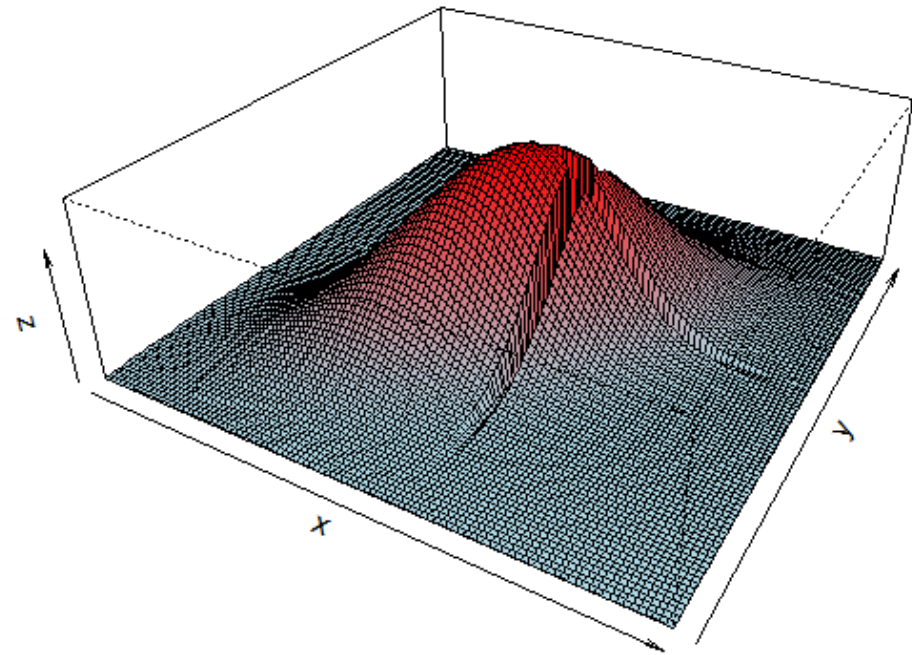
Pearson  $r = 0.529$ ; Tetrachoric  $r = 0.53$  ( $t_1 = 0.507$ ,  $t_2 = 0.156$ )



# Tetrachorická korelace ( $\rho = 0,6$ )



Bivariate density  $\rho = 0.6$



# Ordinální faktorová analýza

Klasická CFA: latentní faktor způsobuje manifestní odpověď.

$$X_i = \lambda_i f + v_i + \varepsilon, \quad \text{var}(\varepsilon) = \theta_i$$

- $f$  – faktor,  $\lambda_i$  - faktorový náboj,  $\theta_i$  - reziduální rozptyl

Ordinální CFA: latentní faktor způsobuje latentní odpověď (LR).

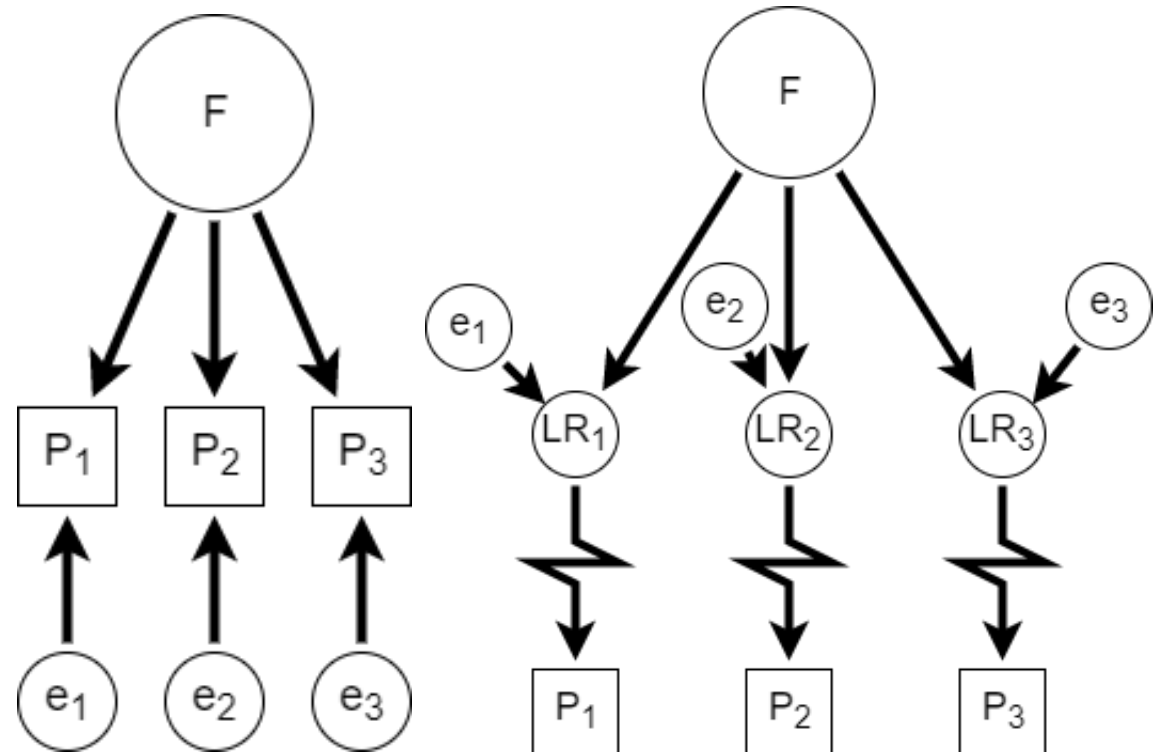
$$LR_i = \lambda_i f + v_i + \varepsilon, \quad \text{var}(\varepsilon) = \theta_i$$

$$LR_i \geq \tau_{i(k-1)} \wedge LR_i < \tau_{ik} \implies X_i = k, \quad \tau_{i0} = -\infty$$

- $\tau_{ik}$  - k-tý práh položky i.

Ordinální CFA je probitový Graded Response Model.

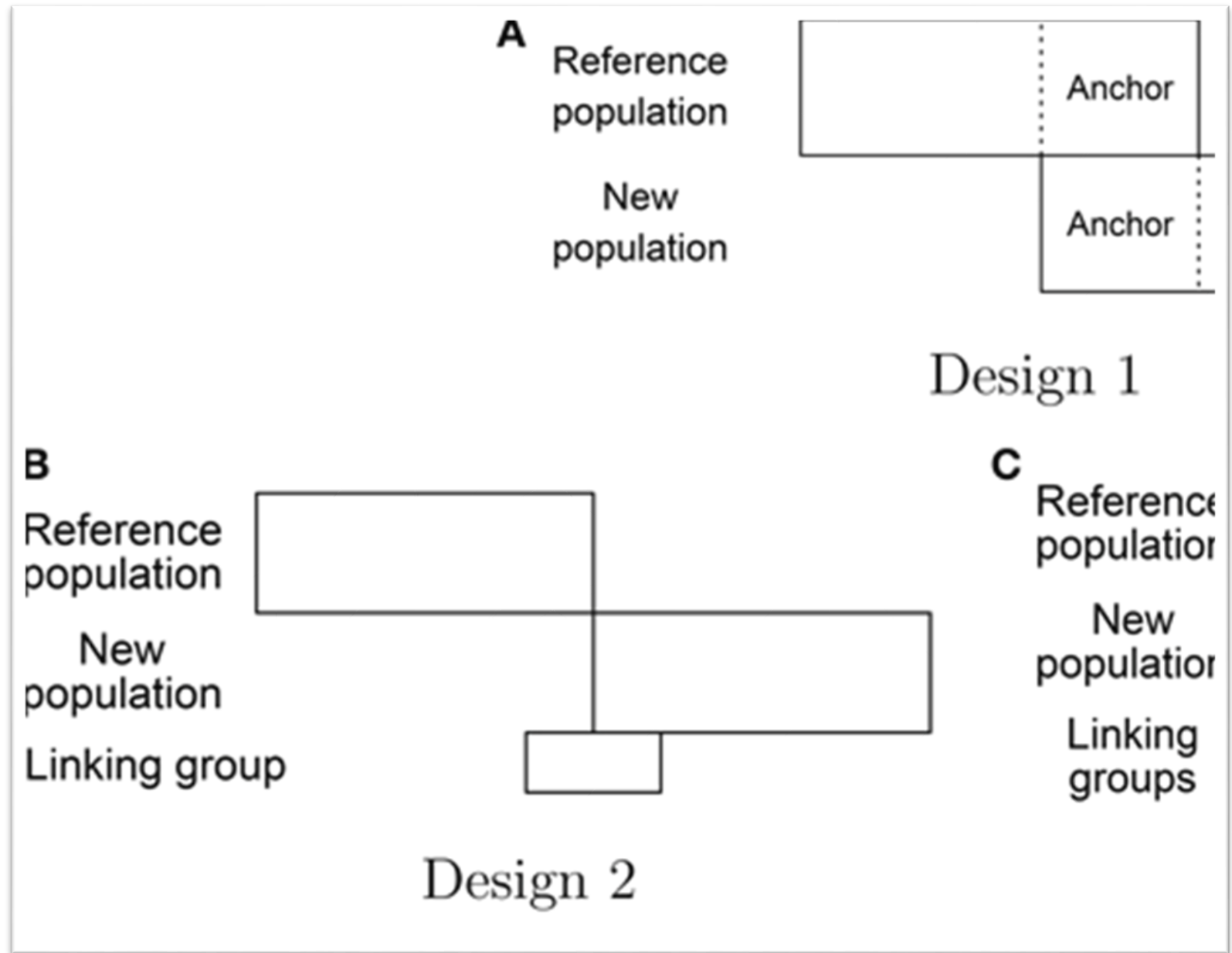
- S nepatrně odlišnou parametrizací.



# Vybrané aplikace IRT:

Počítačové adaptivní testování (CAT)

Equating, linking



# Typická využití IRT

---

Běžné ověření (konfirmační IRT) a explorace (explorační IRT) faktorové struktury.

- Test pak může být skórován klidně s využitím CTT.

IRT jako nástroj pro škálování.

- Zajímají nás právě IRT odhady latentního rysu.

IRT jako výzkumný nástroj (explanační modely).

IRT jako model měření.

DIF analýza a MG IRT (viz přednáška o férovosti).

**Další specifická využití:**

- **Počítačové adaptivní testování (CAT)**
- **Vyvažování paralelních forem testu (linking, equating) - souvisí se škálováním.**



# Počítačové adaptivní testování

---

## Computerized Adaptive Testing (CAT)

1. myšlenka: Nemá smysl administrovat respondentovi takové položky, které nezpřesní odhad jeho latentního rysu.

- Jsou pro něj příliš jednoduché (téměř jistě je odpoví správně)
- Případně příliš těžké (téměř jistě odpoví chybně).
- Takové položky nesou příliš málo informace (nízká hodnota informační funkce).

2. myšlenka: IRT nevdává chybějící data. Pracuje s dílčími položkami, nikoliv celým testem.

**Použití:** TOEFL, GRE, v ČR A3DW či ATAVT od Schufrieda, Invenio od IVDMR 😊.

# Počítačové adaptivní testování: Postup

---

1. Administruji úvodní set položek a odhadnu úroveň latentního rysu.
2. Vyberu a administruji položku, která má pro danou úroveň rysu maximální odpověďovou funkci.
  - Tedy (u 1PL), jejíž obtížnost je nejbližší úrovni odhadnuté schopnosti ( $P(\theta) = 0,5$ ).
  - Případně nepatrně lehčí (typicky  $0,5 < P(\theta) < 0,7$ ), abych respondenta motivoval.
  - Často ještě randomizace, aby se neopakovaly stále tytéž položky (s největším  $a$ -parametrem).
3. Odhadnu znovu rys.
4. Opakuji kroky 2 a 3, dokud nedosáhnu pravidla ukončení.
  - Vyčerpám všechny položky nebo cílového počtu položek/času administrace.
  - Standardní chyba odhadu se sníží pod stanovenou mez.
  - Apod.

# Počítačové adaptivní testování: Výhody

---

Efektivnější testování.

- Zkrácení testu při zachování reliability / zvýšení reliability při zachování délky.

Větší množství položek, každý má trochu jiné položky.

- Redukce možnosti opisovat.
- Snížení rizika a hlavně důsledků případného úniku položek.
- Respondent nemusí odpovídat na neadekvátní položky (příjemnější testování).

Lze využít i při individuální administraci.

- Např. s využitím administrace na tabletu.

# Test equating (vyvažování testů)

---

Vyvážení obtížnosti jednotlivých forem testu.

- V high stakes testech jednorázové vyvážení – sjednocení obtížností a srovnání probandů napříč formami testu.
- V psychologických metodách vyvážení skóru paralelních forem a vyvinutí rovnocenných nástrojů.
- **Linking** (prosté srovnání měřítek) vs. **equating** (zajištění stejné škály).

Předpoklad: Obě formy měří stejný konstrukt (otázka validity).

GRE, SAT: od konce 80./začátku 90. let je (v USA) IRT vyvažování high-stakes testů normou.

Typické kroky: volba designu, sběr dat, samotná transformace.

# Test equating (vyvažování testů)

---

Tři tradiční způsoby založené na pozorovaném skóre:

- **Vyvažování na základě průměru (M)** – testy musí mít stejné rozptyly, data musí být normálně rozdělená.  $x_2 = x_1 + \bar{X}_2 - \bar{X}_1$
- **Lineární vyvažování (M, SD)** – rozptyly se mohou lišit, data musí být normální.  $x_2 = \bar{X}_2 + \frac{\sigma_2}{\sigma_1}(x_1 - \bar{X}_1)$  (transformace přes z-skór)
- **Equipercilové vyvažování** – varianty jsou upraveny tak, aby tentýž skór měl v obou variantách stejný percentil. Výsledkem je stejné rozdělení dat, je silně závislé na vzorku (použitelné jen u velkých souborů).
  - Používá se i pro standardizaci nenormálních skórů na normální.
  - Percentilové vyvažování není vyvažování, percentil z principu ztrácí část informace. Žádné zvláštní požadavky na data.

IRT vyvažování bylo prvními hromadnými aplikacemi IRT do praxe.

# IRT equating: Sběr dat

---

Celá řada různých designů.

Designy s jednou výzkumnou skupinou: **single-group design**.

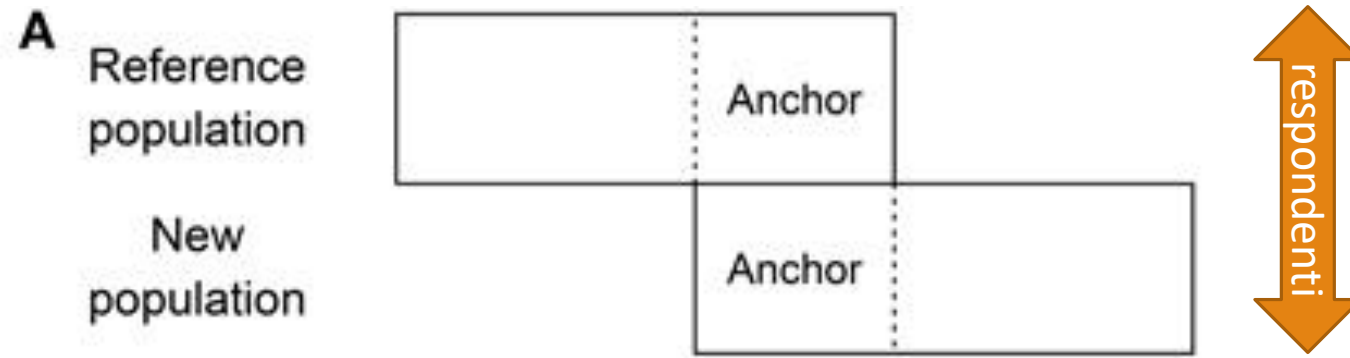
- Každá osoba absolvuje oba testy (counterbalancing = střídání pořadí).
- Případně část respondentů absolvuje oba testy (common-person design).

Designy s náhodnými skupinami: **random-group design, random-equivalent-group**.

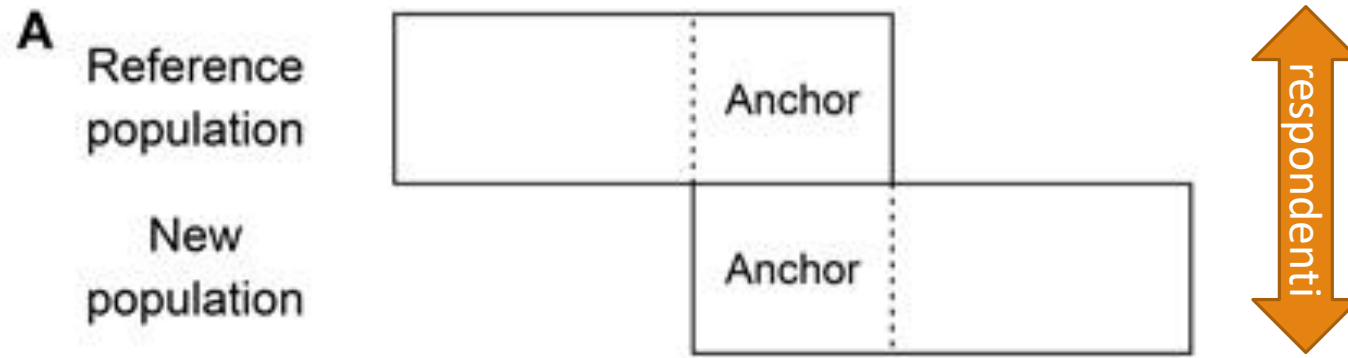
- Respondenty náhodně přiřadíme do výzkumných skupin. Předpokládáme, že jsou ekvivalentní.

Designy se společnými položkami:

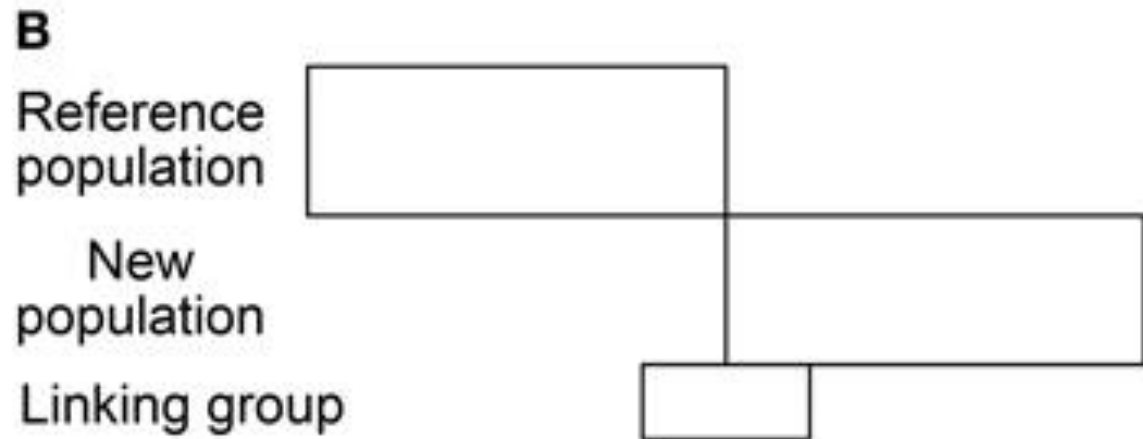
- Dvě nezávislé/nenáhodné skupiny, ale oba testy mají společné položky (tzv. „kotvu“ – **anchor test**), které slouží ke kalibraci. **Největší spolehlivost a hlavní výhoda IRT.**
- Ta může, ale nemusí být zahrnuta pro zjištění celkového skóru.
- Kotev může být více („planned missing data design“).



Design 1: anchor-item design

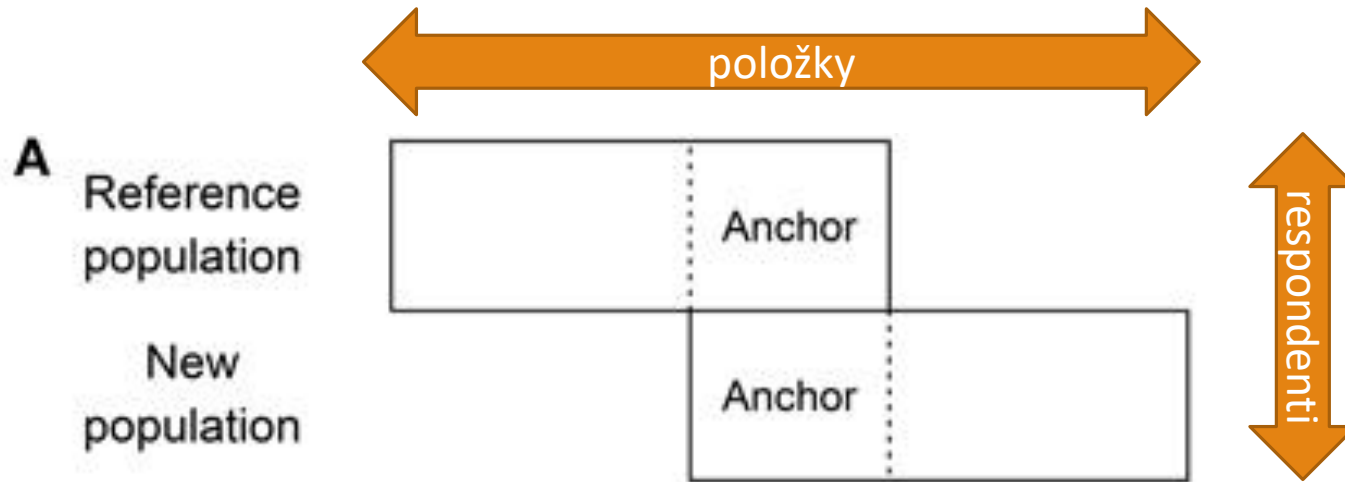


Design 1: anchor-item design

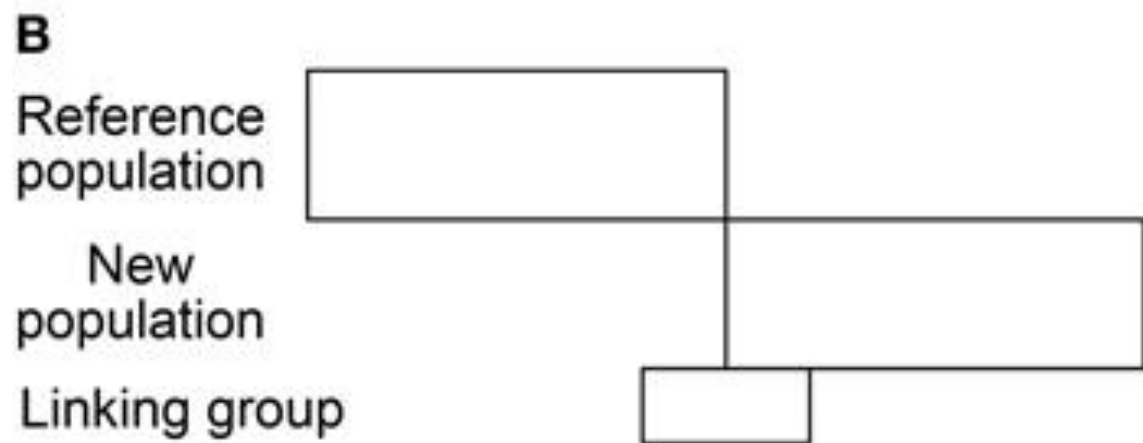


Design 2: post-equating design

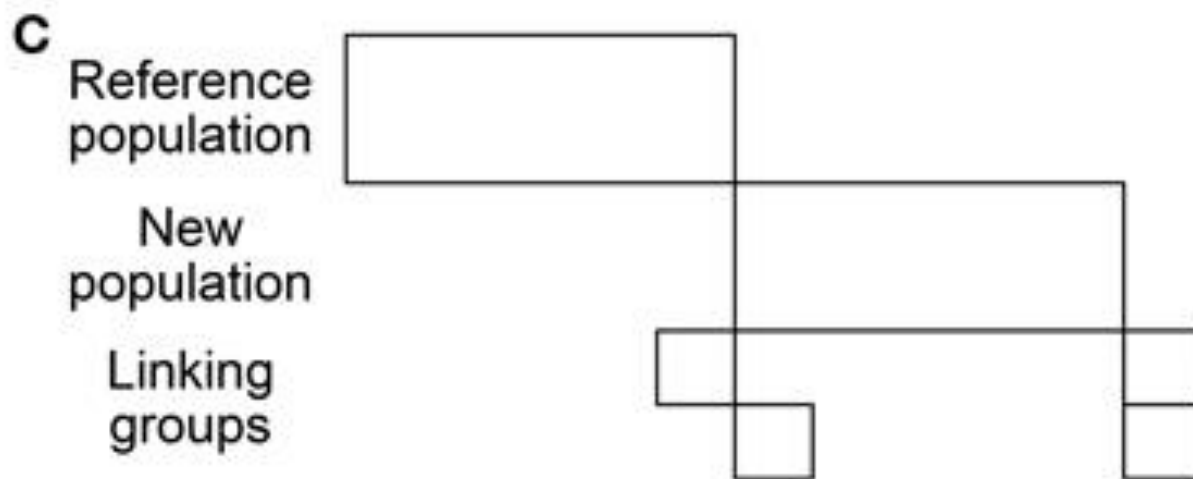




Design 1: anchor-item design



Design 2: post-equating design



Design 3: post-equating design

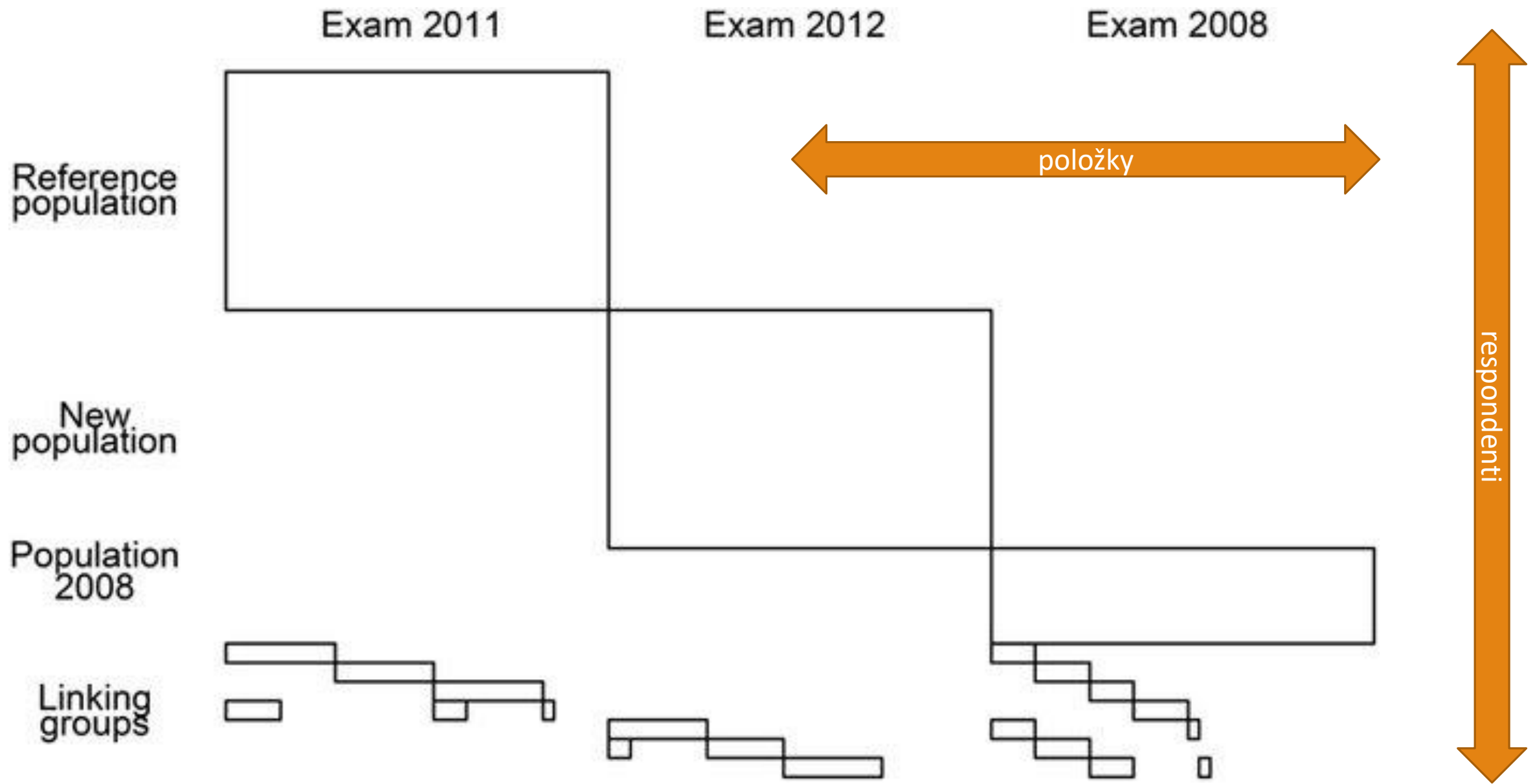


Figure 2-I: Test design CSEC

Session		Testblocks												
January	2014	<b>Block 1 (linking items )</b>	Block 2 (new items)	Block 3 (new items)										
July	2014			<b>Block 3 (linking items )</b>	Block 4 (new items)	Block 5 (new items)								
January	2015					<b>Block 5 (linking items )</b>	Block 6 (new items)	Block 7 (new items)						
July	2015							<b>Block 7 (linking items )</b>	Block 8 (new items)	Block 9 (new items)				
January	2016									<b>Block 9 (linking items )</b>	Block 10 (new items)	Block 11 (new items)		
July	2016											<b>Block 11 (linking items )</b>	Block 12 (new items)	Block 13 (new items)

Design použitý v Caribbean Secondary Education Certificate (Stancel-Piątak, Cígler, Wild, 2018).