MUNI FSS Katedra psychologie

Přednáška 4–5:
Teorie odpovědi na položku

8. a 15. 10. 2019 | PSYn4790 | **Psychometrika: Měření v psychologii**
Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler | 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 intensivní vs. extensivní 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 neznamenají, že jde o „čísla“ v pravém slova smyslu.

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

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

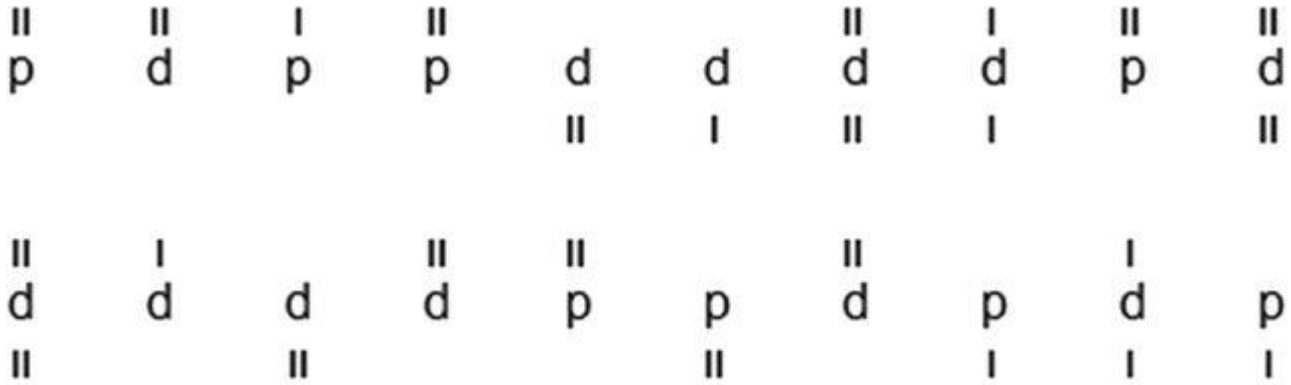
Kdy zejména to vadí?

Jde o měření? | Likertova škála

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

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

Jde o měření? | Měření pozornosti



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**.

https://commons.wikimedia.org/wiki/File:D2-Test.jpg

Měření v rámci CTT

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

- **1.** Pravidelně zvracím, abych si udržela svou váhu.
- **2.** Počítám gramy tuku na jídle, které jím.
- **3.** Tvrdě cvičím, abych spálila 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, $CI_{95\%} = 2{,}94$

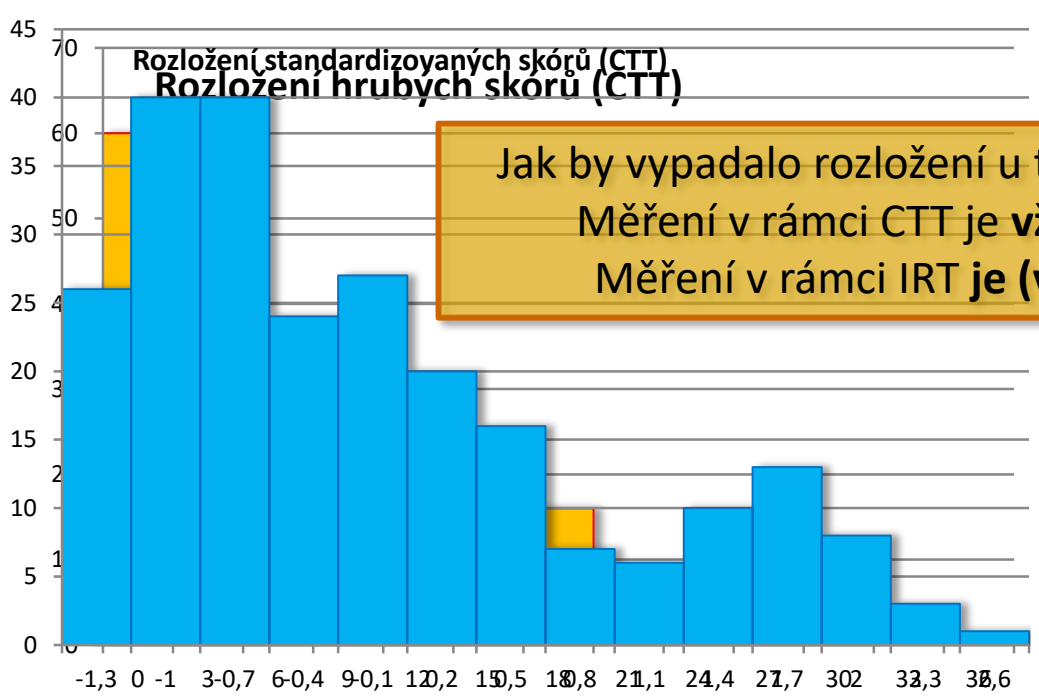
otázka	respondentka 1	respondentka 2
1	spíše nesouhlasím (2)	souhlasím (5)
2	spíše souhlasím (4)	souhlasím (5)
3	souhlasím (5)	nesouhlasím (1)
hrubý skór:	**11 (6,06–11,94)**	**11 (6,06–11,94)**

- CTT: obě dívky 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).

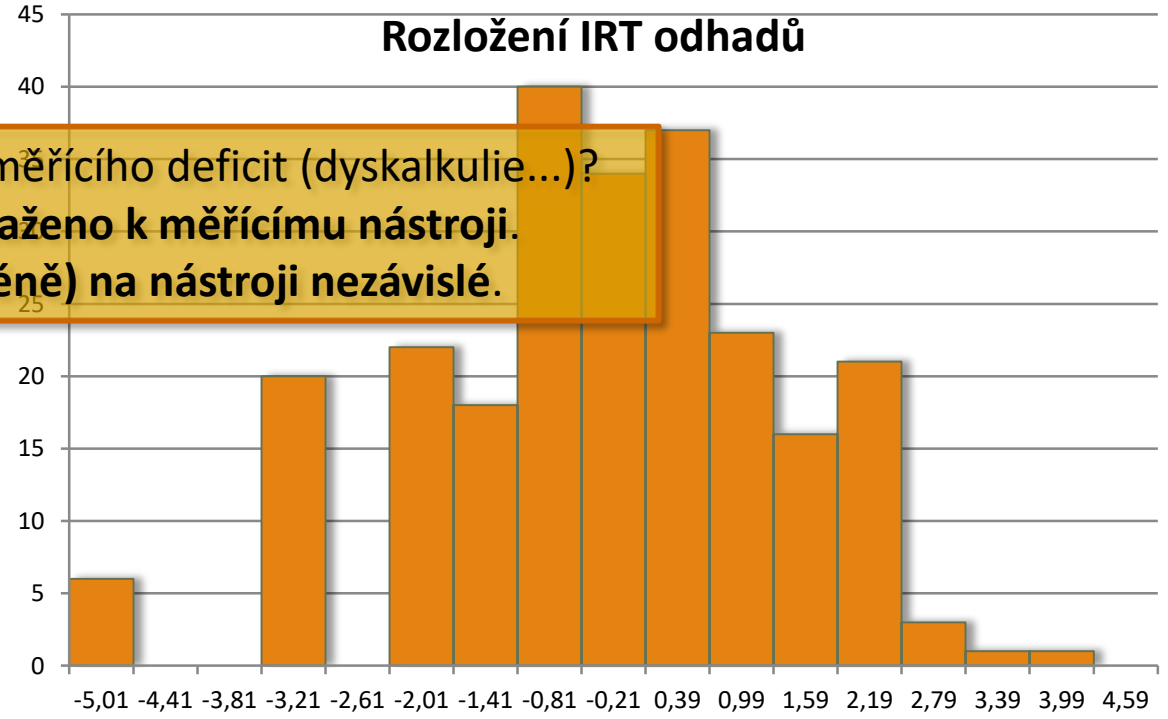
Příklad: Nezávislost měření na nástroji

TIM[3–5]: Test pro identifikaci matematicky nadaných dětí

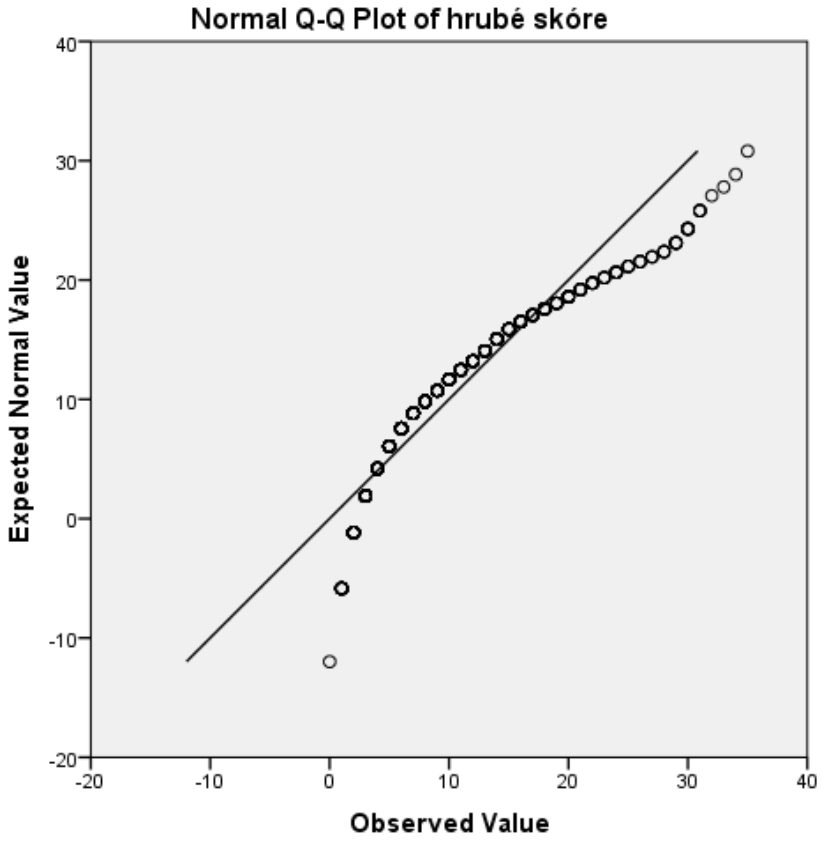
- Test je **velmi obtížný**, aby dobře měřil nadprůměr.
- $r_{xx'}$ = 0,82; M = 8,51; SD = 6,72; min = 0; max = 33
- **Předpoklad:** Rozložení matematických schopností je v populaci normálně rozložené.
- **Závěr:** Jaké budou naměřené skóry?

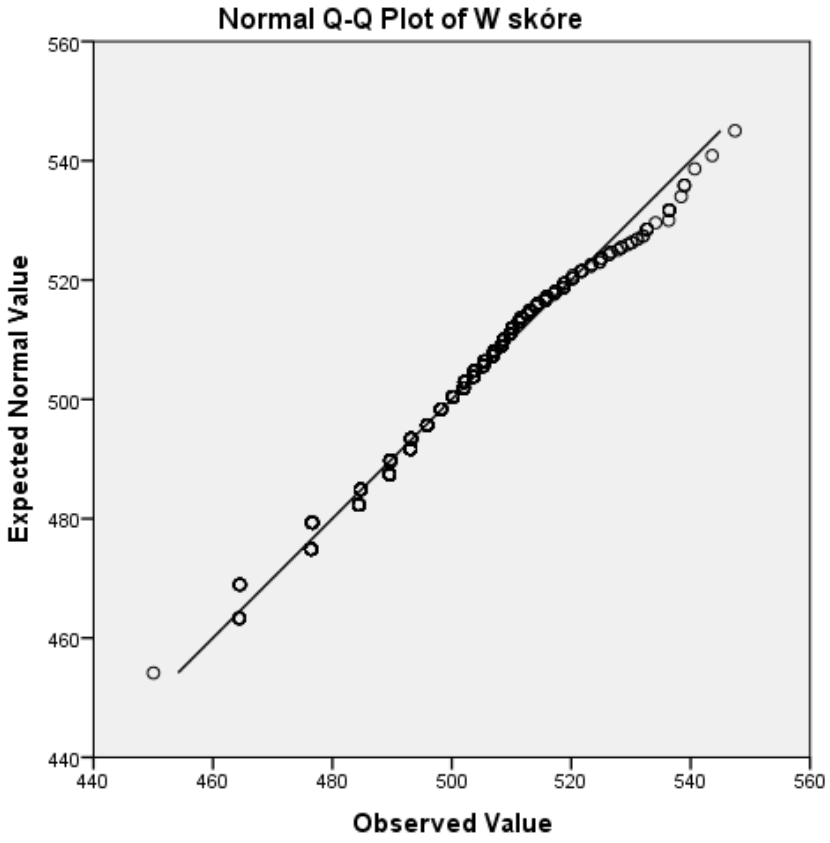


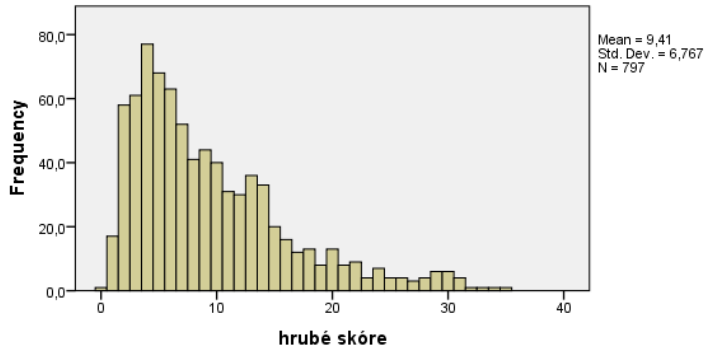
Jak by vypadalo rozložení u testu, měřícího deficit (dyskalkulie...)?
Měření v rámci CTT je **vždy vztaženo k měřícímu nástroji**.
Měření v rámci IRT **je (více méně) na nástroji nezávislé**.



Příklad: Nezávislost měření na nástroji







Kolmogorův-Smirnovův test (MC, p-value)			
ročník	3 (n=243)	4 (n=276)	5 (n=278)
hrubé skóre	,000	,001	,001
W-skóre	,000	,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).

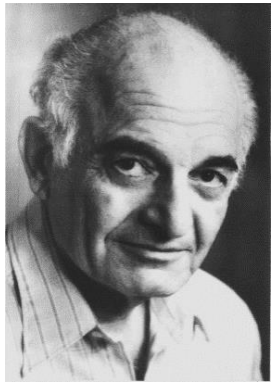
Jde o stochastickou úpravu původně deterministického Guttmanova modelu.

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)

Frederic M. Lord
(1912–2000)



Louis Guttman
(1916–1987)

van der Linden, W. J. (2016). Introduction. In W. J. van der Linden (ed.), *Handbook of Item Response Theory, vol. 1: Models*, pp. 1–10. Boca Raton: CRC Press.

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 matricích.

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

MODELY S LATENTNÍMI PROMĚNNÝMI

Měřený atribut: **Předpokládaný latentní rys**.

Faktorová analýza

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

Teorie odpovědi na položku

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

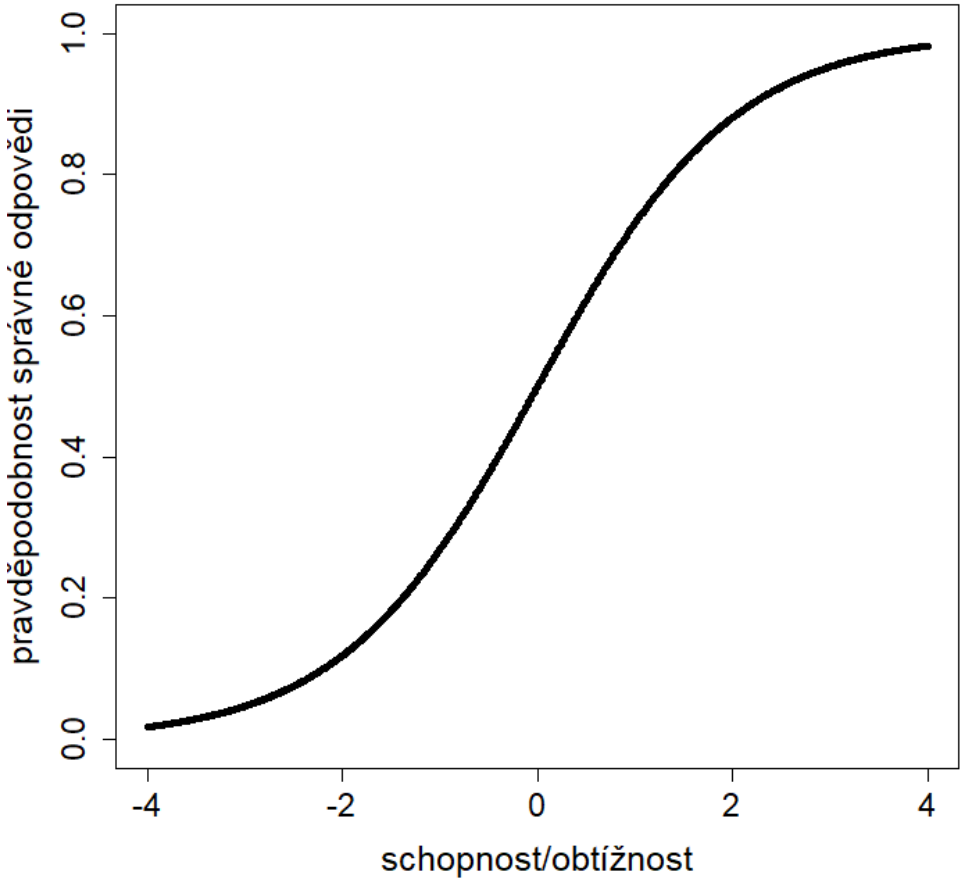
Základy IRT: Charakteristická funkce položky (ICC)

Výkon probanda v položce lze odhadnout pomocí množiny latentních rysů.

- Obtížnost položek a schopnost respondenta na stejné škále.

Item Characteristic Curve (ICC):

- Má (zpravidla) přibližně tvar normální ogivy (kumulativní normální rozdělení).
  - Výjimečně přímo přímo ogiva, tzv. „Φ model“.
- popisuje vztah mezi schopností probanda a jeho výkonem v dané položce;
- Pravděpodobnost správné odpovědi podle parametrů položky a probanda.



Jednoparametrový Raschův model (1PL)

Logistický vztah rysu a odpovědi:

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

Analogicky po úpravě:

$$\ln \frac{P_i(\theta)}{1 - P_i(\theta)} = \theta - b_i$$

- e = Eulerova konstanta
- ln = přirozený logaritmus (se základem e)

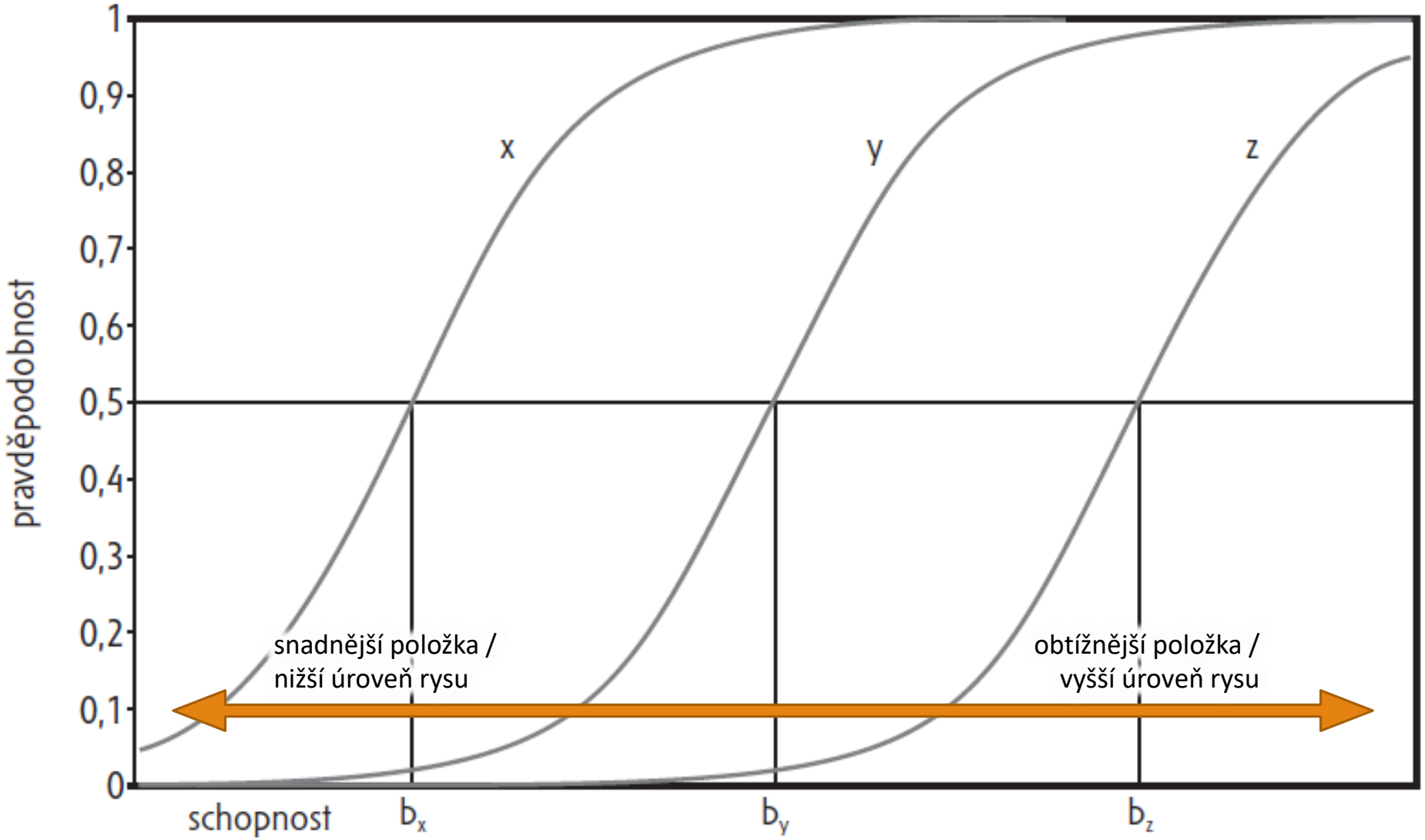
$P_i(\theta)$ je pravděpodobnost správné odpovědi na položku i při schopnosti θ.

- Tato pravděpodobnost Pi se někdy nazývá také „true-score“ respondenta v dané položce (u binárních položek).

Theta (θ) je úroveň schopnosti respondenta

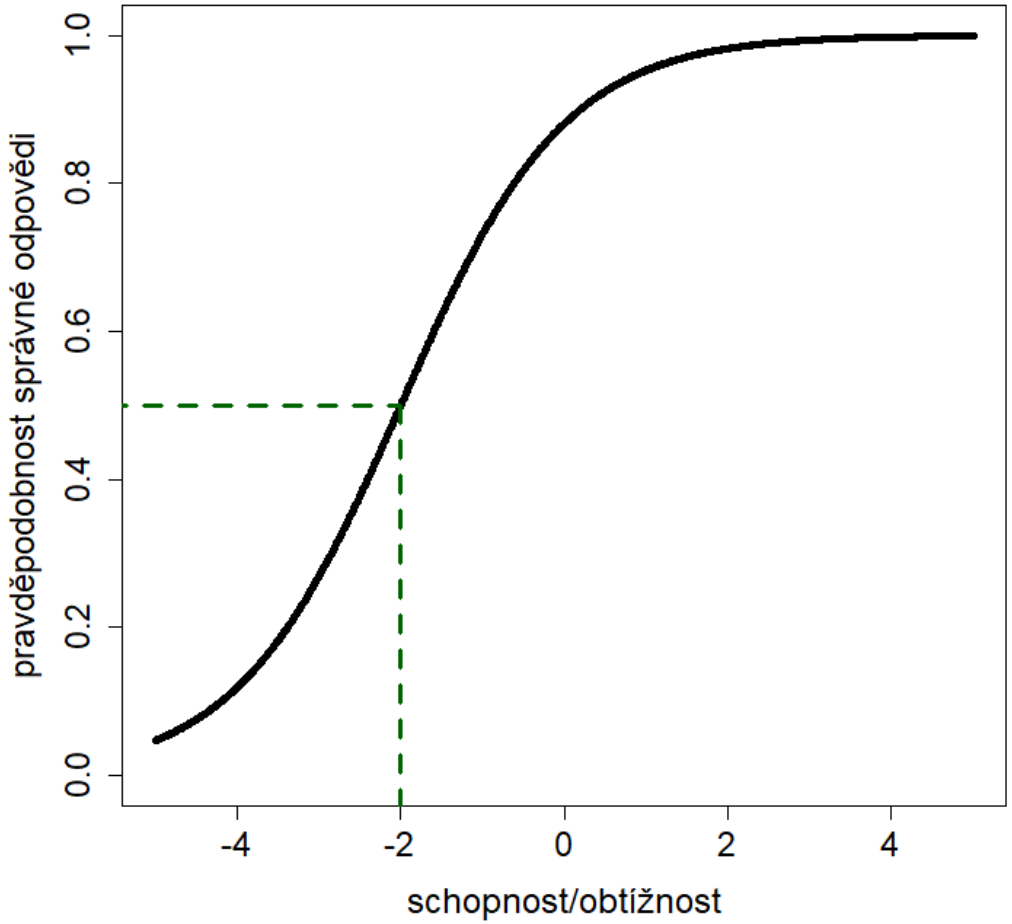
b_i je parametr obtížnosti položky i

- Parametr obtížnosti bi položky i je bod na škále schopnosti, v němž je pravděpodobnost správné odpovědi respondenta j se stejnou mírou schopnosti (θj = bi) na danou položku Pi(θj) = 0,5.



Urbánek, T., Denglerová, D., & Širůček, J. *Psychometrika*. Praha: Portál.

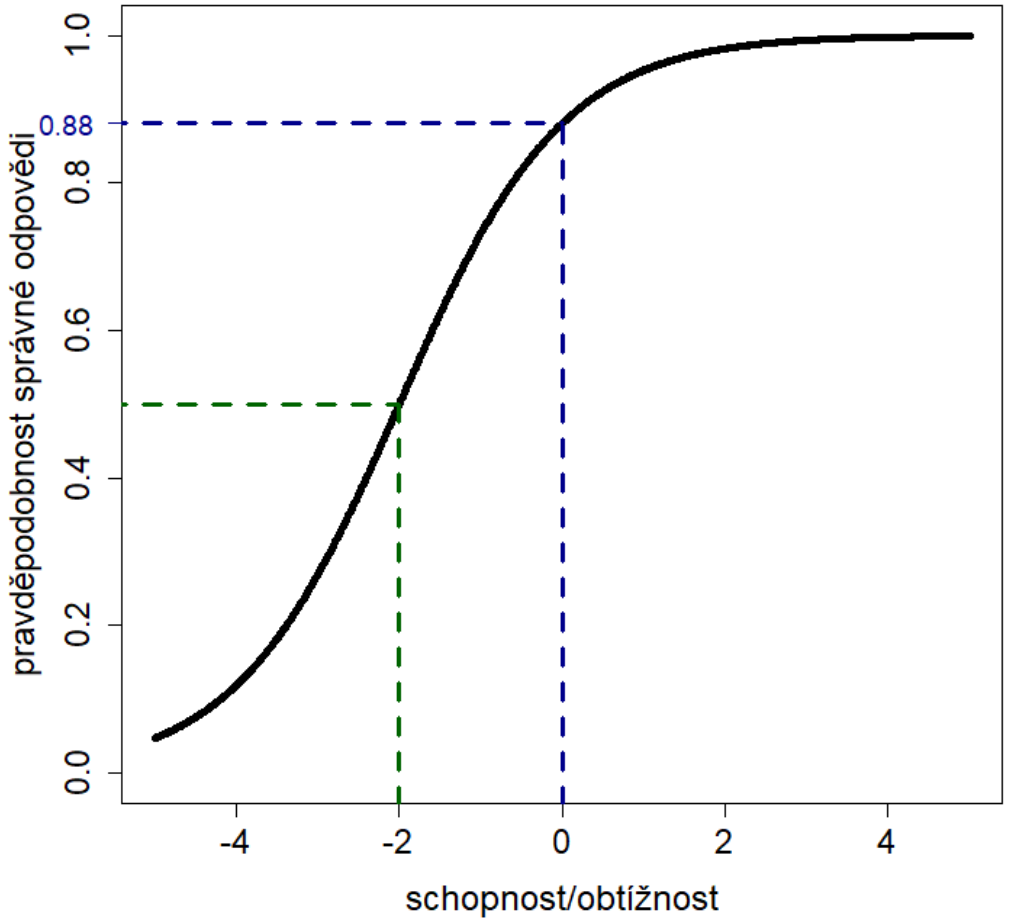
Raschův model (jednoparametrový)



Položka s obtížností $b_i = -2$.

Respondent se schopností θ = -2 má 50 % pravděpodobnost správné odpovědi.

Raschův model (jednoparametrový)



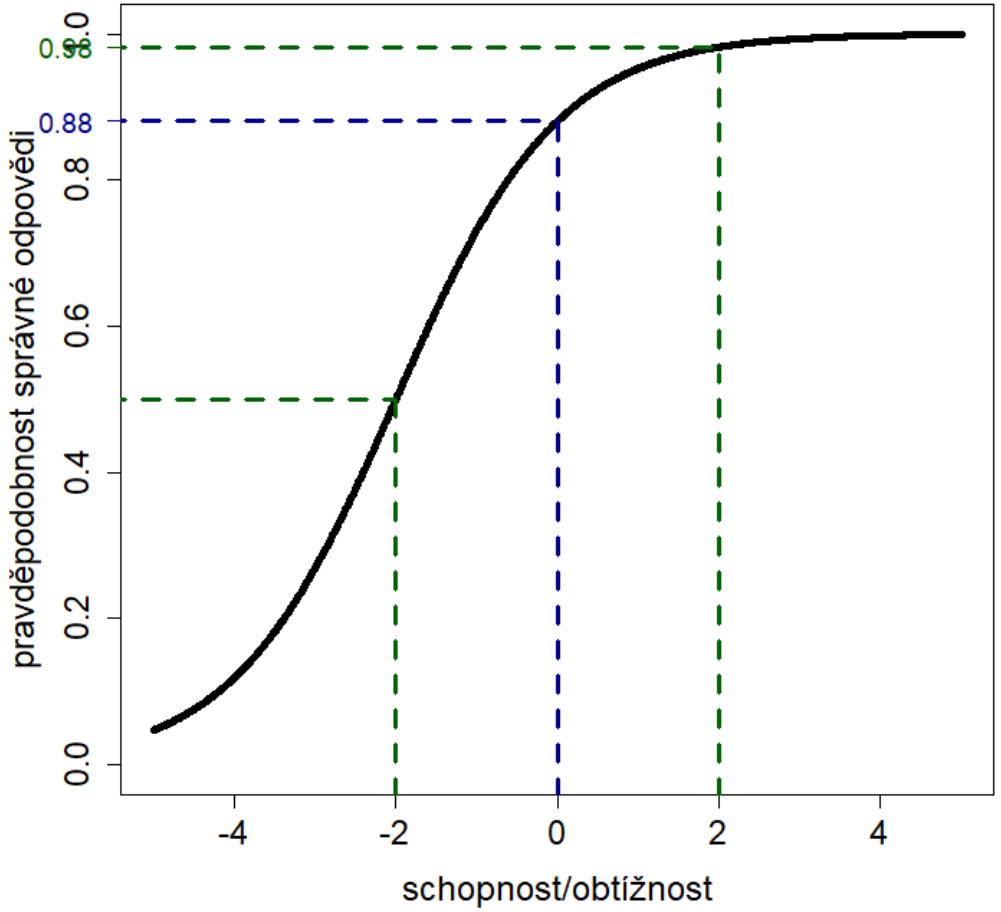
Položka s obtížností $b_i = -2$.

Respondent se schopností θ = -2 má 50 % pravděpodobnost správné odpovědi.

Analogicky respondent s θ=0 odpoví správně s 88% pravděpodobností:

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

Raschův model (jednoparametrový)



Položka s obtížností $b_i = -2$.

Respondent se schopností θ = -2 má 50 % pravděpodobnost správné odpovědi.

Analogicky respondent s θ=0 odpoví správně s 88% pravděpodobností:

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

A respondent s θ=2 → 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.

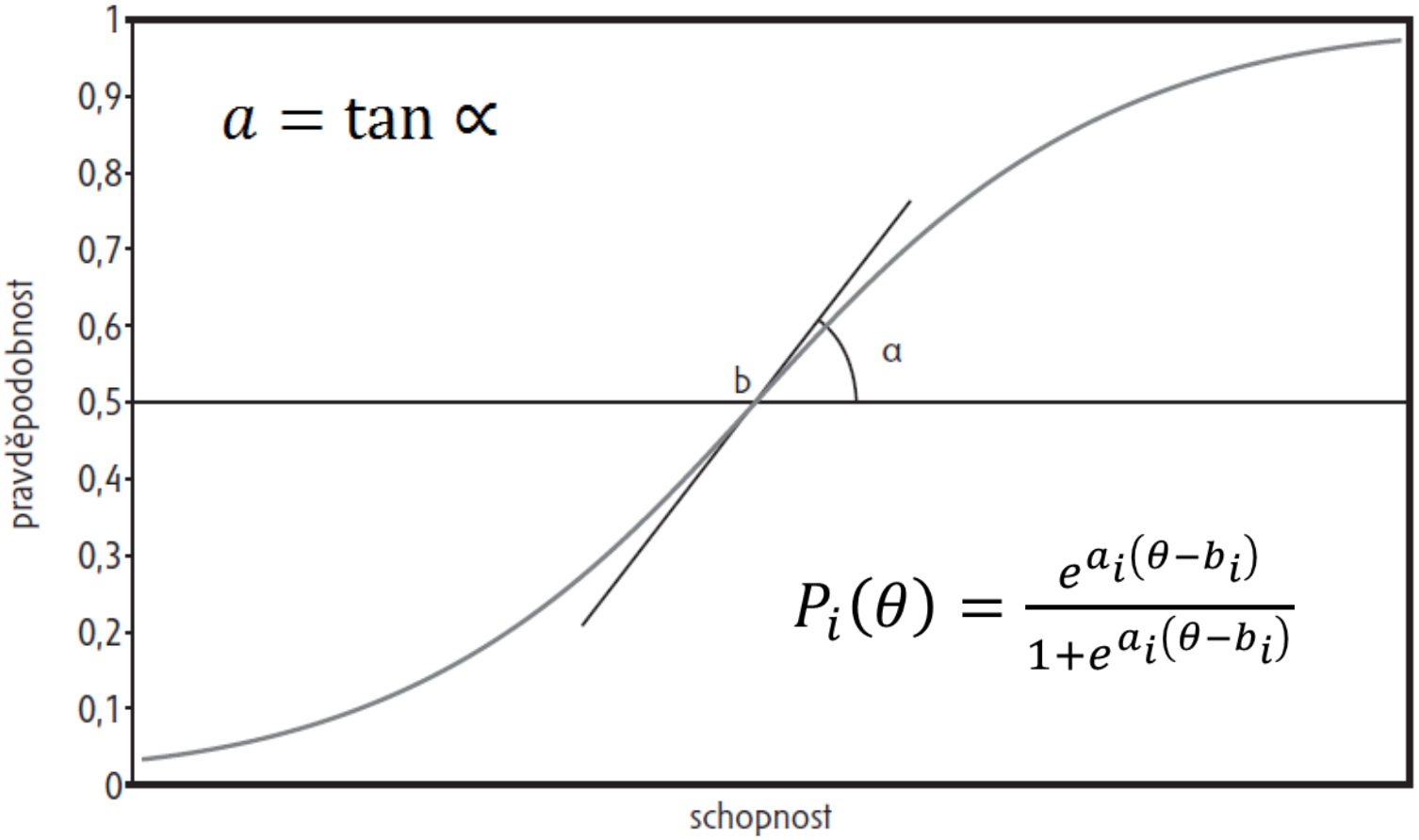
$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

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

- čím je křivka „plošší“, tím méně rozlišuje

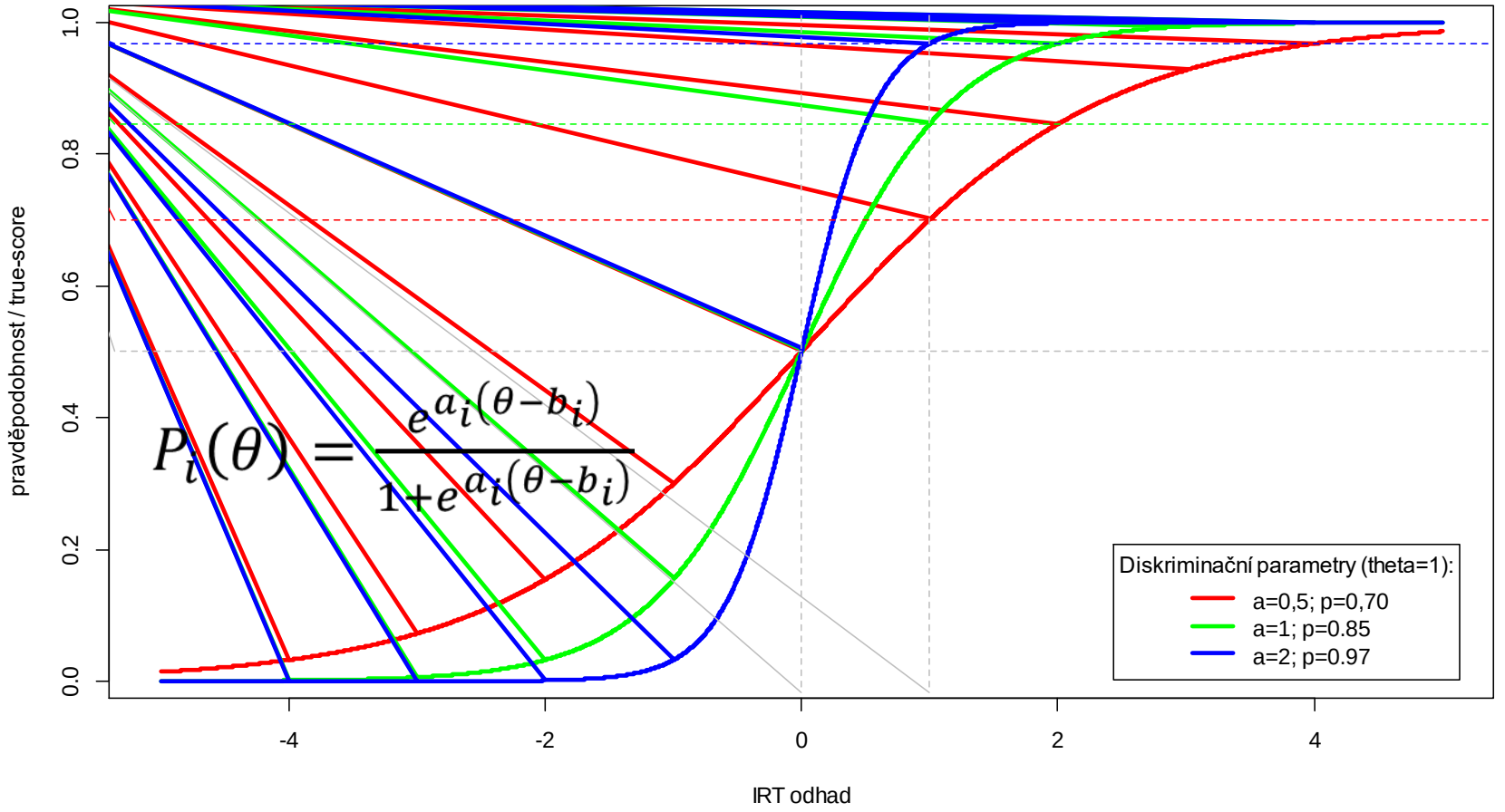
Analogií ve faktorové analýze je faktorový náboj a v CTT položkové analýze korigovaná korelace.

Charakteristická křivka položky 2PL



Urbánek, T., Denglerová, D., & Širůček, J. *Psychometrika*. Praha: Portál.

Charakteristická křivka položky 2PL



Dvě parametrizace 2PL IRT modelu

Tradiční IRT parametrizace

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

Výhody:

- Jednoduchá interpretace parametru obtížnosti.
- Tradiční, dobře známé.

Nevýhody:

- Odlišné od faktorové analýzy, problém s multidimenzionálními modely.

Tzv. slope-intercept parametrizace

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

Výhody:

- Snadná implementace (tzv. Reckaseho) multidimenzionálních modelů, např.

$$P_i(\theta) = \frac{e^{a_{i1}\theta_1 + a_{i2}\theta_2 + \ldots + a_{in}\theta_n + b_i}}{1 + e^{a_{i1}\theta_1 + a_{i2}\theta_2 + \ldots + a_{in}\theta_n + b_i}}$$

  - Kde θ_1–θ_n je množina *n* latentních rysů a
  - b_i je celková „obtížnost“ položky.
- Určité výhody při estimaci, univerzálnější.
  - MIRT package v R má jako default.

Tříparametrový model (3PL)

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

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

- $\boldsymbol{c_i}$ je parametr (pseudo)uhádnutelnosti pro položku *i*.

V multiple-choice testech lze nahradit Bockovým NRM modelem.

- NRM je nejvíce obecný model, jehož specifikací lze nahradit (téměř) cokoliv.

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.

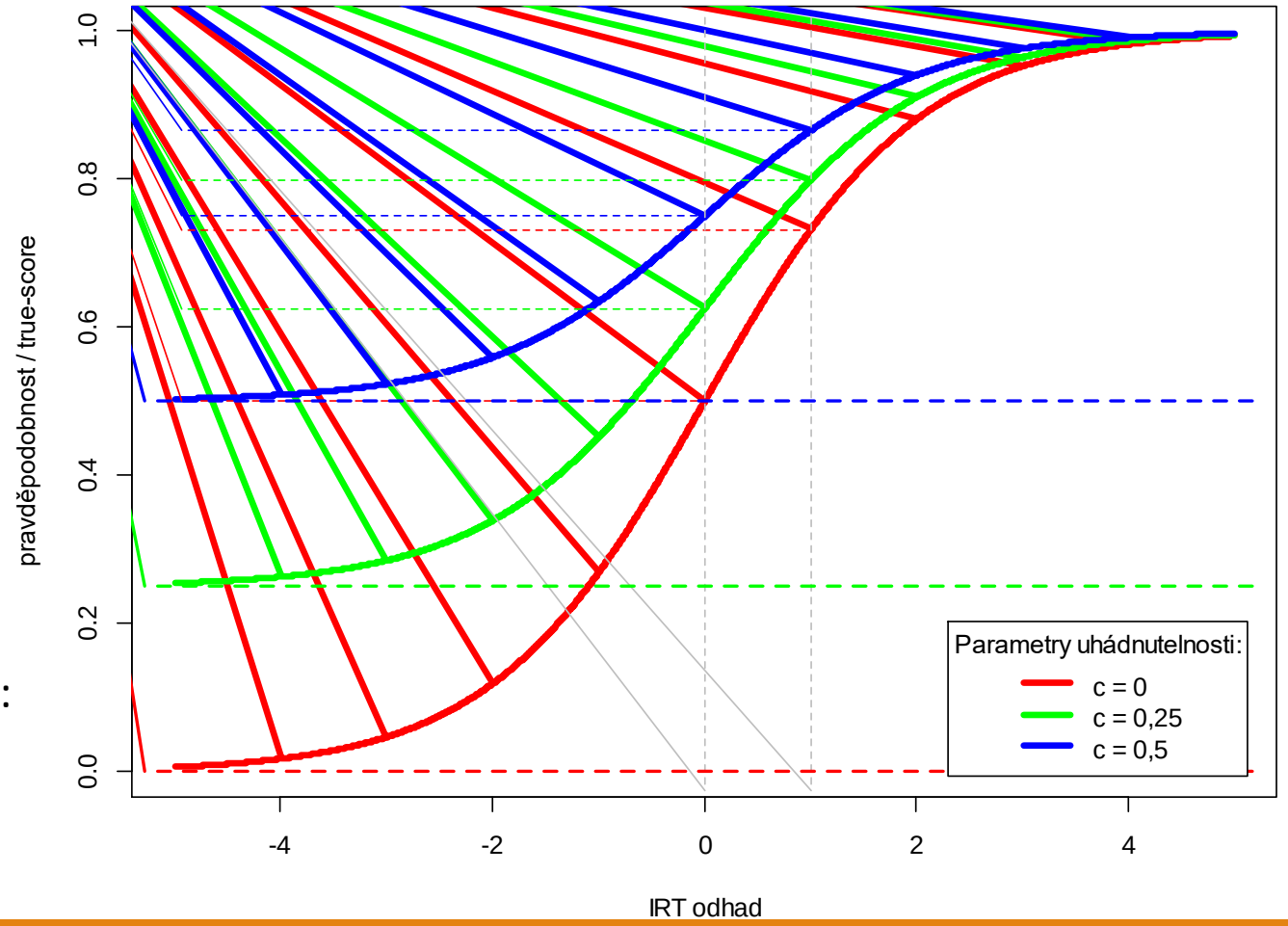
Charakteristické křivky položek 3PL

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

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.

- Např. „projektivní hypotéza“ u TAT (Ťápal, unpublished manuscript☺).

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 %.

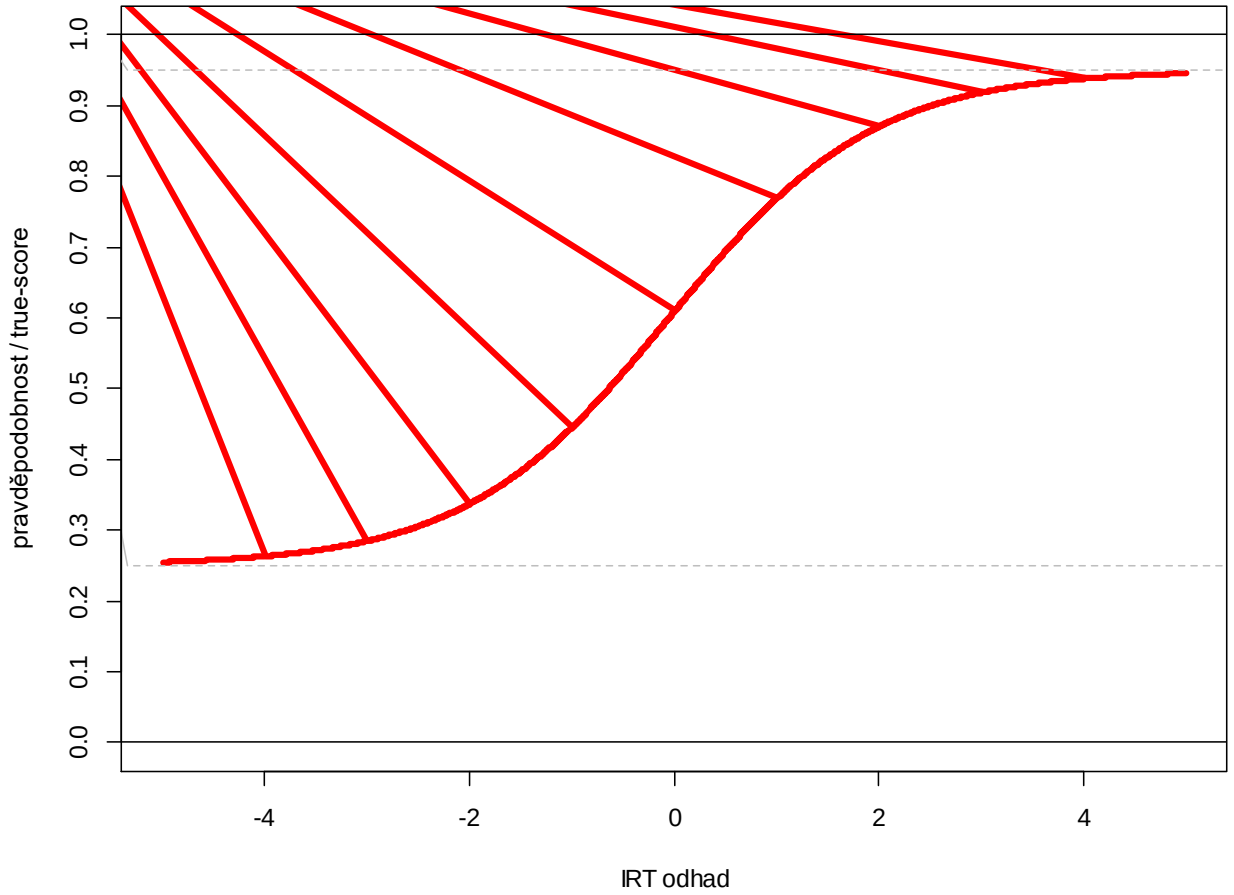
$$P_i(\theta) = c_i + (d_i - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

- 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

jednoparametrový model
- pouze parametr obtížnosti položky b_i

dvouparametrový model
- přidává diskriminační parametr a_i

tříparametrový model
- přidává parametr pseudo-uhádnutelnosti c_i

Ostatní symboly:
- schopnost respondenta: θ
- pravděpodobnost správné odp.: P_i
- i – číslo položky

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

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

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

- 4PL: $d_i=1$ → 3PL
- 3PL: $c_i=0$ → 2PL
- 2PL: $a_i=1$ (nebo $a_i=a$) → 1PL

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).

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:

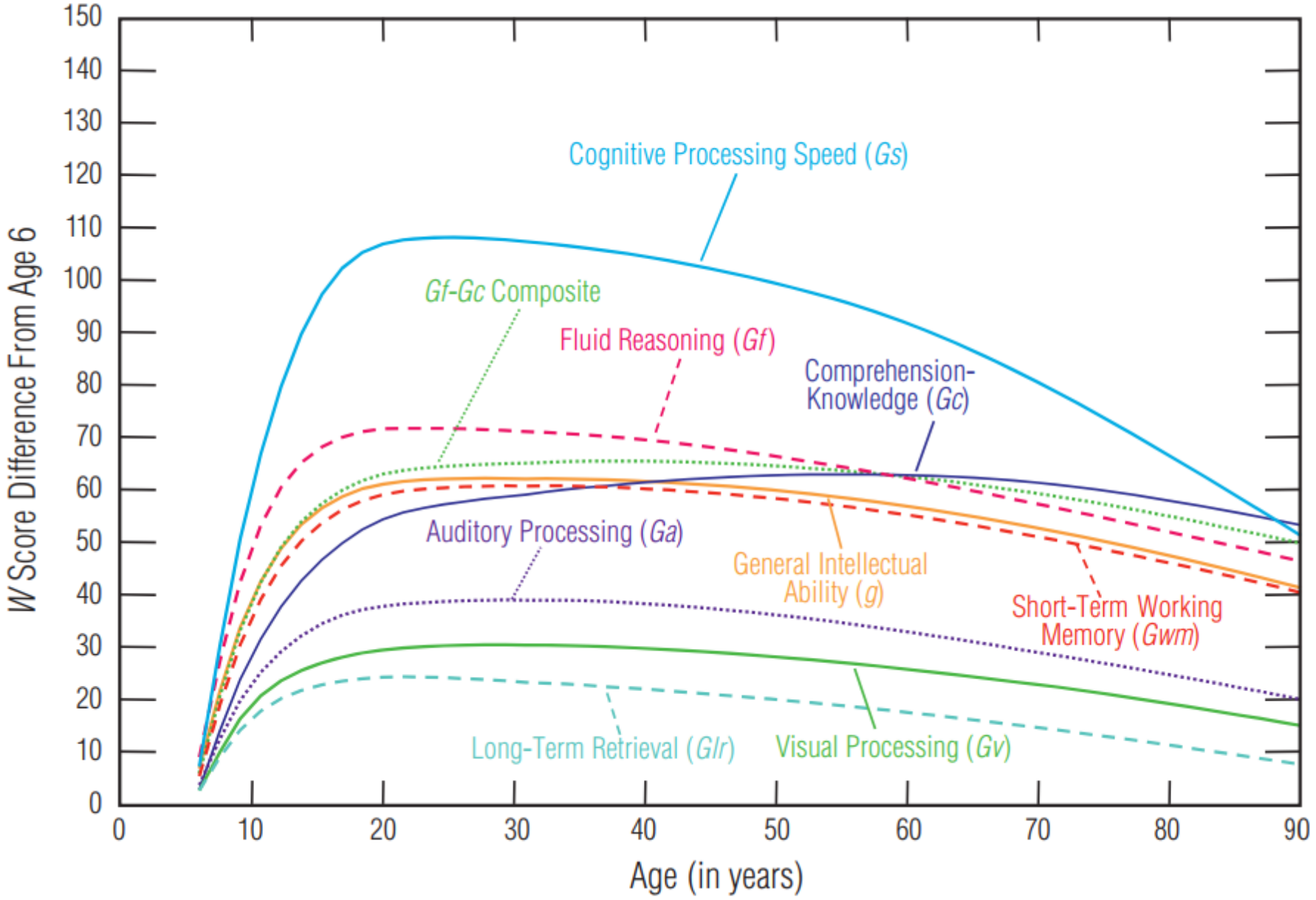
- Z dopředu „nakalibrovaných“ položek usuzujeme na míru rysu, kterou pak převedeme na standardní skóry.

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.

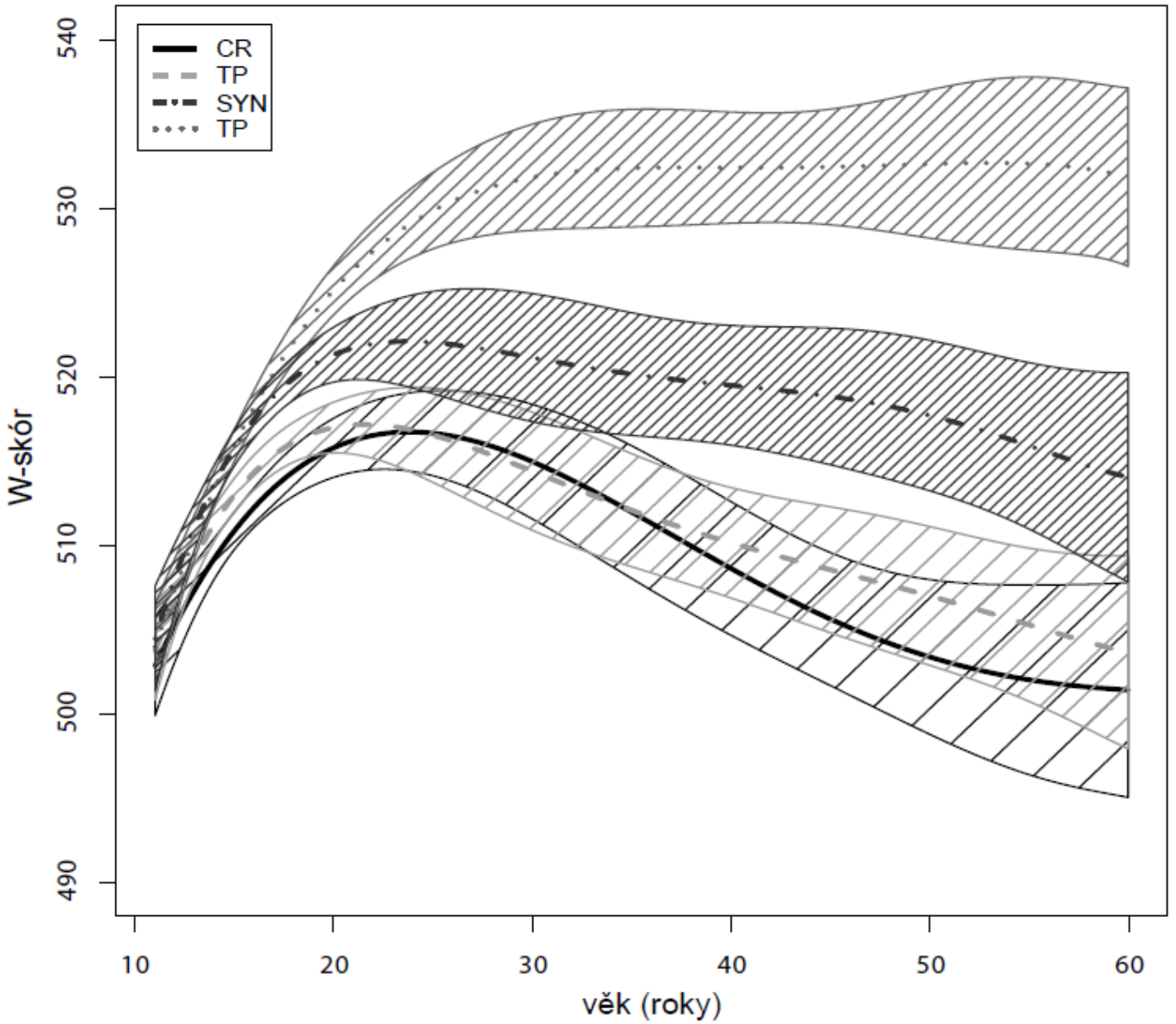
Ve vícePL IRT modelech problematické (nestejná „škála“).



McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical Manual. Woodcock Johnson IV.* Rolling Meadows, IL: Riverside.

Krátký inteligenční test (KIT)

Srovnání vývojových křivek použito jako důkaz konstruktové validity.



Cígler, H. (2017). *Měření matematických schopností.* Brno: Masarykova univerzita.

Charakteristická křivka testu (TCC)

Výhodou Raschova modelu je fakt, že každému hrubému skóre odpovídá právě jeden odhad latentního skóre.

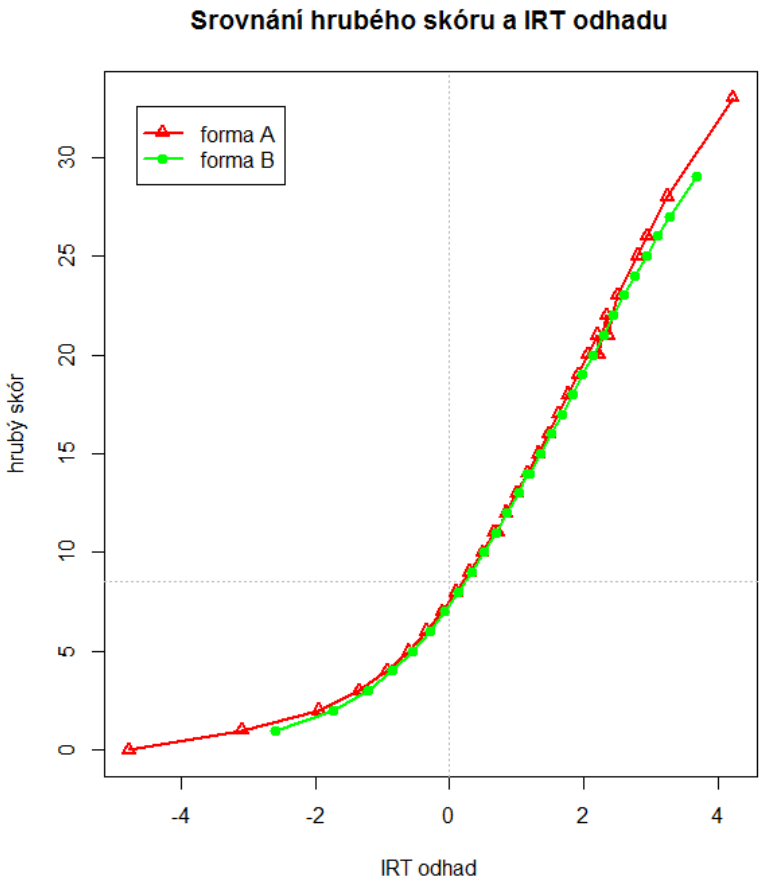
Lze proto definovat Test Characteristic Curve (TCC):

$$TCC(\theta) = \sum_{i=1}^{n} ICC_i(\theta)$$

- Očekávaný HS podle míry latentního rysu (odhad TS v CTT).

Ve 2PL není vztah jednoznačný.

- Diskriminační parametr přikládá jinou „váhu“ každé položce, která má proto jiný vliv na odhad latentního skóre.
- Každému HS odpovídá konečný počet odhadů latentních rysů podle konkrétních správných položek.
- TCC je pak „průměrem“ všech možných křivek (tou nejpravděpodobnější).



Cígler, H., Jabůrek, M., Straka, O., & Portešová, Š. (2017). *Psychometrická analýza TIM3–5 – 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

Předpoklady IRT

Latentní rys existuje a jde o spojitou intervalovou proměnnou.

- Často navíc normálně rozloženou (závisí na estimátoru).
- Ale existují i diskrétní IRT modely, analýza latentích tříd (LCA) atd.

Lokální nezávislost položek.

- Veškeré souvislosti položek lze vysvětlit výhradně modelovanými latentními rysy.
  - Tzn. parciání vztah položek po kontrole úrovně latentního rysu je nulový.
- V případě jediného rysu: Jednodimenzionalita.
  - Na rozdíl od CFA nelze modelovat reziduální kovariance, je nutné zavést specifické faktory.

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.

Přesnost měření v IRT

2. ČÁST PŘEDNÁŠKY

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).

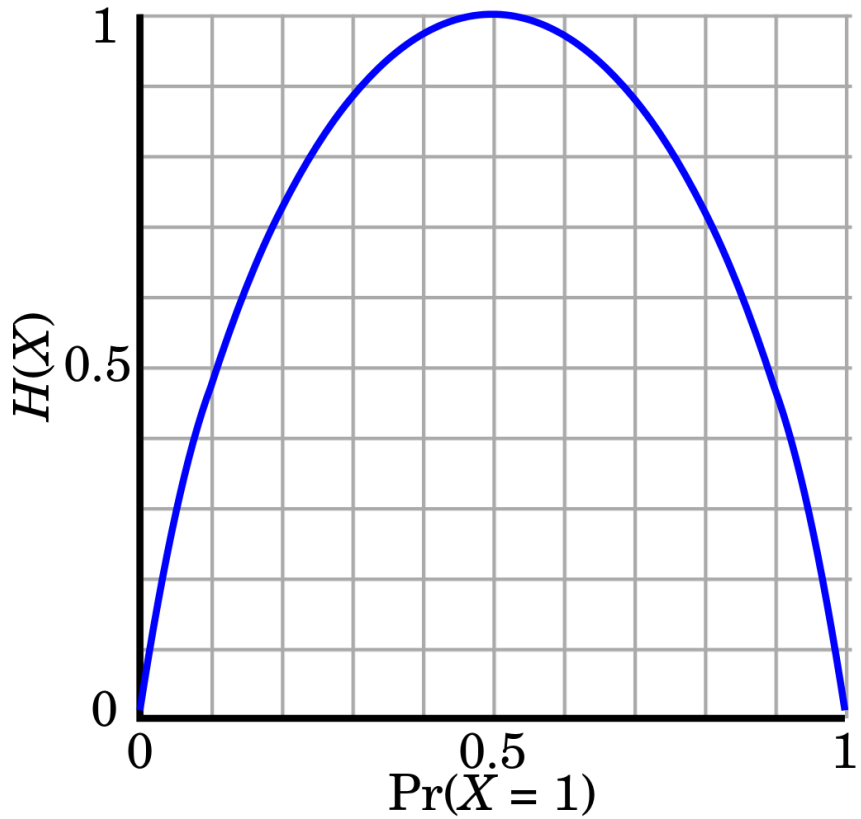
Odbočka: Informační teorie

Množství informace nesené (nejen) diskrétní proměnnou souvisí s obtížností předpovědět daný jev.

- Jinými slovy: Čím nižší souvislost má apriorní očekávání s pozorováním, tím více informace.
  - Př.: Pokud jev může nabývat hodnot 0/1, ale reálně nabývá vždy 1, pozorovaná odpověď nenese žádnou informaci, protože tu 1 očekáváme.

Př.: Lidé odpovídají ano/ne na různé otázky.

- *Ignác vždy odpoví „ano“ nehledě na otázku.*
- *Ignácie se zamyslí a odpoví podle otázky.*
- ***Odpovědi Ignácie nesou více informace, než odpovědi Ignáce.***



Informace Bernoulliho pokusu podle pravděpodobnosti úspěchu.

Informační funkce položky (IIF, IIC)

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

- Pro každou úroveň schopnosti θ jiná.

Binární položky:

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

- $P_i(\theta)$ = Informační funkce položky (pravděpodobnost správné odpovědi při úrovni schopnosti θ).
- P_i' = první derivace této funkce.
- $1 - P_i(\theta)$ = pravděpodobnost jiné než správné odpovědi.

Pro 1PL model platí

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

a lze tedy zjednodušit:

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

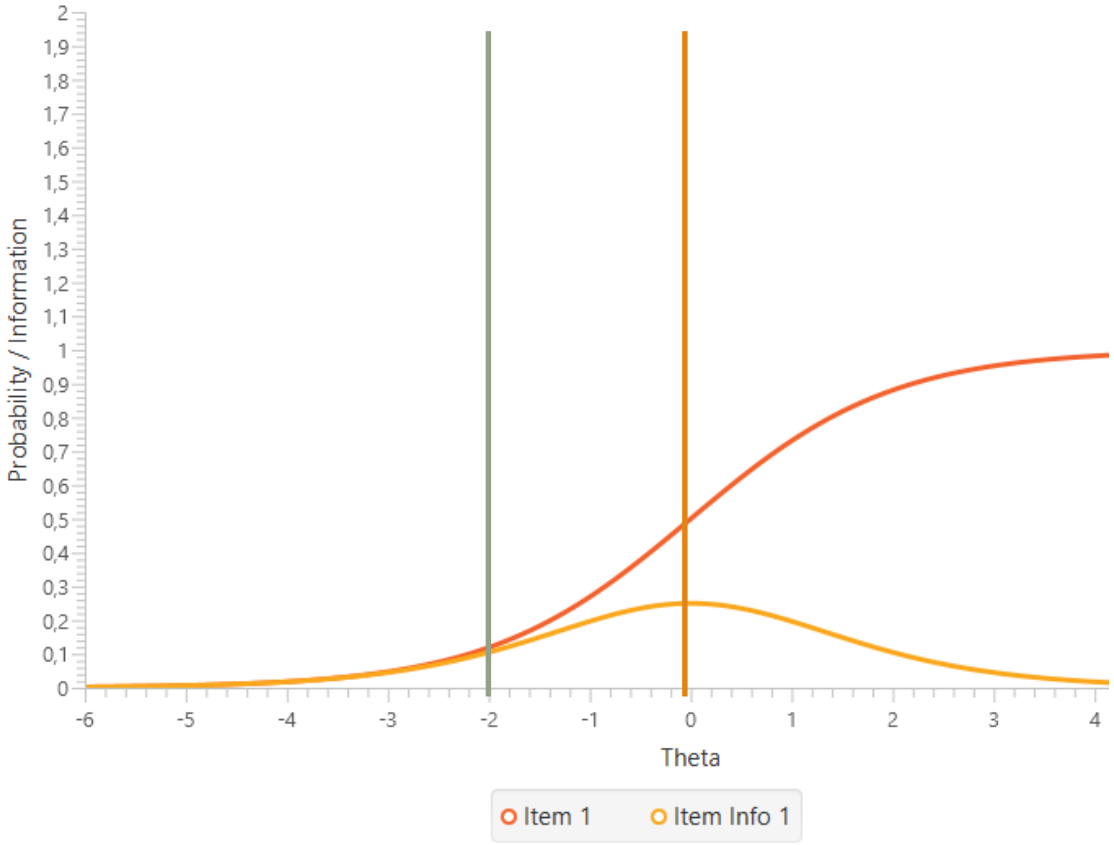
- V Raschově binárním modelu mají všechny položky stejný průběh funkce (diskriminační parametr), liší se jen obtížností.
- Maximum je tedy vždy 0,5 · 0,5 = 0,25.

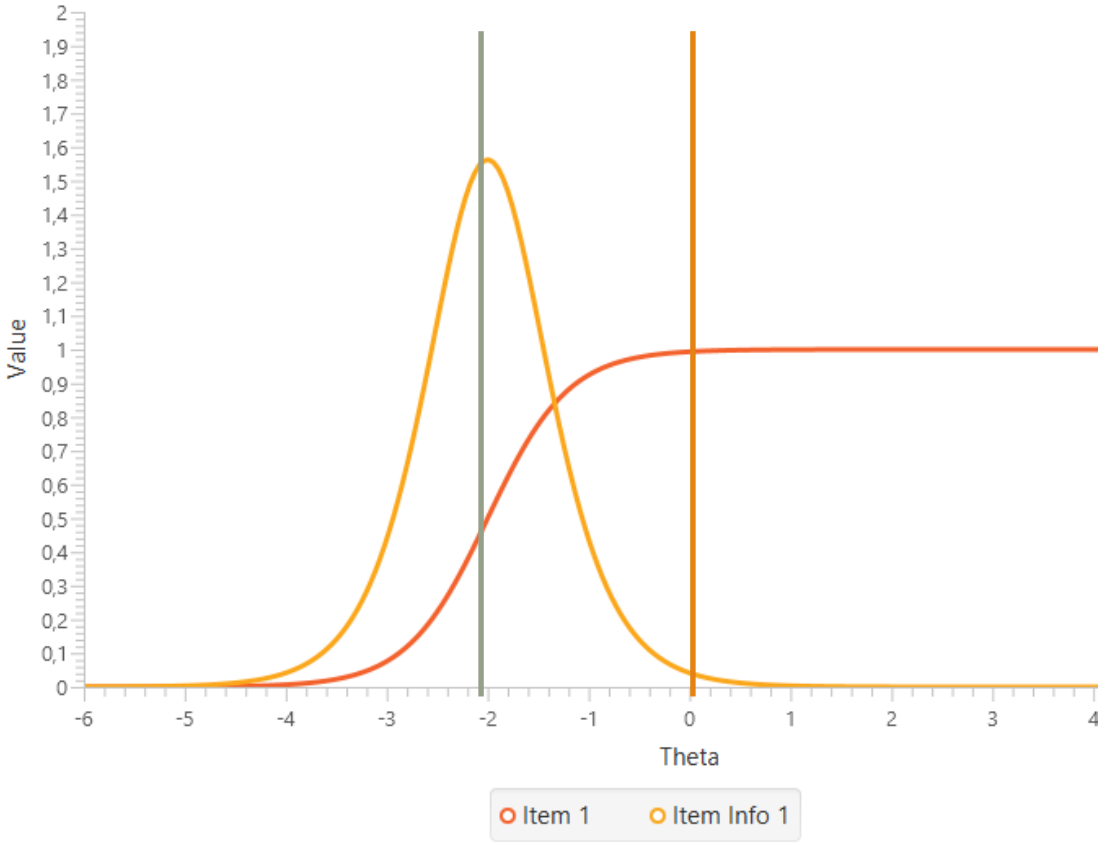
Item Information Function/Curve

- (IIF/IIC)

Informační funkce položky

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

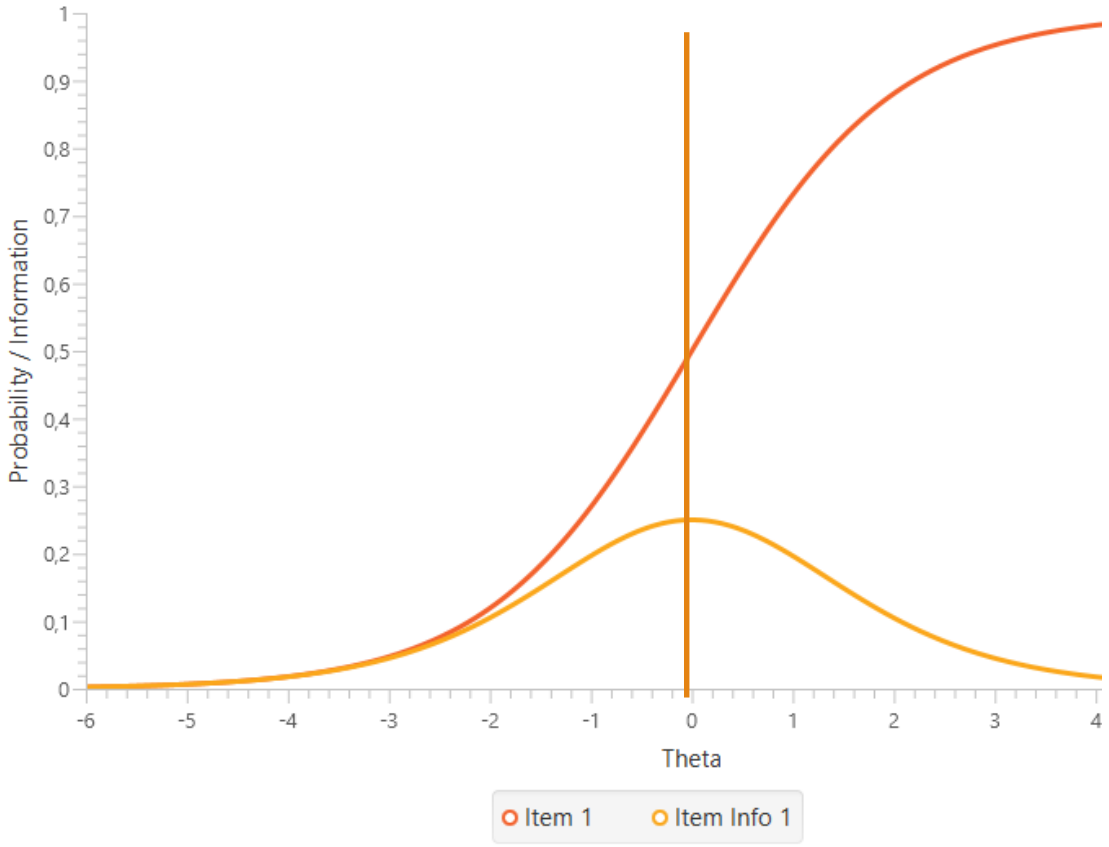


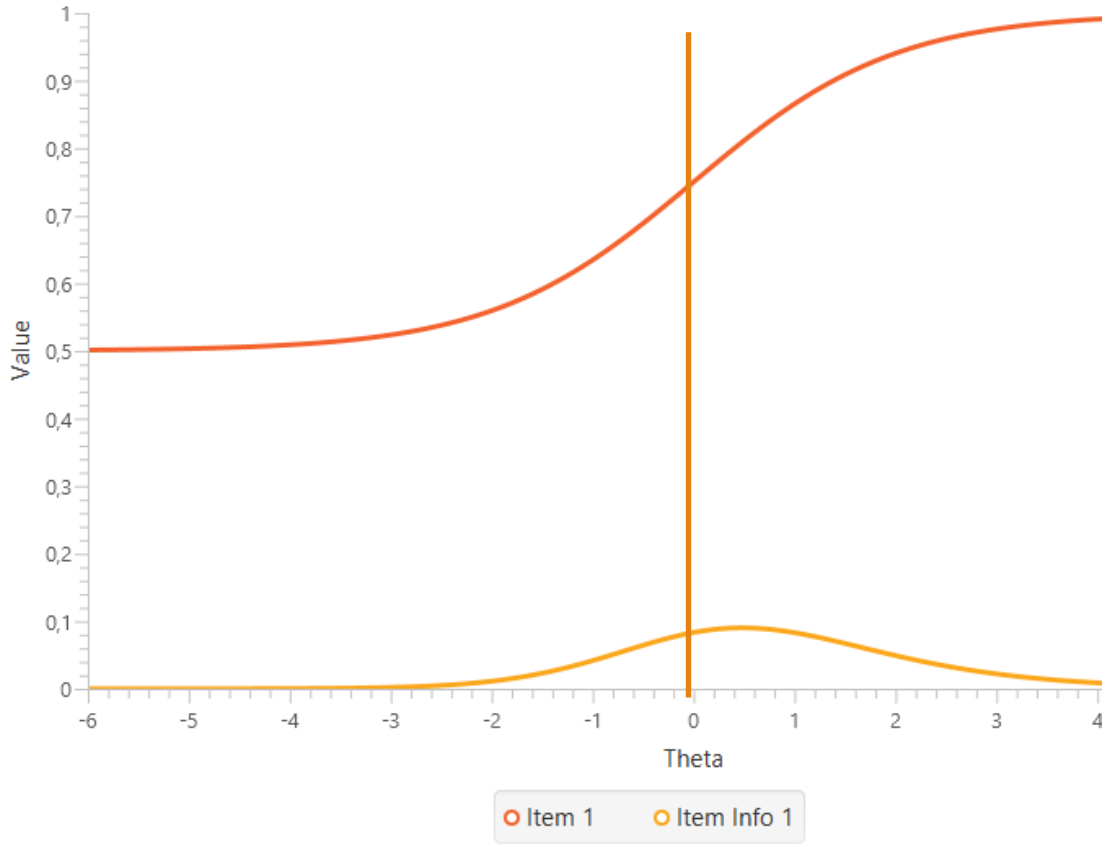


https://itemanalysis.com/irt-illustrator/

Informační funkce položky

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





https://itemanalysis.com/irt-illustrator/ (Pozor, osa y má odlišné měřítko od předchozího snímku.)

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 θ a závisí dále na:

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

Informační funkce testu a standardní chyba měření

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

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

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

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

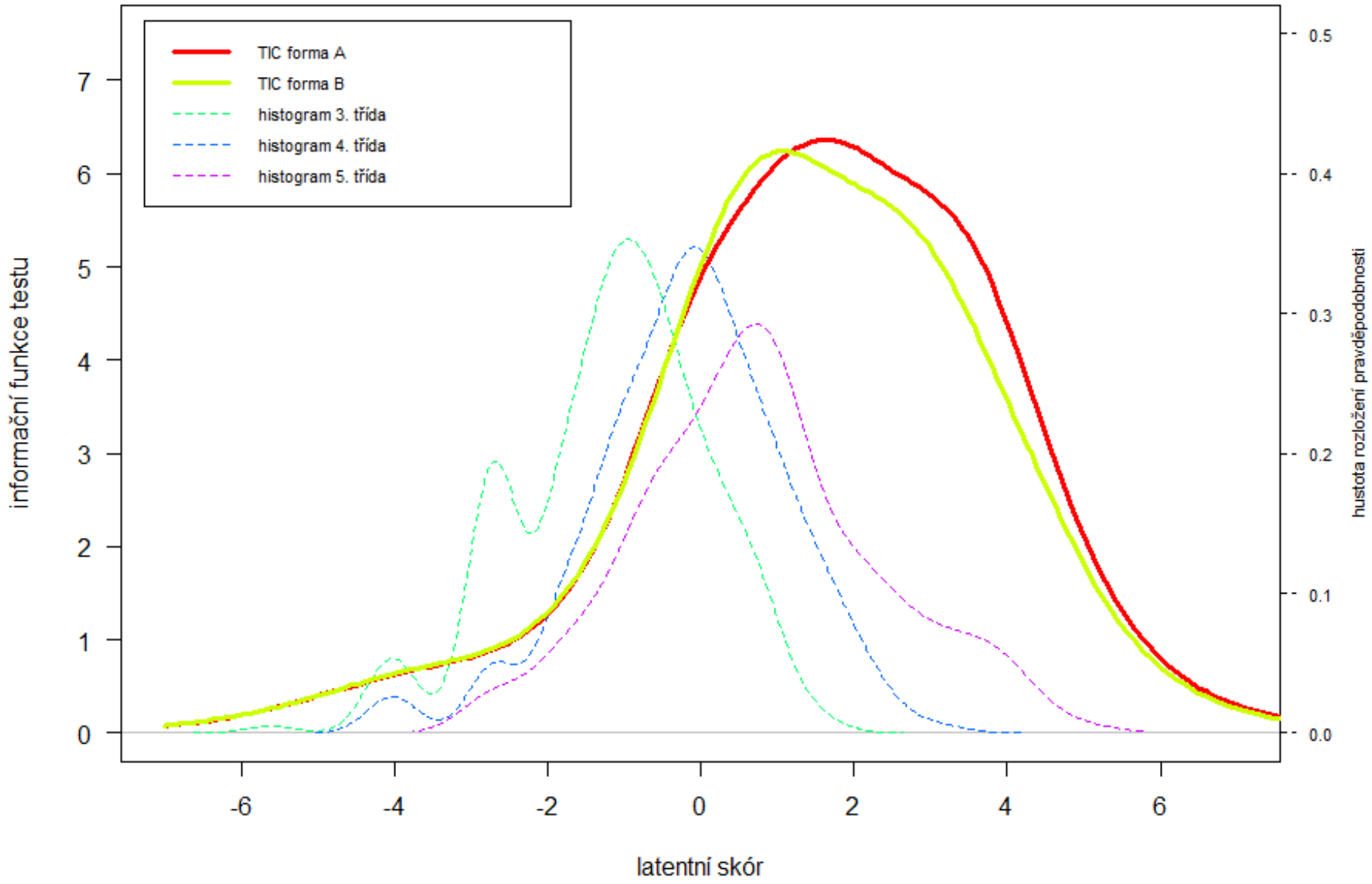
- (tedy čím vyšší informační funkce, tím přesnější měření/menší chyba měření)

Interval spolehlivosti potom získáme obdobně jako v CTT:

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

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

Informační funkce testu a chyba měření



Cígler, H., Jabůrek, M., Straka, O., & Portešová, Š. (2017). *Psychometrická analýza TIM3–5 – 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

Odhad reliability

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 σ_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:

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

Komplikace: Záleží na estimátoru.

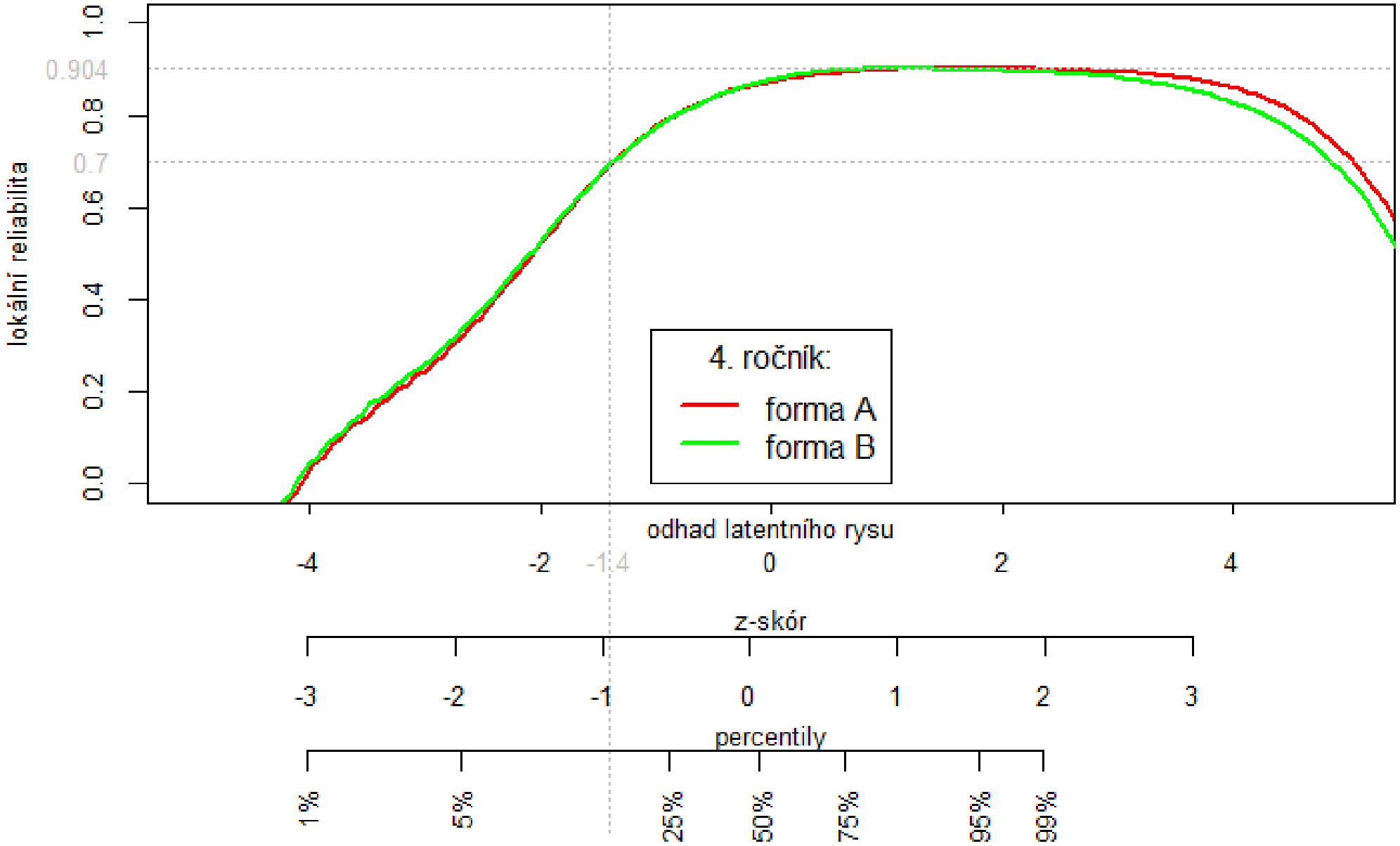
- CML, MML a resp. EAP, MAP odhady pracují s odhadem latentního rysu (regrese k průměru) a tedy je odhadován nikoliv σ_X^2, ale přímo σ_T^2.

A tedy: $r_{xx'} = 1 - \frac{\sigma_T^2+RMSE^2}{\sigma_T^2}$

Lokální reliabilita

Pro reliabilitu měření konkrétního respondenta nebo konkrétní skupiny dosadíme za σ_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.


lokální reliabilita
0.0
0.2
0.4
0.6
0.7
0.8
0.904
1.0
4. ročník:
forma A
forma B
odhad latentního rysu
-4
-2
-1.4
0
2
4
z-skór
-3
-2
-1
0
1
2
3
percentily
1%
5%
25%
50%
75%
95%
99%

Odhad reliability

Lze spočítat pro osoby i pro položky.

Reliabilita osob záleží na:

- rozptylu probandů;
- délce testu;
- počtu kategorií každé položky (zvyšuje se většinou cca do 6, vyšší
  počet totiž zpravidla zhoršuje věrohodnost modelu a fit položky);
- „sample-item targeting“ – jsou položky vhodně těžké pro daný vzorek?
- Je naopak nezávislá na počtu osob.
- Kritéria stejná jako v CTT.

Reliabilita položek závisí na:

- rozptylu obtížnosti položek;
- počtu probandů;
- „item-sample targeting“.
- Je nezávislá na délce testu.
- Odpověď na otázku „jak přesně jsme odhadli obtížnosti položek“?
- Kritéria výrazně přísnější... u běžných testů chceme alespoň 0,99.

Shoda modelu s daty

NA ÚROVNI CELÉHO MODELU

Odpovídají pozorovaná data IRT modelu?

Obdobný přístup jako v konfirmační faktorové analýze

- χ^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

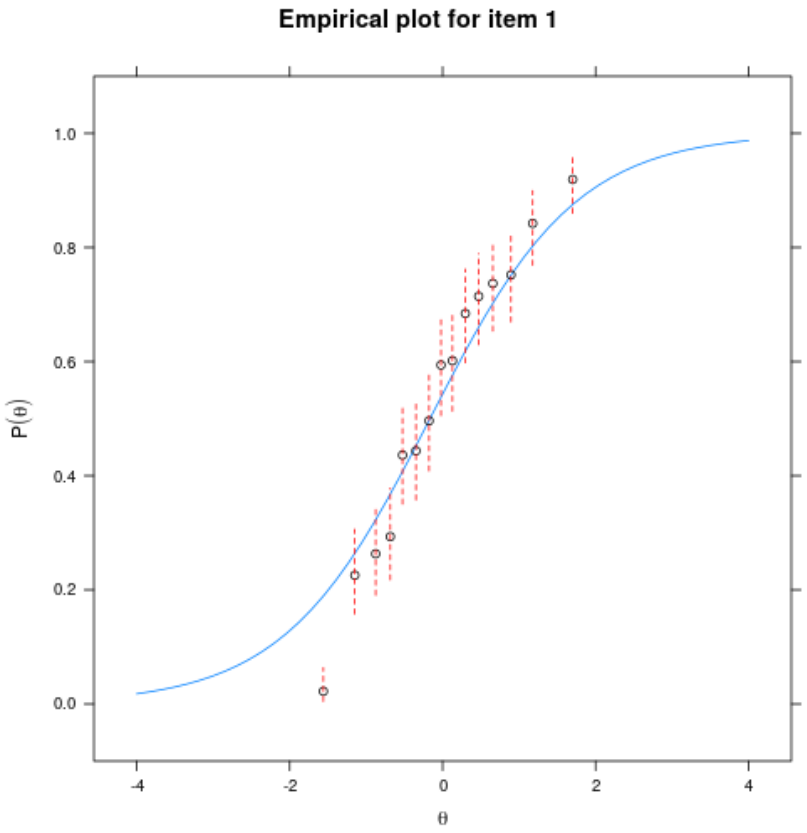
NA ÚROVNI POLOŽKY/RESPONDENTA

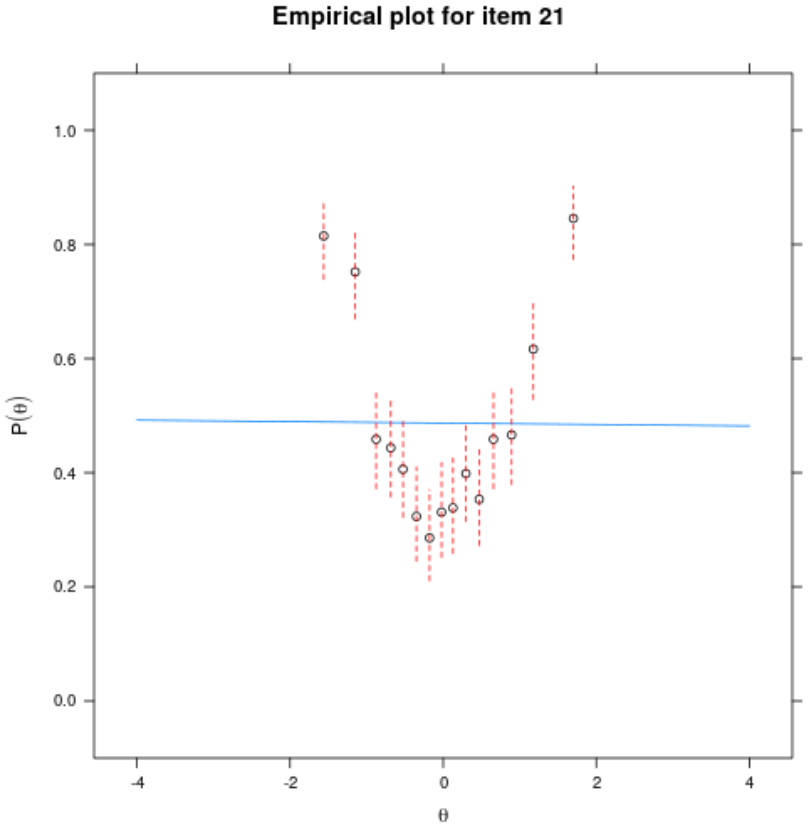
Na kolik dobře odpovídají pozorované odpovědi 1 respondenta nebo odpovědi na 1 položku zvolenému IRT modelu?

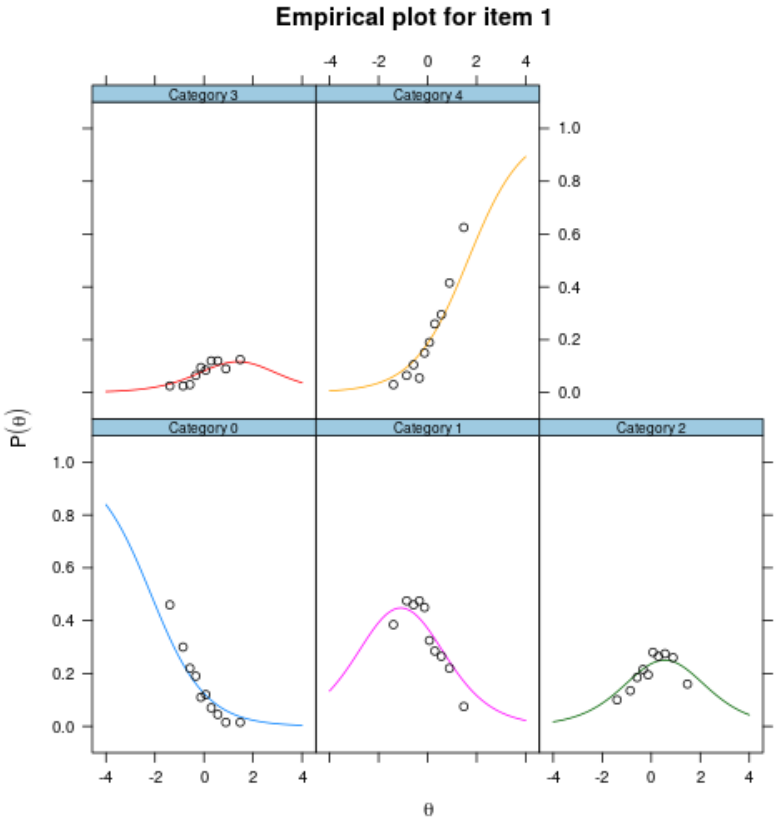
Celá řada indexů.

- **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 položky s daty (item fit)

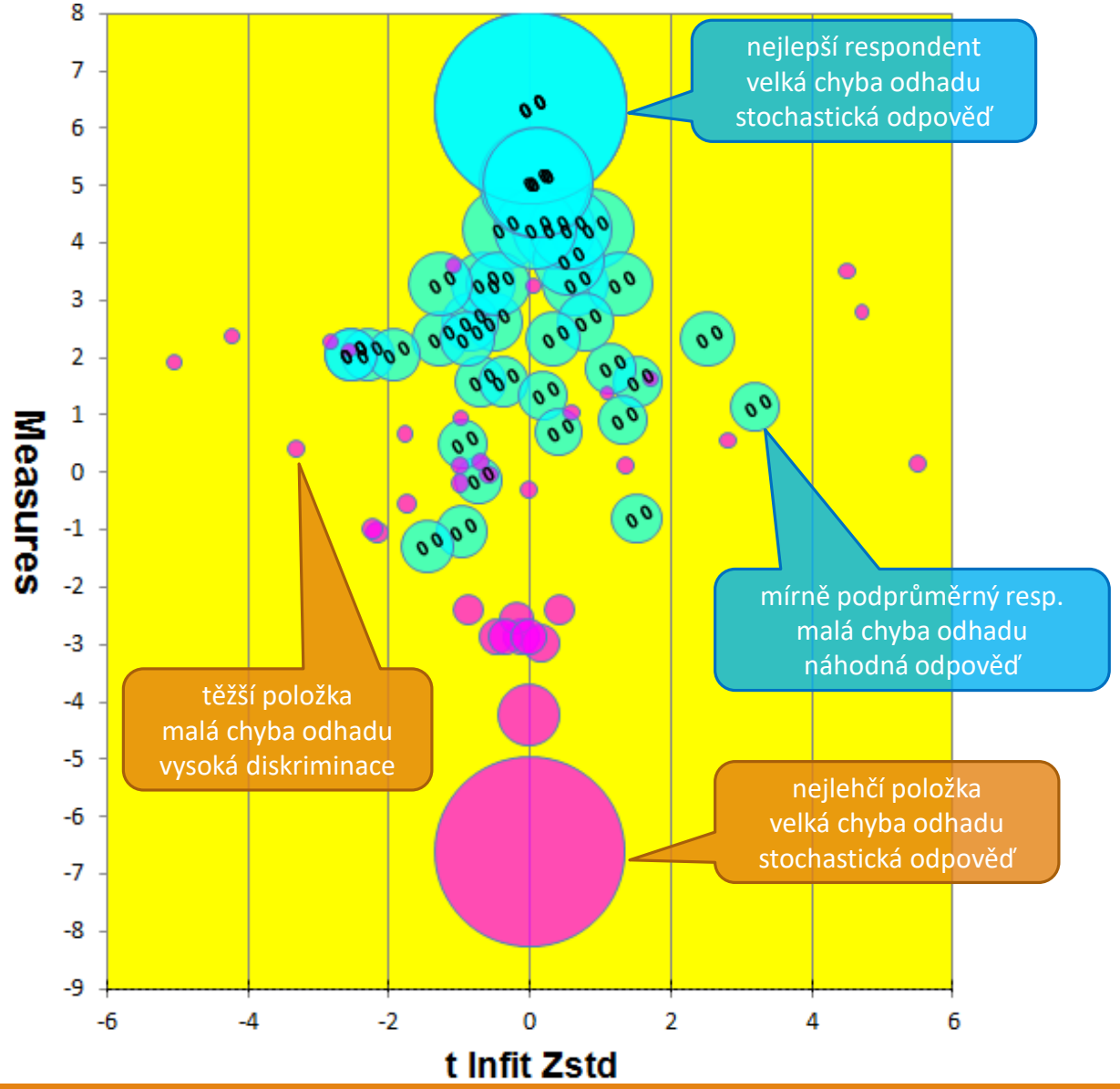


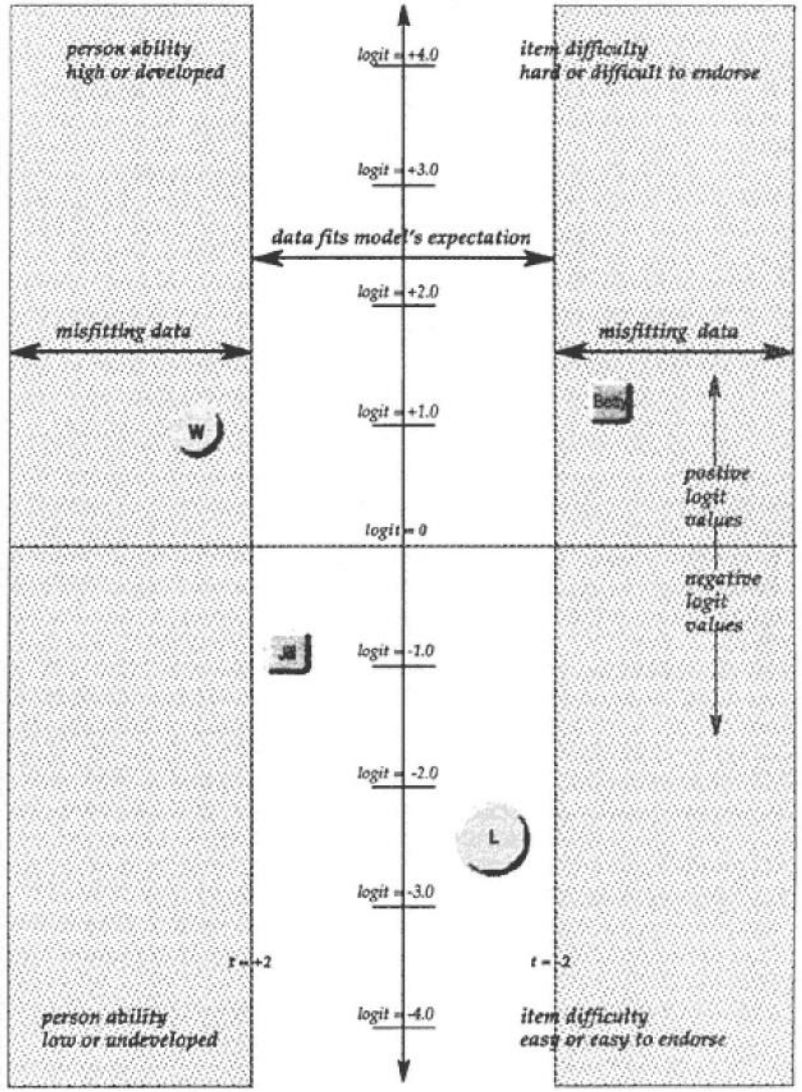




https://philchalmers.github.io/mirt/html/itemfit.html

Raschův model - infit (Příklad využití fitu a obtížnosti položek)





Shoda položky s daty (item fit)

Shodu lze testovat pomocí signifikance odlišnosti od modelu, příp. velikosti efektu.

Raschův model: infit, outfit

Obecně: Signed χ^2 ($_S\chi^2$), případně jen χ^2; G2; Q_1; plausible value Q_1 ($_{PV}Q_1$)

- A jejich bootstrapové varianty s vyšší robustností.

Velikost efektu: velmi často Cramerovo V.

- Jak moc se liší pozorované frekvence odpověďových kategorií od kategorií predikovaných modelem?

Další IRT modely

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.
- V případě škál lze zvažovat stejnou či rozdílnou vzdálenost „prahů“ u jednotlivých položek.
- Zpravidla 1PL či 2PL.

3 hlavní kategorie polytomních modelů:

- kategorie mohou být ordinální (PCM, GPCM, RSM)
- kategorie jsou ordinální (GRM, MGRM)
- kategorie jsou nominální (NRM)

Partial Credit Model (PCM) a RSM

Partial Credit Model (PCM; Masters, 1982): vyvinut v rámci Raschova modelu pro účely položek, kde je nutné provést sérii kroků vedoucích ke správnému řešení

$$P(X_{ni} = x) = \frac{e^{\sum_{k=0}^{x}(\theta_n - (b_i - \tau_k))}}{\sum_{x=0}^{m} e^{\sum_{k=0}^{x}(\theta_n - (b_i - \tau_k))}}$$

- $x \in \{0, 1, 2, \ldots, m_i\}$
- plus dílčí specifikace kvůli identifikaci modelu.
- Použitelný pro jakékoliv položky s více odpověďmi.
- τ_k „obtížnosti“ jednotlivých „prahů“ (zbytek viz dříve)

Rating Scale Model (RSM; Andrich, 1978): prahy τ_k napříč položkami jsou stejné

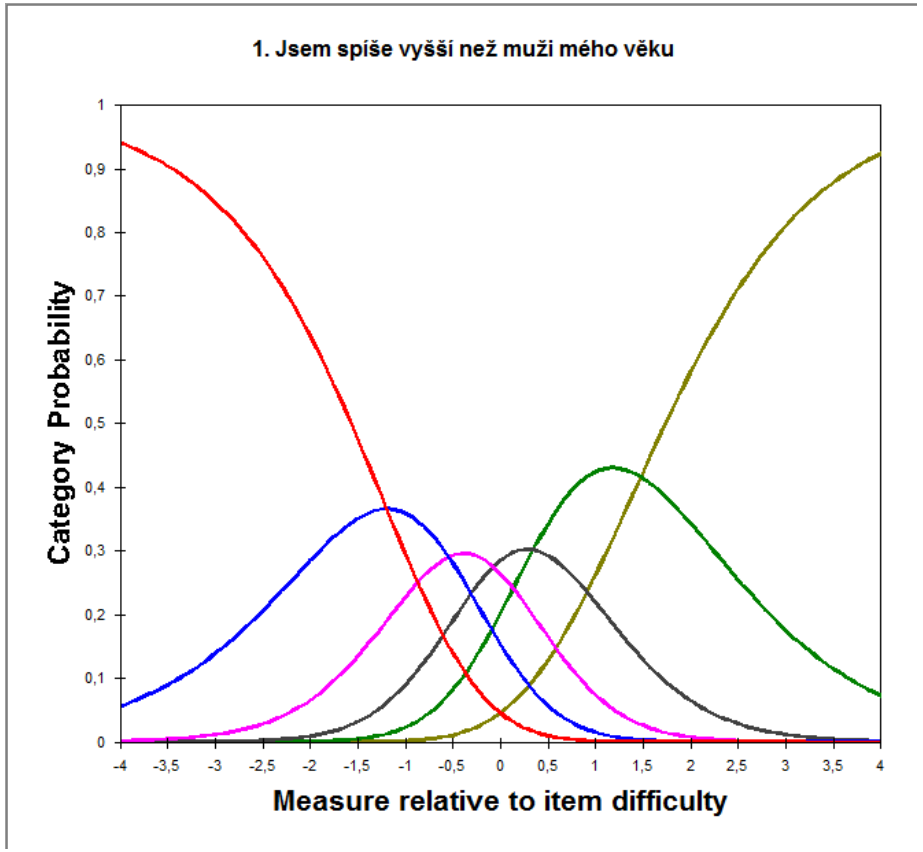
- méně parametrů, menší počet respondentů
- vhodné pro Likertovské škály s podobnými položkami

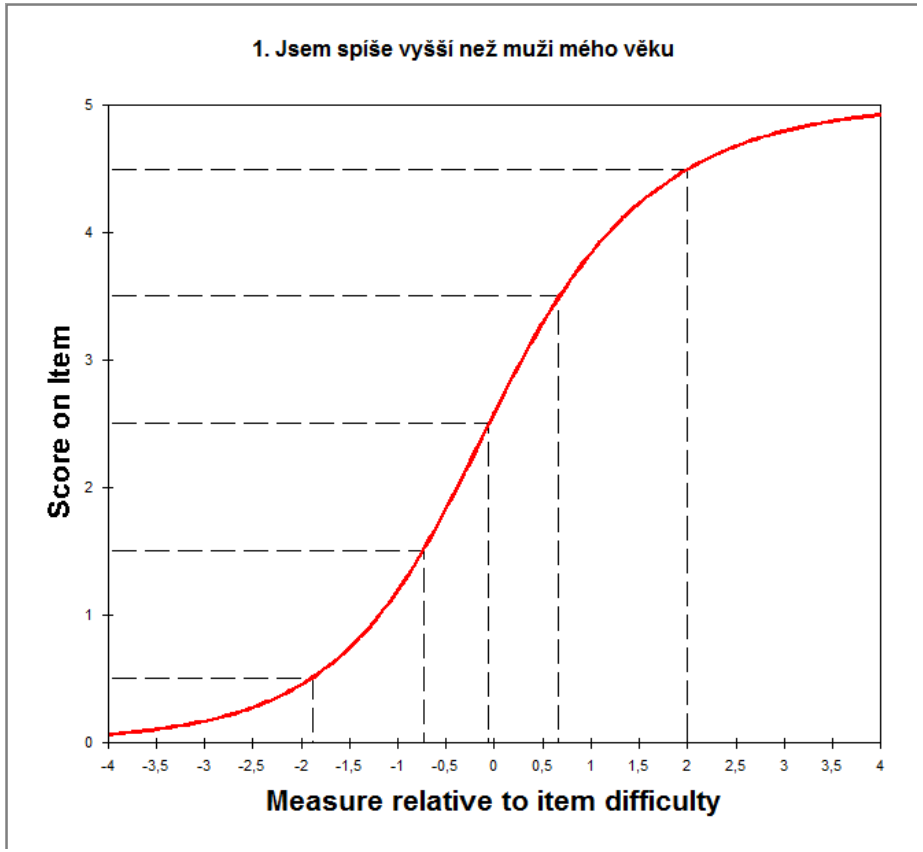
Generalized Partial Credit Model (2PL PCM) (GPCM; Muraki, 1992)

$$P(X_{ni} = x) = \frac{e^{\sum_{k=0}^{x} a_i(\theta_n - (b_i - \tau_k))}}{\sum_{x=0}^{m} e^{\sum_{k=0}^{x} a_i(\theta_n - (b_i - \tau_k))}}$$

- $x \in \{0, 1, 2, \ldots, m_i\}$
- plus dílčí specifikace kvůli identifikaci modelu.
- Položky se liší z hlediska své diskriminace (a_i)

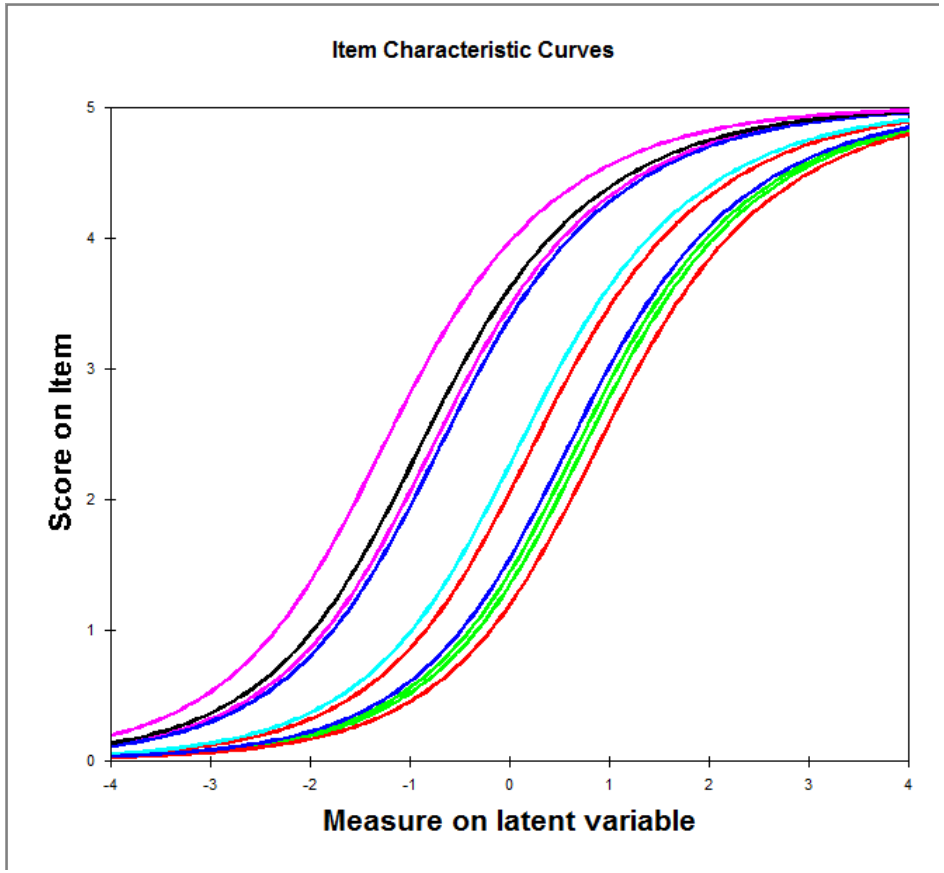
Příklad: PCM (5steps Likert)



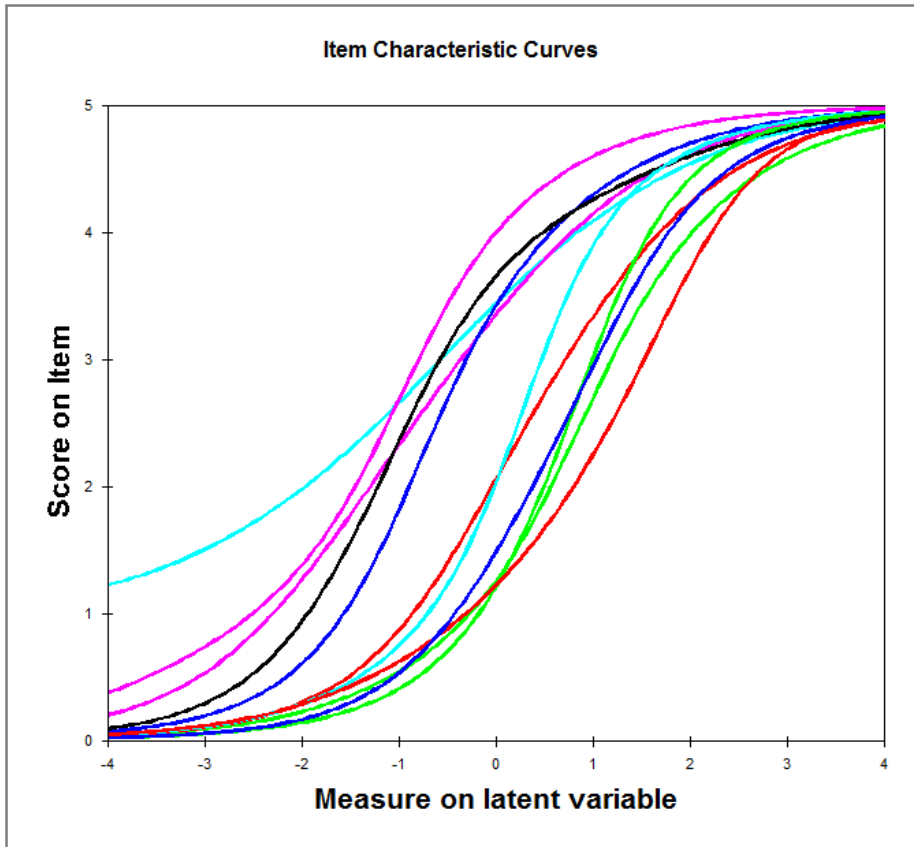


Příklad RSM vs. PCM

RATING SCALE MODEL



PARTIAL CREDIT MODEL



Graded Response Model (GRM)

Zobecnění 2PL modelu (Samejima, 1969): vlastně jen „navrstvení“ 2PL modelů za sebe.

- $P_{ix}^{*}(\theta) = \frac{e^{a_i(\theta - b_i)}}{1+e^{a_i(\theta - b_i)}}$,
- $P_{ix}(\theta) = P_{ix}^{*}(\theta) - P_{i(x+1)}^{*}(\theta)$
- Tzv. dvoukrokový odhad pravděpodobnosti:
  - Pro každou odpověď je odhadnuta pravděpodobnost, že respondent odpoví touto nebo vyšší odpovědí.
  - Výsledná pravděpodobnost konkrétní odpovědi je rozdílem odhadnuté pravděpodobnosti a pravděpodobnosti o jedna „vyšší/těžší“ odpovědi.

Modified Graded Response Model (MGRM, Muraki, 1990)

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

GRM vs. PCM

Výsledky obou modelů jsou velmi podobné.

Přestože predikované pravděpodobnosti a výsledky jsou velmi podobné, logika je diametrálně odlišná.

- PCM: Série navazujících kroků/znalostí nutných pro získání vyššího „skóre“.
  - Musím získat 1 bod, abych mohl získat 2 body; musím získat 2 body, abych mohl získat 3 body...
  - Pokud bych odpověděl správně možnost K, jaká je pravděpodobnost, že zodpovím správně i K+1?
- GRM: latentní kontinuum je rozčleněné na dílčí binární 2PL modely.
  - Určí se pravděpodobnost překročení každého ze „stupňů“ separátně a ty se pak „složí“ dohromady
  - Jaká je pravděpodobnost, že odpovím K vs. K+1? Jaká, že odpovím K+1 vs. K+2? K+2 vs. K+3? ... ?

GRM: Shodný model s ordinální faktorovou analýzou – liší se jen estimátor a (nikoli nutně) tzv. „link funkce“ (logit vs. probit).

Nominal Response Model (NRM)

Bock (1972): Obecný model pro položky s více odpověďmi, které nejsou (nemusí být) ordinálně seřazené.

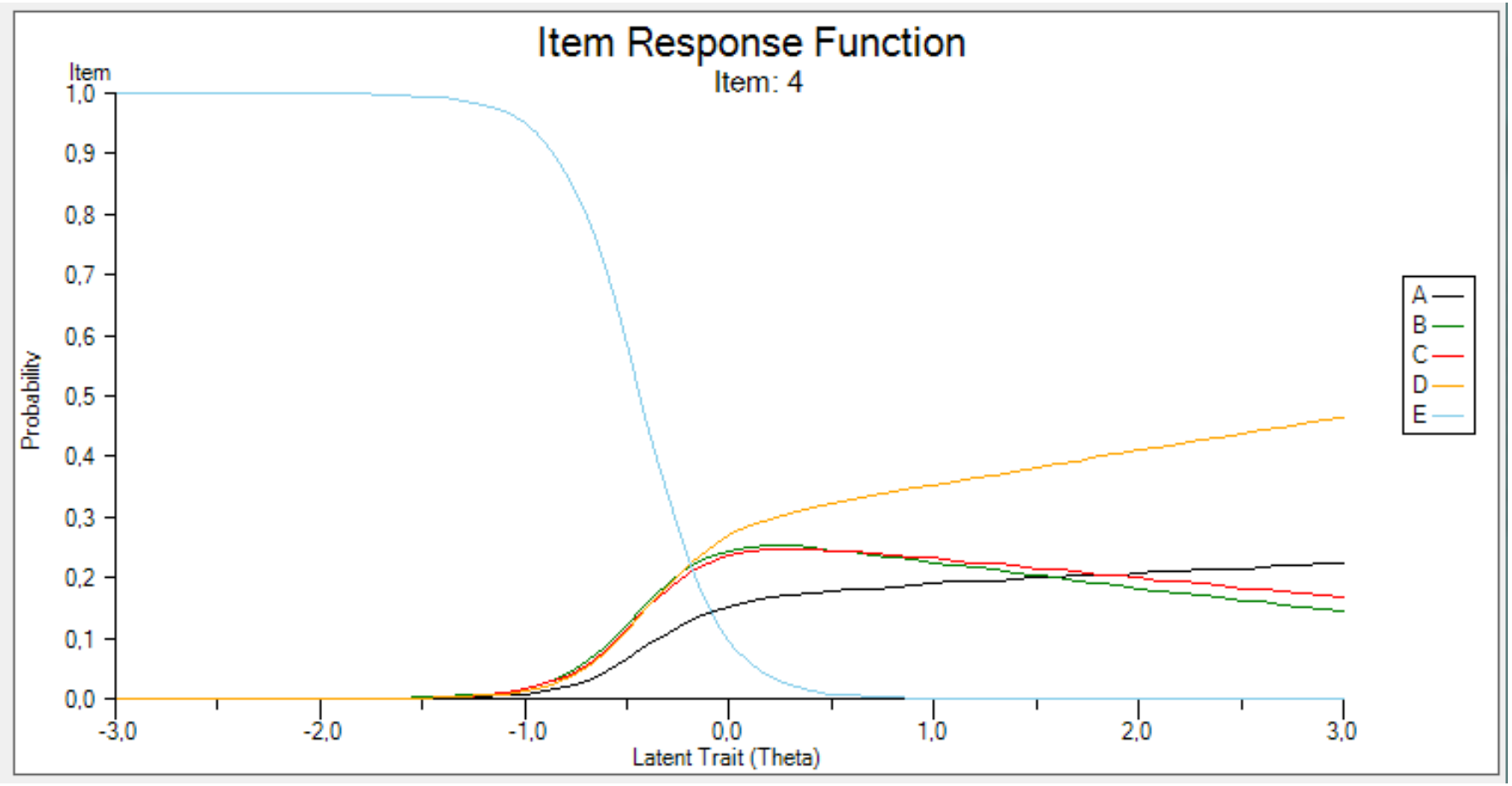
$$P_{ix}(\theta) = \frac{e^{a_{ix}\theta + c_{ix}}}{\sum_{x=0}^{m} e^{a_{ix}\theta + c_{ix}}}$$

- kde pro každou položku $\sum a_{ix} = \sum c_{ix} = 0$.
- Každý práh x položky i má tedy vlastní diskriminační koeficient a_{ix} a vlastní obtížnost c_{ix}.
- Vhodný pro multiple-choice testy (s jednou správnou) či výběr z odpovědí, kdy každá má jiný vztah s latentním rysem (rysy).
  - Výhodou je, že jsou pro odhad latentního rysu využity i chybné odpovědi (ale zase více parametrů...).

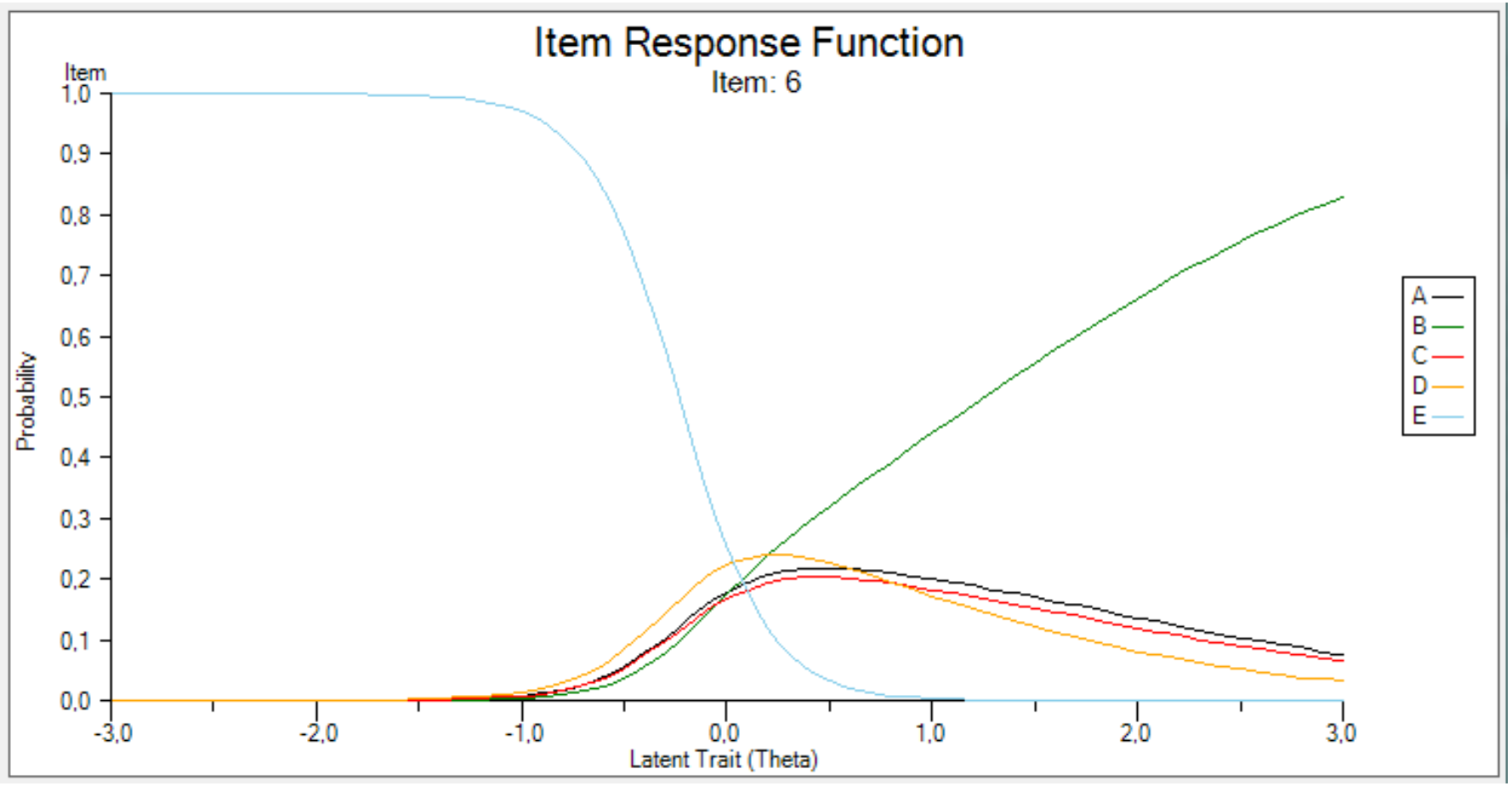
Lze ale použít i pro dotazníková data.

- Obecný model, většina ostatních je specifikací NRM; zvláště silné multidimenzionální verze.

Nominal Response Model (NRM)

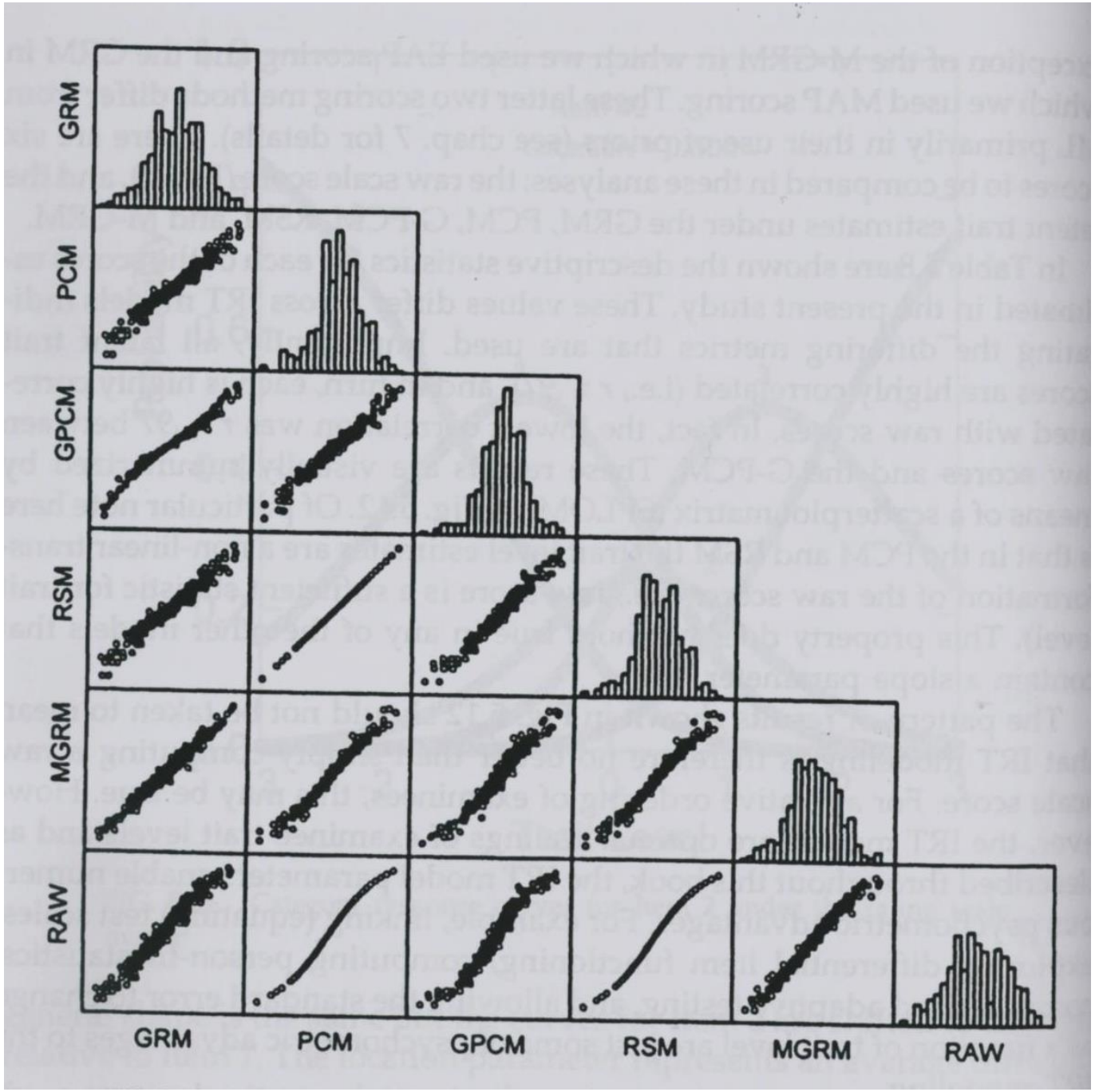


Nominal Response Model (NRM)



Srovnání modelů 1

Embretson a Reise (2009)



Srovnání modelů 2

García-Peréz, M.A. (2017); doporučuji pro mnoho dalších srovnání v různých situacích

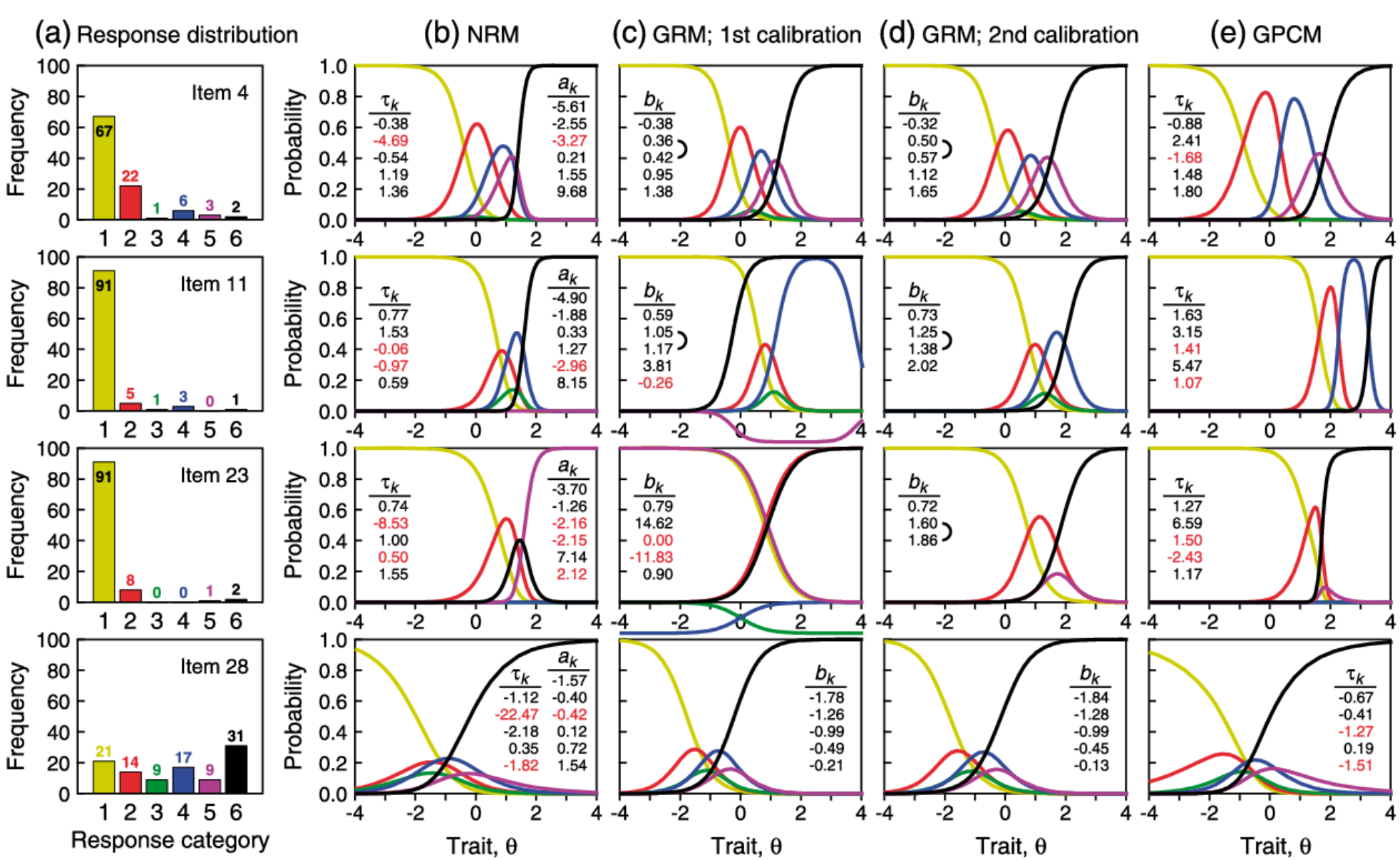


Figure 5. Empirical response distributions for four items (a) and estimated option response functions under the nominal

Další modely

IRT framework je velmi otevřený a existuje řada specifických modelů:

- **Neparametrické IRT modely** (monotónní i nemonotónní ICC).
- **Multidimenzionální modely** (tzv. item-factor analysis).
- **Item Response Time Models** (např. van der Linden, různé modely)
- **Ipsative Item Response Models** (pro položky s nucenou volbou)
- **Multiple-choice modely** (non-extrémní asymptoty pro všechny distraktory)
- Přesah do **kognitivního modelování** (IRT jsou jednoduché kognitivní modely).
- **Explanační IRT modely** (de Boeck), multilevel m. a m. s latentní regresí.
  - Testují hypotézy o vlastnostech položek, nikoliv lidí.
- A mnoho dalších...

Vybrané aplikace IRT:

Škálování, vyvažování,
počítačové adaptivní testování

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 informační funkce).

2. myšlenka: IRT nevadí 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 nejblíže úrovni odhadnuté schopnosti ($\mathrm{P}(\theta) = 0{,}5$).
   - Případně nepatrně lehčí (typicky $0{,}3 < \mathrm{P}(\theta) < 0{,}5$), 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.
   - 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 testu.

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.

IRT škálování

Samotný skór v logitech se pro praktické použití se 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-θ=10 W odpovídá změně pravděpodobnosti z 0,5 na 0,75 (resp. 0,25). 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?

Jednoznačně doporučuji: http://www.assess.nelson.com/pdf/asb-11.pdf

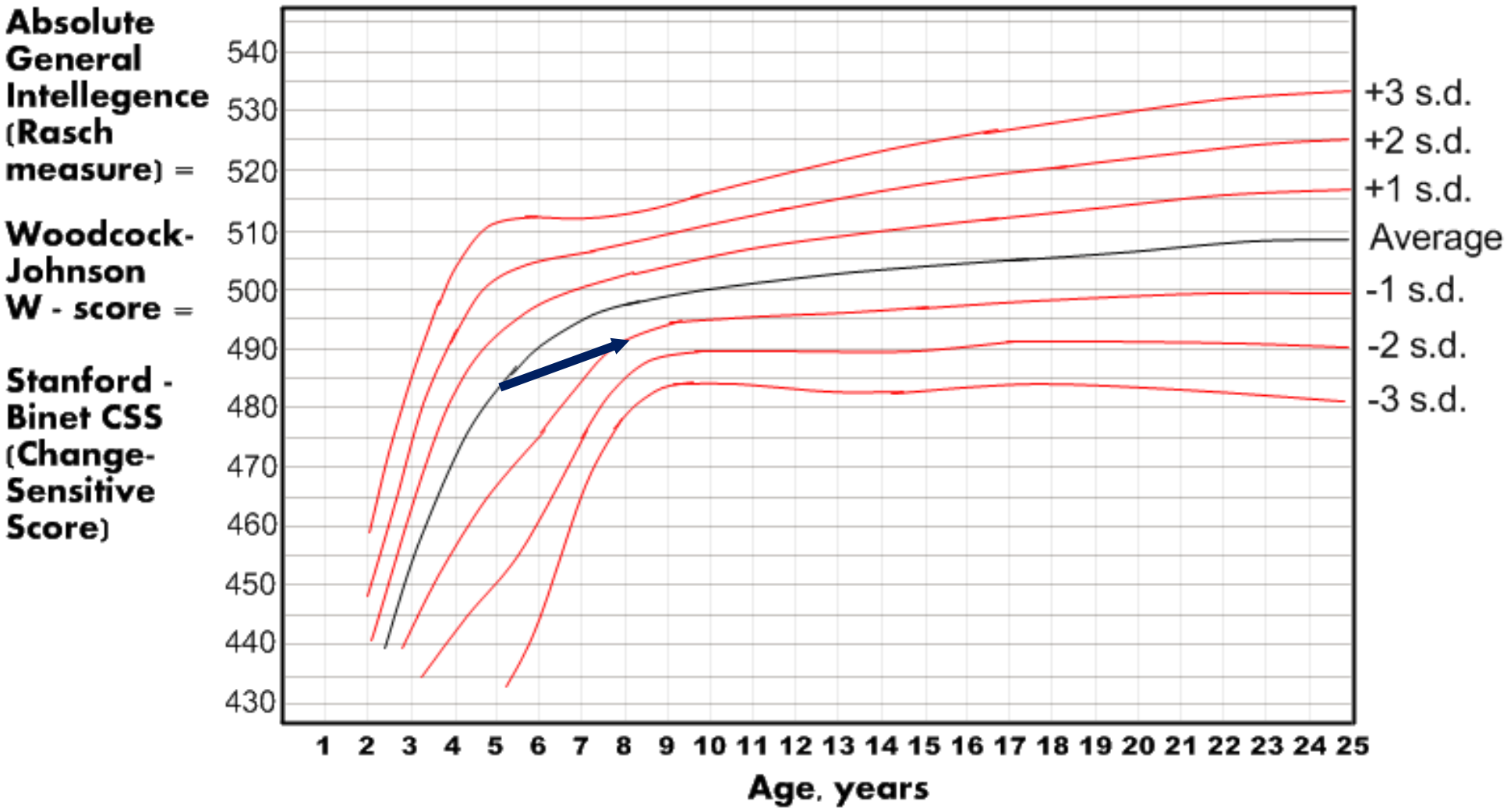
IRT škálování

Příklad z měření fluidní inteligence:

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

Inteligence 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.

http://mindsbasis.blogspot.cz/2016/03/rasch-measure-of-intelligence-age-2-25.html

IRT škálování

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

Ipsativní skóry poskytují intraindividuální srovnání (diagnostika profilu atp.).

- Statisticky, klinicky významný rozdíl...

W-skóry zasazují výkon člověk na fundamentální škálu společnou typu 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. Není ale závislý na SD vzorku/populace.

- Rozdíl 30 IQ v pěti a dvaceti letech znamená velmi odlišný rozdíl v reálném výkonu.
- Rozdíl 30 IQ v CHC faktoru psychomotorické tempo (Gs) znamená daleko vyšší rozdíl než rozdíl 30 např. u krátkodobé paměti (Gsm), protože $SD_{Gs} > SD_{Gsm}$.

W DIFF	RPI	*W* DIFF	RPI	*W* DIFF	RPI
29 and above	100[1]/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

http://www.assess.nelson.com/pdf/asb-11.pdf

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 equating high stakes testů normou.

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

Test equating (vyvažování testů)

Tři klasické (CTT) způsoby:

- Vyvažování **na základě průměru** (M) – testy musí mít stejné rozptyly, data musí být normálně rozdělená. $x_2 = x_1 + \overline{X_2} - \overline{X_1}$
- **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)
- **Equipercentilové 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.

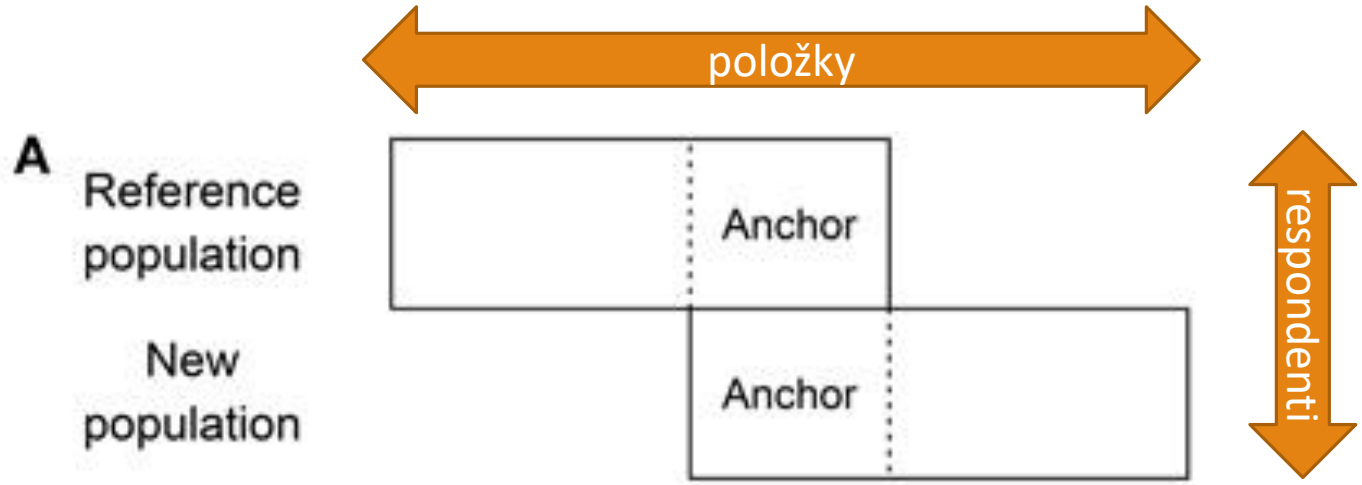
IRT equating: Sběr dat

Designy s jednou výzkumnou skupinou

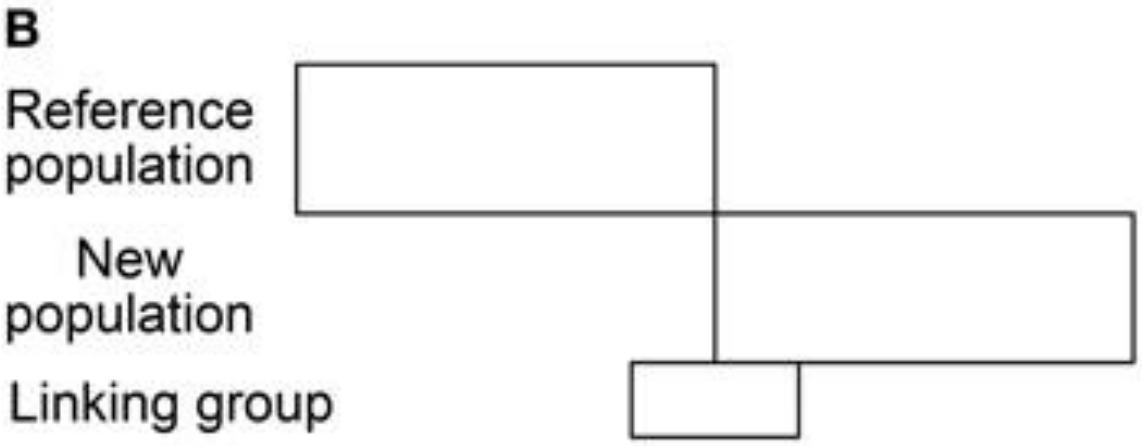
- Skupinu rozdělíme náhodně na dvě (tři...) podskupiny.
- **Counterbalancing** – Jeden test administrujeme jedné skupině dvakrát (střídáme pořadí).
- **Náhodné skupiny** – každé osobě administrujeme test jen jednou.
- Data musejí být sebrána vždy ve stejném čase!

Design s více skupinami:

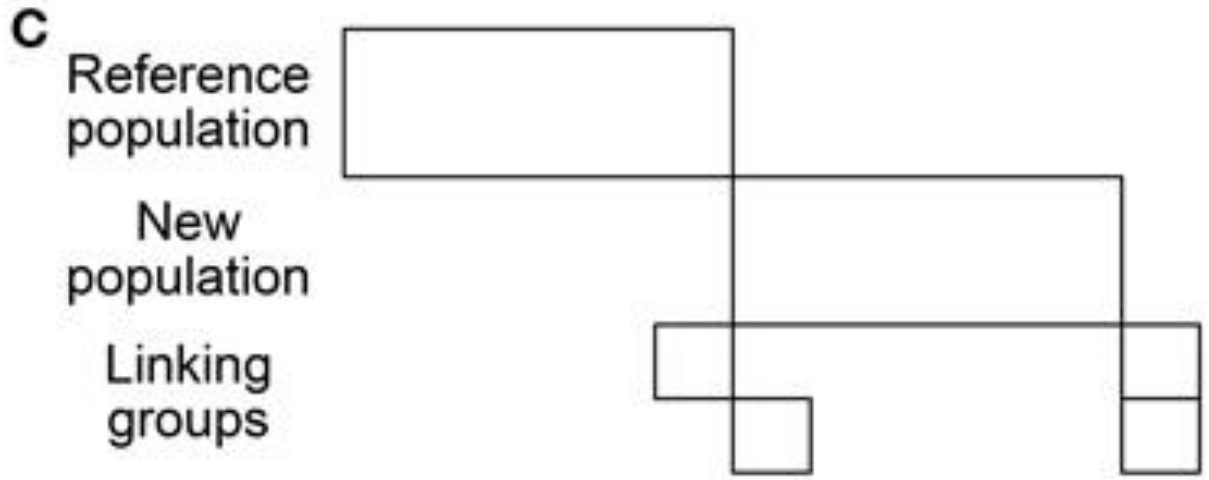
- Dvě nezávislé/nenáhodné skupiny, ale oba testy mají společné položky (tzv. „kotvu“ – **anchor test**), které slouží ke kalibraci.
- Ta může, ale nemusí být zahrnuta pro zjištění celkového skóru.
- Kotev může být více („planned missing data design“).



Design 1: anchor-item design

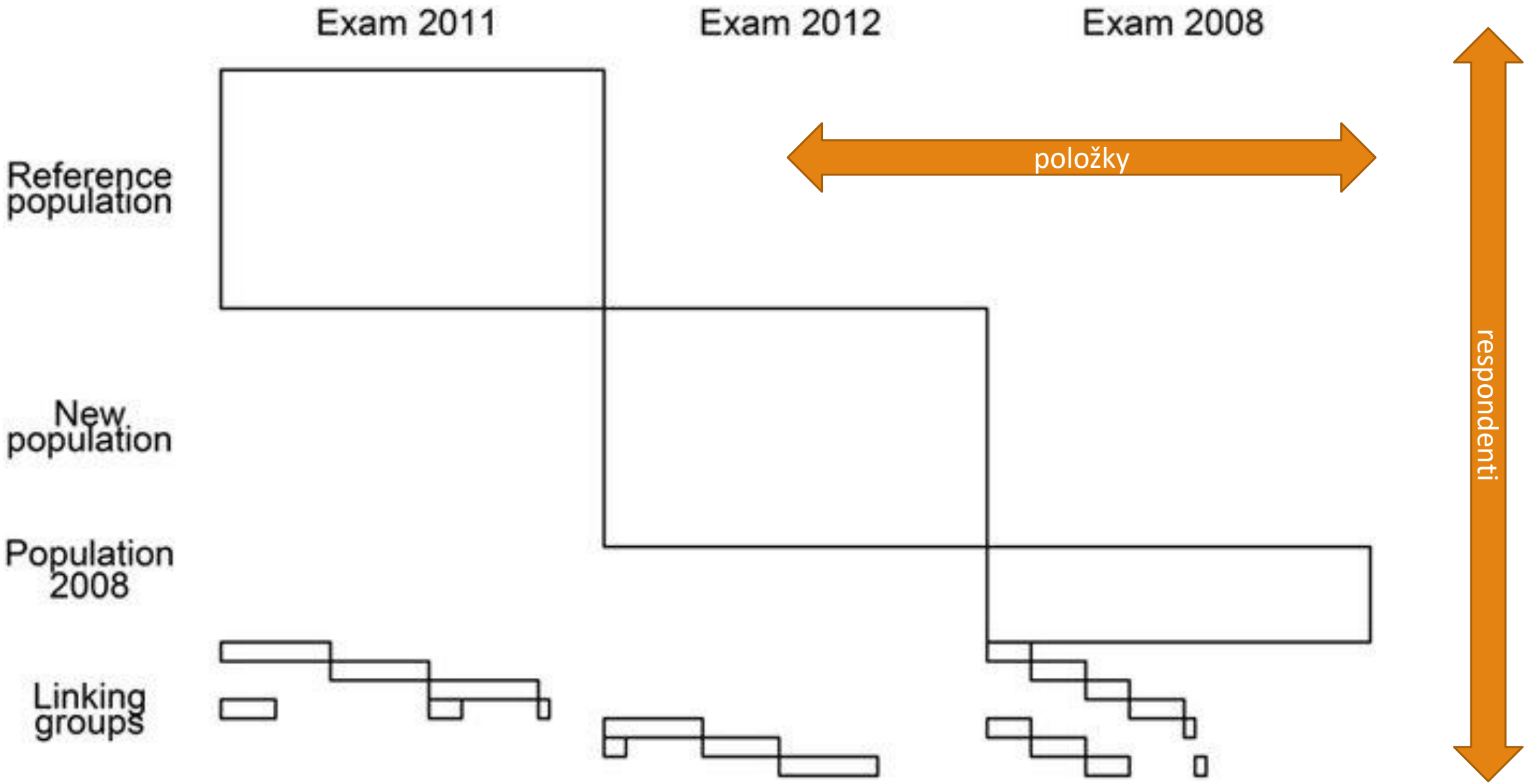


Design 2: post-equating design



Design 3: post-equating design

Bolsinova, M., & Maris, G. (2016; suppl. mat)



Bolsinova, M., & Maris, G. (2016; suppl. mat)

Figure 2-I: Test design CSEC

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

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

Multidimenzionální IRT (MIRT)

Předpoklad lokální nezávislosti zachován, ale rozptyl je vysvětlován více faktory.

- Má tedy méně předpokladů než klasické IRT.
  - Na rozdíl od CFA/SEM neumožňuje reziduální korelace, ty jsou proto řešeny specifickými faktory.

Dva hlavní typy:

- Kompenzatorní – vysoká úroveň jednoho rysu může kompenzovat nízký druhý rys.
  - Thety jsou aditivní (na stejné škále): $\theta_g = \theta_A + \theta_B$.
  - Běžnější, jednodušší
- Non-kompenzatorní – pro správnou odpověď je nutné mít vysokou úroveň všech rysů.
  - Lineární rysy nejsou aditivní (vznikají např. součinem): $\theta_g = \theta_A \theta_B$
  - Málo používané, řada komplikací.
  - Lze použít např. pro parametrizaci teorie vědomostních prostorů.

Multidimenzionální IRT (MIRT)

McDonaldův MIRT založený na normální ogivě

- Technicky vzato faktorová analýza s nelineární parametrizací.
- GRM = kategorická FA.

vs. Reckaseho logistický model.

- Protože normální ogiva je blízká logistické funkci, výsledky jsou v praxi velmi podobné.
- Výpočetně výrazně jednodušší.
- Logistický model dnes jednoznačně vede (McDonaldův model se zpravidla odhaduje prostřednictvím ordinální CFA).

MIRT: Latentní rysy

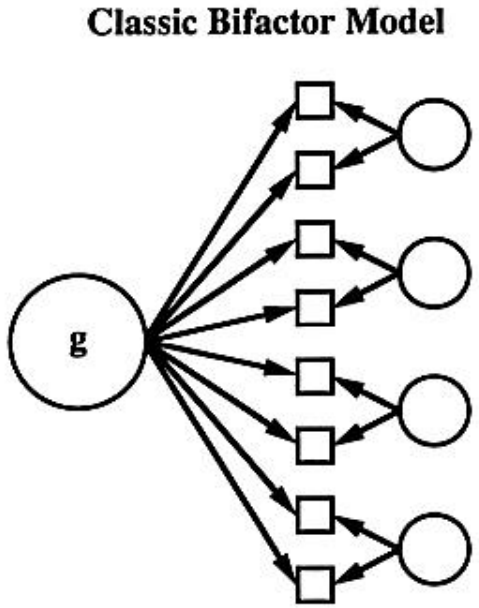
Model může být exploratorní (EFA MIRT) nebo konfirmační (CFA MIRT).

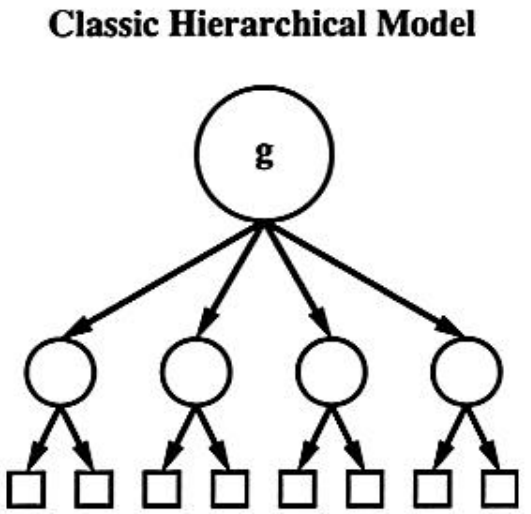
- Rotace u exploračních modelů stejně jako v EFA.

Každé osobě je přiřazen vektor latentních rysů, pro každou dimenzi jeden.

- Mohou být korelované nebo nekorelované.

Namísto hierarchických modelů jako v CFA se používá bifaktorový model.





MIRT: diskriminace

Pro kombinaci každého rysu (1-a položky (*i*) je vlastní diskriminační parametr a_{ki}.

Diskriminační „síla“ položky je zpravidla součtem diskriminačních parametrů:

$$\mu_{ki} = \sqrt{\sum_{k=1}^{N_i} a_{ki}^2}$$

Typicky se používá 2PL MIRT model.

- V případě Raschova MIRT modelu a_{ki}=1, a tedy každou položku sytí právě jeden faktor (a všechny stejně). Jde tedy vlastně jen o souběžný odhad více Raschových modelů najednou.
- Stejný výsledek, jako separátní odhad a následný součet chí-testů, jen korelace faktorů je odhadnuta na úrovni latentní úrovni.

MIRT: Ostatní

Namísto charakteristické křivky testu je definovaná „charakteristická plocha testu“.

- Ale její výpočet je analogický.

Obdobně pak „informační plocha“ testu...

- ... vzniká součtem informačních ploch položek.
- Zajímá nás rovněž, ve směru které dimenze chceme diskriminovat, podle toho se může odhad informační funkce lišit.

Další aplikace IRT

DIF analýza

- Differential item functioning – zjišťujeme, zda položka měří pro všechny respondenty shodně.
- Otázka konstruktové validity a férovosti testu.
- Ukážeme si podrobněji na příslušném setkání.

Multifasetový design, explanatorní IRT modely, modely s odpověďovými kovariáty atd.

- Odpověď je predikována dalšími pozorovanými proměnnými; například příslušností ke skupině, „přísností“ posuzovatele atp.
- Podobné jako teorie zobecnitelnosti.