MUNI
FSS

Teorie odpovědi na položku

PSY028 | JARO 2018 | BLOK 1

TEORIE MĚŘENÍ, VÝCHODISKA RASCHOVA MODELU

Organizace kurzu

HARMONOGRAM

středa 4. 4. 2018, 13:30–18:30 (PC25)

čtvrtek 5. 4. 2018, 13:30–18:30 (PC25)

-

středa 18. 4. 2018, 13:30–18:30 (PC25)

čtvrtek 19. 4. 2018, 13:30–18:30 (PC25)

-

středa 2. 5. 2018, 15:15–18:30 (PC25)

čtvrtek 3. 5. 2018, 15:15–18:30 (**PC26**)

POŽADAVKY NA UKONČENÍ

Přiměřená docházka ☺

Kurz je zakončen kolokviem v podobě individuální konzultace v domluveném termínu.

Ke zvládnutí kolokvia je nutné mít základní přehled v probrané látce a hlavně se seznámit s vybranými studijními zdroji podle vlastní preference (přiměřeného rozsahu)

Část obsahu se překrývá s PSY479.

Teorie měření

Tradiční dělení teorií měření v psychologii: kontrast CTT a IRT.

- To je ale hodně povrchní dělení.

Borsboom:

- Klasická testová teorie (Classical Test Theory, CTT)
- Modely s latentními proměnnými (EFA, CFA, CCFA, RM, IRT)
- Spojité měření (conjoint measurement)
- (Pokud jste tu čekali ještě network modely, tak ty nejsou měření☺).

To, co se dnes vydává za CTT, je určitá smíšenina původní CTT a faktorové analýzy.

- Faktorová analýza je používána jako důkaz konstruktové validity či pro konstrukci testu.
- CTT je využíváno pro vlastní parametrizaci a definici měření.

Teorie měření

Je vhodné konceptuálně oddělit:

- Způsob „škálování“, tedy vztah pozorovaného chování a naměřené proměnné (tedy jakým způsobem z „pozorování“ vytvářím „skóry“).
- Způsob uvažování o tom, jaká je charakteristika měřených jevů.
- Způsob, jakým dokládáme kvalitu měření (validitu, reliabilitu...).

Co to tedy je měření?

Druhy měření

Množství a velikost.

- Základní charakteristika měřeného pojmu/konstruktu/veličiny, kterou chceme měřit. Z principu tedy intervalové.

Už v 18. a 19. století (Kant, Leibniz aj.) definice „intenzity“:

- **Extensivní velikost** – celek se fyzicky skládá z částí (délka). Položením dvou metrových předmětů za sebe vznikne předmět o délce 2 m.
- **Intensivní velikost** – rovněž intervalové, ale „projevují se“ instantně (teplota, barva...). „Sloučením“ dvou 20stupňových předmětů nevznikne předmět 40stupňový.
- Nepostačuje pro definici měření – jsou závislé na měřené veličině.

Z principu jsou všechny psychické rysy intenzivními veličinami.

Druhy měření

Koncept fundamentálního měření.

- **Základní:** Není odvozené z jiného měření, měří se přímo objekt za pomocí stejné veličiny (délka pravítkem, váha závažím).
- **Odvozené:** Je odvozené pomocí aditivních operací z jiných naměřených hodnot (objem, čas, síla zemětřesení na Richterově stupnici).
- Nic jiného není měření v pravém slova smyslu.

Podobné extensivní (fundamentální) a intensivní (odvozené) veličině.

- Fundamentalita však není charakteristikou měřené veličiny, ale měření jako takového.
- Odvozená veličina je ta, pro niž bylo použito odvozené měření; časem se z ní s vývojem jiného měřicího nástroje může stát veličina základní.

Z této definice ale vyplývá, že měření v psychologii tak, jak existuje dnes, zřejmě není fundamentální ☺

Aditivita

Aditivita je předpokladem sčítání; zjednodušeně princip: „celek je součet částí“.

Umožňuje např. převést funkci „+“ do „×“: např. f(a+b) = f(a)+f(b).
Předpokladem aditivity je „řazení“ (ordering) a „řetězení“ (concatenation).

Hodnoty lze sčítat (násobit) a provádět běžné matematické operace.

- Což ale nemusí být smysluplné z hlediska dané veličiny (intenzivní vs. extenzivní).

Základem měření je tedy intervalová/poměrová škála se stejně velkými jednotkami a aditivní strukturou (případně binární škála ano/ne).

Je nutné oddělit aditivitu veličiny a aditivitu měřeného objektu.

- Např. objem je aditivní jednotkou, lze „sčítat“ např. m^3.
- Při smíchání dvou látek ale může být výsledný objem jiný, než odpovídá součtu původních objemů.
  - Intenzivní veličiny tedy umožňují řetězení/aditivitu škály, ale technicky to nemusí dávat smysl.

Klasická testová teorie

vpravo: Cronbach

Klasická testová teorie

Klasická testová teorie stojí na třech pilířích/objevech (Traub, 1997):

- Existence chyby měření I. typu (nezpůsobené ničím jiným).
- Chyba měření je náhodná veličina.
- Koncept korelace.

Spearman (1904) přišel s koeficientem proti oslabení korelace („attenuation coefficient"), chybu měření parametrizoval a umožnil vznik CTT.

Důležitým impulzem byla rovněž Fergusonova komise (1932-1940).

- Striktní požadavek aditivity. Protože psychologové nedokázali „zřetězení" svého měření, psychologie podle závěrů komise neměří.
- Reakcí byla Stevensova „operační teorie měření", která rozšířila definici měření: „...measurement, in the broadest sense, is defined as the **assignment of numerals to objects and events according to rules.**" (Stevens, 1946, s. 677). Klíčový pojem je „matching".
  - Ve skutečnosti zjednodušení konsenzu z přírodních věd: „Measurement is a method of *assigning numbers to magnitudes*" (např. Helmholtz, 1887).
  - Umožnila nezabývat se tím, co jsou naměřené hodnoty, a „jít dál".

Vývoj CTT byl prakticky ukončen do 60. let, vše podstatné z hlediska CTT jako teorie měření je v Lordovi a Novickovi (1968).

CTT: Axiomy

„Dobré“ měření je takové, kdy různí lidé v různých časech dojdou různými nástroji ke stejným naměřeným hodnotám, pokud se míra samotného objektu nezměnila.

Postup fyzikálního měření (např. délky):

- Změřím objekt n-krát a získám n měření délky označených jako d_i.
- Bodový odhad délky je průměr z těchto měření: $\mathrm{E}(d) = \frac{\sum_{i=1}^{n} d_i}{n}$
  - To $\mathrm{E}(d)$ je „expected value“ – odhad měřené hodnoty.
- Chyba tohoto měření (Standard Error /of Measurement/) je:
  - Pro jediné měření: $SE = s_d$, kde s_d je výběrová směrodatná odchylka pozorovaných hodnot d_i .
  - Pro průměr z n měření: $SE = \frac{s_d}{\sqrt{n}}$ (standardní chyba průměru!).
  - (A použijeme Studentovo t-rozložení, protože s_d je pouze pozorovaným odhadem populační σ_d.)

CTT: Paralelní testy

Koncept reliability byl zaveden Spearmanem (1904) za účelem odhadu hodnot korelačních koeficientů nezkreslených nepřesnostmi měření. Od něj pak byla odvozena reliabilita v sociálních vědách.

Pojem reliability je založen na konceptu paralelních testů.

- A na něm je založen celý model měření v CTT.

Paralelní testy jsou takové, u kterých platí:

- A. Pravý skór je v obou testech a pro každý měřený subjekt stejný (exaktněji je průměrem z nekonečně velkého počtu měření).
- B. Rozptyl těchto pravých skórů je v obou testech stejný (platí automaticky, platí-li A).
- C. Chybový rozptyl je v obou testech a pro každý subjekt stejný (exaktněji jde o SD nekonečně velkého počtu měření).

CTT: Paralelní testy

CTT vychází z operacionalismu.

- CTT definuje „pravý skór“ (tj. objekt měření) skrze použitý měřící nástroj. Neřeší, jak tento skór vznikl – zpravidla jde tedy o součet položek.

Základní CTT vztah $X_p = \tau_p + e_p$ lze chápat jako lineární funkci. Standardizovaný regresní koeficient je potom roven korelaci prediktoru a závislé proměnné, tedy $r_{x\tau}$.

CTT: Reliabilita

Reliabilita $r_{xx'}$ testu x je definovaná jako vysvětlený rozptyl pozorovaného skóre pravým skóre:

$$r_{xx'} = (R^2) = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma_e^2} = \frac{\sigma_x^2-\sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

- Úpravy platí, protože dosazujeme podle vzorce $\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$.

Vysvětlený rozptyl je druhá mocnina korelace, tedy:

- $r_{xx'} = r_{x\tau}^2 = (R^2)$
- $\sqrt{r_{xx'}} = r_{x\tau} = (R)$

OK. Ale jak tedy zjistíme to $r_{x\tau}$?

CTT: Paralelní testy

Výpočtu lineární regrese je „jedno“, kterým „jde směrem“ ☺

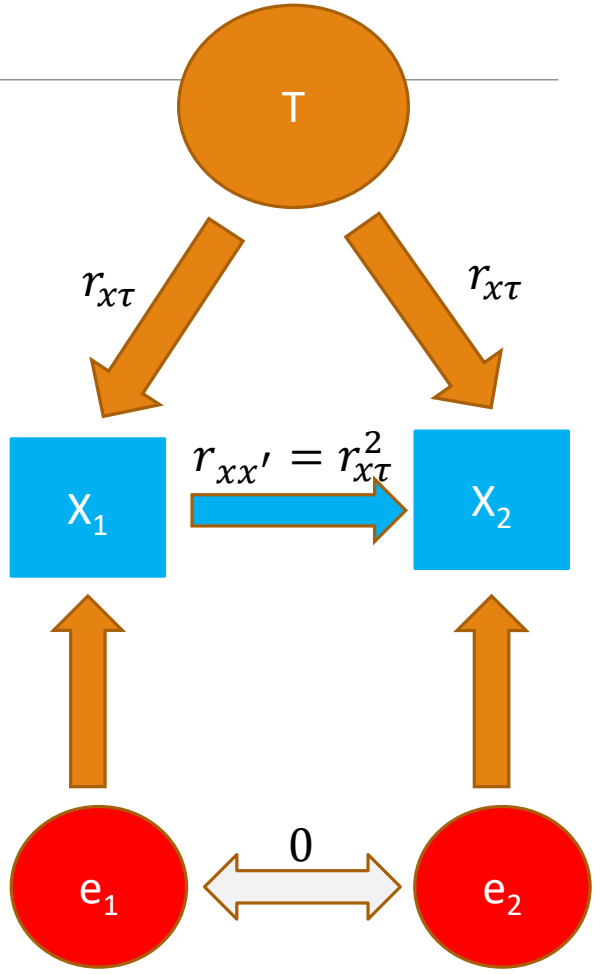
- Pokud tedy pravé skóre vysvětlí např. 80 % pozorovaného skóre, pak to samé pozorované skóre zároveň vysvětlí 80 % skóre pravého.

Protože **chyby měření spolu nekorelují** (v CTT modelu), pravé skóre mediuje veškerý vztah obou testů.

- X_1 vysvětlí 80 % T a to zase 80% X_2. $80\%^2=64\%$.

Korelace dvou měření je tedy $r_{xx'} = r_{x\tau}^2$.

- Což už známe.



CTT: Reliabilita

Reliabilita testu je proto mj. definována jako uvažovaná „korelace dvou paralelních testů“.

- Někdy zjednodušeně uváděno jako korelace metody se sebou samou, proto ten symbol $r_{xx'}$ – korelace měření x s virtuálním paralelním měřením x'.

Hypotéza s paralelními měřeními nebyla nová, paralelní měření byly dlouho známým principem zpřesňování fyzikálních měření. Průlomový byl ale právě Spearmanův (1904) článek o oslabení nereliabilitou, díky kterému z pozorované korelace dokážeme odhadnout reliabilitu.

Základní důkazy reliability v CTT jsou proto postaveny na „paralelní administraci testu“.

- Měří ale různé paralelní administrace stále to samé?

CTT: Attenuation (oslabení)

Hlavním motivem Spearmana (1904) při práci s reliabilitou byl odhad korelace dvou testů nezkreslený chybou měření.

Tzv. „attenuation coefficient“, „korekce proti oslabení“, „korekce proti nereliabilitě“. Odhad korelace pravých skórů:

$$r_{pq}^{*} = \frac{r_{pq}}{\sqrt{r_{pp'} r_{qq'}}}$$

- Kde r_{pq}^{*} je odhad korelace pravých skórů, r_{pq} je pozorovaná korelace testů p a q a $r_{pp'}$, $r_{qq'}$ jsou jejich reliability.
- Protože korelace pravých skórů $r_{pq}^{*} \leq 1$, lze odhadnout maximální možnou pozorovanou korelaci testů: $r_{pq} \leq \sqrt{r_{pp'} r_{qq'}}$

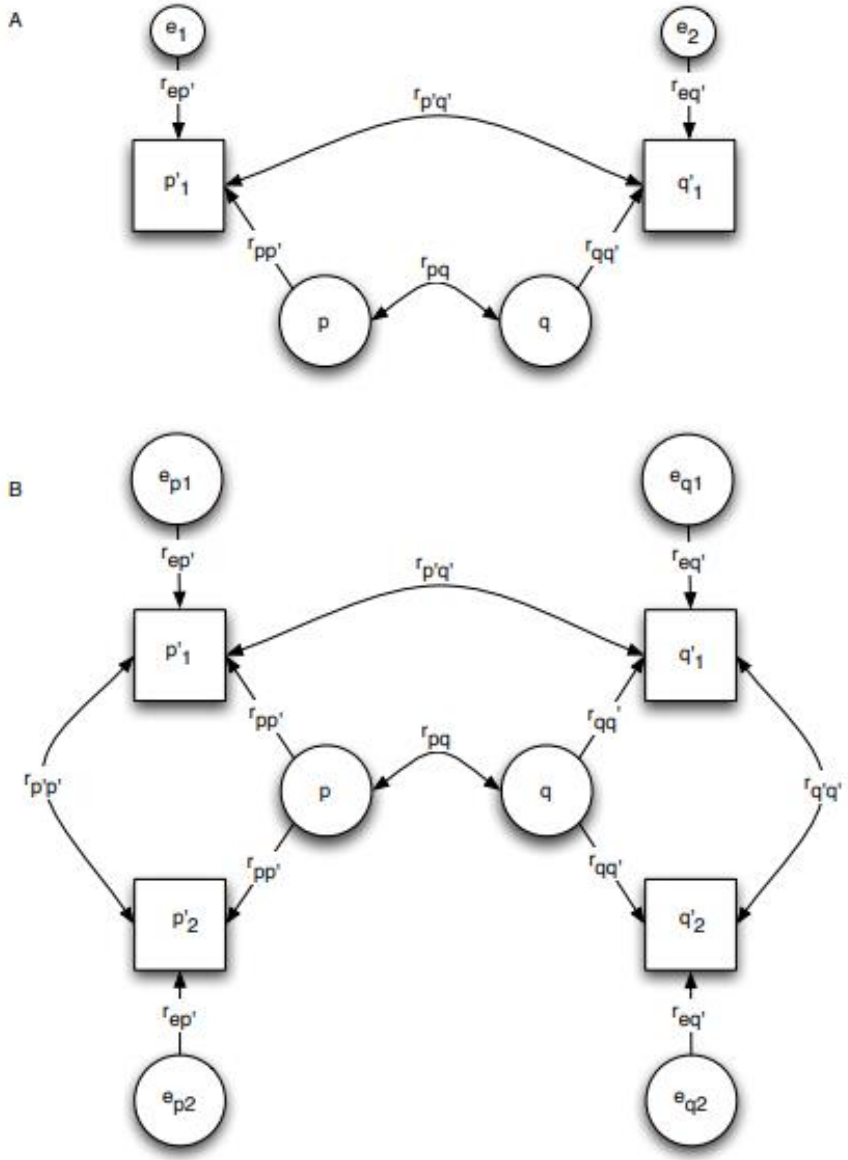


Fig. 7.1 Spearman's model of attenuation and reliability. Panel A: The true relationship between p and q is attenuated by the error in p' and q'. Panel B: the correlation between the latent variable p and the observed variable p' may be estimated from the correlation of p' with a parallel test.

CTT: Odhady reliability

Stabilita v čase, reliabilita typu test-retest

- Paralelním testem (PT) je ten samý test administrovaný jindy.

Shoda posuzovatelů, inter-rater reliabilita.

- PT je ten stejný test administrovaný někým jiným.

Reliabilita paralelních forem.

- PT je jiný test vytvořený tak, aby „byl stejný“.

Vnitřní konzistence

- Split-half: PT jsou jednotlivé půlky testu.
- Alfa: PT jsou jednotlivé položky/půlky testu.

Lze čekat, že všechny koeficienty budou stejné?

Jsou položky paralelními testy?

Mohou být. Ale to je hodně silný předpoklad. Proto koncept „míry“ paralelnosti založený na faktorové analýze.

- $X_{ip} = \tau_i + \lambda_i \theta_p + \varepsilon_{ip}$, kde X_{ip} je pozorovaný skór člověka *p* na položku *i*, τ_i je intercept (průměr všech osob na dané položce), λ_i faktorový náboj (směrnice, „měřítko“ položky) a ε_{ip} je reziduum; ta mají rozptyl σ_i^2 označovaný jako unicita (reziduální rozptyl, chyba měření položky).

Jsou položky paralelními testy?

$$X_{ip} = \tau_i + \lambda_i \theta_p + \varepsilon_{ip}$$

Stupně paralelnosti („vyšší“ obsahuje všechny „nižší“ předpoklady):

- Kongenerické položky – měří stejný rys.
  - Položky jsou pouze vybrány ze stejné domény.
- Tau-ekvivalentní položky – měří na stejné „škále“.
  - „Měřítko“ všech položek je stejné; shodné faktorové náboje λ_i napříč položkami.
- Paralelní položky – měří se stejnou chybou.
  - Reziduální rozptyl σ_i^2 shodný napříč položkami.
- Striktně paralelní položky – stejná obtížnost.
  - Intercepty τ_i shodné napříč položkami.
  - U binárních položek má paralelní a striktně paralelní shodný význam (protože $\sigma_i^2 = \tau_i(1 - \tau_i)$.

CTT: Předpoklady odhadů

Řada odhadů reliability má nějaké předpoklady. Při jejich nedodržení je reliabilita součtového skóre (tedy odhad rozptylu vysvětleného pravým skóre) zkreslena:

Cronbachovo alfa: tau-ekvivalence položek.

Spearman-Brownova korekce split-half korelace: paralelní položky.

Vše výše uvedené: jednodimenzionalita, případně lokální nezávislost

- A tím se dostáváme k FA.

CTT: škálování

Vizuální analogová škála (Hayes and Paterson, 1921)

Thurstonova škála (1928)

- 3 různé typy, např. „metoda stejně se jevících intervalů“.

Likertova škála (1932)

- Metoda sigma vs. zjednodušená metoda

Guttmanova škála (40. léta)

- Rozšíření původní Bogardovy (1925) škály sociální distance.
- Původně deterministický model později rozšířen na stochastický model; **Guttmanova škála jako základ IRT**.

Osgoodův semantický diferenciál (1957)

Doporučuji kap. 5: Price, L. R. (2016). *Psychometric Methods: Theory into Practice.* New York: Guilford Press. Libgen ☺

CTT vs. faktorová analýza

CTT měří pravé skóre testu. *Co to sakra jako je*?

Pravé skóre je očekávaný skór respondenta **v daném testu**.

- Měření je závislé na měřicím nástroji. Délka stolu měřená pravítkem A a pravítkem B je nejen jiná, ale je to *jiná délka*.

Jsou položky paralelními testy z hlediska měřeného pravého skóre?

Z toho důvodu se CTT prolnulo s faktorovou analýzou, která ospravedlňuje CTT postupy (sčítání položek, přijetí konceptu „pravého skóre“), ověřuje „konstruktovou validitu“ atd.

Teorie zobecnitelnosti

Jednotlivé konceptualizace paralelních testů (test-retest, split-half, shoda posuzovatelů) neměří „tu stejnou chybu“. Jde o různé reliability. Jaká je ale „absolutní přesnost měření“?

Cronbach, Nageswari a Gleser (1963) rozvinuli původní CTT koncept do teorie zobecnitelnosti.

„Rozparcelovali“ náhodnou chybu měření do dílčích složek, které jsou odhadovány naráz; namísto pravého skóre měříme „universe score“, tedy pravý skór pro prostor daných kombinací podmínek.

Původně se rozptyl parceloval pomocí rmANOVA, dnes spíš mixed model.

Spojité měření

vpravo: John Tukey

víc vpravo: Gérard Debreu

Spojité měření

Nezávisle na sobě objevili francouzský ekonom Gérard Debreu (1960) a psycholog Duncan Luce s matematikem Johnem Tukey (1964).

Conjoint measurement theory (CM; teorie spojitého měření) definuje, jakým způsobem lze z nominálních pozorování sestavit škálu s aditivními vlastnostmi.

- A tedy vyvrací závěry Fergusonovy komise; resp. poskytují možnost testovat kvantifikovatelnost pozorování psychických jevů.
- Pro nás je podstatné, že Raschův model je jednou ze stochastických specifikací jinak deterministického CM.

V současnosti docela rychlý rozvoj v oblasti dalších stochastických aplikací pro různé účely, např. Karabatsos chrlí jeden model za druhým.

Zajímavost: Tversky z dvojice Kahneman a Tversky (1979), kteří získali jako první psychologové Nobelovu cenu, se zaměřoval právě na CM (např. 1967) a jejich prospektová teorie je na CM založena.

CM: Axiomy

CM je založeno na několika axiomech. Jejich splnění vede k tzv. spojitému měření.

Mějme dvě proměnné A a X.

- Nevíme, zda A, X, nebo A i X jsou kontinuální proměnné.

a, b, c... jsou disjunktní, identifikovatelné úrovně proměnné A;
x, y, z... jsou disjunktní, identifikovatelné úrovně proměnné X.

P je seřazená množina všech možných $A \times X$ párů atributů A a X.

- Buď může být seřazená (přirozená čísla), nebo může jít o hodnoty (reálná čísla).

CM: Single cancellation

Požadavek „nezávislosti“.

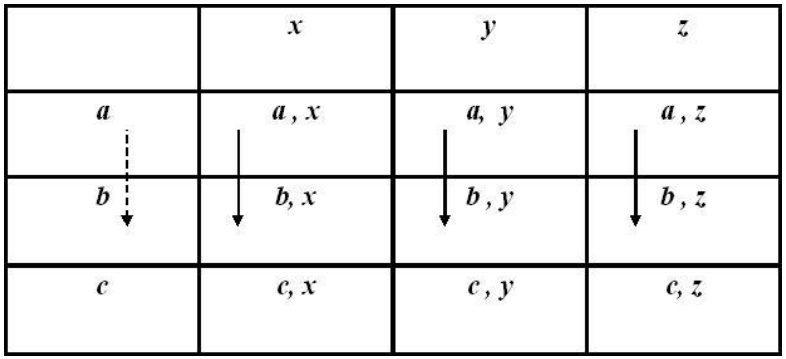
Řazení prvků A je stejné pro všechny úrovně X.

- Vyžaduje „řazení“, tedy první podmínku aditivity.

Pokud $(a, x) < (b, x)$,
pak $(a, w) < (b, w)$ pro všechna $w \in X$.

Platí tranzitivita:

- $[(a, x) > (b, x)] \wedge [(b, x) > (b, y)] \Rightarrow (a, x) > (b, y)$



CM: Single cancellation

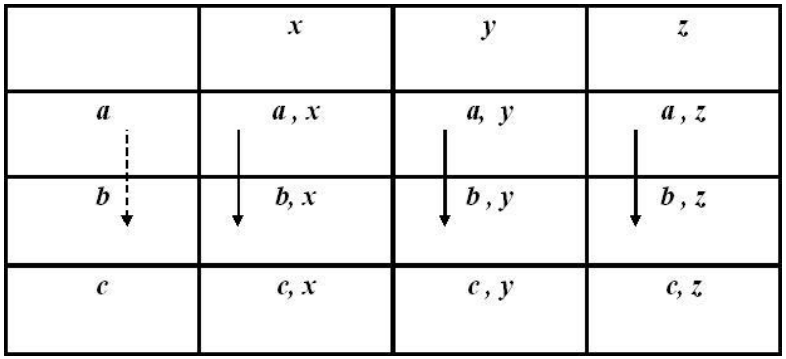
Single cancellation – jednoduché vykrácení.

- $(a, x) < (b, x)$
- $(a, \cancel{x}) < (b, \cancel{x})$
- $a < b$

„left-leaning diagonal“

Pohyb „zpět“ ale není možný:

- $[(a, x) > (b, y)] \Rightarrow [(a, y)?\,(b, x)]$
- Nic nelze vykrátit



CM: Double cancellation

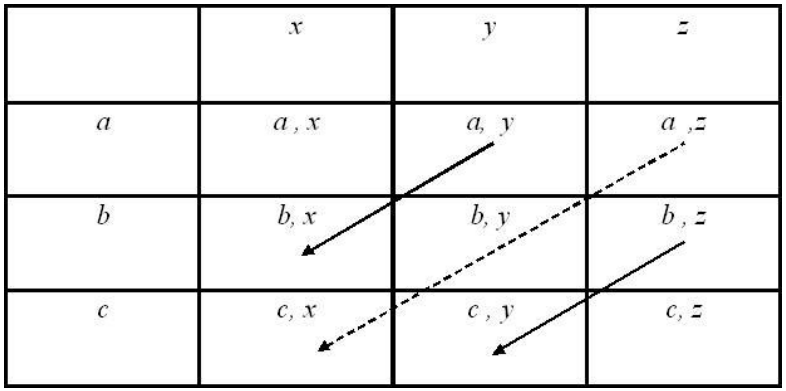
Předpokládejme:

- $(a, y) > (b, x)$
  - a tedy $a + y > b + x$
- a $(b, z) > (c, y)$
  - a tedy $b + z > c + y$

Tedy:

- $a + y + b + z > b + x + c + y$
- $a + \cancel{y} + \cancel{b} + z > \cancel{b} + x + c + \cancel{y}$
- $a + z > x + c$
- $(a, z) > (c, x)$

„right leaning diagonal“



CM: Příklad 1

Délka.

- m > cm > mm
- stůl > kniha > tužka

Jsou tužka-kniha-stůl kvantitami?

Jednoduché vykrácení

- (tužka, mm) < (kniha, mm)
- (tužka, mm) > (tužka, cm)

Dvojité vykrácení

- (tužka, cm) < (kniha, mm)
  - (kniha, mm)/(tužka, cm) = 300/15 = 20
- (kniha, m) < (stůl, cm)
  - (stůl, cm)/(kniha, m) = 150/0,3= 500

	mm	cm	m
tužka	150	15	0,15
kniha	300	30	0,3
stůl	1500	150	1,5

(tužka, m) < (stůl, mm)

- (stůl, mm)/(tužka, m) = 1500/0,15 = 10000

tužka+*cm*+*kniha*+m <
kniha+mm+stůl+*cm*

- tužka+m < stůl+mm
- 20*500 = 1000

Vznikne fundamentální škála:

- tužka=1, kniha=2, stůl=10

CM: Příklad 2

„Ovocná škála“.

- třešně > hrušky > jablka
- červené > žluté > zelené

Je druh ovoce, resp. barva kvantifikovatelná?

	zelené	žluté	červené
jablko	zelené jablko	žluté jablko	červené jablko
hruška	zelená hruška	žlutá hruška	červená hruška
třešeň	zelená třešeň	žlutá třešeň	červená třešeň

CM: Příklad 2

„Ovocná škála“.

- třešně > hrušky > jablka
- červené > žluté > zelené

Je druh ovoce, resp. barva kvantifikovatelná?

Ohodnoťte chutnost každého ovoce na škále 0-100.

ANO

- třešeň = 5, hruška = 2, jablko=1

	zelené	žluté	červené
jablko	1	10	20
hruška	2	20	40
třešeň	5	50	100

CM: Příklad 2

„Ovocná škála“.

- třešně > hrušky > jablka
- červené > žluté > zelené

Je druh ovoce, resp. barva kvantifikovatelná?

Ohodnoťte chutnost každého ovoce na škále 0-100.

NE

- Není dodrženo pořadí (single cancelation).

	zelené	žluté	červené
jablko	1	50	60
hruška	1	80	80
třešeň	1	10	100

CM: Příklad 2

„Ovocná škála“.

- třešně > hrušky > jablka
- červené > žluté > zelené

Je druh ovoce, resp. barva kvantifikovatelná?

Ohodnoťte chutnost každého ovoce na škále 0-100.

NE

- dodrženo double cancelation.

	zelené	žluté	červené
jablko	1	3	4
hruška	2	10	20
třešeň	8	11	100

CM: Příklad 2

„Ovocná škála“.

- třešně > hrušky > jablka
- červené > žluté > zelené

Je druh ovoce, resp. barva kvantifikovatelná?

Ohodnoťte chutnost každého ovoce na škále 0-100.

ANO?

- Existuje $3! \times 3! = 36$ možností dvojitého vykrácení. 30 z nich platí automaticky v případě jednoduchého vykrácení a pokud z těch 6 platí jediná, pak platí všech 6.
- Jedna platí.

	zelené	žluté	červené
jablko	1	3	19
hruška	2	10	20
třešeň	8	11	100

Ne!

- Krácení není jedinými předpoklady v případě, kdy obsahem tabulky není jen řada přirozených čísel (prosté pořadí).

CM: Další podmínky

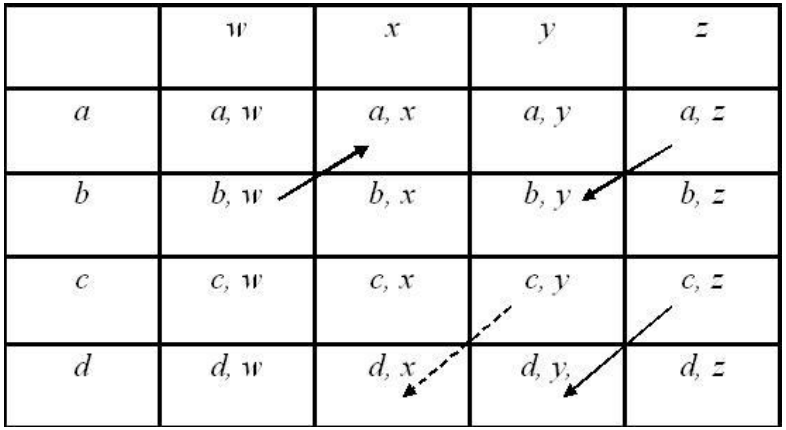
Jednoduché a dvojité vykrácení nestačí plně pro kvantifikaci.

Řešitelnost:

- Pokud známe tři ze čtyř úrovni a, b, x, y, lze čtvrtou dopočítat tak, aby $(a, x) \sim (b, y)$.
- Jinými slovy: Každá úroveň P má hodnotu jak z X, tak z A.

Archimedovská podmínka:

- „Hodnoty jsou rovnoměrně rozprostřeny.“
- „To vše platí až do nekonečna.“
- Neexistují příliš malé či příliš velké hodnoty, které by už nešlo srovnávat.



Jinými slovy obsahem tabulky (*P*) musí být:

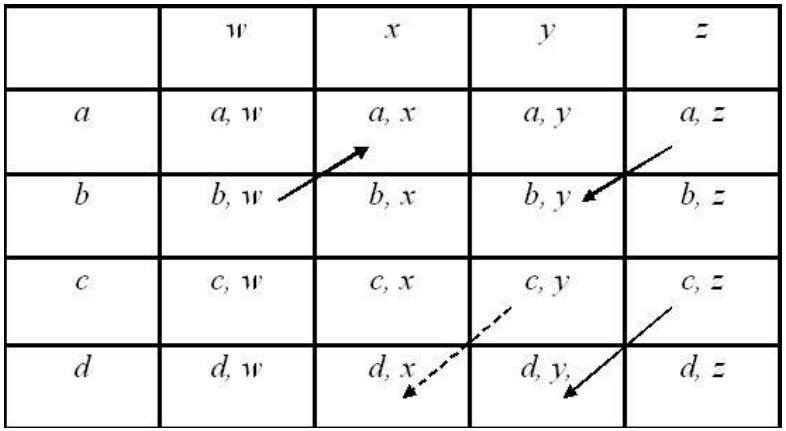
- Prostá pořadí bez vynechání (tedy přirozená čísla, pěkná škála),
- Dobře definovaná „škála“.

CM: Posloupnost kancelací

Bohužel, řešitelnost a archimedovská podmínka není přímo testovatelná.

Lze ale řešit nepřímo pomocí „posloupnosti kancelací“.

- Pokud A=X=3, stačí dvojité krácení pro důkaz spojitého měření.
- Pokud A=X=4, je nezbytné trojité krácení.
- („Tranzitivita“ rozdílů).



CM: vztah k psychologii

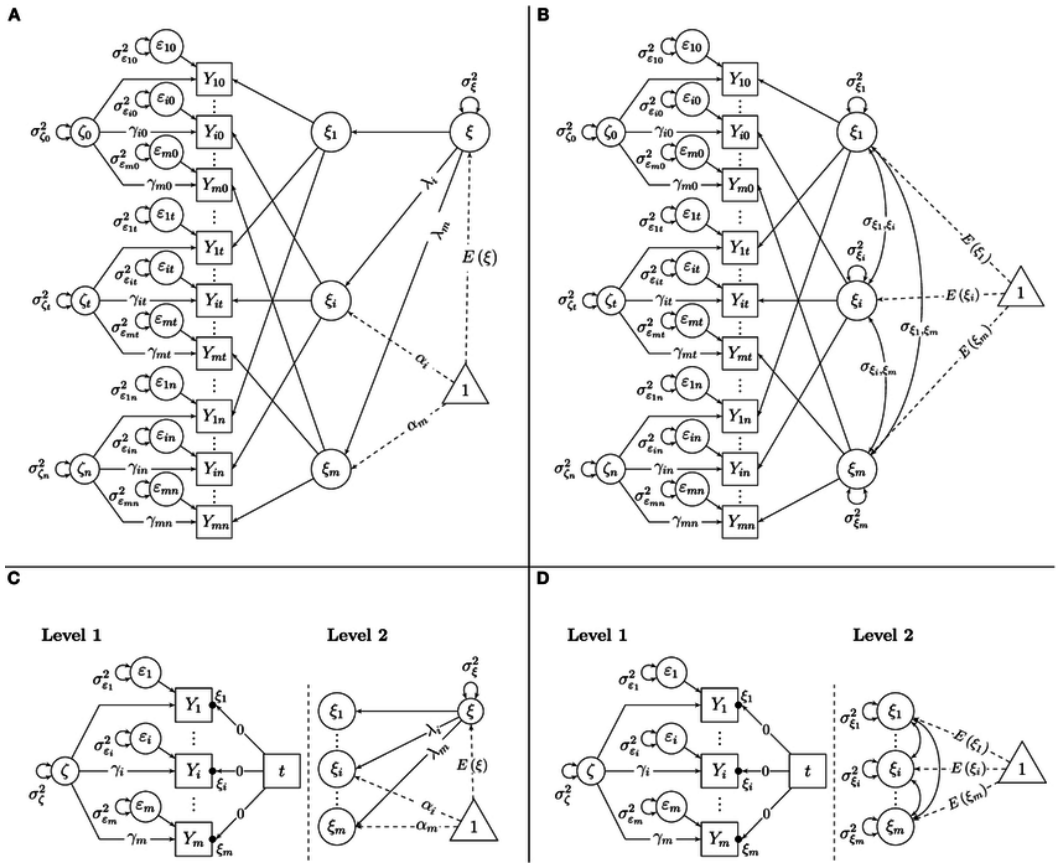
Pokud A jsou například respondenti (které lze ordinálně seřadit podle míry schopnosti) a X položky (řaditelné podle obtížnosti), lze použít spojité měření.

Pak by každý člověk měl být umístitelný na nějakou škálu obtížností položek a naopak.

- Šikovnější člověk vyřeší obtížnější příklady než méně šikovný člověk.
- Lehčí položku vyřeší i lehčí respondenti.
- Implicitní předpoklad Guttmanovy škály.

Kvůli chybě měření (na úrovni položky) však přímo neplatí a platit nemůže.

Modely s latentními rysy



[Random článek.]

Modely s latentními rysy

Předpokládají, že existuje latentní, nepozorovaný rys, který kauzálně „způsobuje“ pozorované chování (odpovědi v dotazníku/testu).

Vychází z realismu; proměnná musí existovat, aby mohla něco způsobovat (ale předmětem diskuze).

Příkladem modelu může být stará dobrá faktorová analýza: latentní faktor je lineárně spjatý s pozorovanými proměnnými.

- Hahaha, smáli jsme se už hodně dávno (omezené rozpětí pozorovaných vs. neomezené rozpětí latentních proměnných).

Navíc víme, že je to celé složitější.

- Položkou škály depresivity je špatný spánek. Léky zkvalitňující spánek však působí na ostatní indikátory depresivity při retestu.
- Vztahy proměnných jsou komplexnější, při jediném měření ale může jít o vhodné zjednodušení.

Modely s latentními rysy

	faktorová analýza	ordinální faktorová analýza	latent class analysis	IRT a Raschův model
latentní proměnná (prediktor)	intervalová	intervalová	nominální	intervalová
manifestní proměnná (závislá)	intervalová	ordinální	nominální, ordinální, intervalová	nominální, ordinální
vztah	lineární	komplikovaný ☺ (lineární s probit. SC)	lineární	logistický

Ordinální faktorová analýza

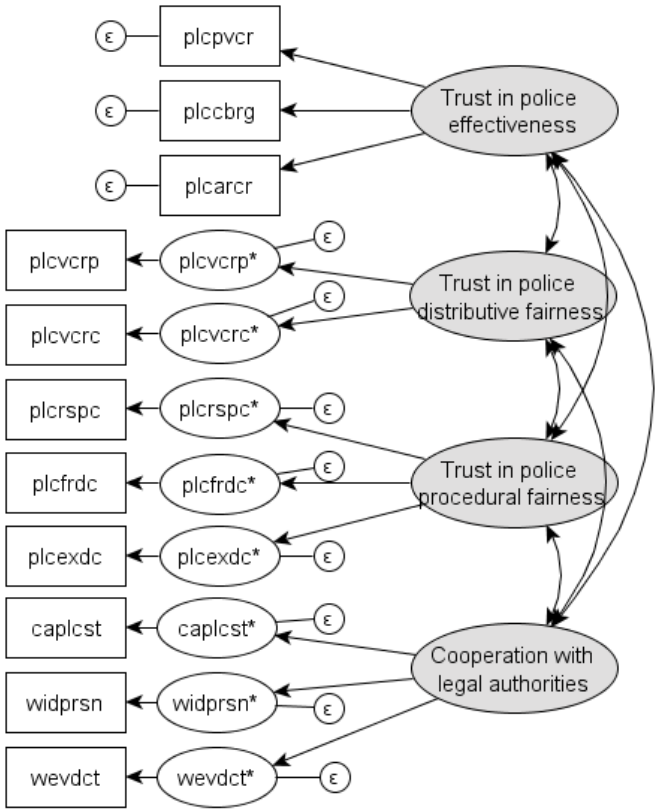
Odhad na polychorické korelační matici.

- Jaká je korelace dvou spojitých intervalových proměnných, z nichž pozorujeme jen určité „rozmezí“?

Latentní proměnná nepredikuje přímo manifestní, ale tzv. „item latent response“.

- Ta se pomocí tzv. „skórovací funkce“ manifestuje v ordinálních odpověďových kategoriích.

2PL IRT model s binární odpovědí ekvivalentní nCFA; u delší odpověďové škály jen obdobný.



Raschův model

Benjamin Wright

s fotografií George Rasche

Vývoj teorií odpovědi na položku

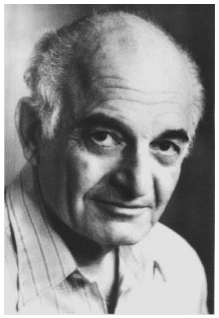
50. a 60. léta, další rozvoj v 80. letech (počítače).

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

Jde o stochastickou úpravu původně deterministického Guttmanova modelu.

- Přelom zejm. Rasch (1960).

Paul Felix Lazarsfeld (1901–1976)



Louis Guttman (1916–1987)

Řada různých modelů, např.:

- počtu parametrů +
  Raschův model: 1PL–3PL (a 4/5 PL)
- binární/ordinální/nominální: GRM, RSM, PCM, GPCM, NRM
- jedno- vs. multidimenzionální: kompenzatorní vs. nonkompenzatorní
- další rozšíření: IRTree, modely pro „VAŠ“, modely pro ipsativní položky a další s volným přelivem do kognitivního modelování, modely zahrnující čas...
- Rasch, Novick, Lord, Andrich, Samejima, Hambleton, De Boeck, van der Linden a další

Vztah rysu a odpovědi

Jak, faktorová analýza, tak CTT předpokládají lineární vztah mezi pozorovaným skóre v testu (položce) a měřenou veličinou.

- $X_{ip} = \tau_i + \lambda_i \theta_p + \varepsilon_{ip}$

To není realistické.

- Při určitých konstelací úrovně respondenta a obtížnosti položky to může vést k predikcím mimo možný rozsah pozorovaných skórů.
- Predikovaný skór je zpravidla reálné číslo; možné pozorované skóry jsou ale na úrovni položky zpravidla celá čísla (0/1, 0-1-2-3).
- V případě binární položky by predikovaný skór z intervalu <0 ; 1> bylo možné chápat jako pravděpodobnost správné odpovědi. V takovém případě by ale neměl být vztah pravděpodobnosti a schopnosti lineární.
  - Vztah predikovaných hodnot a pravého skóre byl proto řešen dávno před vznikem IRT.

Vztah rysu a odpovědi

Předpoklady základního Raschova modelu:

- Existuje spojitý, intervalový latentní rys, který „způsobuje“ pozorované binární odpovědi.
- Tyto odpovědi záleží dále na parametrech položky.
- Odpověď lze predikovat prostřednictvím tzv. **charakteristické funkce položky**.
- Pozorované odpovědi jsou navzájem lokálně nezávislé (jsou způsobeny výhradně úrovní latentního rysu, parametry položek a náhodné chyby).
  - To nemusí být pravda u vícedimenzionálního IRT.

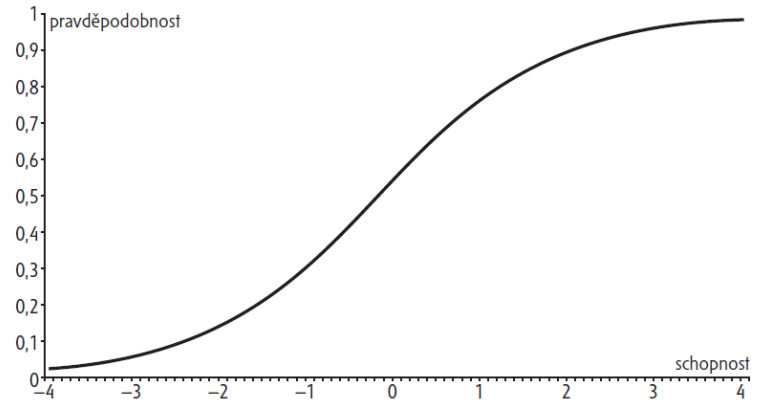
Jaký je vztah spojitého intervalového prediktoru a binárné závislé proměnné?

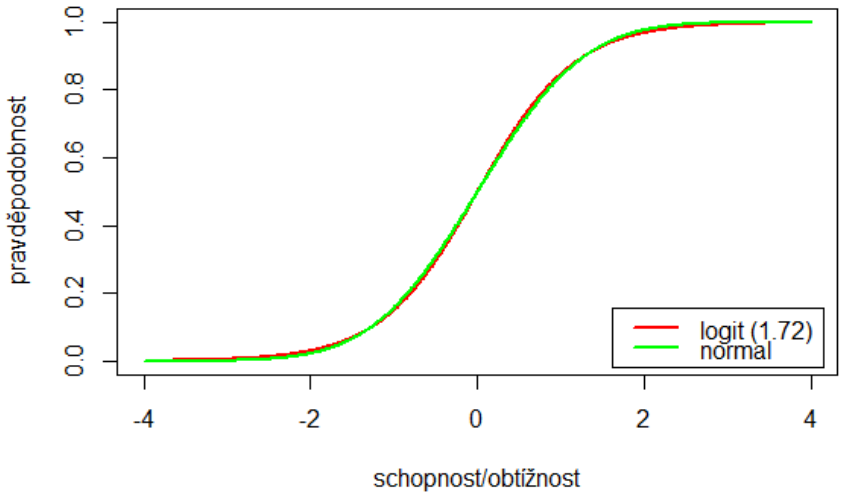
- Jakou analýzu byste použili?

Vztah rysu a odpovědi

Rasch navrhl normálně rozdělenou kumulativní distribuční funkci (tzv. „ogiva“).

- Dnes se označuje jako tzv. probit model.
- Častěji se používá logit model, který je prakticky identický, ale používá logistickou funkci.
  - Má řadu výhod, lépe se derivuje, atd.
    - $\Phi(x) = \frac{1}{\sqrt{2x}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt$
  - $\Phi(x) \sim \frac{e^{1,72x}}{1+e^{1,72x}} = \frac{1}{1+e^{-1,72x}}$
  - kde $\Phi(x)$ je kumulativní normální distribuce



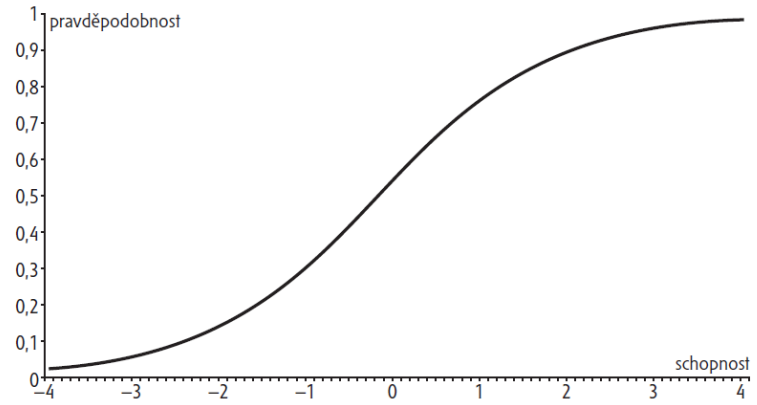


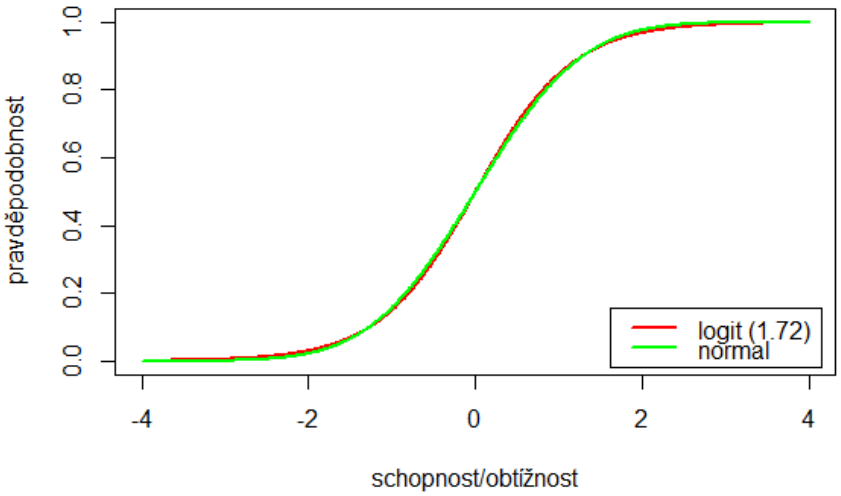
Vztah rysu a odpovědi

Funkce, která popisuje pravděpodobnost (správné) odpovědi, se nazývá:

- **charakteristická křivka položky**
- item characteristic curve (**ICC**)
- item response function (IRF)
  - Technicky vzato jde o funkci; křivka je jen její zobrazení.
- V případě polytomních IRT modelů se používá ještě označení scoring function.

A jednotlivé IRT modely se odlišují právě touto ICC.





Charakteristická funkce položky

Charakteristická funkce Raschova modelu:

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

- θ – míra latentního rysu daného člověka (správně by měla být notace θ_p, tedy míra rysu pro osobu p, ale zjednodušuji to).
- b_i – tzv. parametr obtížnosti; obtížnost položky i.
- $P_i(\theta)$ – pravděpodobnost správné odpovědi na položku i při úrovni latentního rysu θ. Pravděpodobnost špatné odpovědi je $Q_i(\theta) = 1 - P_i(\theta)$.

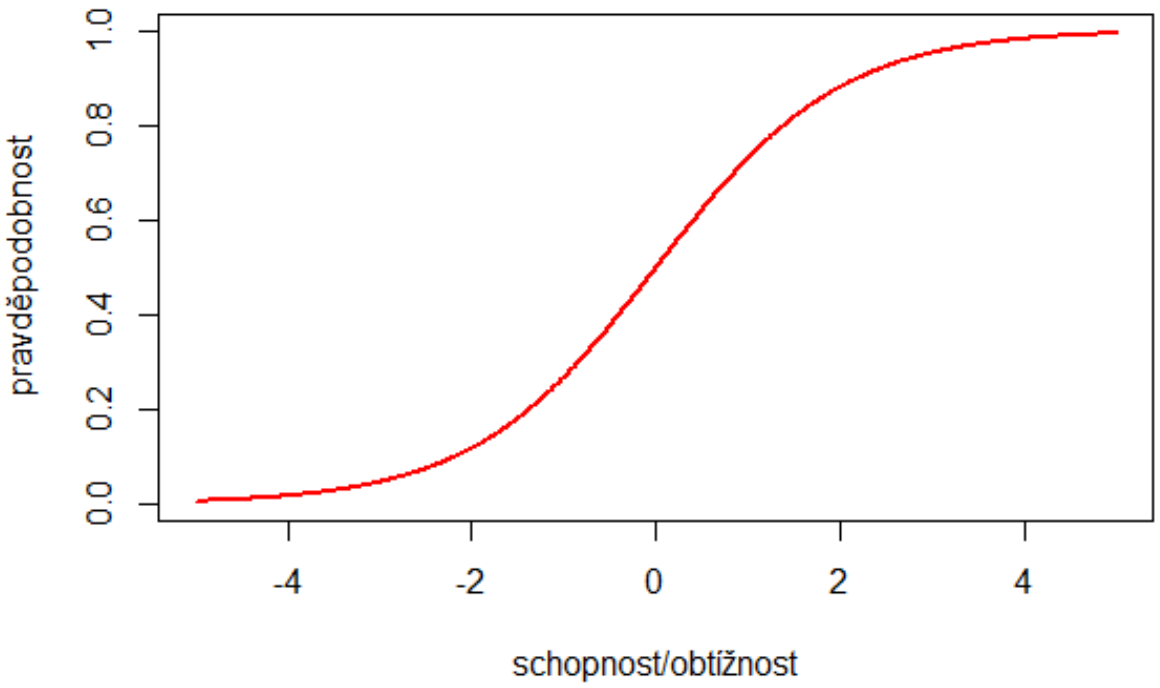
Lze upravit na $\ln \frac{P_i(\theta)}{1-P_i(\theta)} = \theta - b_i$.

- Interpretace: logaritmus šance (log-odds, viz logistická regrese!) je roven rozdílu schopnosti člověka a obtížnosti položky.

Pravděpodobnost není lineární. Log-odds ji linearizuje.

$\theta - b_i$	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%

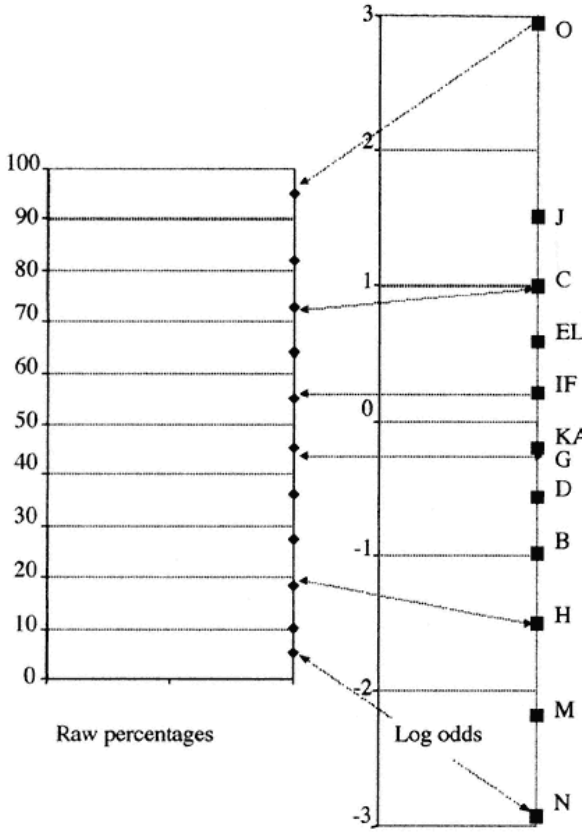
Charakteristická funkce položky



$\theta - b_i$	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%

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

Charakteristická funkce položky



Stejná logika odhadu lze použít pro skóre celého test.

- Vpravo je skór osob/položek jednoduše $\ln \frac{P_i(\theta)}{1-P_i(\theta)}$.
- Toho využívá tzv. PROX estimator.
- Jde o iterativní postup, který updatuje parametry podle algoritmu:
- $\mathrm{E}(\theta) = ln \frac{X-\min(x)}{\max(x)-X} \sqrt{1+\frac{\sigma_{diff}^2}{2{,}9}}$
  - první část je logaritmus šance, protože jde o podíl správné ku chybné části testu; pod odmocninou je korekce, kde σ_{diff}^2 jde rozptyl obtížností položek, aby to nekonvergovalo k nule.
  - Iterativně se opakuje nastřídačku pro položky a osoby.

RM vs. spojité měření

	Anička (-2)	Béďa (-0,5)	Cyril (0)	Draha (1,5)
W (-1)	0,27	0,62	0,73	0,92
X (0,5)	0,08	0,27	0,38	0,73
Y (2)	0,02	0,08	0,12	0,38
Z (3)	0,01	0,03	0,05	0,18

	Anička (-2)	Béďa (-0,5)	Cyril (0)	Draha (1,5)
W (-1)	-1	0,5	1	2,5
X (0,5)	-2,5	-1	-0,5	1
Y (2)	-4	-2,5	-2	-0,5
Z (3)	-5	-3,5	-3	-1,5

Máme čtyři položky W-Z a čtyři osoby A-D. Pomocí RM každé byl odhadnut skór (rys, obtížnost).

- Nahoře: pravděpodobnost správné odpovědi $P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}}$.
- Dole: prostý rozdíl $\theta - b_i$ na spojité intervalové škále.

Double cancelation (-!):

- (B-W)+(C-X)>(A-X)+(B-Y) = C-W>A-Y
- (B-W)-(A-X) = 3; (C-X)-(B-Y) = 2

Další podmínky

- 3+2 = 5 = (W-C)-(A-Y)

Proč jsme u délky násobili/dělili a zde sčítáme/odčítáme? ☺

Charakteristická funkce testu

Výhodou Raschova modelu je fakt, že je „plně identifikován“.

- Každému hrubému skóre odpovídá právě jeden odhad latentního skóre.

Lze proto definovat charakteristickou křivku testu.

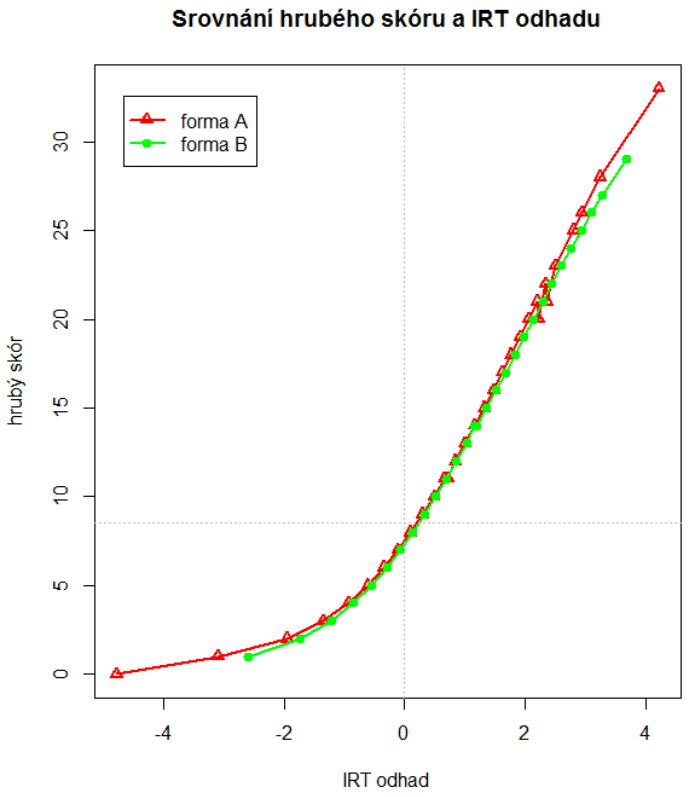
- test characteristic curve (TCC)

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

- Očekávaný hrubý skór podle míry latentního rysu (odhad pravého skóre v CTT).

Využívá se při skórování testu.

- Součet položek nese „všechny“ informace o latentním rysu.



Informační funkce položky

Doteď jsme mluvili o vztahu latentního rysu a pravděpodobnosti (správné) odpovědi.

Jaká je ale těsnost tohoto vztahu?

Odpovědí na tuto otázku je **informační funkce položky** $I_i(\theta)$ (item information function/curve). Pro dichotomické pol.:

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

- Pro každou úroveň schopnosti jiná.
- $P_i(\theta)$… pravděpodobnost správné odpovědi při úrovni θ schopnosti respondenta (tzv. pravděpodobnostní funkce, viz modely dříve).
- P_i` … první derivace této pravděpodobnosti
- $1 - P_i(\theta)$ … je pravděpodobnost jiné, než správné odpovědi.
  - Ve jmenovateli je tedy rozptyl hrubého skóru. Proč?

Informační funkce položky

Raschův model snadno derivuje.

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

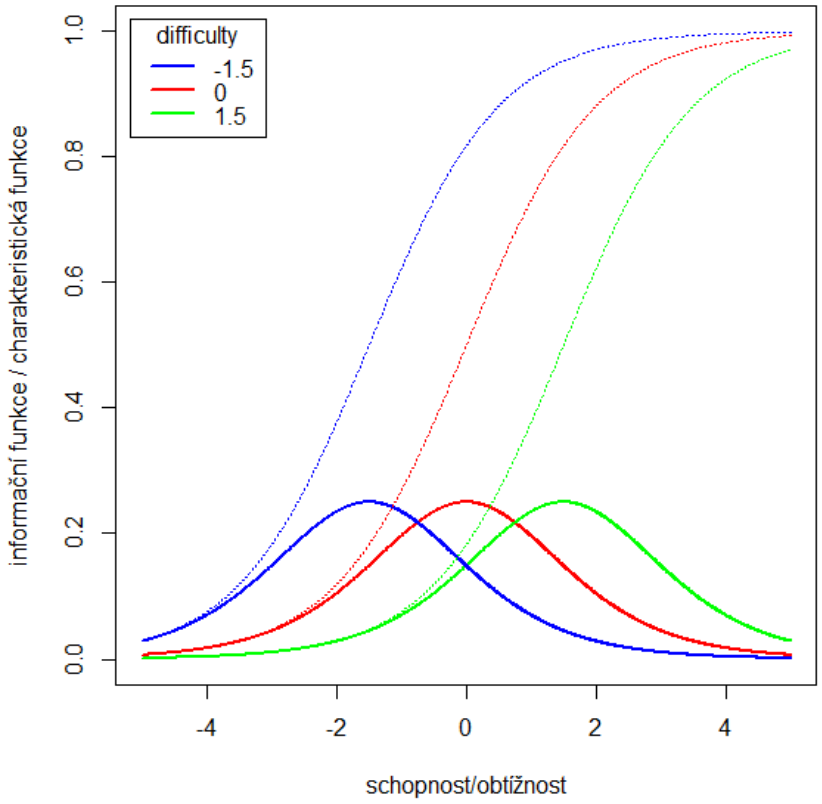
ICC Lze tedy zjednodušit:

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

- Informační funkce je tedy přímo rovna rozptylu predikovaného pravého skóru na položce.

Maximum je vždy tam, kde je položka nejstrmější – a to je v bodě obtížnosti položky.

- V RM tedy:
  - $I_i(\theta = b_i) = 0{,}5(1 - 0{,}5) = 0{,}25.$



Informační funkce testu

Informační funkce testu je součtem informačních funkcí položek:

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

Informační funkce položek/testu je reciprokou („převrácenou“) funkcí k chybovému rozptylu: $\sigma^2_{e,\theta} = \frac{1}{I(\theta)}$.

Z toho důvodu standardní chyba měření je:

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

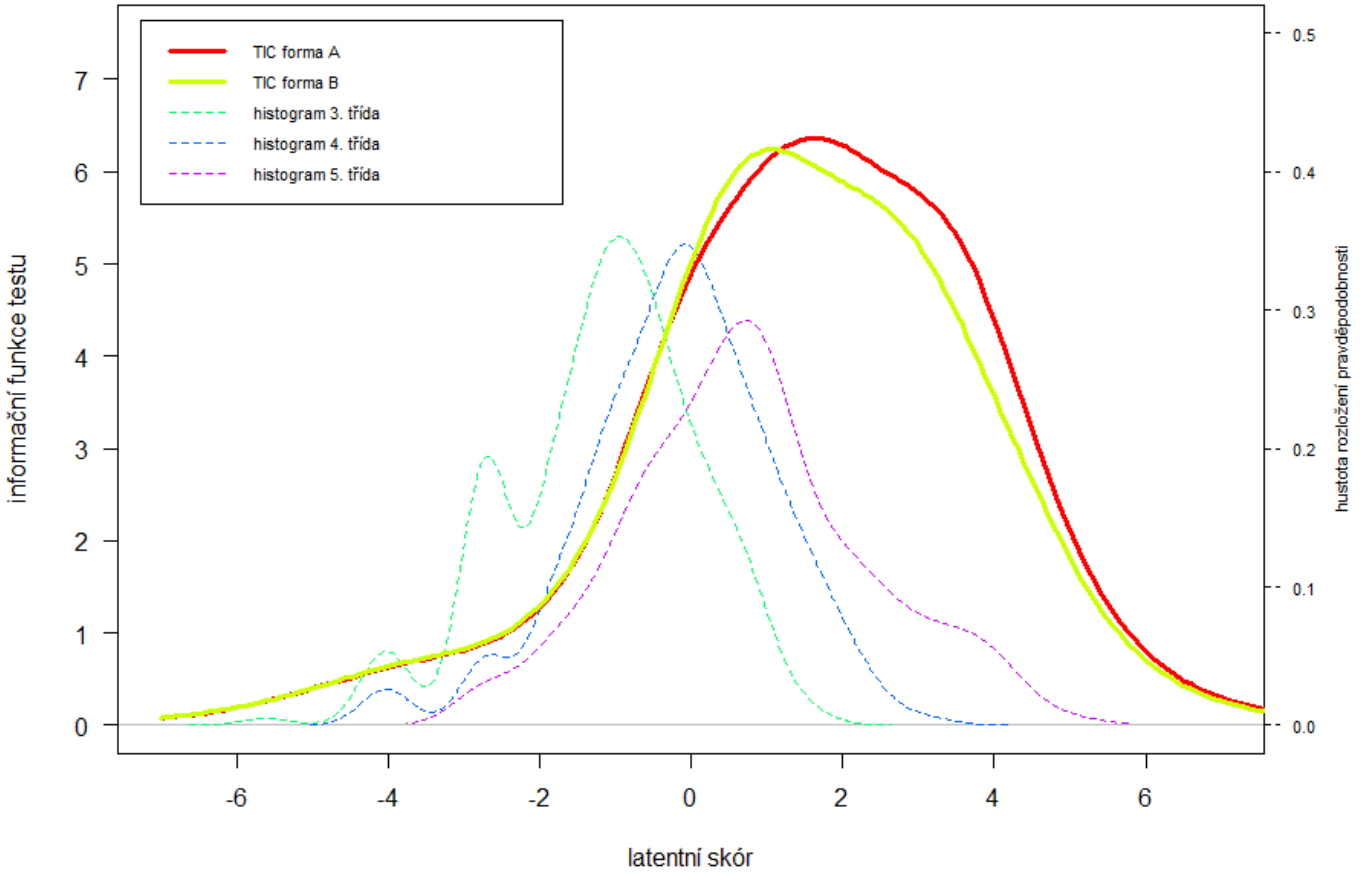
- Tedy čím vyšší informace, tím menší chyba měření.

Interval spolehlivosti potom získáme vynásobením kvantilem normálního rozdělení (stejně, jako v CTT):

$$CI_{95\%}(\hat{\theta}) = \theta \pm z \cdot SE_{\hat{\theta}}$$

- Jde nicméně o chybu latentního rysu, nikoliv jeho odhadu (CI kolem pravého vs. pozorovaného skóre. Reálně se proto používají různé bootstrapové techniky.

Informační funkce testu



Reliabilita v IRT

Definice reliability je v IRT naprosto stejná, jako v CTT, tedy virtuální korelace paralelních testů. Bohužel je porušen předpoklad homoskedascity, protože každá úroveň rysu má jinou chybu měření.

Lepší je tedy o reliabilitě uvažovat jako o rozptylu pozorovaných odhadů vysvětleném latentními rysy.

- $r_{xx'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2} = \frac{\sigma_X^2-\sigma_e^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$

Bohužel nemáme σ_e^2. Zde se používá „průměrná“ chyba měření, tzv. root mean-square error (RMSE):

$$\sigma_e = RMSE = \sqrt{\frac{\sum_{p=1}^{N} SE_p^2}{N}}$$

Reliabilita v IRT

Po dosazení:

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

V případě Raschova modelu a tzv. JMLE estimátoru.

- Jiné estimátory používají jiné odhady latentních rysů, například EAP (expected a-posteriori estimates) atd.; pak se rozptyl těchto odhadů dosazuje za rozptyl pravých σ_T^2, nikoliv pozorovaných σ_X^2 skórů.

Tzv. empirický odhad reliability: za σ_X^2 je dosazen pozorovaný rozptyl latentních rysů.

- Většina IRT estimátorů má předpoklad normálního rozdělení, proto se občas používá tzv. marginální odhad reliability, kam se za σ_X^2 (resp. σ_T^2) dosazuje apriorní rozptyl, se kterým estimátor počítal, zpravidla 1.

Stejně jako jiné odhady vnitřní konzistence je to „spodní mez“ reliability.

- Ne vždy!

Lokální reliabilita

Daniel (1999) navrhl používat tzv. lokální reliabilitu: odpověď na otázku, jaká by byla reliabilita testu, když by pro všechny respondenty/skupiny měřila jako pro daného respondenta/skupinu.

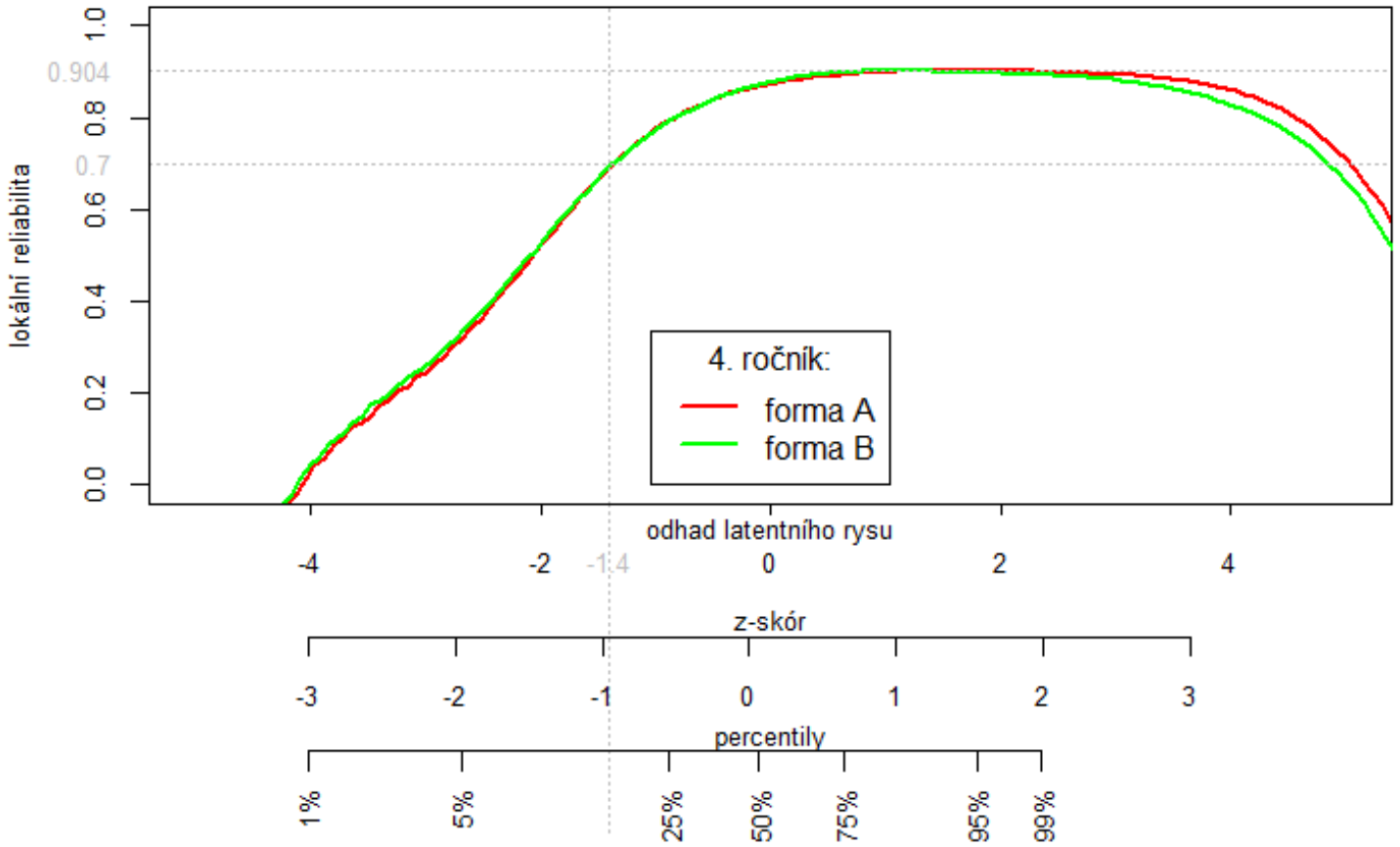
Namísto RMSE se do vzorce výše dosadí chyba měření daného respondenta, chyba pro danou úroveň skórů, RMSE dané skupiny atp.:

$$r_{xx'}(\theta \in M) = 1 - \frac{RMSE^2(\theta \in M)}{\sigma^2_{E(\theta)}}$$

Celková reliabilita je pak váženým průměrem všech možných lokálních reliabilit (Cígler, 2017☺):

- Pro 2 skupiny: $r_{xx'} = \frac{ar_{aa'} + br_{bb'}}{a+b}$, kde *a* a *b* jsou počty respondentů.
- Pro N skupin: $r_{xx'} = \frac{\sum_{i=1}^{N} n_i r_{ii'}}{\sum_{i=1}^{N} n_i}$ a $r_{ii'}$ je reliabilita ve skupině *i*.

Lokální reliabilita



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 dat s modelem

I když si to Linacre a Bond s Foxovou nemyslí ☺ , předpokladem toho, aby Raschův model byl fundamentálním měření, je potřeba, aby dobře popsal data.

Respektive aby data dobře vyhovoval Raschovu modelu.

Shoda dat na úrovni:

- Položky / respondenta
- Celého modelu

Shoda dat s RM: položky

Jak moc dobře pozorovaný pattern odpovědí (1101100100…) odpovídá predikovaným odpovědím (0,98; 0,84; 0,32; …)?

Východiskem je tzv. standardizované reziduum respondenta *p* na položku *i* (rozdíl predikované pravděpodobnosti a pozorované binární odpovědi, dělené chybou predikce):

$$z_{pi} = \frac{x_{pi} - P_{pi}}{\sqrt{I_{pi}}} = \frac{x_{pi} - P_{pi}}{\sqrt{P_{pi}\left(1 - P_{pi}\right)}}$$

Celková chyba v datech lze vyjádřit jako $\chi_i^2 = \sum_{p=1}^{N} z_{pi}^2$, které má chí rozdělení o N_i počtu stupňů volnosti (počet respondentů, kteří odpovídali na danou položku).

RM: Outfit

Prvním ukazatelem fitu je tzv. outfit, který se tradičně vyjadřuje dvěma způsoby.

Mean-square outfit: celková chyba dělená počtem stupňů volnosti (průměrná hodnota z-standardizovaného rezidua):

$$u_i = \frac{\sum_{p=1}^{N} z_{pi}^2}{N}$$

- Optimální fit je 1 (protože SD=1). Vyšší hodnotu značí nižší shodu s daty (underfit), nižší hodnoty pak tzv. guttmanovský patter – vyšší shodu s daty (overfit).

Mean-square nám neříká nic o signifikanci. Proto se převádí na tzv. **z-standardizovanou hodnotu**, tedy z-skór o stejné p-hodnotě jako původní chí s daným počtem stupňů volnosti.

- Provádí se buď analyticky, nebo empiricky.
- Nula znamená optimální fit, nižší overfit, vyšší underfit; hodnoty mimo rozsah -1,96–1,96 jsou ukazatelem neshody s daty na hladině $p < 0,05$.

RM: Infit

Tím, že outfit zvažuje všechny respondenty/položky stejně, je outfit náchylný na náhodný šťastný tip špatného respondenta, resp. na náhodné selhání dobrého respondenta.

Proto se používá infit, kde každý case je vážený hodnotou jeho informační funkce:

$$u_i = \frac{\sum_{p=1}^{N} z_{pi}^2 I_{pi}}{\sum_{p=1}^{N} I_{pi}} = \frac{\sum_{p=1}^{N} z_{pi}^2 \left[P_{pi}\left(1 - P_{pi}\right)\right]}{\sum_{p=1}^{N} P_{pi}\left(1 - P_{pi}\right)}$$

- Jde tedy o vážený průměr standardizovaného rezidua.

Tento mean-square se převádí na z-standardizovanou hodnotu stejně, jako v případě outfitu.

Interpretace fitu položky

Ukazatel, jak položka/respondent odpovídá Raschovu modelu.

- Položky: Odpovídali respondenti na položku dle předpokladu?
- Respondenti: Odpovídal respondent na položky dle předpokladu?
- Je založená na průměru sumy čtverců standardizovaných reziduí probanda/položky s df=n-1.
- Pozor: vysoká hodnota se neintuitivně označuje jako „underfit“, nízká „overfit“!

Vysoká hodnota (underfit): respondent/i odpovídal/i více náhodně.

- Méně „guttmanovská“ škála, než jsme předpokládali.

Nízká hodnota (overfit): respondent/i odpovídal/i méně náhodně.

- Více „guttmanovská“ škála, než jsme předpokládali.

Fit položek je základem položkové analýzy v RM.

Interpretace fitu položky

Příklad:

◦ obtížnost položek:	snadné ..... střední ..... těžké.
◦ stochastická předpověď (průměrný fit):	111...1101100100...000.
◦ deterministická odpověď (overfit):	111...**1111100000**...000.
◦ nahodilá odpověď: (underfit):	1**0**1...1010101010...0**1**0.
◦ šťatný tip (vliv na outfit):	111...1101100100...00**1**.
◦ nepozornost (vliv na outfit):	**0**11...1101100100...000.
◦ náhodná znalost (vliv na infit):	111...11011**11**100...000.

Využití infitu: Korigovaná reliabilita

Modelová reliabilita, kterou jsme si ukázali, je „unbiased“ pouze tehdy, pokud model popisuje data dobře.

- Například v případě porušení lokální nezávislosti přestává být spodní mezí stejně, jako Cronbachovo alfa.

Proto se občas využívá tzv. „real reliability“, která koriguje oproti neshodě s daty funkcí $\max(1; u_p)$, kde u_p je infit respondenta *p*:

$$RMSE_{korig.} = \sqrt{\sum_{i=1}^{n} \sigma_e^2(\theta_i) \max(1; u_i)} \,,$$

- Korigujeme tedy je underfitující respondenty; overfitující fit nezlepšují.
- RMSE se dosadí do výpočtu reliability úplně stejně jako u nekorigované rel.

Shoda dat s RM: model

Často nás ale zajímá, jak data jako celek vyhovovala RM.

Výstupem z ML estimátoru je tzv. log-likelihood estimační funkce (alternativně pak suma všech standardizovaných reziduí v modelu).

- Ten má přibližně chí rozdělení.

Počet stupňů volnosti:

$$df = N_i N_p - NA - \left[N_i + N_p - 1 + \sum_{j=1}^{N_c} (N_j - 2) \right]$$

- N_i, N_p počet položek, respondentů v modelu; *NA* – počet chybějících dat.
- V závorce počet tzv. „volných“, tj. odhadovaných parametrů.
- Ta suma platí pro polytomické položky, v binárním RM je 0.
  - (N_j je počet odpověďových kategorií v celkem N_c položkách s různou strukturou).

Shoda dat s RM: model

Tento výpočet je velmi striktní a stejně jako v CFA je výsledek zpravidla signifikantní.

Proto se používají jiné ukazatele fitu: CFI, TLI, RMSEA…

Výpočet z log-likelihood funkce je ale zkreslující a vede k odlišně interpertovatelným výsledkům oproti CFA.

- Ale používá se (např. Tennant a Pallant, 2012).

Proto se ukazatele počítají na základě korelační matice standardizovaných reziduí.

- Maydeu-Olivares, Cai a Hernández (2011) a jejich vytuněné M2 a M2* ukazatele (Maydeu-Olivares a Joe, 2006).

A další analýzy nad reziduální korelační maticí (PCA…).