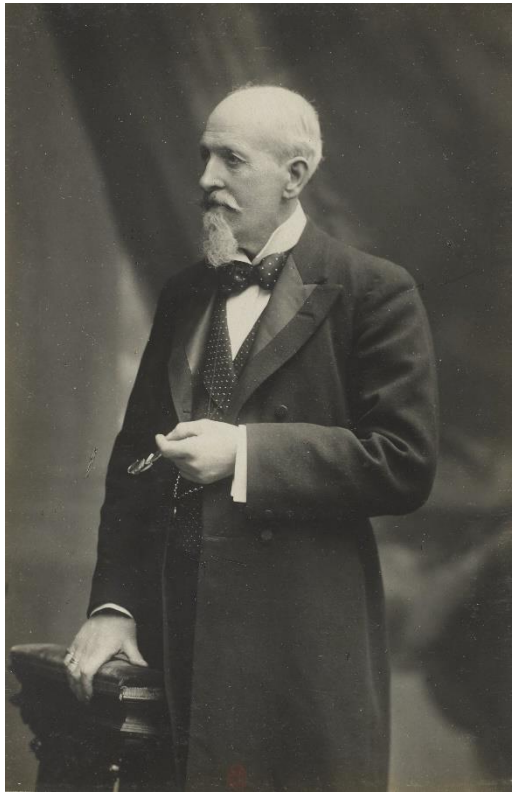
# Reliabilita
# Klasická testová teorie



PSYb2590: Základy psychometriky | **Přednáška 2**

28. 2. 2022 | Hynek Cígler

# Cíle přednášky (a semináře)

Hodnocení „přesnosti“ měření psychologického testu skrze reliabilitu.

Interpretace modelu měření klasické testové teorie.

Postupy různých odhadů reliability.

Pochopení důsledků (ne)reliability na praktické použití testu.

Práce s chybou měření při praktickém použití testu (seminář).

# Chcete měřit výšku postavy

Jak poznáte, že měření výšky je „dobré“?

- A co to znamená „dobré měření“?

Co jsou všechny možné zdroje chyby měření?

Jak můžete nepřesnost měření vyjádřit?

Jakým způsobem můžete měření zpřesnit?

Measuring children's height and weight, The University of Iowa, ca. 1920
https://www.flickr.com/photos/uiowa/8037651744

# 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“), čímž umožnil vznik CTT.

- Motivací byl odhad korelace nezkreslené chybou měření.

CTT „imituje“ opakované měření v přírodních vědách.

# Měření délky

*„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 délky $d$ pomocí „paralelních testů“:

- 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 $d$.
- Standardní chyba tohoto měření (SE, SEM, Standard Error of Measurement, $\sigma_e$):
  - 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, viz Statistika 1!).
  - $d$ = latentní proměnná, kterou měřím; $d_i$ = manifestní proměnná; $\mathrm{E}(d)$ = odhad latentní proměnné.
    - Někdy se pro odhad používá ještě symbol $\hat{d}$ (to proto, že odhad může být definovaný i jinak než průměrem).

# Měření délky

Analogie v sociálních vědách: N položek dotazníku?

- Chyba měření by měla být SD naměřených hodnot na N položkách ($SD/\sqrt{N}$).

Tento postup není dost dobře použitelný.

- Málokdy intervalové měření → předpoklady při výpočtu M i SD.
- Malý počet pozorování (položek) → nepřesný odhad SD.
- Velká míra chyby vzhledem k odlišnostem osob → nepřesný odhad SD vadí.
- Problém s extrémními hodnotami (nulová chyba při max./min. odpovědi).

Výsledek: Extrémně nepřesný odhad chyby.

Řešení: Přidání pár realistických předpokladů, které odhad zpřesní.

# Paralelní testy

Na konceptu paralelních testů Spearman založil koncept reliability.

- A na reliabilitě stojí zase CTT.

**Paralelní testy/měření jsou takové, pro které platí:**

- A. Pravý skór je ve všech testech a pro každý měřený subjekt stejný
  - $T = \mathrm{E}(X) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n}$.
- B. Rozptyl pravých skórů je v obou testech stejný (důsledek A).
- C. Chybový rozptyl je v obou testech a pro každý subjekt stejný.
- D. Shodný rozptyl pozorovaných skórů obou testů (důsledek A a C).
- Jinými slovy: *„Lidé se **nemění** a test měří pořád **‚stejně‘**.“*

Tyto předpoklady jsou v sociálních vědách zpravidla příliš striktní.

- Proto později budeme pracovat spíše s *„mírou paralelnosti“* (podrobněji přednáška o FA).

# Výlet do algebry

Mějme dvě náhodné, normálně rozložené proměnné $A$, $B$: $A \sim \mathrm{N}(\mu_A, \sigma_A^2)$ a $B \sim \mathrm{N}(\mu_B, \sigma_B^2)$. $\mu_A$, $\mu_B$ – průměry; $\sigma_A^2$, $\sigma_B^2$ – rozptyly.

Mějme proměnnou *C*, která je jejich součtem: $C = A + B$.

Potom platí, že $C \sim \mathrm{N}(\mu_A + \mu_B, \sigma_A^2 + \sigma_B^2 + 2cov_{AB})$,

- kde $cov_{AB} = r_{AB}\sigma_A\sigma_B$ je kovariance a $r_{AB}$ korelace.
- Pomůcka: $(a + b)^2 = a^2 + b^2 + 2ab$

Jinými slovy:

- Průměr součtu je součet průměrů: $\mu_{A+B} = \mu_A + \mu_B$
- Rozptyl součtu je součet rozptylů a 2× kovariance: $\sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + 2cov_{AB}$

# Výlet do algebry (simulace v Excelu)

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | A | B | C |
| 2 | | 2 | 3 | 5 |
| 3 | | 1 | 2 | 3 |
| 4 | | 4 | 3 | 7 |
| 5 | | 3 | 2 | 5 |
| 6 | | 1 | 2 | 3 |
| 7 | **M** | **2,2** | **2,4** | **4,6** |
| 8 | **SD** | **1,1662** | **0,4899** | **1,4967** |
| 9 | korelace | 0,5601 | | |
| 10 | E(C) | 4,6 | | |
| 11 | E(SD_C) | 1,4967 | | |

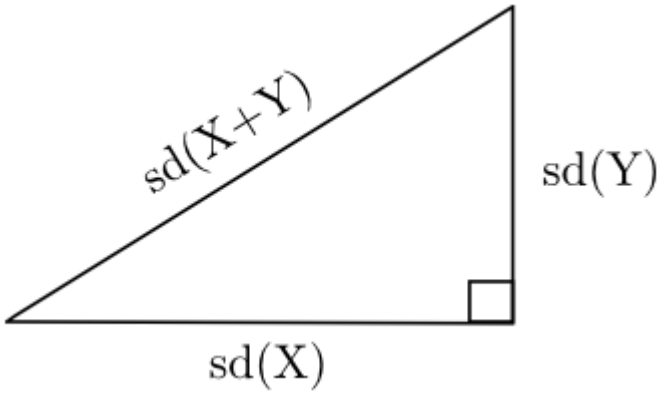
| | A | B | C | D |
|---|---|---|---|---|
| 1 | | A | B | C |
| 2 | | 2 | 3 | =SUMA(B2:C2) |
| 3 | | 1 | 2 | =SUMA(B3:C3) |
| 4 | | 4 | 3 | =SUMA(B4:C4) |
| 5 | | 3 | 2 | =SUMA(B5:C5) |
| 6 | | 1 | 2 | =SUMA(B6:C6) |
| 7 | **M** | **=PRŮMĚR(B2:B6)** | **=PRŮMĚR(C2:C6)** | **=PRŮMĚR(D2:D6)** |
| 8 | **SD** | **=SMODCH(B2:B6)** | **=SMODCH(C2:C6)** | **=SMODCH(D2:D6)** |
| 9 | korelace | =PEARSON(B2:B6;C2:C6) | | |
| 10 | E(C) | =B7+C7 | | |
| 11 | E(SD_C) | =ODMOCNINA(B8*B8+C8*C8+2*B9*B8*C8) | | |

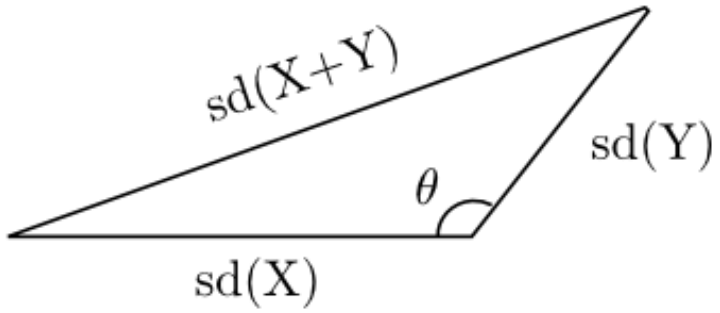
Viz studijní materiály.

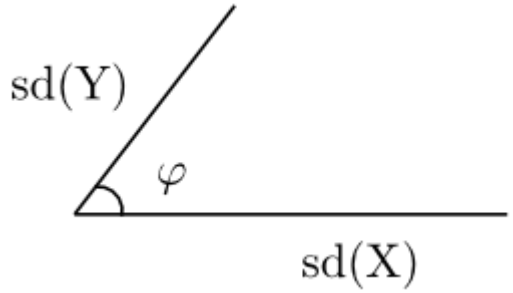
# Výlet do algebry (grafická ilustrace)

https://hynekcigler.shinyapps.io/covariance

- Využívá geometrického významu korelace: $\cos\varphi = \rho = r_{AB}$
- Jiný příklad: https://www.johndcook.com/blog/2010/06/17/covariance-and-law-of-cosines/



# Výlet do algebry

Lze zobecnit na korelační/kovarianční matice.

| $\Sigma_{AB}$ | A | B |
|---|---|---|
| A | $\sigma_A^2$ | $cov_{AB}$ |
| B | $cov_{AB}$ | $\sigma_B^2$ |

| $\Sigma_{ABC}$ | A | B | C |
|---|---|---|---|
| A | $\sigma_A^2$ | $cov_{AB}$ | $cov_{AC}$ |
| B | $cov_{AB}$ | $\sigma_B^2$ | $cov_{BC}$ |
| C | $cov_{AC}$ | $cov_{BC}$ | $\sigma_C^2$ |

$$\sigma_{A+B}^2 = \sigma_A^2 + \sigma_B^2 + 2cov_{AB} = \Sigma_{AB}$$

$$\sigma_{A+B+C}^2 = \sigma_A^2 + \sigma_B^2 + \sigma_c^2 + 2(cov_{AB} + cov_{AC} + cov_{BC}) = \Sigma_{ABC}$$

Vážený součet $C = w_A A + w_B B$: $C \sim \mathrm{N}(w_A\mu_A + w_A\mu_B, w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_A w_B cov_{AB})$

# Klasická testová teorie

Základní teorém CTT: Pozorovaný skór $X$ (manifestní proměnná) se skládá z pravého skóre $\tau$ a chyby měření $e$ (obě jsou latentní proměnné):

$$X = \tau + e$$

Chyba je nezávislá na měřeném.

- Jinak by nebyla chybou.
- $r_{\tau e} = 0$ , tedy korelace pravého skóre a chyby měření je nulová → $0 = 2r_{\tau e}\sigma_\tau\sigma_e$.

Chyba měření (i pravý skór) jsou normálně rozložené, průměr chyby $\mathrm{E}(e) = 0$.

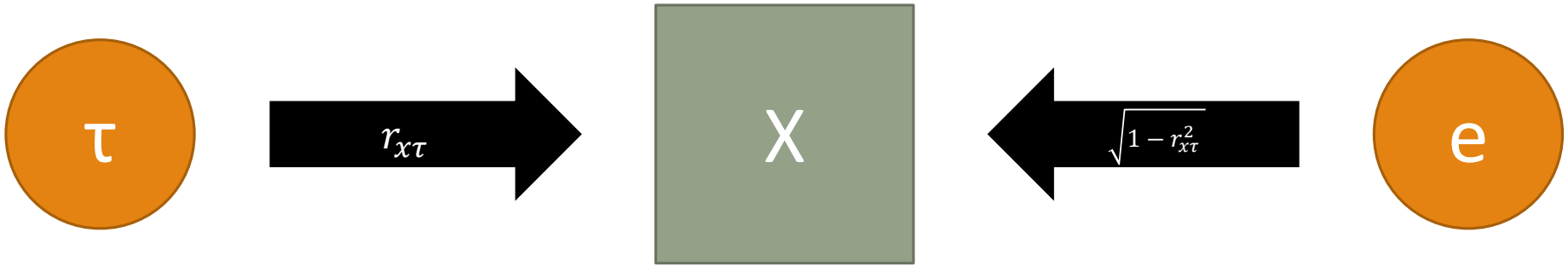
Uvedený vztah proto platí i pro rozptyly obou proměnných:

$$\sigma_X^2 = \sigma_\tau^2 + \sigma_e^2$$

# Klasická testová teorie

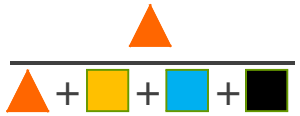
Základní teorém $X = \tau + e$ lze chápat jako lineární funkci.

Standardizovaný regresní koeficient je tedy roven korelaci prediktoru (τ) a závislé proměnné (X), tedy $r_{x\tau}$.




- Protože platí $1 = r_{x\tau}^2 + r_{e\tau}^2$ (celkový standardizovaný rozptyl, 1, je součtem rozptylů vysvětlených TS a chybou), korelace (standardizovaný regresní koeficient) chyby měření a OS je $r_{e\tau} = \sqrt{1 - r_{x\tau}^2}$.

# Reliabilita: metaforicky

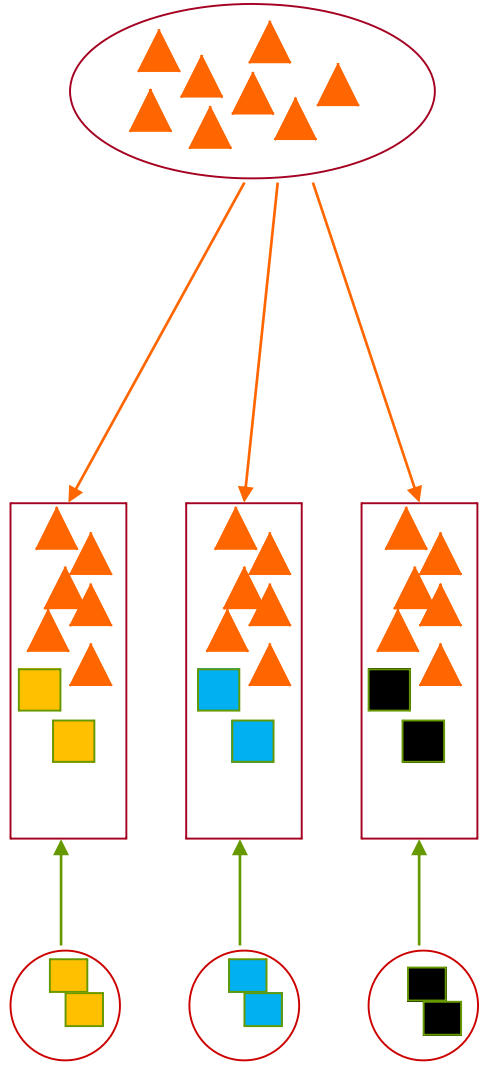
Reliabilita: podíl společného a celkového rozptylu: $r_{xx'} =$ 

*Čím více „společné variability“ sdílejí paralelní testy, tím vyšší je reliabilita jednoho každého z nich.*

Více společného rozptylu → vyšší reliabilita.

Více chybového (specifického) rozptylu → nižší reliabilita.

(pro notaci k obrázku viz S1)



# Reliabilita: technicky

Reliabilita je definovaná jako podíl rozptylu pozorovaného skóre (manifestní proměnné) vysvětleného pravým skóre (latentní proměnnou):

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

- Jaký je vztah korelace a vysvětleného rozptylu (lineární regrese...)?

Vysvětlený rozptyl je druhá mocnina korelace, tedy:

- $r_{xx'} = r_{x\tau}^2 = (R^2)$
- $\sqrt{r_{xx'}} = r_{x\tau} = (R)$
- Jinými slovy: reliabilita je umocněná korelace pravého a pozorovaného skóre.

**OK. Ale jak tedy zjistíme to $r_{x\tau}$ (korelaci OS a TS)?**

# Reliabilita

Lineární funkci můžeme otočit: pokud TS (true score) vysvětlí $r_{xx'}$ rozptylu OS (observed score), pak musí platit, že OS vysvětlí $r_{xx'}$ rozptylu TS.

Kolik rozptylu jednoho paralelního měření ($X_2$) vysvětlí jiné paralelní měření ($X_1$)?

- Pokud jsou chyby měření nezávislé?

$X_1 \rightarrow TS$: $r_{xx'}$ rozptylu. TS $\rightarrow X_1$: rovněž $r_{xx'}$ rozptylu.
Dohromady tedy $X_1 \rightarrow (\mathrm{TS}) \rightarrow X_2$ vysvětlí $r_{xx'} \cdot r_{xx'} = r_{xx'}^2$ rozptylu.

- To odpovídá korelaci $r_{xx'}$.

Pokud je reliabilita testu $r_{xx'}$ , pak korelace dvou paralelních měření bude rovněž $r_{xx'}$.

**Odhady reliability jsou proto založeny na korelaci paralelních testů.**

# 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 $x'$.

**Významy reliability:**

- Korelace paralelních testů.
- Vysvětlený rozptyl měření měřeným $\frac{\sigma_\tau^2}{\sigma_x^2}$.
- Relativní nepřítomnost chyby měření $1 - \frac{\sigma_e^2}{\sigma_x^2}$.

**Předpoklady:**

- Chyba měření je náhodná proměnná.
- Chyby měření paralelních testů navzájem nekorelují.
- OS se skládá výhradně z TS a chyby, neexistuje jiný systematický rozptyl (jinak tento další systematický rozptyl nelze odlišit od TS).
- Veškeré vztahy jsou lineární, proměnné jsou normálně rozložené.
- Homoskedascita vztahu TS a OS.

# Attenuation

Spearmanovou (1904) motivací byl odhad korelací pravých skórů nezkreslených chybou měření.

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

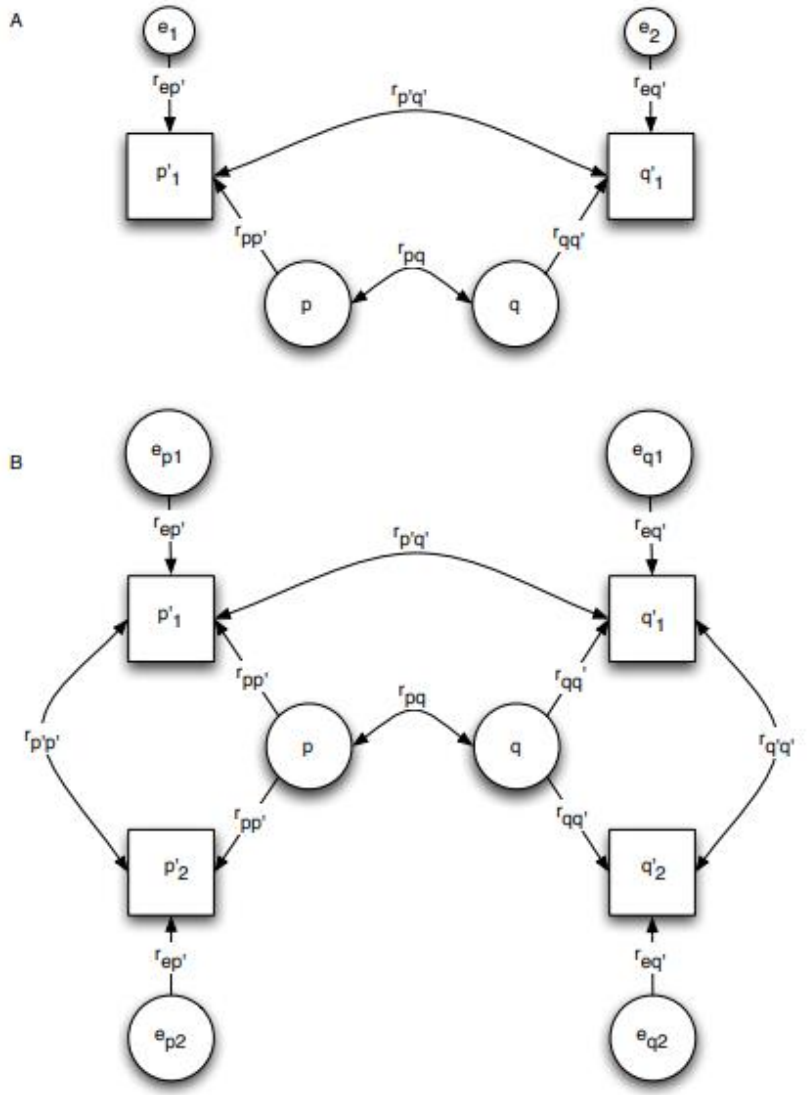
$$r^*_{pq} = \frac{r_{pq}}{\sqrt{r_{pp'} r_{qq'}}}$$

- Kde $r^*_{pq}$ je odhad korelace pravých skórů *p*, *q*, $r_{pq}$ 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 2 testů jako:

$$r_{pq} \leq \sqrt{r_{pp'} r_{qq'}}$$

- **Korelace nemůže být vyšší než odmocnina součinu reliabilit!**




(Pozor, notace na diagramu je atypická a neodpovídá rovnicím.)

**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.

https://www.personality-project.org/r/book/Chapter7.pdf

# Odhady reliability

Tradiční způsoby odhadu:

1. Stabilita v čase (test-retest)
2. Shoda posuzovatelů
3. Paralelní formy testu
4. Vnitřní konzistence

Lee Cronbach (1916–2001)
autor koeficientu alfa

# Reliabilita: typické postupy ověření v CTT

**Stabilita v čase, reliabilita typu test-retest**

- Měří test stále stejně? Paralelním testem (PT) je ten samý test administrovaný jindy.

**Shoda posuzovatelů, inter-rater reliabilita.**

- Docházejí administrátoři ke stejným závěrům? PT je stejný test administrovaný někým jiným.

**Reliabilita paralelních forem.**

- Měří obě/všechny formy testu to stejné? PT je jiný test vytvořený tak, aby „byl stejný“.

**Vnitřní konzistence a split-half**

- Měří položky to stejné? PT jsou jednotlivé položky/půlky testu.
- Cronbachovo alfa, split-half a další.

Lze čekat, že všechny koeficienty/odhady reliability budou stejné?

# Metoda test-retest

ODHAD RELIABILITY

# Stabilita v čase, test-retest reliabilita

**Poskytuje test při opakovaném měření shodné odhady atributu?**

Metoda: Korelace dvou měření (rank-order stability).

Předpoklady:

- Rys je (dostatečně) stabilní v čase.
- Měření jsou na sobě nezávislá. Zapamatování položek? Únava?

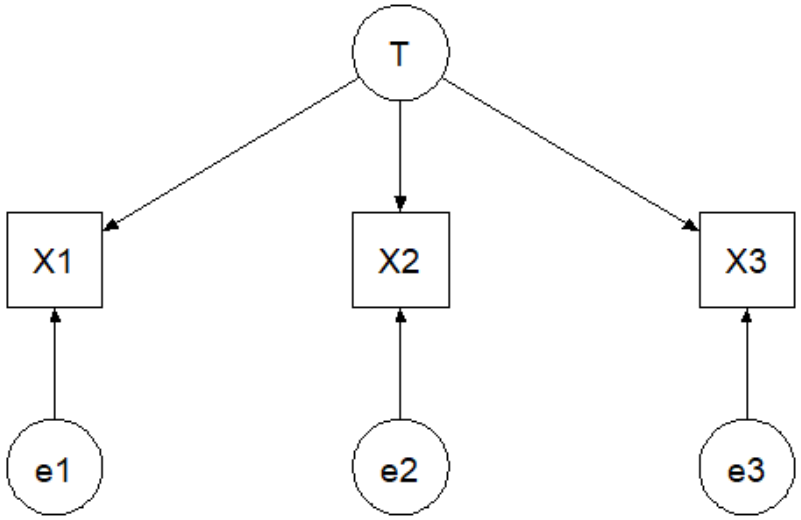
Problém: reálná fluktuace rysu v čase je považována za chybu měření.

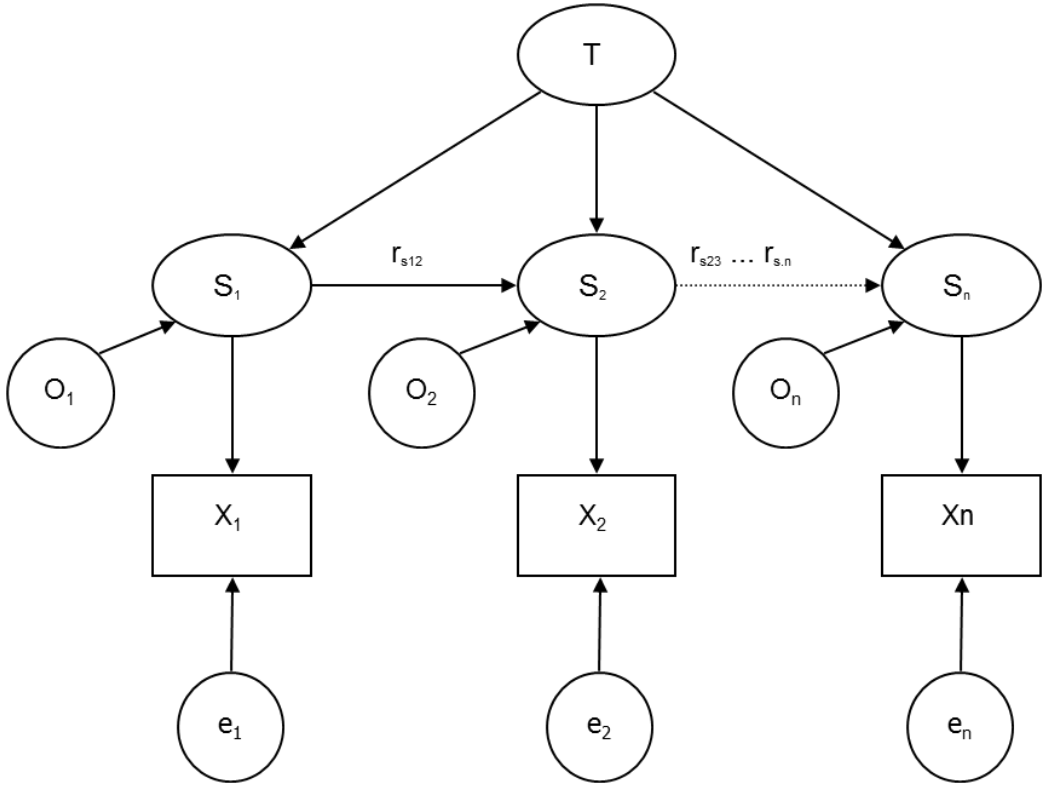
Stabilita rysu (korelace TS) vs. stabilita metody (korelace OS|TS).

Někdy se rozlišuje:

- **Dependabilita měření** – krátký interval, nepředpokládá se změna úrovně rysu.
- **Stabilita měření** – dlouhý  interval, zahrnuje přirozené rysu fluktuace rysu v čase.

# Test-retest vs. individuální rozdíly




Cole, D.A., Martin, N.C., & Steiger, J.H (2005). Empirical and conceptual problems with longitudinal trait-state models: introducing a trait-state-occasion mode. Psychological Methods 10(1), 3–20. https://doi.org/10.1037/1082-989X.10.1.3

# Test-retest vs. individuální rozdíly

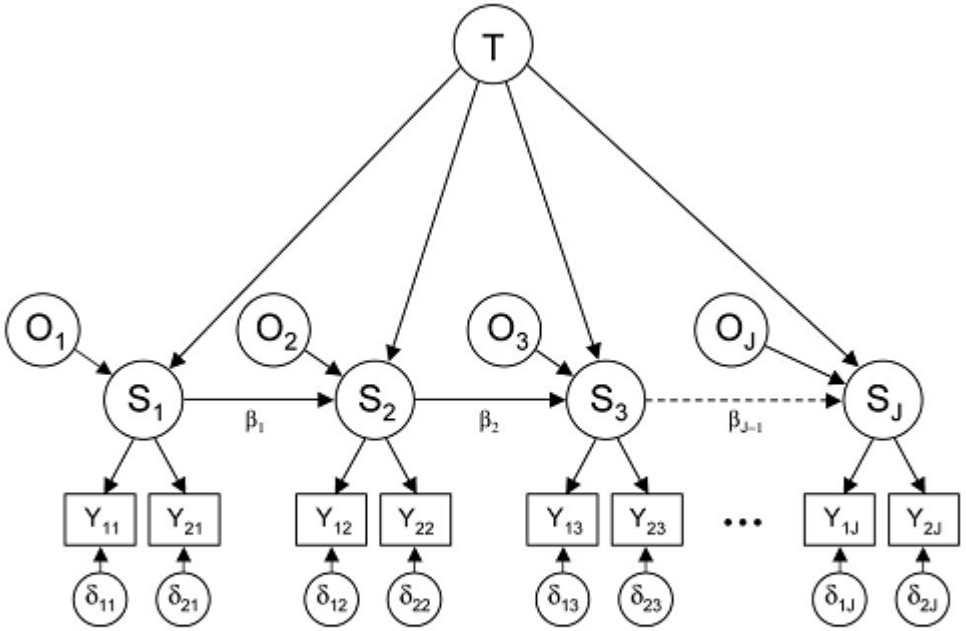



*Figure 5.* Steyer and Schmitt's (1994) latent state–trait autoregressive model. T = trait, O = occasion, and S = state for *J* waves and any manifest variable *Y*.

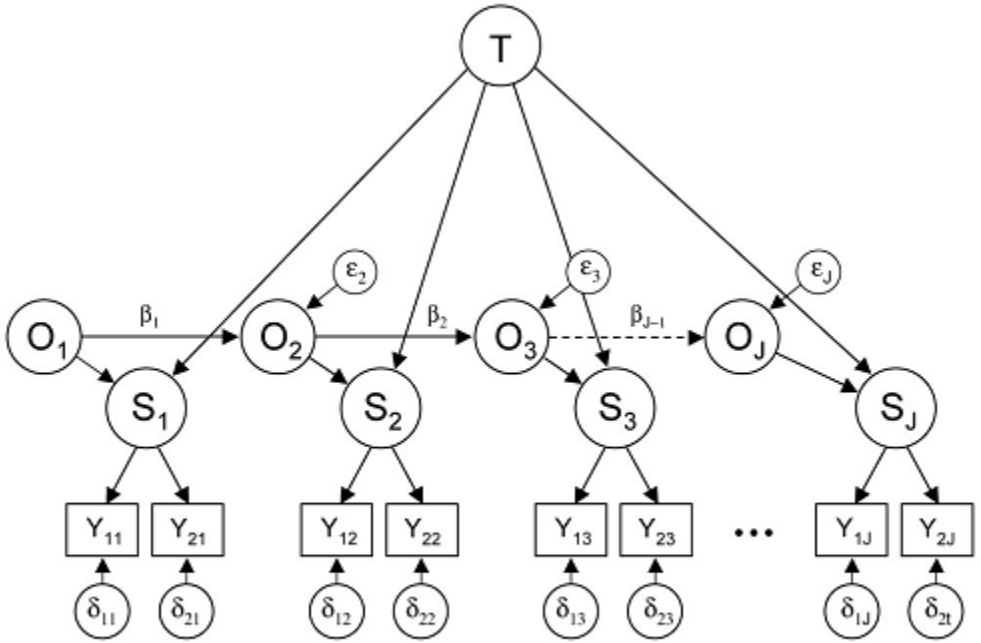
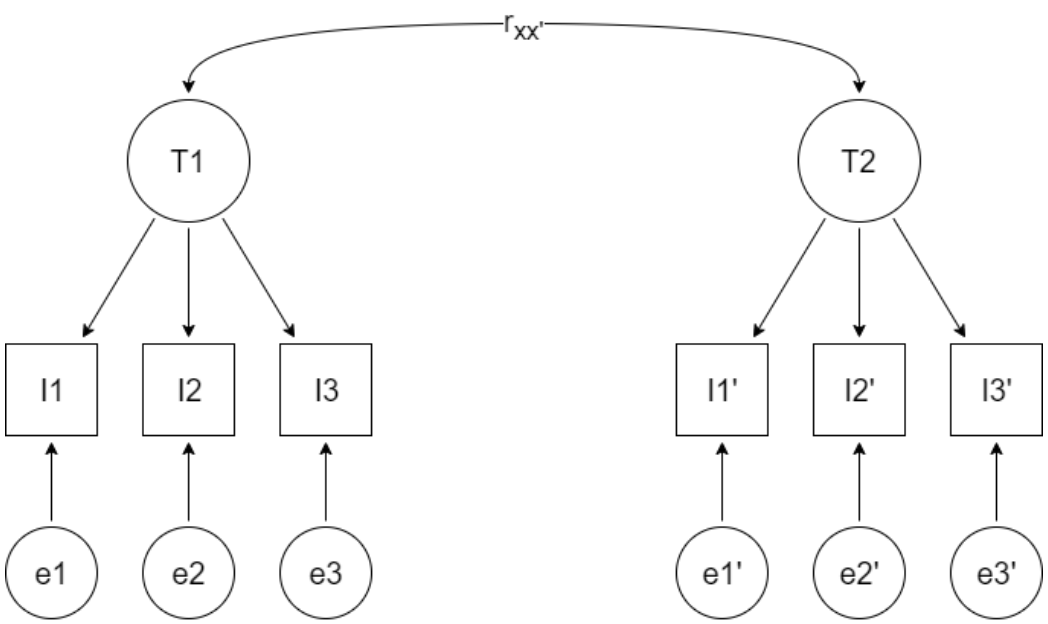



*Figure 7.* Trait–state–occasion model. T = trait, O = occasion, and S = state for *J* waves and any manifest variable *Y*.

Cole, D.A., Martin, N.C., & Steiger, J.H (2005). Empirical and conceptual problems with longitudinal trait-state models: introducing a trait-state-occasion mode. Psychological Methods 10(1), 3–20. https://doi.org/10.1037/1082-989X.10.1.3

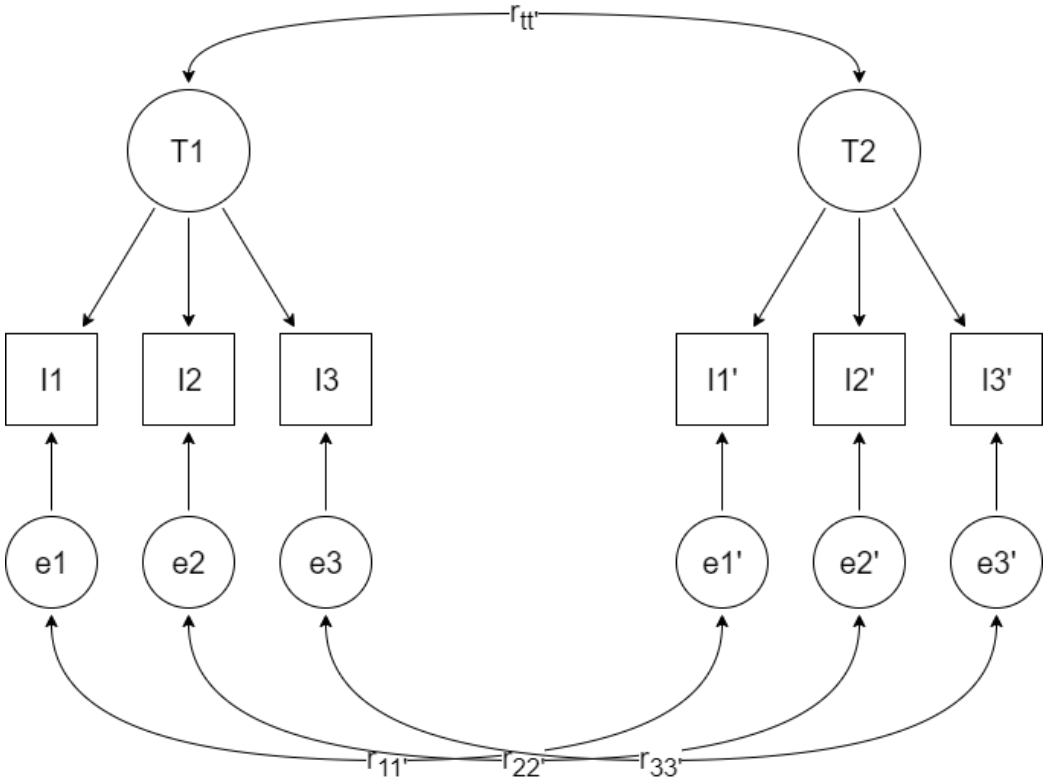
# Nezávislost chyb měření




Chyby měření nebývají zcela náhodné, ale obsahují systematickou složku stabilní v čase.

- U výkonových testů méně, u dotazníků více.

# Nezávislost chyb měření




Chyby měření nebývají zcela náhodné, ale obsahují systematickou složku stabilní v čase.

- U výkonových testů méně, u dotazníků více.

Co to udělá s korelací celého testu?

- Tedy $r(\sum I_i , \sum I_i')$?

Jakou informaci ponese tato korelace?

Jaký bude vztah reliability a korelace?

**Dvě pojetí reliability:**

- reliabilita jako vztah atributu a měření ($r_{x\tau}^2$)
- reliabilita jako stabilita měření ($r_{xx}'$)
- nejsou-li chyby měření nezávislé, $r_{x\tau}^2 \neq r_{xx}'$

# Reliabilita paralelních forem

ODHAD RELIABILITY

# Reliabilita paralelních forem

**Poskytují dva testy shodné odhady atributu?**

Metoda: Korelace paralelních forem testu.

Účel používání paralelních testů:

- Zabránit opisování při hromadné administraci.
- Zabránit zapamatování položek při opakované administraci a retestování (PPP).
- Umožnit sběr data ve více nezávislých termínech (SCIO, TSP...).

**Problém:** I když jsou testy vytvořené stejným způsobem, málokdy měří zcela ten samý pravý rys.

**Je nutné odlišit reliabilitu paralelních forem od existence paralelních forem jako takových.**

- Vyvažování paralelních forem je celkově velmi náročné.

Více stupňů ekvivalence dvou testů:

- Alternativní: pouze podobné.
- Srovnatelné: srovnatelné standardní skóry.
- Ekvivalentní: srovnatelné hrubé skóry.
- (Striktně) paralelní: shodné pravé skóry.
- *Souvisí s problematikou paralelních testů, viz přednáška o faktorové analýze.*
  - *Terminologie není jednoznačná*

# Paralelní formy prakticky

Pokud neaspirujeme na „vyvážené“, striktně paralelní testy…

… postupujeme stejně, jako v případě test-retest.

Převedeme na stejné jednotky (T-skóry atp.) pro každou formu zvlášť a ověříme:

- shodu pořadí (korelace);
- shoda průměrů (t-test; standardizace to zajistí);
- shodu rozptylů = homoskedascitu (Levenův test; standardizace to zajistí);
- případně i linearitu skórů (scatter-plot, kvadratická/polynomická regrese).

(Vnitrotřídní) korelace je potom koeficientem reliability.

Co vše může způsobovat rozdíl průměrů obou forem?

- Jak se vyhnout těm vlivům, které „nechceme“?

# Vyvažování paralelních forem

Spíš otázka norem a pedagogického testování (nikoli reliability).

**Linking** (skóry dvou forem testu jsou srovnatelné) vs. **equating** (testy měří to samé)

Jedno z typických využití *teorie odpovědi na položku* (IRT).

Samostatné obsáhlé publikace, specifická expertíza.

- Kolen, M. J., & Brennan, R. l. (2014). *Test equating, scaling and linking: methods and practices*. Springer.

Raw-score equating, ekvipercentilové vyvažování, linking functions, mapping functions.

Vyvažují se nejen formy, ale i jazykové mutace (PISA, TIMSS, TALIS).

# Shoda posuzovatelů

ODHAD RELIABILITY

# Shoda posuzovatelů

**Docházejí dva hodnotitelé/administrátoři ke shodným závěrům?**

Druhy neshody:

- Shoda administrátorů (např. WISC).
- Shoda posuzovatelů (např. ROR) – inter-rater, intra-rater reliabilita.
- V diagnostické praxi obtížně odlišitelné.

Korelace napravo: $r_{AB} = 0{,}93$. Opravdu se hodnotitelé shodují?

**Komplikace 1:** rozdílná „přísnost“ hodnotitelů.

- Je nutné vzít v úvahu i rozdílnou přísnost (Cohenovo $d = 1{,}3$).
- Používá se proto tzv. vnitrotřídní korelace (intra-class correlation), která bere v úvahu shodu pořadí, průměrů a lze použít pro libovolný počet hodnotitelů. Existuje 2×(3+2) variant ICC.
- V tomto případě $ICC(2{,}1) = 0{,}51$.
  - Pozn.: *ICC*(3,k) pro průměrné hodnocení je ekvivalentní s pojetím reliability podle Hoyta [URB, s. 112-114] a tedy s Cronbachovým α, v tomto případě $ICC(3{,}2) = 0{,}96$.

| | rater A | rater B |
|---|---|---|
| ID1 | 4 | 7 |
| ID2 | 2 | 4 |
| ID3 | 6 | 7 |
| ID4 | 1 | 3 |
| ID5 | 3 | 5 |
| ID6 | 5 | 6 |
| M | 3,00 | 5,67 |
| SD | 2,19 | 1,97 |
| $r_{AB}$ | 0,93 | |

# Shoda posuzovatelů: **komplikace 2**

Až příliš často nás zajímá shoda jednotlivých kritérií: **Úroveň měření**.

Reliabilita kódování na úrovni položky.

- Používá se i jako ukazatel interní validity v kvalitativním výzkumu.

Položky bývají nominální nebo ordinální, nelze proto použít *ICC* a korelace.

- A nelze použít podíl shody (např. „shodli se v 90 % případů“) kvůli nahodilé shodě.

Proto velké množství různých statistik:

- Cohenovo kappa – absolutní shoda 2 hodnotitelů vážená proti nahodilé shodě.
  - $\kappa = \frac{P_o - P_e}{1 - P_e}$, kde $P_o$ je pozorovaná shoda a $P_e$ zcela náhodná shoda (očekávaná)
- Vážené kappa – shoda 2 hodnotitelů v případě ordinálních položek.
- Fleissovo (vážené) kappa – shoda N hodnotitelů u nominálních (ordinálních) položek.
- Kendallův koeficient konkordance – analogie Spearmanovy korelace pro N hodnotitelů (jen pořadí).

# Shoda posuzovatelů: Co si pamatovat?

V nouzi: shoda průměrů (např. t-test, ANOVA) plus pořadí (alfa, korelace)

- Nebo ordinální ekvivalenty (Mann-Whitney, Kruskal-Wallis, Spearmanova korelace).

V případě nominálních proměnných **za žádných okolností nepoužívat % shody!**

Zpravidla o dost jiná informace, než zbylé koeficienty.

Specifické koeficienty. Některé stojí pamatovat si podle jména:

- (Cohenova) kappa; vnitrotřídní korelace; Kendallův koeficient konkordance; Krippendorfova alfa.

Další zdroje:


- Hallgren, K. A. (2012). Computing Inter-Rater reliability for observational data: An overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23–34. doi:10.20982/tqmp.08.1.p023
- Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., … Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International Journal of Nursing Studies, 48(6), 661–671. doi:10.1016/j.ijnurstu.2011.01.016

# Vnitřní konzistence

ODHAD RELIABILITY

# Vnitřní konzistence

Často máme ale jedinou formu testu bez vlivu posuzovatele (dotazník) a nezajímá nás stabilita v čase nebo nemáme prostředky na dvě administrace (nebo to není možné).

Prostě je k dispozici jediné měření jednou metodou.

**Dva hlavní postupy:**

- Split-half reliabilita.
- Vnitřní konzistence.

# Split-half

**Postup: Test rozdělíme na dvě půlky a pracujeme jako s reliabilitou paralelních forem.**

**Problém 1: Jak test rozdělit?**

- Poloviny by měly být paralelní.
- Zpravidla tedy nějaké pseudo-náhodné rozdělení (sudá–lichá).
- Existuje velmi mnoho různých rozdělení a každé poskytne poněkud jiný odhad split-half reliability.

**Problém 2: Odhad založen jen na jediné korelaci.**

- Při srovnání s jinými koeficienty vnitřní konzistence (alfa, omega) menší přesnost odhadu (širší *CI*).

**Problém 3: Zkrácení testu.**

- Reliabilita je závislá na délce testu. Delší testy → vyšší reliabilita.
- Rozpůlením testu zjistíme reliabilitu jedné poloviny, reliabilita celého testu je nutně vyšší.

**Problém 4: Lichý počet položek.** Podstatný není počet položek, ale rozptyl půlek testu.

- U delších testů proto nehraje roli.

# Split-half: Spearmanův-Brownův postup

„**Spearmanův-Brownův věštecký vzorec**“ (Spearman-Brown prophecy formula):

$$r^*_{xx'} = \frac{N r_{xx'}}{1 + (N-1) r_{xx'}}$$

- *N* – poměr délek testů; $r_{xx'}$ – původní reliabilita; $r^*_{xx'}$ odhad reliability po změně délky.
- *„Jaká bude reliabilita $r^*_{xx'}$ při N-násobné změně délky testu?“*

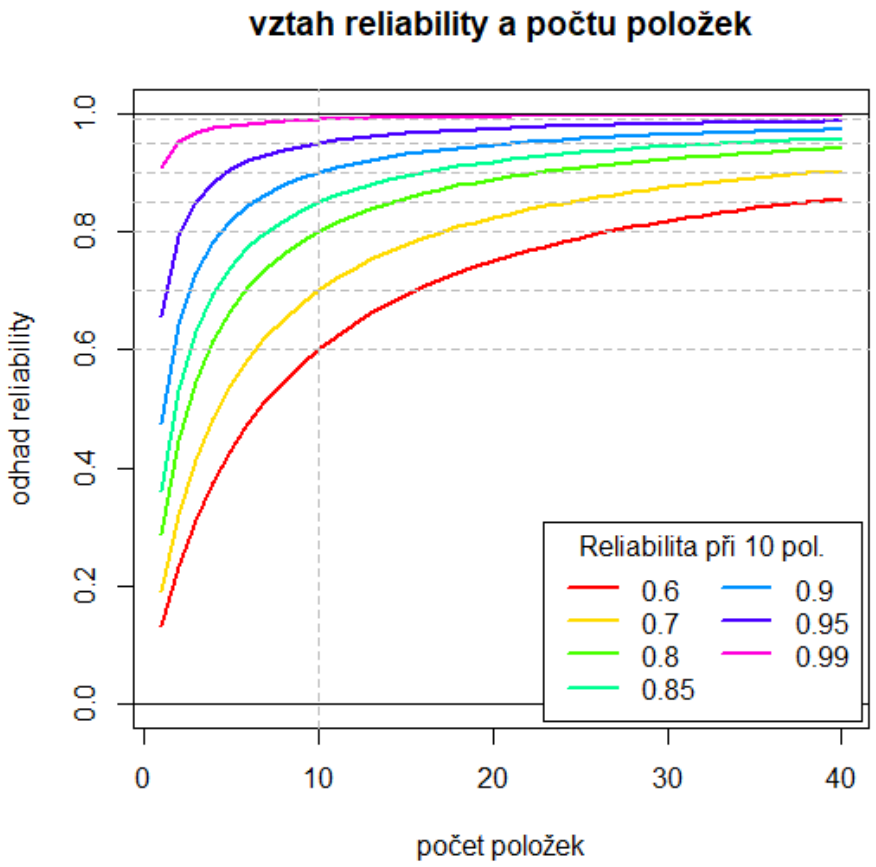
V případě split-half reliability *N* = 2 (test je dvakrát delší než polovina):

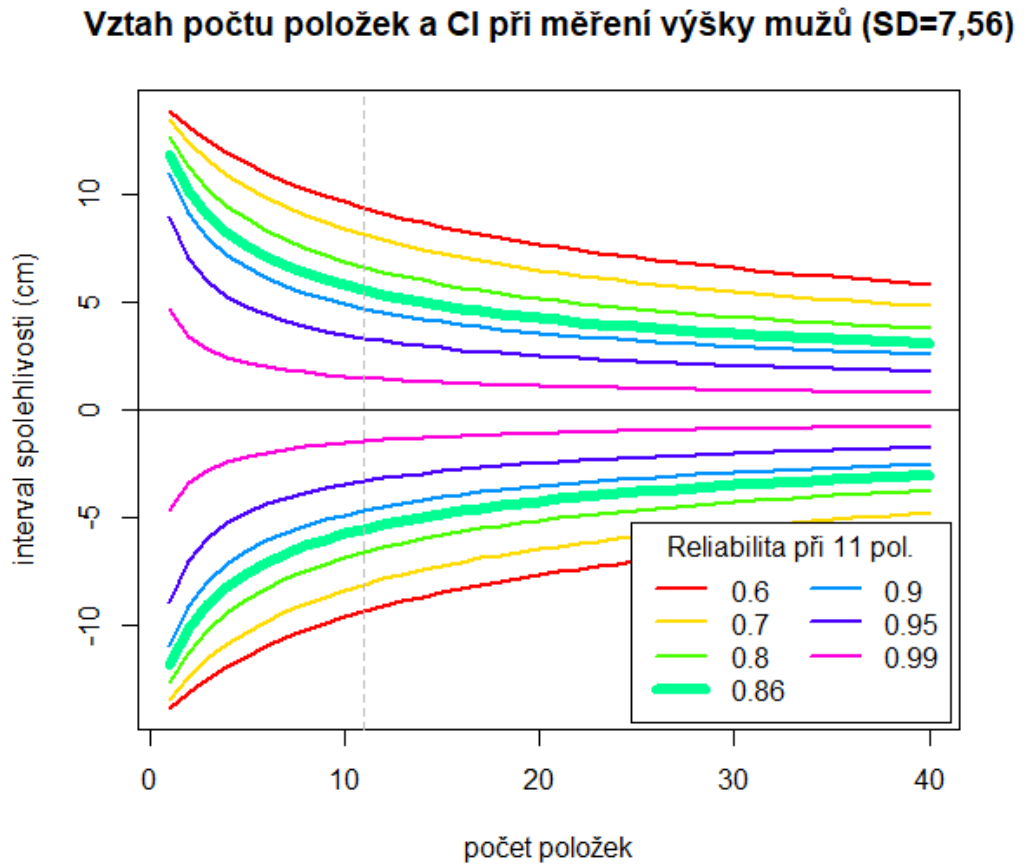
$$r_{SB} = r^*_{xx'} = \frac{2 r_{xx'}}{1 + r_{xx'}}$$

Slouží i k odhadu požadovaného počtu položek pro dosažení určité reliability.

- Předpokladem jsou striktně-paralelní položky (viz přednáška o FA).

# Vztah reliability testu a jeho délky




V případě 11položkového dotazníku výšky ze začátku semestru $r$ = 0,86.

# Split-half: Guttmanova lambda 4

Guttman (1945) publikoval 6 různých odhadů reliability $\lambda_{1-6}$. Podstatné jsou dva z nich:

$$\lambda_4 = \frac{4\sigma_{pq}^2}{\sigma_x^2}$$

- kde $\sigma_{pq}^2$ je kovariance polovin testu a $\sigma_x^2 = \sigma_p^2 + \sigma_q^2 + 2\sigma_{pq}^2$ je rozptyl celého testu.
- $\lambda_4$ je shodná s Cronbachovou alfou u dvoupoložkových testů.
- $\lambda_3$ je určena pro vícepoložkové testy a je shodná s Cronbachovou alfou (viz dále).

Spearman-Brown vs. lambda 4:

- SB může při porušení předpokladů reliabilitu nadhodnotit, $\lambda_4$ je vždy nižší než skutečná reliabilita.
- Pokud se poloviny testu výrazně liší svou délkou či rozptylem, $\lambda_4$ může výrazně podhodnotit.
- Jsou-li poloviny standardizovány, pak platí $\lambda_4 = r_{SB} = \alpha$.
- U dlouhých testů oba postupy vedou k podobným odhadům.

Poloviny testů by při jakémkoli split-half přístupu měly být „stejně dlouhé“.

- Pokud nejsou, lze využít jiné postupy (Cígler a Chvojková, preprint; Warrens, 2016).

# Split-half: specifické použití

**Greatest-Lower Bound of reliability.**

- Řada rozdílných postupů a algoritmů.
- Anotace jako *GLB*, *glb*, $\sigma_+$, $\rho_{glb}$ apod.

V poslední době je Guttmanova $\lambda_4$ chápána jako synonymum pro GLB.

Položky jsou rozděleny tak, aby byla korelace polovin testu maximalizovaná.

- Může být analyticky náročné.
- Na malých vzorcích a krátkých testech vede k nadhodnocení z důvodu výběrové chyby („příliš dobré“ rozpůlení).
- Doporučení: N > 1000. Vyhnout se N < 200.

# Cronbachovo alfa

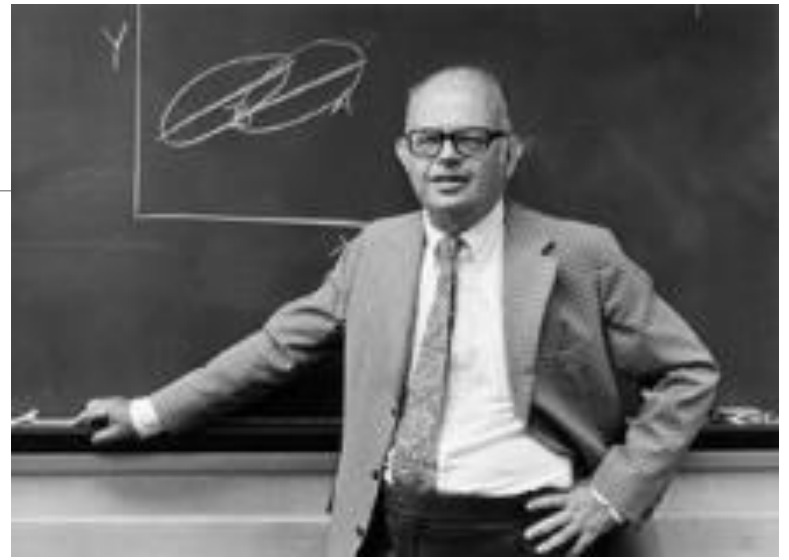
Co když jsou paralelními testy jednotlivé položky?

- Pokud měří všechny to samé, pak by spolu měly hodně korelovat – být vnitřně konzistentní.
- Položky měří totéž, pokud mají hodně sdíleného rozptylu.

Cronbachova (1951) alfa:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_x^2}\right)$$

- $\sigma_i^2$ – rozptyl položky *i*, $\sum_{i=1}^{k} \sigma_i^2$ je diagonála variančně-kovarianční matice (jedinečný/chybový rozptyl položek)
- $\sigma_x^2$ – rozptyl celého testu, tedy suma var-covar matice
- $k$ – počet položek (ne celý jedinečný rozptyl položek je chybou, proto korekce $\frac{k}{k-1}$, aby reliabilita mohla být 1)
  - Bez této korekce jde o Guttmanovu $\lambda_1$.



| | A | B | C |
|---|---|---|---|
| A | 1 | 0,514 | 0,477 |
| B | 0,514 | 1 | 0,662 |
| C | 0,477 | 0,662 | 1 |

Část korelační matice Holzinger a Swineford (1937):

$$\alpha = \frac{3}{2}\left(1 - \frac{1+1+1}{1+1+1+2(0{,}514+0{,}477+0{,}662)}\right) = 0{,}786$$

# Cronbachovo alfa: předpoklady

**Tau-ekvivalentní položky**

- Stejná lineární souvislost položky s pravým skóre...
- ... a tedy shodné faktorové náboje ve faktorové analýze (viz přednáška o FA).
- Při nedodržení podhodnocuje.

**Unikátní rozptyl je celý chybovým rozptylem.**

- A proto tzv. „spodní hranice reliability“.

Lokální nezávislost položek (**jednodimenzionalita**).

- Nedodržení může nadhodit i podhodnotit.

**Alfa není ukazatelem jednodimenzionality!**

- I vícedimenzionální testy mohou mít vysokou vnitřní konzistenci, viz např. Marko (2016).

# Cronbachovo alfa: varianty

**Standardizovaná Cronbachova alfa:**

- Korelační, nikoliv kovarianční matice.
- Vnitřní konzistence *standardizovaných* položek.
- Robustnější při výrazně rozdílné obtížnosti položek (slabší předpoklad tau-ekvivalence).

**Kuderův-Richardsonův (1931) vzorec 20 a 21**

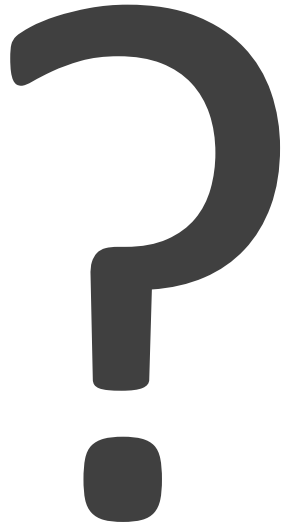
$$KR_{20} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} P_i(1-P_i)}{\sigma_x^2}\right]$$

- V případě binárních položek, kdy $P_i(1-P_i)$ je rozptyl dichotomické položky.
- $KR_{20}$ = $\alpha$, $KR_{21}$ pro položky stejné obtížnosti.
- Spíše historické kvůli snadnosti výpočtu.

# Psychometrický paradox

Hypotetický dotazník extroverze:

- Rád se vídám s lidmi.
- Rád jsem v kontaktu s lidmi.
- Vyhledávám společnost lidí.
- Jsem rád mezi lidmi.
- Dělá mi dobře společnost lidí.
- ...



# Psychometrický paradox

Reliabilita testu je funkcí korelací mezi položkami a jejich počtem.

Čím více spolu položky korelují, tím „ostřeji“ se zaměřují na specifický rys.

„*Alfa tuning*“ škál: výběr nejvíce korelujících položek a zvýšení reliablity.

- Měříme stále přesněji stále méně (menší výsek konstruktu) – ztráta (výběrové) validity.
- Někdy i jako cílená aktivita; de facto je to podvod (synonymní páry položek...).

Nikoli vždy!

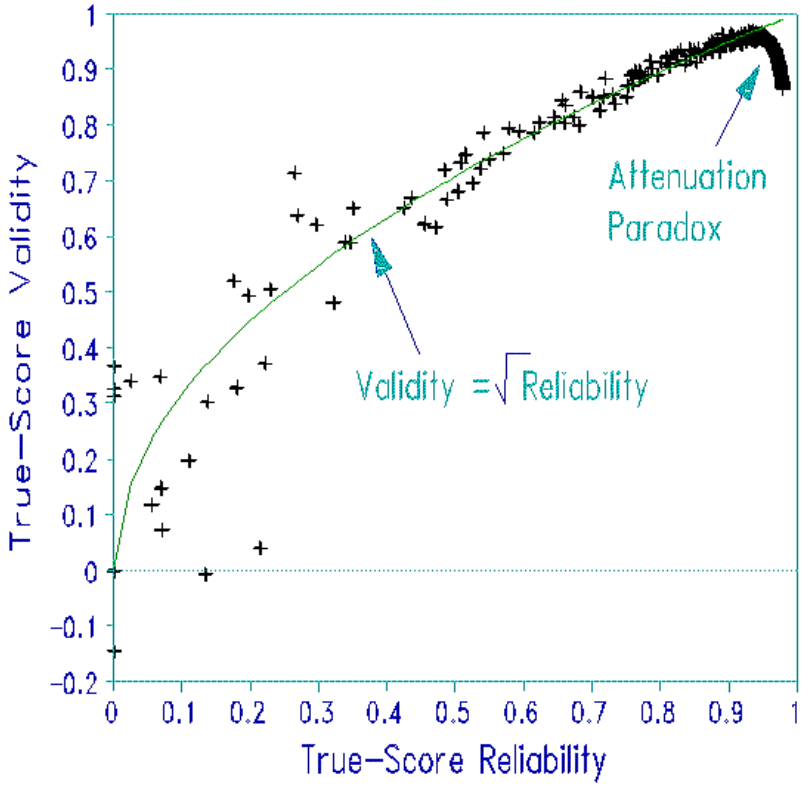



Figure 1. The Attenuation Paradox

https://www.rasch.org/rmt/rmt94a.htm

# Kdy použít split-half?

Vnitřní konzistence (alfa, omega...) bývá výhodnější než split-half.

- Přesnější a robustnější odhad.

Výjimka: časované/rychlostní testy nebo testy s pravidlem ukončení.

- Počet správně vyřešených položek za 1 minutu (např. Test pozornosti d2).
- „Ukončete administraci po 5 chybných odpovědích“ (např. Wechslerovy testy).
- Datová matice obsahuje řadu chybějících dat na koncích řádků.

Určité výhody i u velkých datasetů (N>1000, ideálně N>5000).

- GLB, menší statistické předpoklady (např. ve srovnání s binárními pol.).

# Specifické příklady vnitřní konzistence

Reliabilita celkového skóre v multidimenzionálních testech.

- Např. reliabilita celkového skóre v inteligenčním testu (WISC, WAIS).

Reliabilita váženého skóre.

- Celkové skóre je váženým součtem dílčích položek/subtestů.

Reliabilita rozdílového skóre.

- Např. reliabilita rozdílu rychlosti a správnosti v testu pozornosti d2.

Reliabilita v testech založených na jiné teorii měření.

- Typicky IRT, kde položky nejsou jednoduše sčítány.

**V těchto případech je pro odhad vnitřní konzistence použít jiné postupy.**

- Postupů pro odhad reliability je mnoho – představili jsme jen nejzákladnější postupy.

# Kompozitní reliabilita

Někdy též reliabilita lineárních kombinací.

Jaká je reliabilita (vícedimenzionálního) skóre založeného na součtu více škál?

- Běžné odhady typu Cronbachova alfa zpravidla vedou k podhodnocení.

Více přístupů, užitečná je zejm. stratifikovaná Cronbachova alfa (1965):

- $\alpha_{strat} = 1 - \frac{\sum_{i=1}^{k}[\omega_i^2 \sigma_i^2 (1 - r_{ii'})]}{\sigma_Z^2}$
- $\omega_i$ – „váha“ testu *i* (zpravidla 1); $\sigma_i^2$ – rozptyl testu *i*; $r_{ii'}$ – reliabilita testu *i*
- Pro výpočet stačí kovarianční matice a reliability subtestů.
- Předpoklady tau-ekvivalence položek v testech, tau-ekvivalence testů, ortogonální disturbance.

Dále pak koeficienty omega (viz přednáška o FA).

# Reliabilita rozdílu

Jak reliabilní je používání rozdílu mezi dvěma testy?

- Například VIQ a PIQ ve WAIS-III?

$$r_{x-y} = \frac{\sigma_x^2 r_{xx'} + \sigma_y^2 r_{yy'} - 2 r_{xy} \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 - 2 r_{xy} \sigma_x \sigma_y},$$

- kde $\sigma_x^2$ a $\sigma_y^2$ jsou rozptyly obou testů, $r_{xx'}$ a $r_{yy'}$ jejich reliability a $r_{xy}$ je jejich korelace.
- jmenovatel je roven rozptylu výsledných rozdílů.

Pokud $\sigma_x^2 = \sigma_y^2 = \sigma_{xy}^2$ (v případě standardizovaných testů), pak:

- $r_{x-y} = \sigma_{xy}^2 \frac{r_{xx'} + r_{yy'} - 2 r_{xy}}{2 - 2 r_{xy}}$

# Reliabilita rozdílu

| $r_{xx'}$ | $r_{yy'}$ | $r_{xy}$ | $r_{x-y}$ | $SD_{x-y}$ | $SE_{x-y}$ | $CI_{95\%}$ |
|---|---|---|---|---|---|---|
| 0,7 | 0,8 | 0 | **0,75** | 21,2 | 10,6 | 20,8 |
| 0,7 | 0,8 | 0,2 | **0,69** | 19,0 | 10,6 | 20,8 |
| 0,7 | 0,8 | 0,4 | **0,58** | 16,4 | 10,6 | 20,8 |
| 0,7 | 0,8 | 0,6 | **0,38** | 13,4 | 10,6 | 20,8 |
| 0,7 | 0,7 | 0,6 | **0,25** | 13,4 | 11,6 | 22,8 |
| 0,9 | 0,9 | 0,8 | **0,50** | 9,5 | 6,7 | 13,1 |
| 0,9 | 0,9 | 0,45 | **0,82** | 15,7 | 6,7 | 13,1 |
| 0,6 | 0,6 | 0,5 | **0,20** | 15,0 | 13,4 | 26,3 |
| 0,7 | 0,7 | 0,65 | **0,14** | 12,5 | 11,6 | 22,8 |

Standardní chybu rozdílu lze spočítat s pomocí SD a SE vlevo, nebo prostřednictvím vzorce.

- Viz seminář.

Toto je důvod, proč je problematická interpretace rozdílu vysoce korelovaných subtestů.

- Téměř u nikoho se neliší...

**Table 3.** Names of Reliability Coefficients Currently Used in the Literature.

| | Unidimensional | | Multidimensional |
|---|---|---|---|
| | Split-Half | General | General |
| Parallel | Spearman–Brown formula | Standardized alpha | (Not yet published) |
| Tau-equivalent | Flanagan–Rulon formula<br>Flanagan formula<br>Rulon formula<br>Guttman's $\lambda_4$ | Cronbach's alpha<br>Coefficient alpha<br>Guttman's $\lambda_3$<br>Hoyt method<br>KR-20 | Stratified alpha |
| Congeneric | Raju (1970) coefficient<br>Angoff–Feldt coefficient<br>Angoff coefficient | Composite reliability<br>Construct reliability<br>Congeneric reliability<br>Omega<br>Unidimensional omega<br>Raju (1977) coefficient<br>Classical congeneric reliability coefficient | Omega<br>Omega total<br>McDonald's omega<br>Multidimensional omega |

Cho (2016): http://journals.sagepub.com/doi/abs/10.1177/1094428116656239

# Klasická testová teorie (CTT): overview

CTT je špatným modelem měření. Není jasné, co to je pravý skór.

- Pravý skór je definovaný skrze samotné měření.
- Pravý skór je neoddělitelný od měřicího nástroje.
- CTT je založena na operacionalismu: definice měření je operacionální.
- CTT nepopisuje „data generating process“.

CTT je historicky spojená s faktorovou analýzou.

Protože CTT předpokládá paralelní položky, celkový skór testu je součtem/průměrem položek.

- Ale co když položky nejsou paralelní?

Přesto je CTT jednoznačně nejvíce používanou teorií měření v sociálních vědách.

- I bodování v psychometrice je založené na součtu správných odpovědí v testech...

# Reliabilita: overview

**Reliabilita je ukazatelem kvality testu.**

- Řada doporučení ohledně minimální hranice přípustné reliability. Typicky Klineovo pravidlo: $r_{xx'} > 0{,}7$.
- Záleží ale na účelu testu: nižší nároky pro výzkumné metody, vyšší nároky pro metody určené do praxe, nejvyšší nároky na high-stakes testy (SCIO, inteligenční test...).
- V případě výzkumu záleží i na způsobu využití (SEM vs. pozorované skóry).

Doporučené hodnoty reliability:

- „Nejlepší“ metody (celkový skór IST-2000-R) nebo testy základních kognitivních funkcí (Bourdonova zkouška): $r_{xx'} > 0{,}95$.
- Dobré testy: $r_{xx'} > 0{,}90$. Ve výzkumu výjimečně i $r_{xx'} > 0{,}70$.
- Osobně považuji testy s $r_{xx'} < 0{,}80$ za problematické. **Vždy ale záleží na účelu měření!**

**Reliabilita jako podklad pro práci s chybou při praktické psychologické diagnostice.**

- Viz seminář.

# Nelineární vztah reliability a chyby měření