

# Přednáška 3: Teorie zobecnitelnosti

---

20. 10. 2020 | PSYn4790 | Psychometrika: Měření v psychologii  
Katedra psychologie, Fakulta sociálních studií MU

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

# CTT: Hodně chyb, hodně reliabilit...

---

Mnoho způsobů odhadů reliability a druhů chyby v rámci CTT:

- **stabilita v čase = test-retest:** korelace, regrese, ICC...
- **vnitřní konzistence:** alfa, omega, split-half, GLB...
- **ekvivalence = reliability paralelních forem:** korelace, regrese, ICC...
- **shoda posuzovatelů:** Cohenovo/Fleissovo kappa, Krippendorfova alfa, ICC...

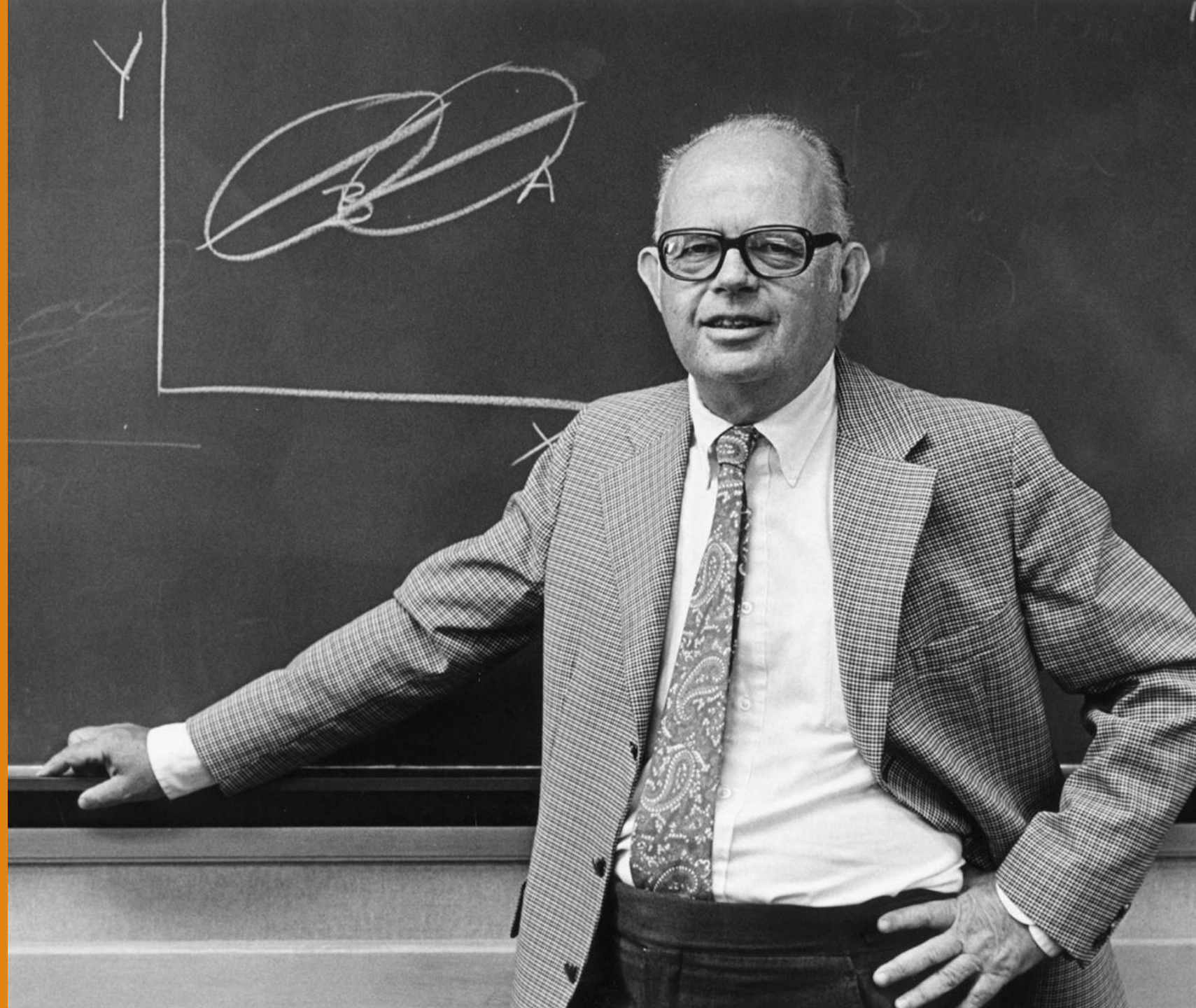
Co ale s tím? Kterou „reliabilitu“ si vybrat?

- Pro různé účely?

Teorie  
zobecnitelnosti

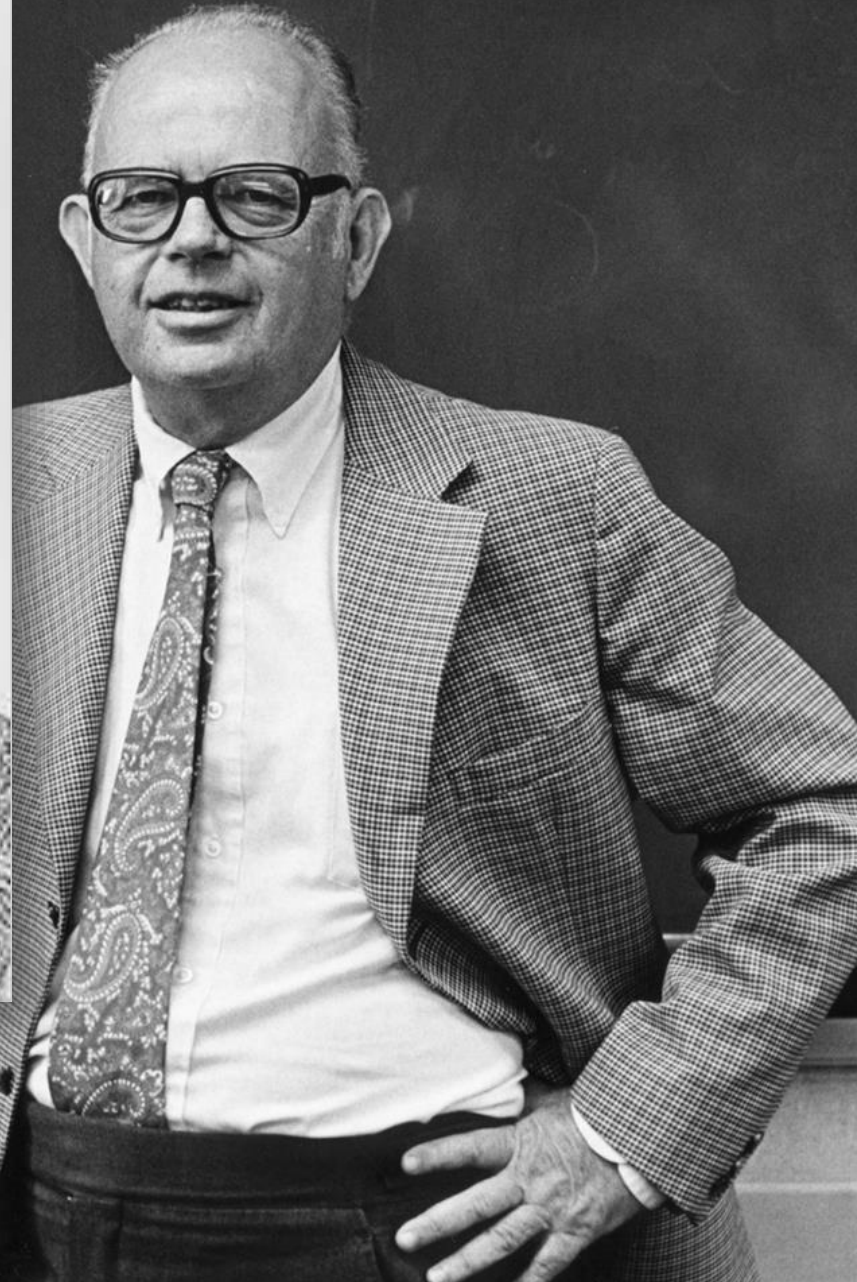
Generalizability  
Theory (GT)

Lee Cronbach (1916–2001)



Teorie  
zobecnitelnosti

Generalizability  
Theory (GT)



Lee Cronbach (1916–2001)

Cronbach, L.J., Rajaratnam, N.,  
& Gleser, G.C. (1963).

In 1957 I obtained funds from the National Institute of Mental Health to produce, with Gleser's collaboration, a kind of handbook of measurement theory. ... "Since reliability has been studied thoroughly and is now understood," I suggested to the team, "let us devote our first few weeks to outlining that section of the handbook, to get a feel for the undertaking." We learned humility the hard way—the enterprise never got past that topic. Not until 1972 did the book appear ... that exhausted our findings on reliability reinterpreted as generalizability. Even then, we did not exhaust the topic.

When we tried initially to summarize prominent, seemingly transparent, convincingly argued papers on test reliability, the messages conflicted.

# Cronbachovo alfa

---

Cronbachova alfa (1951) není tak docela Cronbachova:

- Rulonův vzorec (1939); Guttmanova korekce pomocí  $\lambda_3$  (1945); Hoytův vzorec (1941).

Cyril Hoyt ([1941](#)) – odhad reliability pomocí ANOVA:

$$r_{xx'} = \frac{\sigma_{\tau}^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = \frac{MS_{\tau}}{MS_x} = \frac{MS_x - MS_e}{MS_x}$$

- $MS_x$  - mean-square, tj. průměr sumy čtverců, tj. rozptyl.

ANOVA umí „parcelovat“ pozorovaný rozptyl (**AN**alysis **Of** **V**ariance).

- Typická ANOVA: jakou část pozorované variability mezi lidmi ( $MS_x$ ) lze přičíst rozdílům mezi skupinami ( $MS_{\tau}$ ) a jaká je způsobena rozdílům uvnitř skupin ( $MS_e$ )?
- Resp. pomocí F-testu ověřujeme, zda je  $MS_{\tau} > 0$ .

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99–103.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6(3), 153–160. <https://doi.org/10.1007/BF02289270>

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>

# Hoytův postup

---

Hoyt použil ANOVA k parcelování pozorované rozptylu odpovědí lidí na paralelní testy (položky).

Postup výpočtu (přibližně, bez korekcí):

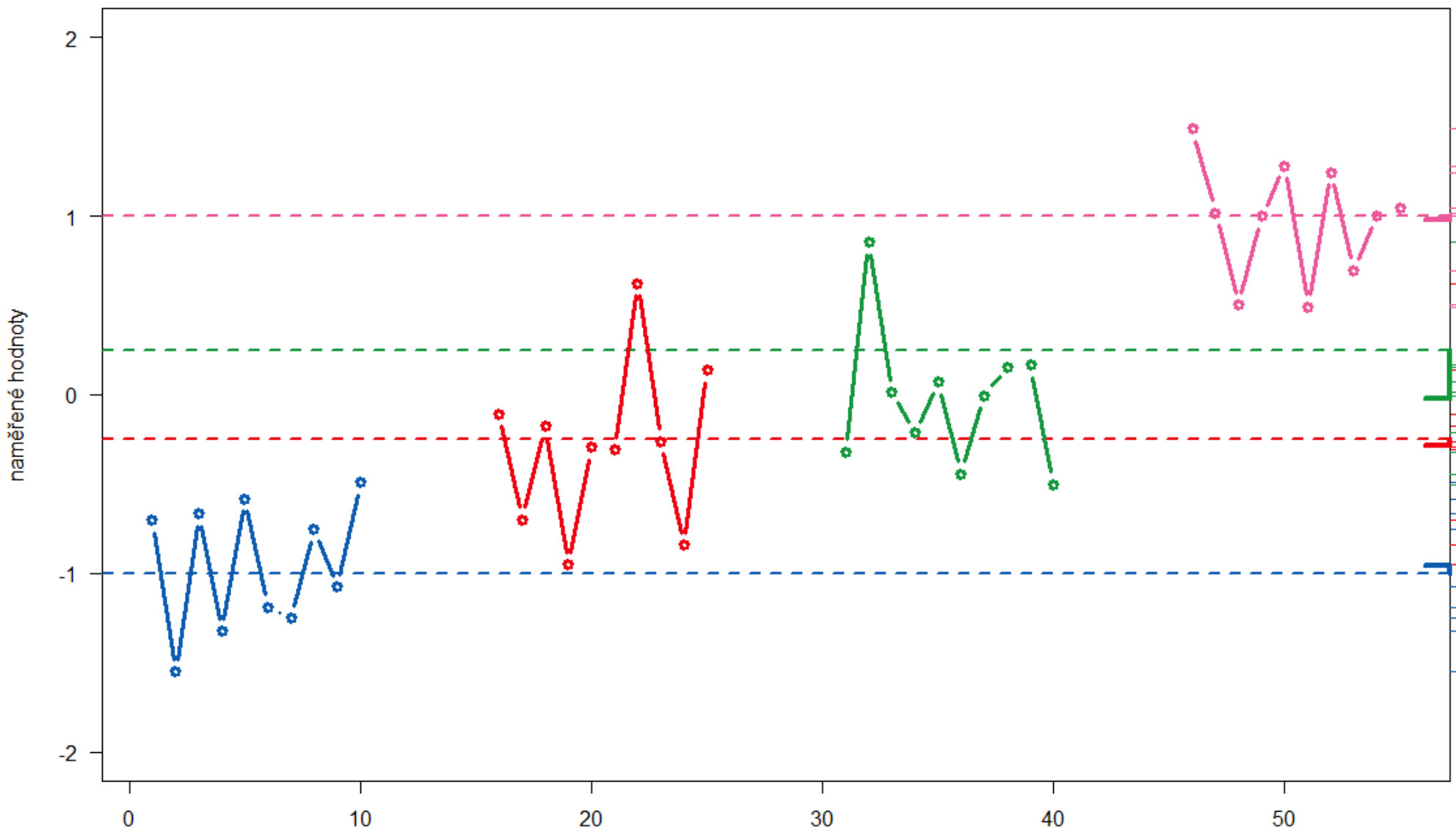
- 1. Rozptyl průměrů osob ( $\bar{X}_p$ ):  $\sigma_x^2$ .
- 2. Rozptyl odchylek jednotlivých pozorování  $x_{ip}$  osob  $p$  na pol.  $i$  od jejich průměrů  $\bar{X}_p$  jako:

$$\sigma_{res}^2 = \frac{\sum_{p=1}^N \sum_{i=1}^I (x_{ip} - \bar{X}_p)^2}{NI}$$

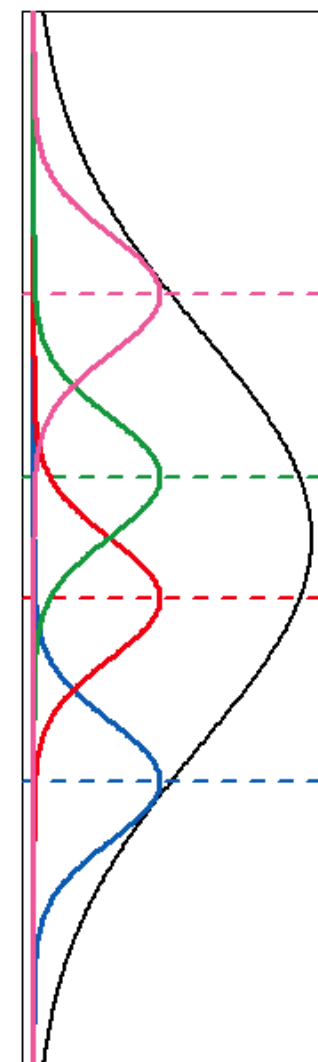
- 3. Standardní chyba odhadu průměru z deseti položek jako  $\sigma_e = \sqrt{\frac{\sigma_{res}^2}{I}} = \frac{SD}{\sqrt{N}}$ 
  - Protože jde o výpočet „přes všechny osoby“, obchází potíže s opakovaným měřením jedné osoby.
- 3. Reliabilita jako  $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ .
- 4. Standardní chyba měření: Buď z reliability, nebo přímo jako  $SE = \sigma_e = \sqrt{\frac{\sigma_{res}^2}{I}}$

Výsledek je ekvivalentní Cronbachovu alfa, asymptoticky průměru všech split-half reliabilit a rovněž i dalším odhadům s předpokladem tau-ekvivalence (**Cronbach, 1951**).

### 1Fasetový design, $r = 0.92$



40 měření (4 osoby, každá měřena desetkrát)  
SE = 0.247; SS = 0.061



Rozložení pozorování



# Hoytův postup

---

Tau-ekvivalence (všechna pozorování mají stejnou váhu,  $E\left(\frac{\sum_{i=1}^I E(X_{pi})}{I}\right) = \tau_p$ ).

- Paralelnost položek (shodné reziduální rozptyly) lze obejít.

Neexistence jiného zdroje rozptylu, než:

- Variabilita ve schopnostech lidí ( $\sigma_p^2$ ).
- Variabilita v obtížnostech položek ( $\sigma_{pi}^2$ ).
- Variabilita v tom, jak různí lidé odpověděli na různé položky ( $\sigma_{pi,e}^2$ ).

Co když je ale zdrojů více?

- Situace a změna rysu v čase, okolnosti testování, hodnotitel, dílčí oblast znalostí...

# Teorie zobecnitelnosti (Generalizability theory)

---

Řešením CTT problému „mnoho chyb, mnoho reliabilit“ je teorie zobecnitelnosti.

- Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137-163.
- Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York: Wiley.

**CTT:**  $X = T + e$

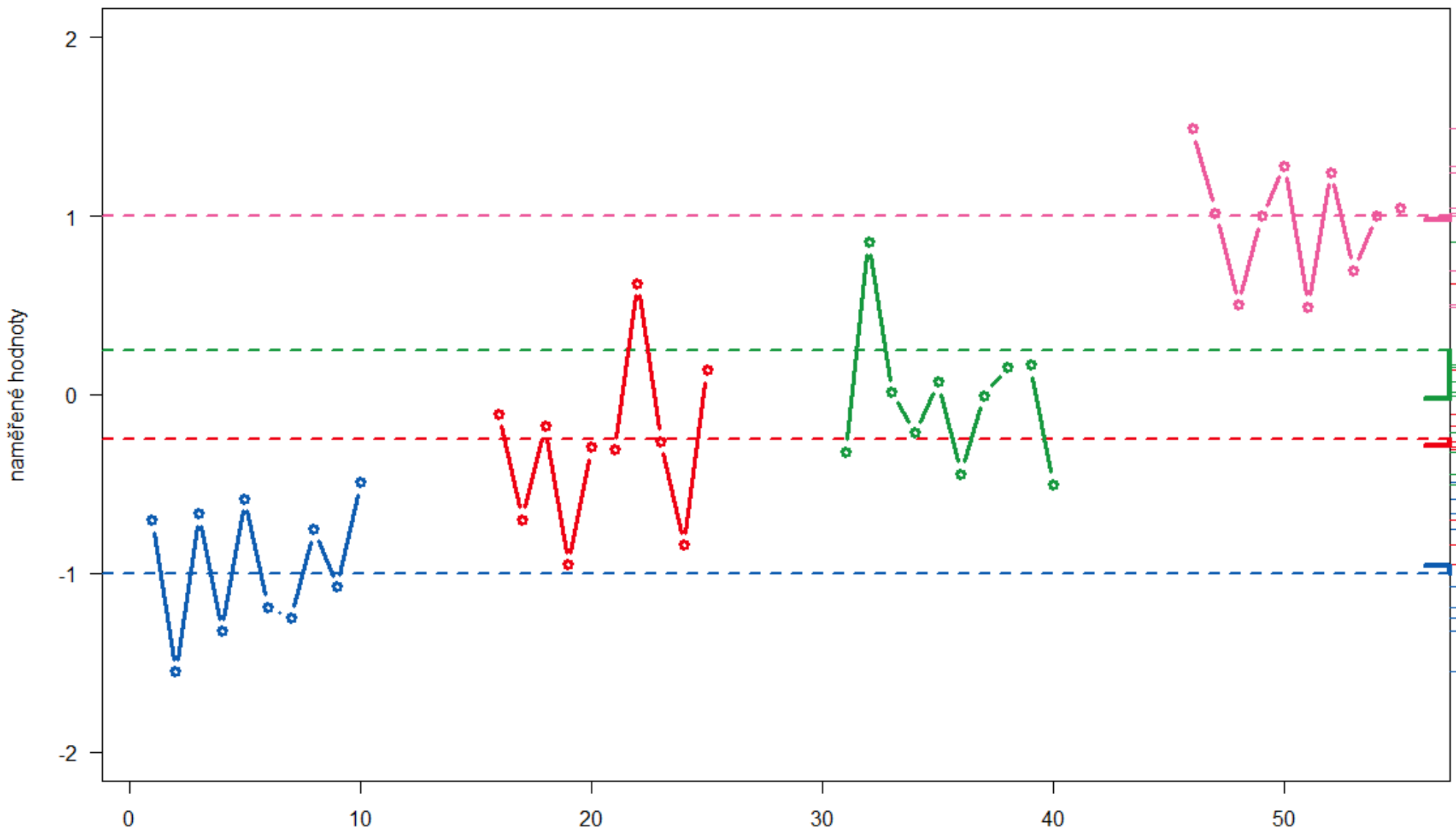
**GT:**  $X = T + e_1 + e_2 + e_3 + \dots + e_k$

- Kde např.  $e_1$  je specifický skór v daném čase (*test-retest*),  $e_2$  rozdílnost posuzovatelů (*shoda posuzovatelů*),  $e_3$  rozdílnost položek (vlastní „nepřesnost metody“, *vnitřní konzistence*) atd.
- Pro různé účely může  $T$  zahrnovat i některé chyby (např. nás zajímá výkon v daném čase a nikoliv stabilita napříč časem, přestože víme, že výkon není stabilní).

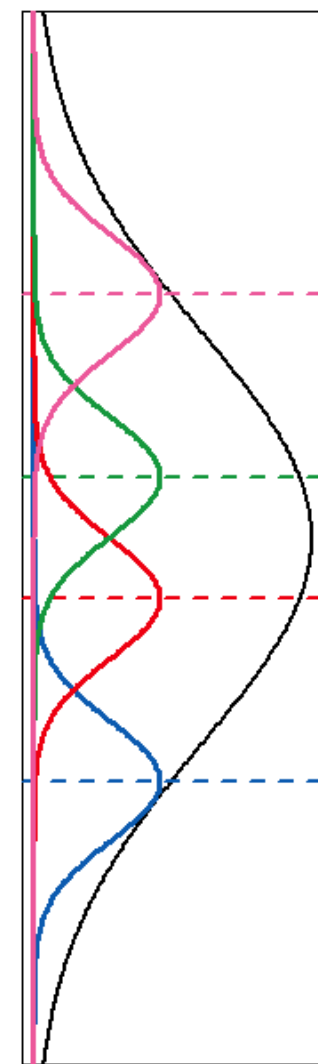
Protože ale např. i ten stejný hodnotitel může hodnotit různě v různých situacích, ve vzorci výše tedy chybí interakce:

- $X = T + e_1 + e_2 + e_3 + e_{12} + e_{13} + e_{23} + e_{123} \dots$

### 1Fasetový design, $r = 0.92$

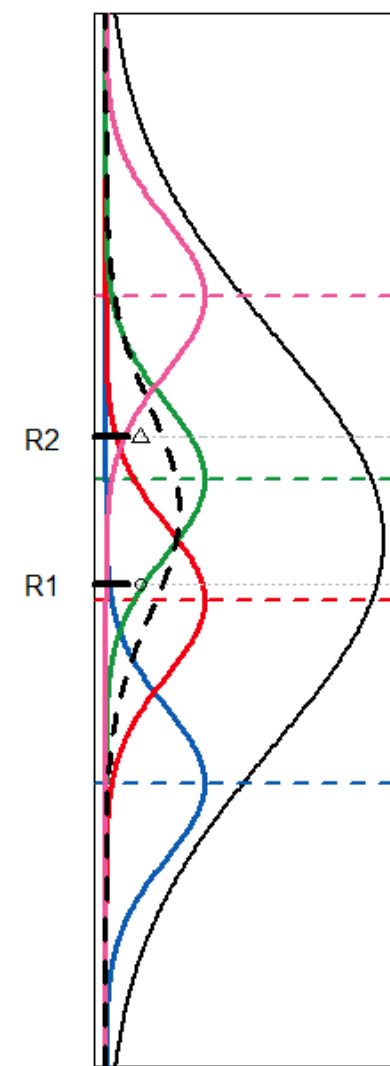
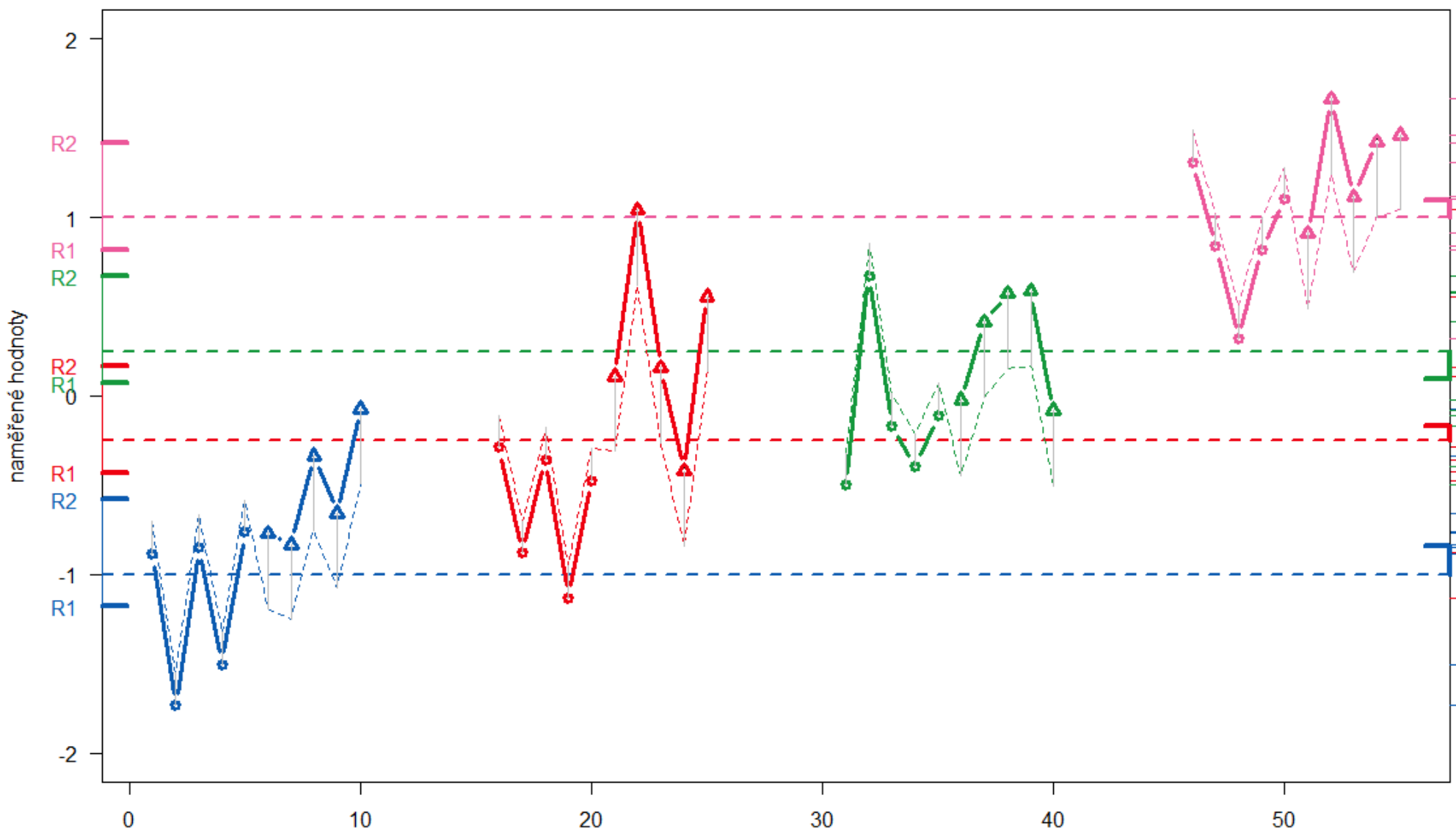


40 měření (4 osoby, každá měřena desetkrát)  
SE = 0.247; SS = 0.061

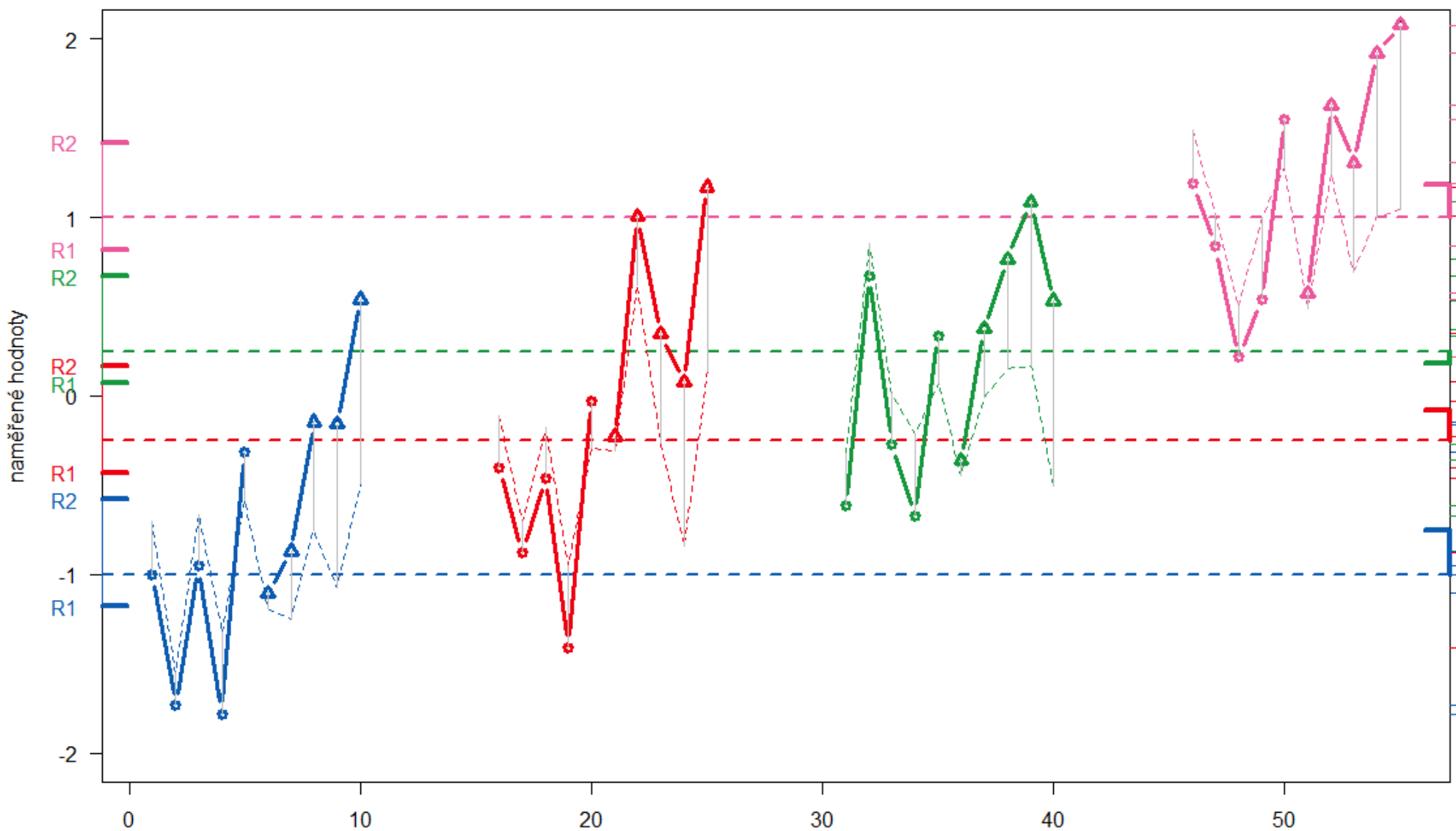


Rozložení pozorování

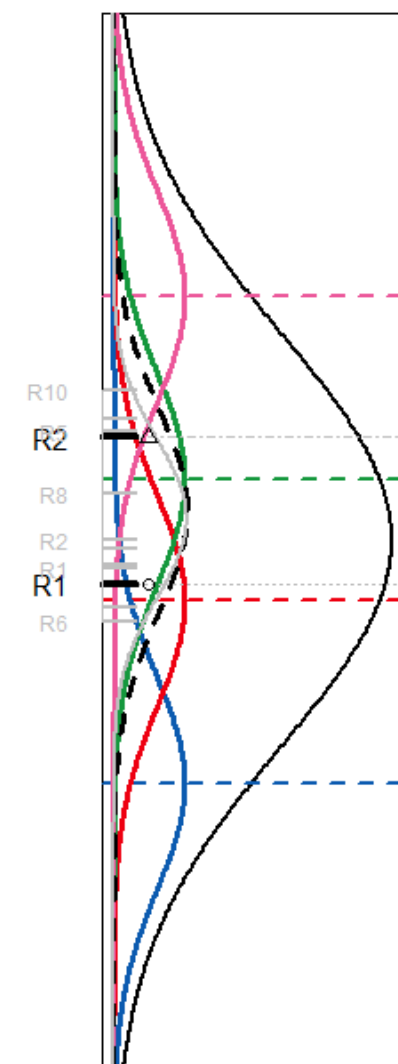
### 2Fasetový design bez interakce, $r = 0.877$



### 2Fasetový design s interakcí, $r = 0.785$



40 měření (4 osoby, každá měřena 5 dvěma hodnotiteli)  
SE = 0.44; SS = 0.194



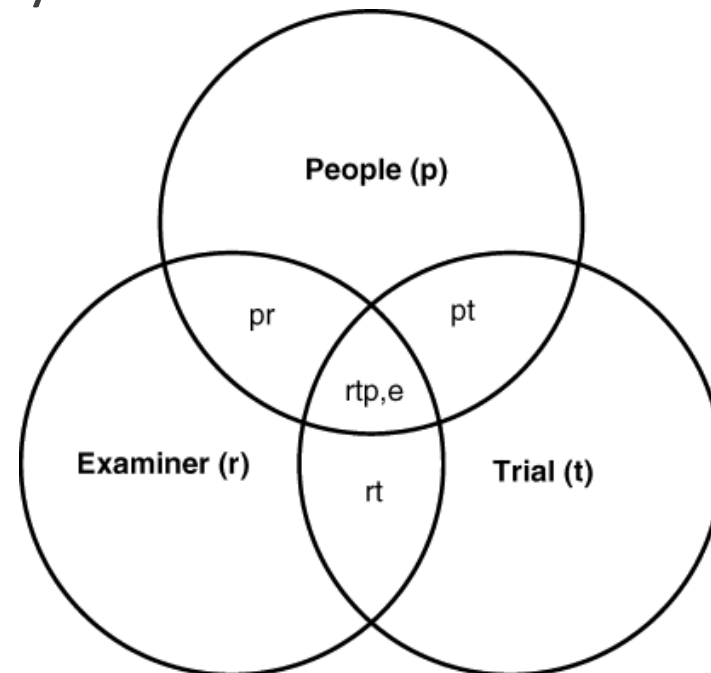
Rozložení pozorování

# Jinými slovy...

---

## CTT: Pouze dva zdroje variability

- systematický = pravý skór
- náhodný = chyba měření



## GT: Neomezeně zdrojů variability

- všechny jsou náhodné ve smyslu výběru z populace
- některé mohou být systematické napříč pozorováními
  - Pravý skór člověka napříč položkami.
  - Přísnost hodnotitele napříč osobami.

$$\begin{aligned}t &= \sigma^2 \text{ trials} \\r &= \sigma^2 \text{ examiner} \\p &= \sigma^2 \text{ people} \\e &= \sigma^2 \text{ error}\end{aligned}$$

# Jinými slovy...

---

„The theory describes the **dependability** (reliability) **of generalizations** made from a person's **observed score** on a test to the score he or she would obtain in the broad **universe of admissible observations**—her “universe score” (true score in classical test theory). Hence the name, *Generalizability Theory*.” ([Shavelson & Webb, 2006](#))

Universe score – problém s překladem. Proto malá [anketa na FB](#):

- 42 (J. Brojáč, osobní komunikace 30. 9. 2019)
- *globální skór, ideální skór* (V. Pišl, osobní komunikace 30. 9. 2019)
- *vesmírnej skór* (A. J. Kšiňan, osobní komunikace 30. 9. 2019), *skór veškera* (J. Štipl, osobní komunikace 30. 9. 2019), *skór veškerenstva* (H. Cígler & R. Modré, osobní komunikace 30. 9. 2019)
- *všeobecný skór* (M. Čadek, osobní komunikace 30. 9. 2019)
- *skór univerza, skór v univerzu, obecný skór* (A. Ťápal, osobní komunikace 30. 9. 2019)

# Princip a účel GT

---

GT zpravidla nepracuje se součtovým skóre, ale s průměrným skóre.

- „Průměrná odpověď“ napříč „prostorem“ všech možných respondentů, položek, situací...
- Reliabilita průměrného a součtového skóre je stejná (lineární transformace).
- Tato průměrná odpověď pro konkrétního respondenta se označuje jako **universe score**.
- Jednotlivé zdroje rozptylu (kromě rozdílů mezi respondenty) se označují jako fasety.

Dvě klíčové části GT:

- **G-studie:** Jak velká část rozptylu odpovědi na jednu položku v jedné situaci jedním respondentem (atd.) je „vysvětlena“ jednotlivými fasetami a rozdíly mezi respondenty samotnými?
- **D-studie:** Jaká bude chyba měření při využití „opakovaného měření“ v konkrétních fasetách – např. měření 10 položkami při 3 administracích?
  - Využívá výsledků G-studie.



# Princip a účel GT

---

Podobné předpoklady jako CTT, jde o její rozšíření.

- Zpravidla (nikoliv nutně) předpoklad náhodného výběru prvků dané fasety z nekonečně velkého doménového prostoru.
  - Existují ale i úpravy pro „finite universe“.

Další běžné předpoklady CTT.

- Jednodimenzionalita (ale existují multivariate úpravy), normální rozdělení (ale...), odpovědi na intervalové škále (ale jistá robustnost proti ordinálním položkám) atd.
- Tau-ekvivalence, resp. paralelnost položek (konkrétně stejný reziduální rozptyl, ale relativně vysoká robustnost zvláště při větším počtu položek.
  - Vícedimenzionalita možná při dodržení tau-ekvivalence faset na univerzu; analogie k hierarchické  $\rho_{SOF}$  ([Cho, 2016](#))

Některé postupy GT „zobecněly“ v běžných CTT postupech.

- Hoyt ([1941](#)), vnitrotřídní korelace (ICC, intra-class correlation; [Shrout & Fleiss, 1979](#)).

# Princip a účel GT

---

Stejně jako CTT, i GT vychází z operacionalismu.

- Měřeným atributem je universe score, nikoli psychický rys jako takový.
- Měření je tedy definováno skrze měřicí nástroj; v tomto případě spíše skrze způsob tvorby položek a popis „univerza položek“, nikoliv konkrétně vybrané položky v daném testu.

Jde tedy společně s CTT o „slabou teorii měření“, na rozdíl třeba od IRT.

- Příliš mnoho nespelnitelných předpokladů.

Logika GT je nicméně využívána i v jiných teoriích měření, kde je rozptyl měřeného rysu „parcelován“ na dílčí složky.

- Multifasetové Raschovy modely.
- Hierarchické (multilevel) IRT modely a hierarchická (multilevel) faktorová analýza.
- Explanační IRT modely (zde je parcelována obtížnost/diskriminace položek).

# G-studie

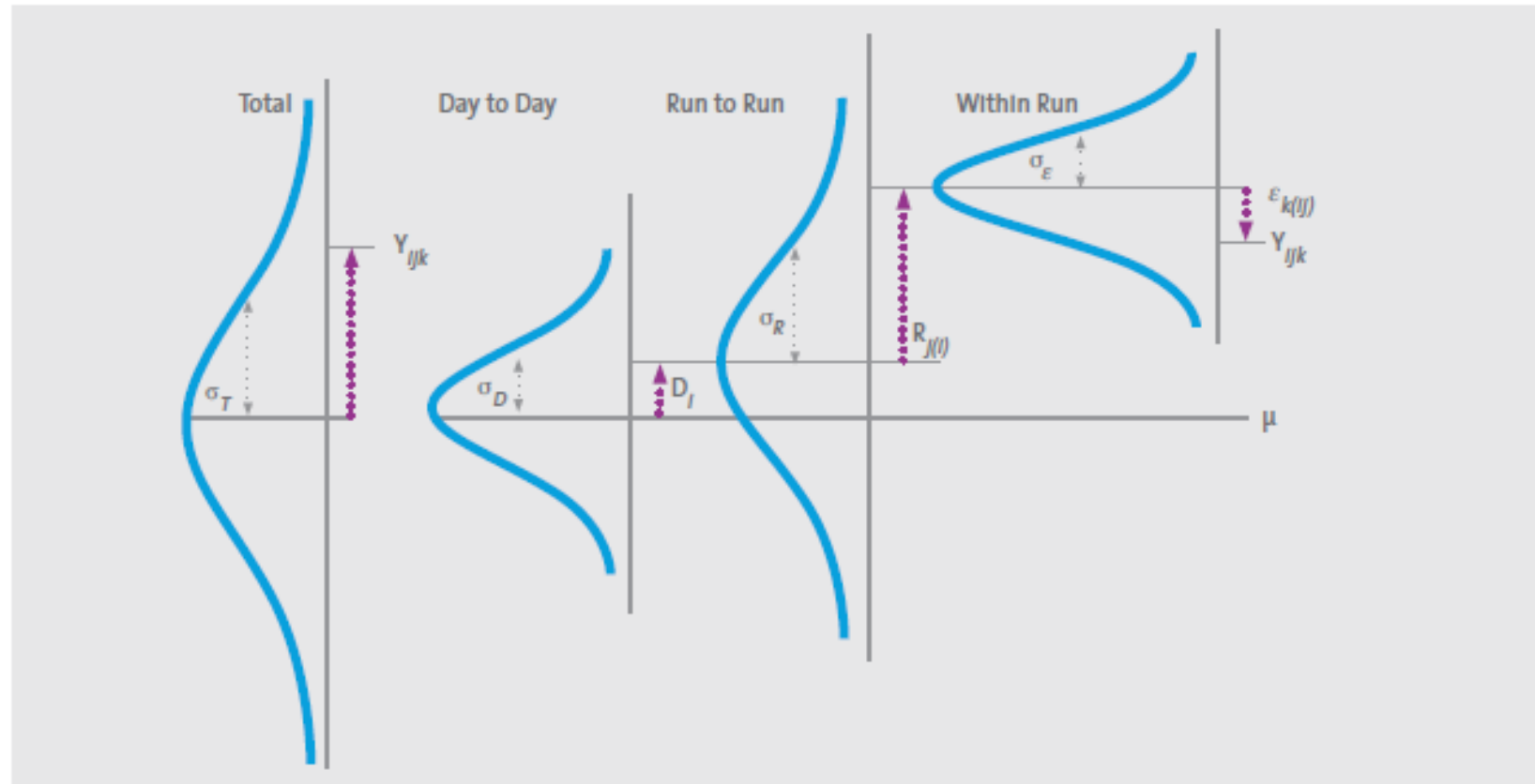
Studie zobecnitelnosti  
Generalizability study

Dekompozice rozptylu

Odhad rozptylových komponent

ANOVA

Smíšený lineární model  
(linear mixed model, LMM)



**Figure 4. Variance Decomposition.** An individual result  $Y_{ijk}$  (ie, the result for a single replicate) in a  $20 \times 2 \times 2$  study, modeled as the vector sum of deviations— $D_i + R_{k(j)} + \epsilon_{k(j)}$ —from the sample's mean,  $\mu$ , due to the combined effect of three composite sources of random variation—namely, day-to-day ( $D$ ), run-to-run within-day ( $R$ ), and within-run or residual ( $\epsilon$ ) variance components—with distributions characterized by SDs of  $\sigma_D$ ,  $\sigma_R$ , and  $\sigma_\epsilon$ , respectively. The heavy, single-headed arrows represent the magnitude and direction of the overall (total) and three source-specific deviations (shifts). The thin, double-headed arrows represent the magnitudes of the respective SDs.

# G-studie

---

## **G-studie = generalizability study (studie zobecnitelnosti)**

- Odhaduje chybový rozptyl pojící se s jednotlivými faktory a jejich interakcemi, resp. chybu pojící se s jednou položkou/jedním měřením/apod. (a jejich interakcemi).
- Jinými slovy – jakou část rozptylu jednoho pozorování (interakce respondentaxpoložkyxsituacexhodnotitelex...) tvoří specifický rozptyl respondentaxpoložky/situace/...

Zobecňuje z měření na prostor (universum).

- Na základě měření odhaduje rozptylové komponenty v prostoru.
- Tohle je ta výpočetně náročnější část GT.

# G-studie: Rozptylové komponenty

---

## KLASICKÁ TESTOVÁ TEORIE

Složení pravého skóru:

$$X = T + e$$

Rozptylové komponenty:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

- pravý skór a chyba jsou ortogonální, proto chyby jejich kovariance („+2 $\sigma_{te}^2$ “)

## TEORIE ZOBECNITELNOSTI

Složení obecného skóru – např. 2fasetový design:

$$X = T + e_1 + e_2 + e_{\tau_1} + e_{\tau_2} + e_{12} + e_{\tau_{12},e}$$

Rozptylové komponenty:

$$\sigma_x^2 = \sigma_t^2 + \sigma_1^2 + \sigma_2^2 + \sigma_{\tau_1}^2 + \sigma_{\tau_2}^2 + \sigma_{12}^2 + \sigma_{\tau_{12},e}^2$$

- Veškeré rozptylové chyby (jak obecné skóry, tak chyby) jsou na sebe ortogonální, proto chybí jejich kovariance.
- Pokud jsou zahrnuty všechny možné složky včetně interakcí, jsou ortogonální z principu.
- Z toho důvodu se zahrnují i nesignifikantní efekty.

# G-studie: příklad

Příklad: 2fasetový design  $p \times i \times o$ .

- N respondentů  $p$  (persons)
  - Osoby jsou tam vždy, proto se nepočítají do počtu faset
- 3 položky  $i$  (items)
- 2 administrace/situace  $o$  (occasions)

Pozorovaný skór  $X$ :

- $X_p = \text{mean}(X_{pio}); E(X_{pio}) = T_p$

Pozorovaný skór je součtem všech komponent:

$$X_{pio} = T_p + e_i + e_o + e_{p \times i} + e_{p \times o} + e_{i \times o} + e_{p \times i \times o}$$

Celkový rozptyl pozorovaného skóre (prvků datové matice):

$$\sigma_{X_{pio}}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{io}^2 + \sigma_{pio,e}^2$$

**TABLE 36-1**  
**Crossed Person  $\times$  Item  $\times$  Occasion G Study of Self-Concept Scores**

<i>Person</i>	<i>Occasion</i>					
	<i>I</i>			<i>II</i>		
	<i>Item 1</i>	<i>Item 2</i>	<i>Item 3</i>	<i>Item 1</i>	<i>Item 2</i>	<i>Item 3</i>
1	4	2	5	4	3	4
2	3	1	4	4	2	3
3	2	3	3	3	2	4
...						
<i>p</i>	4	5	4	3	4	2
...						
<i>N</i>	3	4	4	3	3	3

**Table 36–2**  
**Estimated Variance Components in the Example  $p \times i \times o$  design**

<i>Source</i>	<i>Variance Component</i>	<i>Estimate</i>	<i>Percent of Total Variability</i>
Person (p)	$\sigma_p^2$	1.108	30
Item (i)	$\sigma_i^2$	0.102	03
Occasion (o)	$\sigma_o^2$	0.030	01
$p \times i$	$\sigma_{pi}^2$	0.810	22
$p \times o$	$\sigma_{po}^2$	0.230	06
$i \times o$	$\sigma_{io}^2$	0.001	00
$p \times i \times o, e$	$\sigma_{pio,e}^2$	1.413	38

# G-studie: Odhad rozptylových komponent

---

Historicky GT využívala ANOVA.

- Fasety – „faktory“ v tradiční ANOVA terminologii.
- Proměnné jsou uvažovány jako random (např. náhodný výběr času) nebo fixed effect (např. test stabilně složený ze stejných položek).
  - Random modely jsou častější.

Aktuálně se zpravidla používá LMM (linear mixed model).

- Výhody při odhadu.
  - Unbalanced designy, chybějící data apod.
- Menší předpoklady, vyšší flexibilita.
- Výsledek by se neměl lišit (při dodržení předpokladů), reálně jsou odlišnosti malé.



# GT: SW pro odhad G-studie

---

## Tradiční SW:

- GENOVA, mGENOVA (staré DOSovské aplikace)
- Různé podivné malé programky (G String V)

## SPSS (lze ručně upravit syntax pro mixed-modely).

- Mushquash, C. and O'Connor, B.P. (2006). SPSS and SAS programs for generalizability theory analyses, *Behavior Research Methods*, 38(3), 542–547. doi: [10.3758/bf03192810](https://doi.org/10.3758/bf03192810)

## **R, zejména balíček lme4 (mixed modely) a případně gtheory (nástavba lme4).**

- Případně pak [hemp](#) dostupný na githubu (doplněk ke knize [Desjardins & Bulut, 2018](#)).

## Přehled dostupných programů:

- Taşdelen Teker, G., Güler, N. and Kaya Uyanık, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for Generalizability theory. *Educational Sciences: Theory & Practice*, 15(3). doi: [10.12738/estp.2015.3.2278](https://doi.org/10.12738/estp.2015.3.2278)
- Yelboga, A. (2015). Estimation of Generalizability coefficient: An application with different programs. *Archives of Current Research International*, 2(1), 46–53. doi: [10.9734/acri/2015/17409](https://doi.org/10.9734/acri/2015/17409)



# GT: Způsob odhadu G-studie v R

V předchozím případě by syntax pro R byl:

- Předpokladem je převedení na tzv. dlouhý formát, kde jeden řádek = 1 odpověď, a další proměnné jsou person (1-N), item (1-3), occasion (1-2)

```
require(lme4)
require(gtheory)
```

```
model <- "response ~ (1 | person) + (1 | item) + (1 | occasion) + (1 | person:item) + (1 | person:occasion) + (1 | item:occasion)"
```

Pozn.: poslední chybovou fasetou je (1 | person:item:occasion) – ta reprezentuje „zbytek“ a je proto chybou v klasickém slova smyslu (vše, co není vysvětleno ničím předchozím), proto ji není nutné do modelu zadávat.

```
gstudy <- gstudy(data = data, formula = model)
```

```
print(gstudy)
```

odp.	P	I	O	odp.	P	I	O
4	1	1	1	3	1	2	2
2	1	2	1	4	1	3	2
5	1	3	1	3	2	1	1
4	1	1	2	1	2	2	1

**TABLE 36-1**  
Crossed Person × Item × Occasion G Study of Self-Concept Scores

Person	Occasion					
	I			II		
	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
1	4	2	5	4	3	4
2	3	1	4	4	2	3
3	2	3	3	3	2	4
...						
<i>p</i>	4	5	4	3	4	2
...						
<i>N</i>	3	4	4	3	3	3

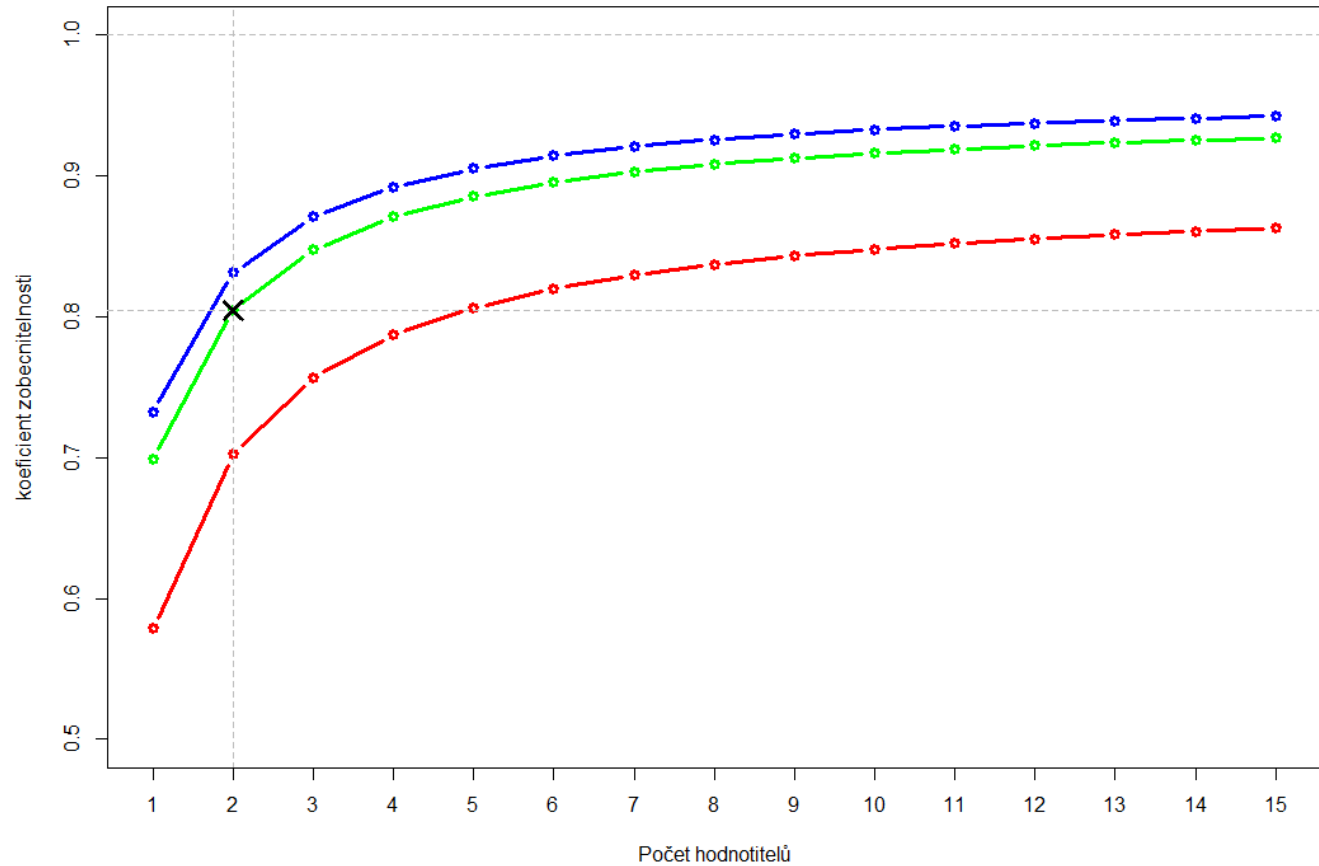
# D-studie

Rozhodovací studie  
Decision study

Koeficient dependability  
a zobecnitelnosti

Absolutní a relativní D-studie

Přijímací zkoušky do NMGR psychologie PS2020



# D-studie

---

Rozhodovací (Decision) studie slouží k odhadu chyby měření pro konkrétní design s využitím informací z G-studie.

Definuje tzv. „prostor zobecnění“ (počtem pozorování, počtem položek atp.), pro který bude naše měření platit.

- V rámci tohoto prostoru má každý respondent tzv. U-skór (universe).

Odhad chyby odhadu universe skóru pro zvolený hypotetický design – např.  $p \times l \times O$ .

# D-studie: Obecný postup

---

1. Volba jednotky/subjektu měření (nemusí být respondent).
2. Volba designu, resp. prostoru/prostorů zobecnění.
3. Identifikace chybových složek.
4. Volba počtu prvků faset (nemusí se shodovat s G-studií).
5. Výpočet chyby odhadu.
6. Výpočet koeficientu reliability.

# D-studie: Dva typy zobecnění

---

**Relativní (norm-referenced)** – zobecnění v rámci vybraných prvků fasety.

- Všechny fasety jsou zafixovány napříč jednotkami měření.
- Např. test složený z pevného setu položek.
- Díky fixaci se jejich prvky stanou konstantou.
- Reliabilita odhadována pomocí **koeficientu zobecnitelnosti**.
- Přímo srovnatelný s různými druhy CTT reliability.

**Absolutní (kriteriální)** – zobecnění na celou fasetu.

- Tento odhad nese více nejistoty.
- Reliabilita odhadována pomocí **koeficientu spolehlivosti** (dependability coef.).
- Lze uvažovat pravděpodobnost překročení absolutního kritéria.

Spíše než otázka celého designu otázka dílčích faset (smíšený design).

# D-studie: Odhad chyby měření

---

Celková chyba odhadu obecného skóru = suma čtverců chyb odhadu komponent.

- Chyba odhadu dílčí komponenty = standardní chyba průměru<sup>1</sup>.
- Tedy rozptylová komponenta z G-studie dělená počtem pozorovaných prvků dané fasety:

$$\sigma_e^2 = \frac{\sigma_{e1}^2}{n_1} + \frac{\sigma_{e2}^2}{n_2} + \frac{\sigma_{e3}^2}{n_3} + \dots + \frac{\sigma_{ek}^2}{n_k}$$

Reliabilita se potom spočítá dle obecného vzorce pro vysvětlený rozptyl:

$$r_{xx'} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2}$$

- $\sigma_\tau^2$  - rozptyl jednotek měření, tedy universe skóru
- $\sigma_e^2$  - chybový rozptyl, tedy součet všech chybových komponent

<sup>1</sup> standardní chyba průměru SEM:  $SEM = \frac{SD}{\sqrt{N}} \rightarrow SEM^2 = \frac{SD^2}{N}$ ; SD – směrodatná odchylka; N – velikost vzorku/počet pozorování

**Table 36–2**  
**Estimated Variance Components in the Example  $p \times i \times o$  design**

<i>Source</i>	<i>Variance Component</i>	<i>Estimate</i>	<i>Percent of Total Variability</i>
Person (p)	$\sigma_p^2$	1.108	30
Item (i)	$\sigma_i^2$	0.102	03
Occasion (o)	$\sigma_o^2$	0.030	01
$p \times i$	$\sigma_{pi}^2$	0.810	22
$p \times o$	$\sigma_{po}^2$	0.230	06
$i \times o$	$\sigma_{io}^2$	0.001	00
$p \times i \times o, e$	$\sigma_{pio,e}^2$	1.413	38



# Relativní D-studie: Příklad

---

Jaká bude chyba průměrného skóre ze 2 administrací 10položkového testu?

**Relativní chybový rozptyl  $\sigma_{\delta}^2$ :**

$$\sigma_{\delta}^2 = \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} = \frac{.810}{1 \times 10} + \frac{.230}{1 \times 2} + \frac{1.413}{1 \times 10 \times 2} = .267$$

Podíl chybového rozptylu (reliabilita): **koeficient zobecnitelnosti:**

$$G = E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\delta}^2} = \frac{1,108}{1,108 + 0,267} = \mathbf{0,806}$$

Koeficient zobecnitelnosti je přímo srovnatelný s reliabilitou v CTT ( $N_p = 1$  vynecháno):

- Vnitřní konzistence 1 měření v 1 okamžik:

$$\sigma_{\delta}^2 = \frac{\sigma_{pi}^2}{N_i} + \frac{\sigma_{pio,e}^2}{N_i \times N_o} = \frac{.810}{10} + \frac{1.413}{10 \times 1} = .222 \rightarrow G = \frac{1,108}{1,108 + 0,222} = 0,833$$

# Absolutní D-studie: Příklad

Jaká bude chyba průměrného skóre 10 položek a 2 měření, pokud každý absolvuje jiné položky?

- Nebo jde o kriteriální test, kdy nás zajímá skóre napříč všemi přípustnými položkami (admissible observation).

**Absolutní chybový rozptyl**  $\sigma_{\Delta}^2$ :

$$\begin{aligned}\sigma_{\Delta}^2 &= \frac{\sigma_i^2}{N_i} + \frac{\sigma_o^2}{N_o} + \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{io}^2}{N_i \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} \\ &= \frac{.102}{10} + \frac{.030}{2} + \frac{.810}{1 \times 10} + \frac{.230}{1 \times 2} + \frac{.001}{10 \times 2} + \frac{1.413}{1 \times 10 \times 2} = .292\end{aligned}$$

Podíl chybového rozptylu: **koeficient spolehlivosti**  $\Phi$  (dependability):

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} = \frac{1,108}{1,108 + 0,292} = 0,791$$

Pokud zjišťujeme spolehlivost překročení absolutního kritéria  $\lambda$ :  $\Phi_{\lambda} = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_{\Delta}^2}$

- $\Phi_{\lambda}$  je vyšší, čím dále je kritérium  $\lambda$  od průměru  $\mu$ .

# Absolutní D-studie: Příklad

Jaká bude chyba průměrného skóre 10 položek a 2 měření, pokud každý absolvuje jiné položky?

- Nebo jde o kriteriární test, kdy nás zajímá skóre napříč všemi přípustnými položkami (admissible observation).

**Absolutní chybový rozptyl**  $\sigma_{\Delta}^2$ :

$$\begin{aligned}\sigma_{\Delta}^2 &= \frac{\sigma_i^2}{N_i} + \frac{\sigma_o^2}{N_o} + \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{io}^2}{N_i \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} \\ &= \frac{.102}{10} + \frac{.030}{2} + \frac{.810}{1 \times 10} + \frac{.230}{1 \times 2} + \frac{.001}{10 \times 2} + \frac{1.413}{1 \times 10 \times 2} = .292\end{aligned}$$

Podíl chybového rozptylu: **koeficient spolehlivosti**  $\Phi$  (dependability):

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} = \frac{1,108}{1,108 + 0,292} = 0,791$$

Pokud zjišťujeme spolehlivost překročení absolutního kritéria  $\lambda$ :  $\Phi_{\lambda} = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_{\Delta}^2}$

- $\Phi_{\lambda}$  je vyšší, čím dále je kritérium  $\lambda$  od průměru  $\mu$ .

# Smíšená D-studie: Příklad

---

Jaká bude test-retest reliabilita?

- 10 položek: relativní faseta (zobecňujeme na těchto 10 položek, ne na všechny možné).
- 1 situace: absolutní faseta (zobecňujeme na všechna možná pozorování).

Chybový rozptyl:

$$\sigma_{\delta}^2 = \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} = \frac{.810}{1 \times 10} + \frac{.230}{1 \times 1} + \frac{1.413}{1 \times 10 \times 1} = .452$$

Koeficient zobecnitelnosti:

$$G = \frac{1,108}{1,108 + 0,452} = 0,710$$

# D-studie: absolutní

---

Uvažuje veškeré faktory jako náhodné, přičemž tyto faktory se mohou lišit mezi respondenty.

Případně nás zajímá skór napříč všemi potenciálními prvky všech faset (typicky u kriteriálních výkonových testů):

- Relativní: 70 % správně z daných 10 položek.
- Absolutní: 70 % správně ze všech možných položek.

Zobecňuje tedy na universe score napříč celým (nejvyšším) prostorem.

- Náhodný výběr položek, časů, hodnotitelů ze všech možných atd.
- Tento universe score bude mít tedy vyšší chybu než universe score.

# Srovnání designů

---

- Relativní D-studie ze 2 měření  $p \times (I=10) \times (O=2)$ :  $G = 0,806$
- I, O fixed
- Relativní D-studie z 1 měření  $p \times (I=10)$ :  $G = 0,833$
- I fixed, O vynecháno
- Absolutní D-studie ze 2 měření  $p \times (I=10) \times (O=2)$ :  $\Phi = 0,791$
- I, O random
- Smíšená D-studie, test-retest z 1 měření  $p \times (I=10) \times (O=1)$ :  $\Phi = 0,710$
- I fixed, O random

# Využití GT

---

Odhad reliability/chyby měření.

Vývoj testu: jak se změní reliabilita, pokud použiju jiný počet prvků z domény?

- S minimální finanční/časovou náročností maximalizovat reliabilitu testu.
- Obdoba Spearman-Brownova věšteckého vzorce, ale pro více zdrojů chyb než „počet testů“.

GT je velmi cenná v případě, že máme skutečně paralelní položky.

- Např. tzv. škrtačí testy pro měření reakčního času, kde jsou dílčí položky řazené do bloků (a třeba testované opakovaně).

# Využití GT: Optimální počet prvků faset

Seminární práce. Variují:

- počtem hodnotitelů;
- počtem hodnocených prací.

Pokud např. chci investovat na každého studenta max. čtyři hodnocené práce, co je nejvýhodnější?

- A) 4 pokusy, 1 hodnotitel
- B) 2 pokusy, 2 hodnotitelé
- C) 3 pokusy, 1 hodnotitel
- D) 1 pokus, 4 hodnotitelé

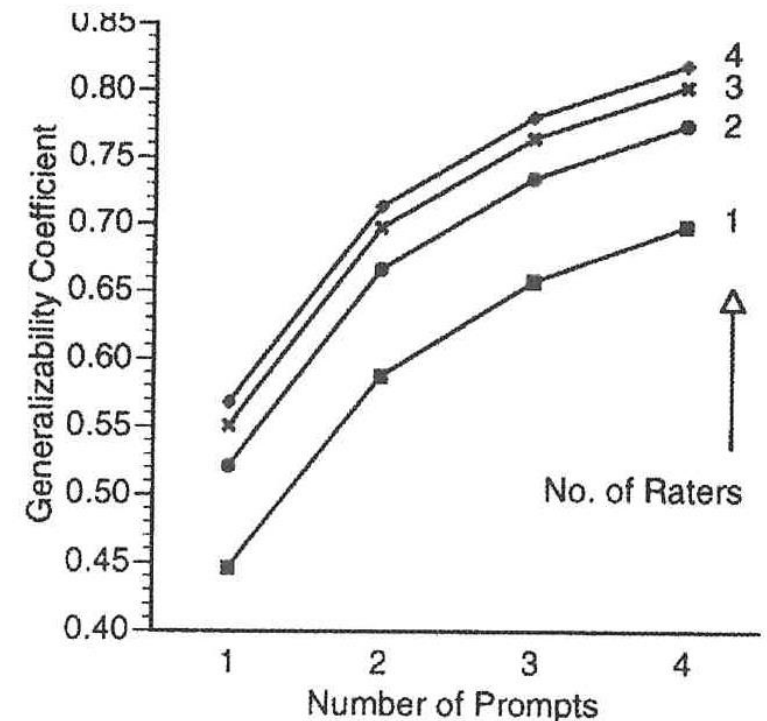
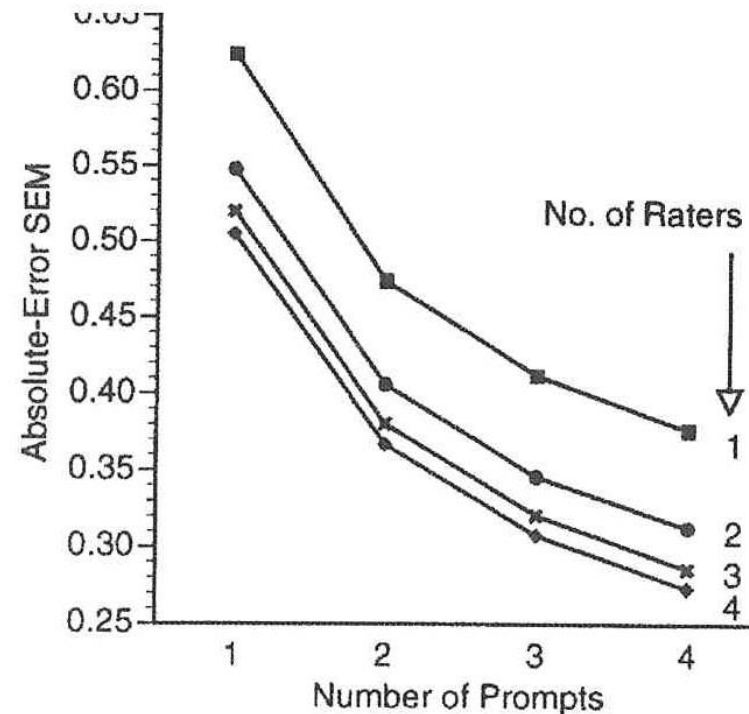


FIGURE 1.2.  $\hat{\sigma}(\Delta)$  and  $E\hat{\rho}^2$  for scenario with  $p \times T \times R$  design.



# Využití GT: Multilevel design

---

Prvkem měření nemusí být respondent, ale např. školní třída (pak je faseta „žáci“ chybou).

Občas nejsou prvky „crossed“, ale „nested“. Např. žáci patří právě do jedné třídy, nepozorujeme je ve více třídách (c=class, S=student, I=item):

- G-studie:  $(s:c) \times i$
- D-studie pro žáka *uvnitř* třídy:  $(s:C) \times i$  (C je fixed)
- D-studie pro žáka *napříč* třídami:  $(s:C) \times i$  (C je random)

Pokud byl design G-studie rozsáhlejší než design D-studie, může se stát, že se rozptyl universe skóru skládá z více rozptylových komponent.

- V příkladu výše zobecnění výkonu žáka uvnitř vs. napříč třídami.
- Doporučuji držet stejný design D a G studií, jinak se vše značně komplikuje (ale specifikační chyba v G-studii...).

# G-studie vs. D-studie

---

## G-STUDIE

Zaměřuje se na rozptylové komponenty.

- Odhad jejich velikosti.

Design např.:  $p \times t \times r$

- Malá písmena značí rozptylové komponenty.

Vychází z dat.

- Zobecňuje z měření na prostor, tvoří model.
- Nejlépe cross-design.

## D-STUDIE

Zaměřuje se na odhad chyby měření.

- A reliability.

Design např.:  $p \times T \times R$

- Velká písmena značí pozorování.

Vychází z modelu G-studie.

- Zobecňuje z prostoru na měření.
- Volíme design dle účelu.

# GT: závěrem

---

Při zobecnění na více položek shodné výsledky s S-B vzorcem.

Lze mít také více závislých proměnných (multivariate analysis of variance, MANOVA):

- Odhad reliability kompozitu, rozdílových skóřů, profilu apod.
- Analogie k velmi zjednodušenému strukturnímu modelu.

Výhodné při standardizaci testů, kde je přítomno více zdrojů chyb

- Např. examinátor-retest-položky.
- Minimum výhod při využití prostého odhadu test-retest reliability pomocí korelace celkových skóřů, GT poskytne více informací.

Nepříliš doceněná (člověk musí rozumět, aby mohl použít).

Doporučuji: Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

- Drobné texty viz studijní materiály.

# Vnitrotřídní korelace pro P×I design

	Shrout a Fleiss (nejběžnější)	McGraw a Wong (občasně používané)	GT design
Ukazatel shody posuzovatelů. Reliabilita při hodnocení 1 posuzovatelem.	ICC(1,1)	One-way random, single score ICC(1)	p (jediná faseta plus error, $N_e=1$ ) <i>Hodnotitelé se neopakují.</i>
	ICC(2,1)	Two-way random, single score ICC(A,1)	p×I (absolutní, $N_i = 1$ ) <i>Stejní hodnotitelé, vybrání náhodně.</i>
	ICC(3,1)	Two-way mixed, single score ICC(C,1)	p×I (relativní, $N_i = 1$ ) <i>Stejní hodnotitelé, nezobecňují na všechny možné.</i>
Reliabilita celkového hodnocení, tj. průměru všech posuzovatelů.	ICC(1,k)	One-way random, average score ICC(k)	p (jediná faseta plus error, $N_e=k$ )
	ICC(2,k)	Two-way random, average score ICC(A,k)	p×I (absolutní, $N_i = k$ )
	ICC(3,k)	Two-way mixed, average score ICC(C,k)	p×I (relativní, $N_i = k$ ) ICC(3,k) = Cronbachovo $\alpha$

A=agreement (shoda hodnocení), C=consistency (konzistence pořadí), k=počet hodnotitelů/skupin.