MUNI FSS Katedra psychologie

Přednáška 6: Teorie zobecnitelnosti

17. 10. 2021 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler | 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 = reliabilita paralelních forem:** korelace, regrese, ICC...
- **shoda posuzovatelů:** Cohenovo/Fleissovo kappa, Krippendorfova alfa, ICC...

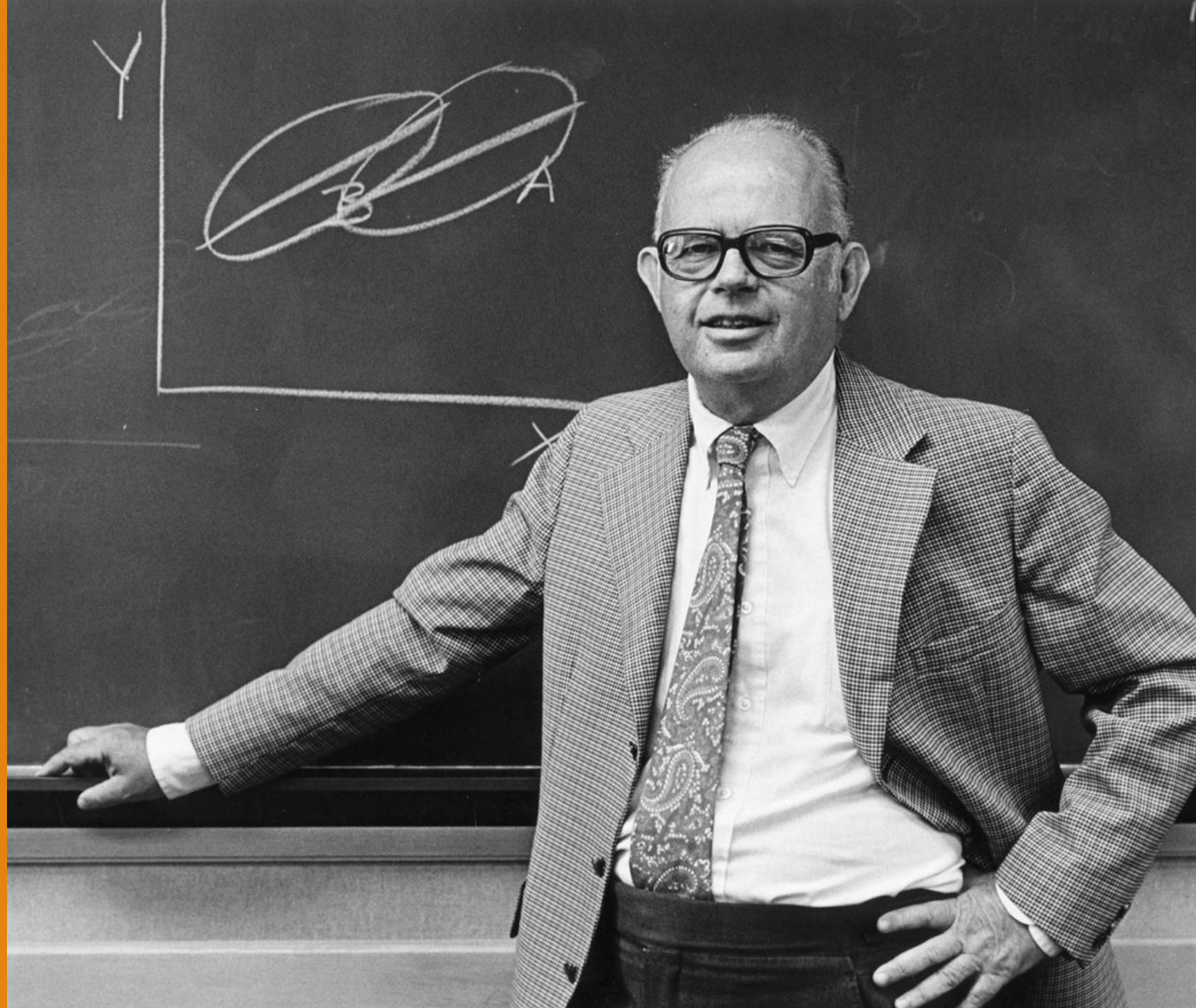
  - To „mnoho“ znamená něco jiného než „mnoho“ vnitřních konzistencí na předchozí přednášce o CTT. V tomto ohledu více viz Ellis, J. L. (2021). A Test Can Have Multiple Reliabilities. *Psychometrika*, *86*(4), 869–876. https://doi.org/10.1007/s11336-021-09800-2

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

- Pro různé účely?

Teorie zobecnitelnosti

Generalizability Theory (GT)

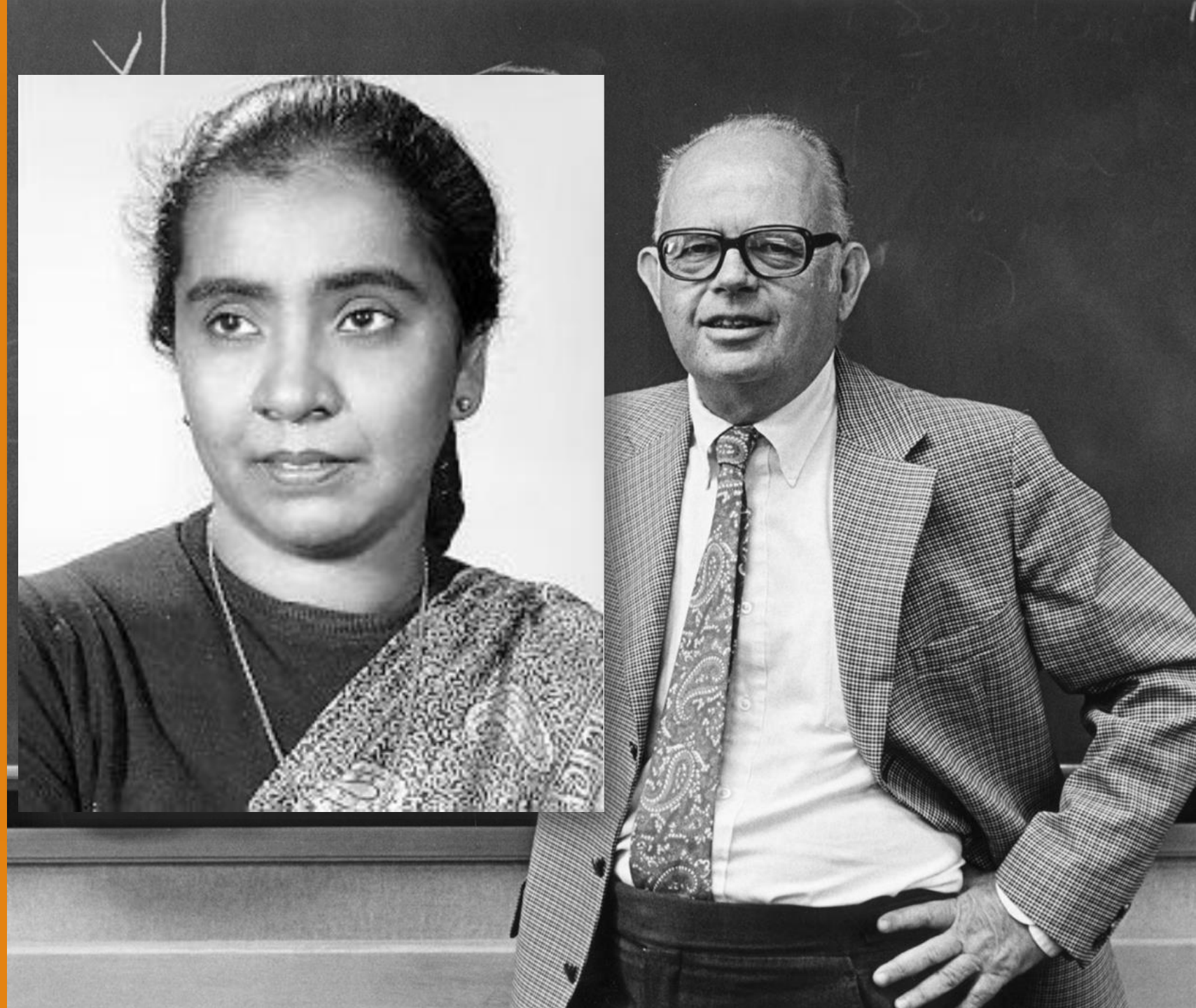


Lee J. Cronbach (1916–2001)

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

Teorie zobecnitelnosti

Generalizability Theory (GT)



Lee J. 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.

Cronbach, 1991, cit. dle Brenan (2001, s. 3)

Klíčové publikace

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. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* John Wiley.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Primer.* Sage.

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

- Tohle je nejvíce recentní a komprehenzivní zdroj, který je aktuálně k dispozici.

Cronbachovo alfa

Cronbachova alfa (1951) není tak docela Cronbachova:

- KR-20 (Kuder-Richardson, 1937); Rulonův vzorec (1939); Guttmanova korekce s λ_3 (1945); Hoytův vzorec (1941).

Cyril Hoyt (1941) – odhad reliability pomocí ANOVA:

$$r_{xx'} = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = \frac{(n-1)(\sigma_x^2 - \sigma_e^2)}{(n-1)\sigma_x^2} = \frac{MS_\tau}{MS_x} = \frac{MS_x - MS_e}{MS_x}$$

- MS_x - mean-square, tj. průměr sumy čtverců, tj. nepodělený rozptyl ($\mathrm{var}(x) = \frac{MS_x}{n-1}$).

ANOVA umí „parcelovat“ pozorovaný rozptyl (**AN**alysis **O**f **Va**riance).

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

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160. https://doi.org/10.1007/BF02288391
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 analyzint 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).

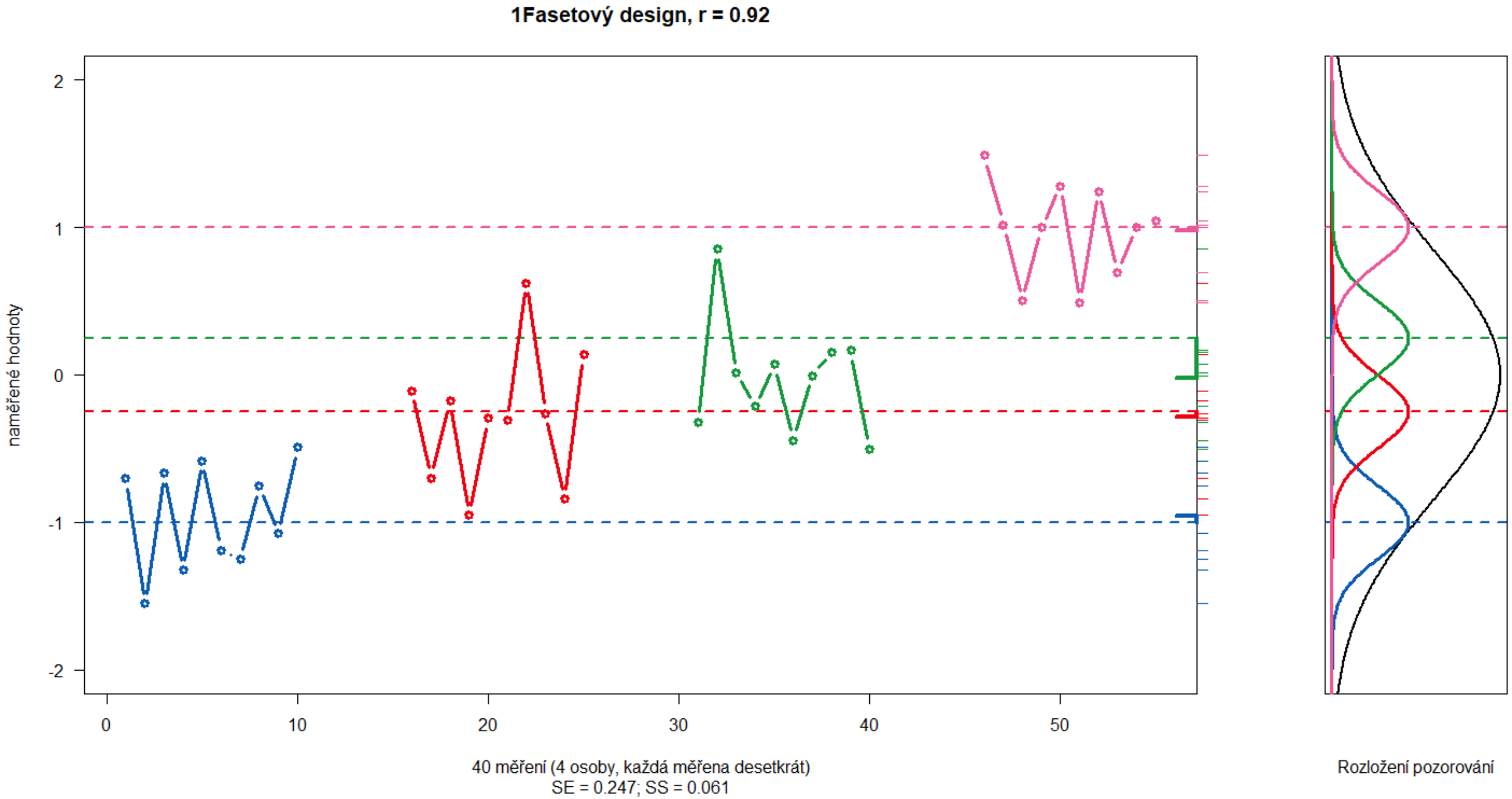
Postup výpočtu (přibližně, bez korekcí):

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

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

  - Kde NI je počet *osob* (N) krát počet *položek* (I; celkový počet pozorování = kusů informace).
- 3. Standardní chyba odhadu průměru z I položek jako $\sigma_e = \sqrt{\frac{\sigma_{res}^2}{I}} \sim \frac{SD}{\sqrt{N}}$
  - Výpočet „přes všechny osoby“, obchází tak potíže s opakovaným měřením jedné osoby – chyba je stejná pro všechny (viz minulá přednáška).
- 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, asymptomaticky se rovná průměru všech možných split-half reliabilit (**Cronbach, 1951**).


1Fasetový design, r = 0.92
naměřené hodnoty
2
1
0
-1
-2
0
10
20
30
40
50
40 měření (4 osoby, každá měřena desetkrát)
SE = 0.247; SS = 0.061
Rozložení pozorování

Hoytův postup

Připomenutí: postup je ekvivalentní koeficientu alfa.

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

- Paralelnost položek (shodné reziduální rozptyly) lze obejít skrze „průměrnou chybu“.

Neexistence jiného zdroje rozptylu, než:

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

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.

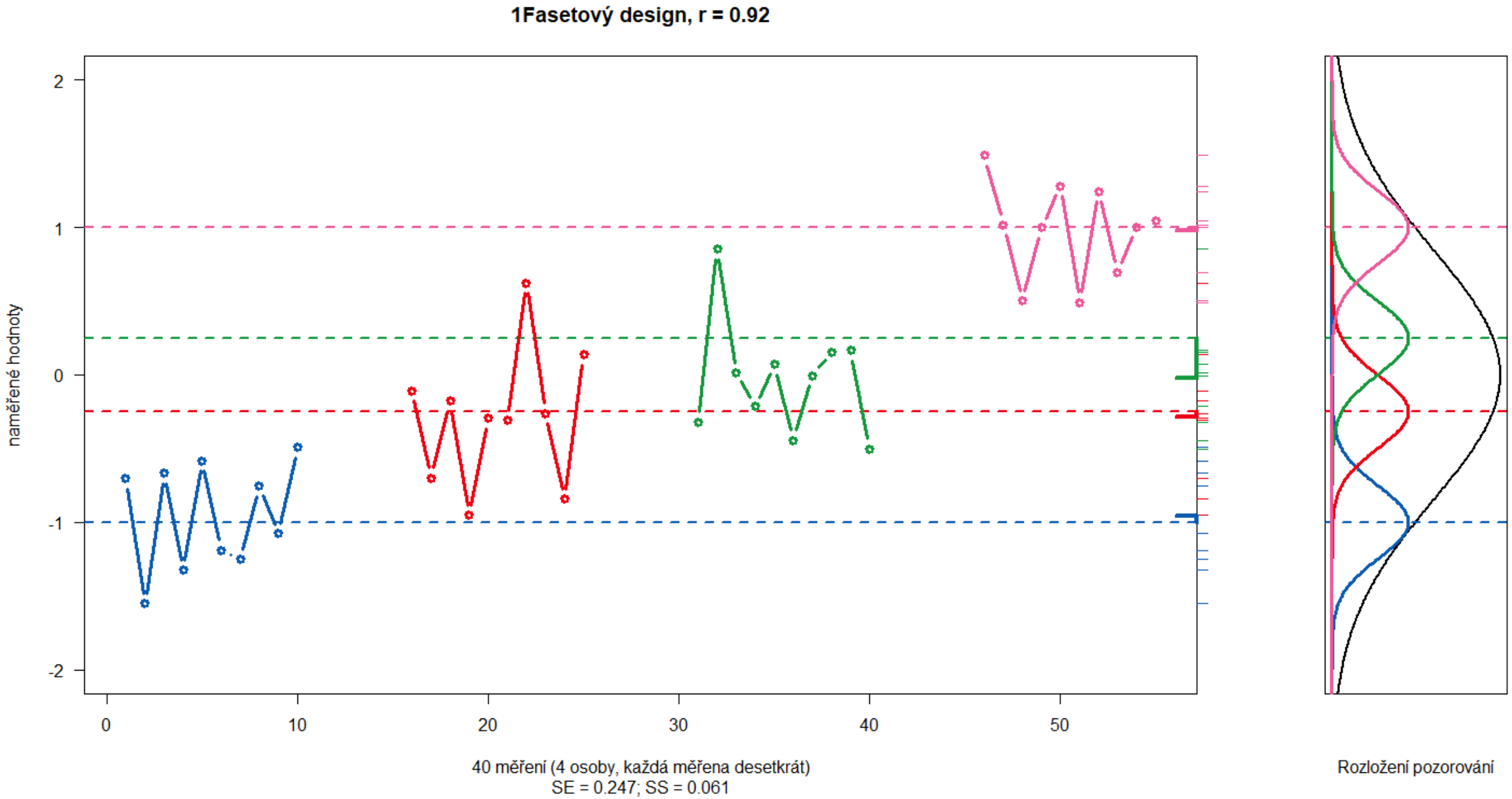
CTT: $X = T + e$

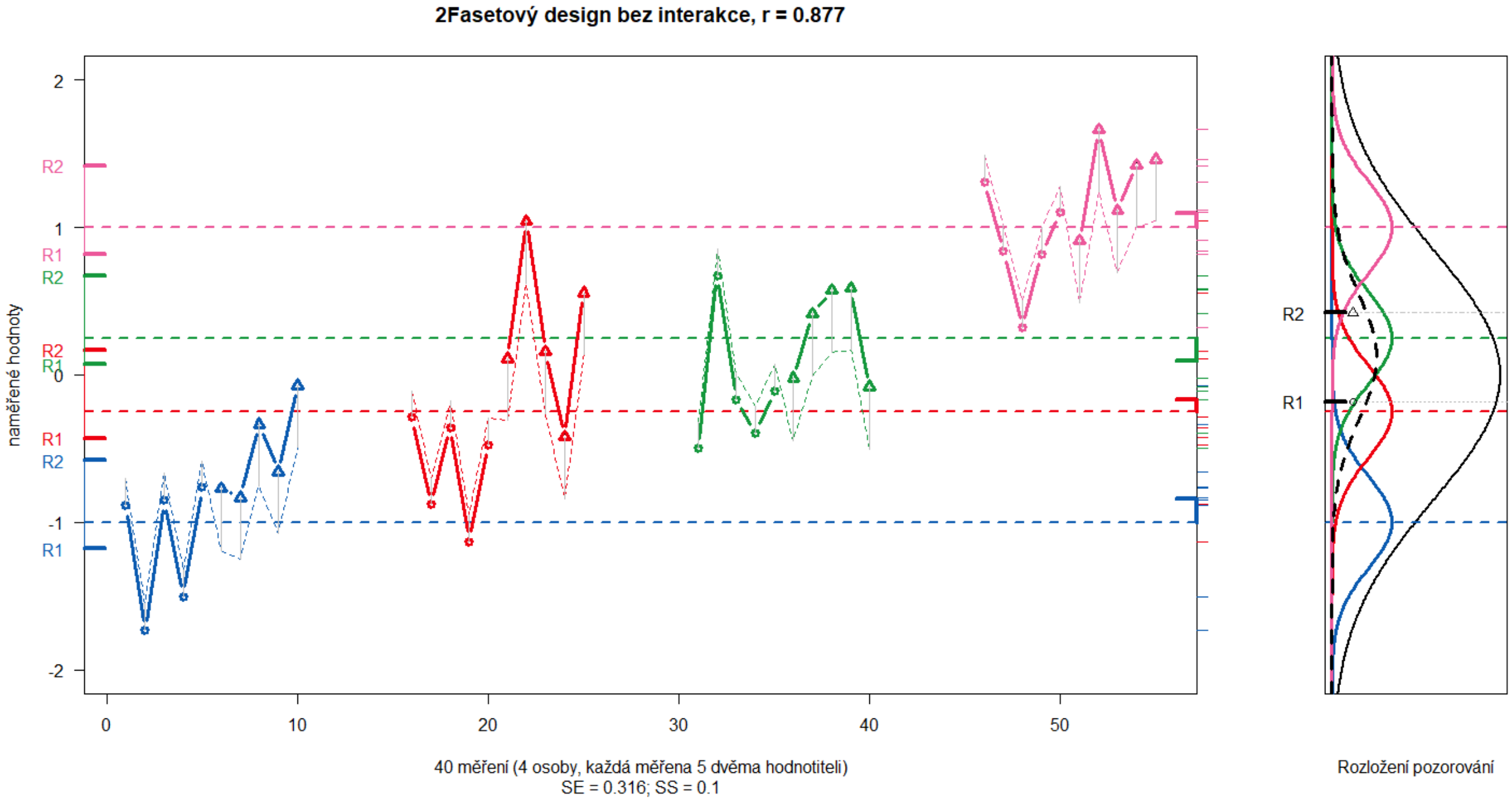
GT: $X = T + e_1 + e_2 + e_3 + \ \ldots + e_k$

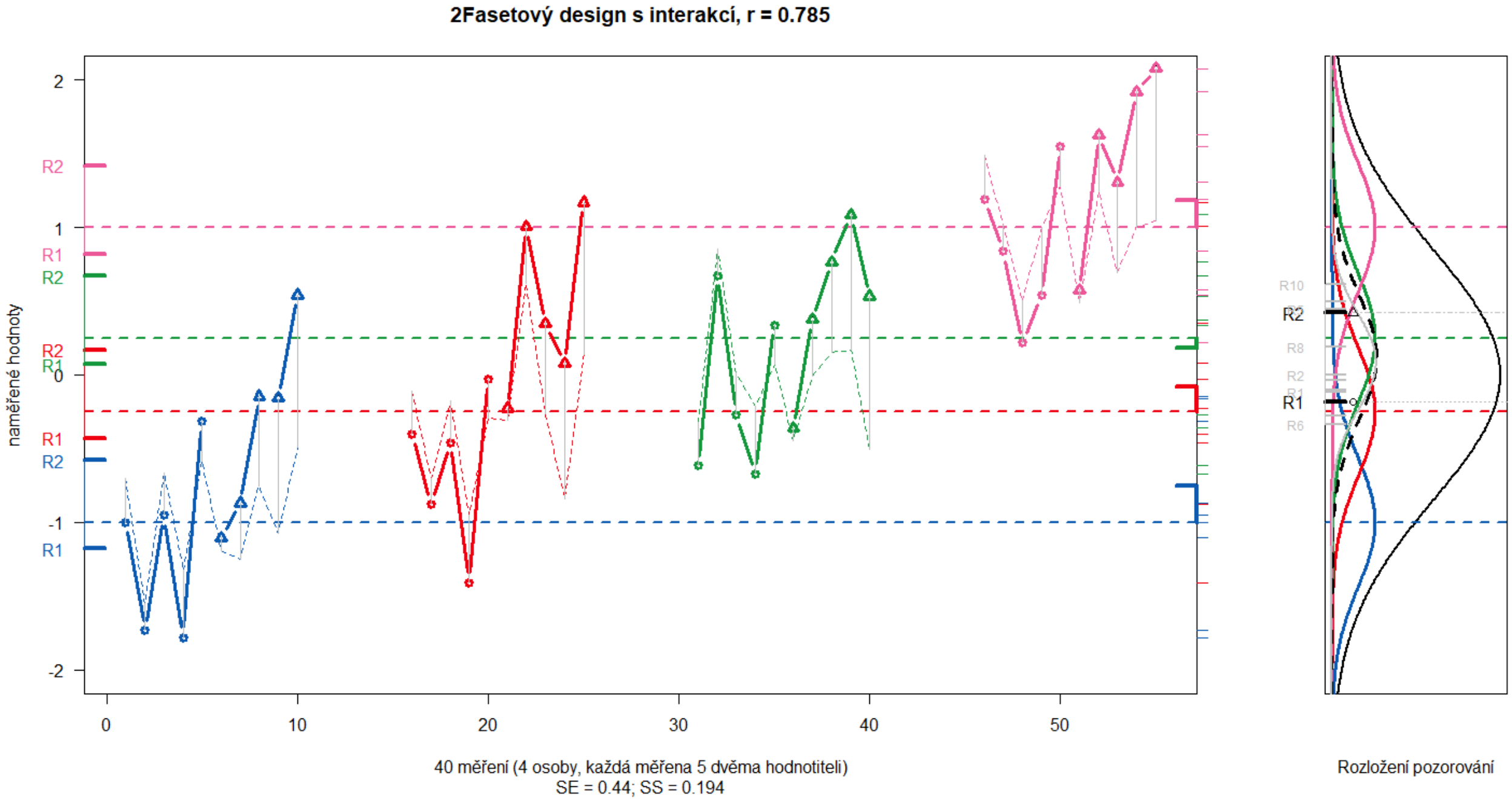
- Kde např. e_1 je specifický skór v daném čase (*test-retest*), e_2 rozdílnost posuzovatelů (*shoda posuzovatelů*), e_3 rozdílnost položek (vlastní „nepřesnost metody“, *vnitřní konzistence*) atd.
- Pro různé účely může *T* zahrnovat i některé chyby (např. nás zajímá výkon v daném čase a nikoliv stabilita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

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


1Fasetový design, r = 0.92
naměřené hodnoty
2
1
0
-1
-2
0
10
20
30
40
50
40 měření (4 osoby, každá měřena desetkrát)
SE = 0.247; SS = 0.061
Rozložení pozorování


2Fasetový design bez interakce, r = 0.877
naměřené hodnoty
2
1
0
-1
-2
R2
R1
R2
R2
R1
R1
R2
R1
0
10
20
30
40
50
40 měření (4 osoby, každá měřena 5 dvěma hodnotiteli)
SE = 0.316; SS = 0.1
R2
R1
Rozložení pozorování


2Fasetový design s interakcí, r = 0.785
naměřené hodnoty
2
1
0
-1
-2
R2
R1
R2
R2
R1
R1
R2
R1
0
10
20
30
40
50
40 měření (4 osoby, každá měřena 5 dvěma hodnotiteli)
SE = 0.44; SS = 0.194
R10
R2
R8
R2
R1
R1
R6
Rozložení pozorování

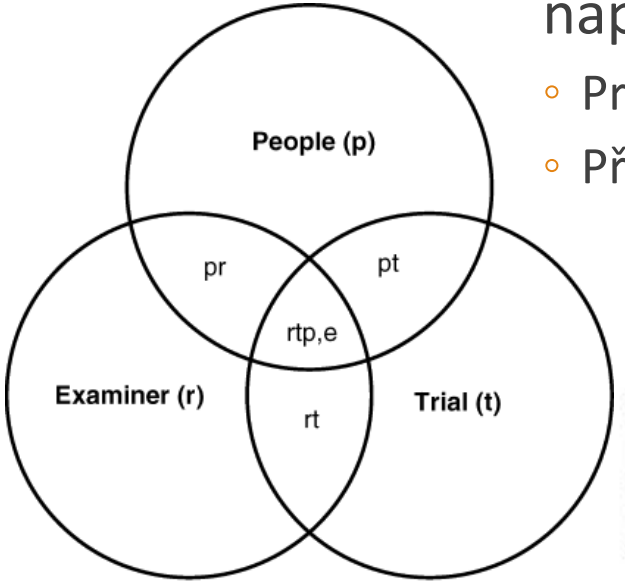
Jinými slovy...

CTT: Pouze dva zdroje variability

- **systematický** = pravý skór
- **náhodný** = chyba měření
- **zanedbaný** = obtížnost položek

GT: Neomezeně zdrojů variability

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



$t = \sigma^2$ trials
$r = \sigma^2$ examiner
$p = \sigma^2$ people
$e = \sigma^2$ error

Heitman et al. (2009)

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 zobecnění“** (všemi možnými respondenty, položkami, situacemi...). **Universe of generalization.**
  - Reliabilita průměrného a součtového skóre je ale 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. při měření 10 položkami při 3 administracích, hodnocených 2 hodnotiteli?
  - Využívá výsledků G-studie.

Princip a účel GT

Podobné předpoklady jako CTT, jde o její rozšíření.

- **Náhodný výběr** prvků dané fasety **z nekonečně velkého** doménového prostoru.
  - Existují ale i úpravy pro „finite universe“.
  - Náhodný výběr lze obejít „zafixováním“ určité fasety – „fixed effects“ namísto „random effects“.

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 a přibližně stejný reziduální rozptyl, což je zajištěno předpokladem náhodného výběru. Relativně vysoká robustnost zvláště při větším počtu položek.
  - Vícedimenzionalita možná při dodržení tau-ekvivalence faset na univerzu; analogie k hierarchické ρ_{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.

- „Weak true-score theory“. Příliš mnoho nesplnitelných 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).

Teorie zobecnitelnosti

Dva klíčové postupy/kroky/součásti GT:

1. Studie zobecnitelnosti (G-studie; generalizability study)
2. Rozhodovací studie (D-studie; decision study)

G-studie

Studie zobecnitelnosti
Generalizability study

Dekompozice rozptylu

Odhad rozptylových komponent

ANOVA

Smíšený lineární model
(linear mixed model, LMM)

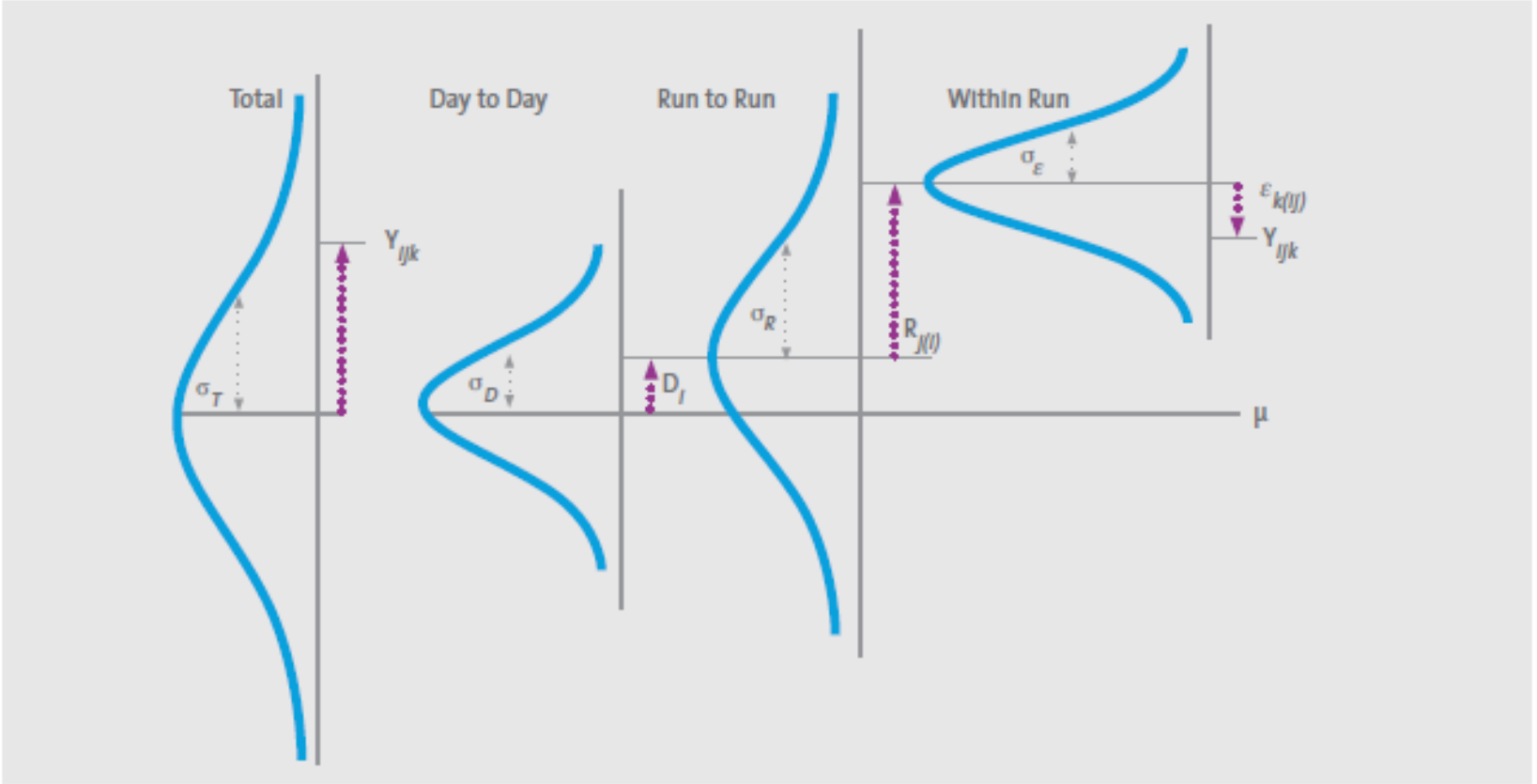


Figure 4. Variance Decomposition. An individual result Y_{ijk} (ie, the result for a single replicate) in a 20×2×2 study, modeled as the vector sum of deviations—$D_i + R_{j(i)} + \varepsilon_{k(ij)}$—from the sample's mean, μ, due to the combined effect of three composite sources of random variation—namely, day-to-day (D), run-to-run within-day (R), and within-run or residual (ε) variance components—with distributions characterized by SDs of σ_D, σ_R, and σ_ε, 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.

https://community.jmp.com/t5/Discussions/How-to-implement-variance-decomposition/td-p/272070

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 respondenta×položky×situace×hodnotitele×...) tvoří specifický rozptyl respondenta/položky/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:

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

Reliabilita:

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

- pravý skór a chyba jsou ortogonální, proto chybí jejich kovariance („$+2\sigma_{\tau e}^2$)

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:

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_1^2 + \sigma_2^2 + \sigma_{\tau 1}^2 + \sigma_{\tau 2}^2 + \sigma_{12}^2 + \sigma_{\tau 12,e}^2$$

Reliabilita:

$$r_{xx'} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \left(\sigma_1^2 + \sigma_2^2 + \sigma_{\tau 1}^2 + \sigma_{\tau 2}^2 + \sigma_{12}^2 + \sigma_{\tau 12,e}^2\right)}$$

- Všechny rozptylové komponenty jsou ortogonální (protože jsou zahrnuty všechny), proto též bez kovariance.
  - Z toho důvodu se zahrnují i nesignifikantní efekty.
  - Jde o obecný vzorec; některé komponenty nemusí být chybou.

G-studie: příklad

Příklad: 2fasetový design p × i × o.

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

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

	Occasion					
	I			*II*		
Person	*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
…						
p	4	5	4	3	4	2
…						
N	3	4	4	3	3	3

Pozorovaný skór *X* a obecný skór *T*:

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

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):

$$\sigma^2_{X_{pio}} = \sigma^2_p + \sigma^2_i + \sigma^2_o + \sigma^2_{pi} + \sigma^2_{po} + \sigma^2_{io} + \sigma^2_{pio,e}$$

Převzato z http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ^2_p	1.108	30
Item (i)	σ^2_i	0.102	03
Occasion (o)	σ^2_o	0.030	01
p × i	σ^2_{pi}	0.810	22
p × o	σ^2_{po}	0.230	06
i × o	σ^2_{io}	0.001	00
p × i × o, e	$\sigma^2_{pio,e}$	1.413	38

Převzato z http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf

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.
- LS estimátor (least-squares).

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.
- ML estimátor (maximum-likelihood).

GT: SW pro odhad G-studie

Tradiční SW:

- GENOVA, mGENOVA (staré DOSovské aplikace)
- Různé podivné malé prográmky (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
- A tedy vlastně jakýkoli SW s modulem pro lineární smíšené (mixed) modely (JAMOVI, JASP)...

R, zejména balíček lme4 (mixed modely) a případně gtheory (nástavba lme4 pro GT).

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

Přehled dostupných programů (poněkud outdated):

- 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
- 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

Teorie zobecnitelnosti

Skript a data pro ilustraci teorie zobecnitelnosti v R. Data pocházejí z přijímací zkoušky do NMGR psychologie FSS MU do podzimního semestru 2020 (hodnocení bakalářských prací).

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), ocassion (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, je tam implicitně).

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

print(gstudy)
```

Dlouhá data:

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

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

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

	Occasion					
	I			*II*		
Person	*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
…						
p	4	5	4	3	4	2
…						
N	3	4	4	3	3	3

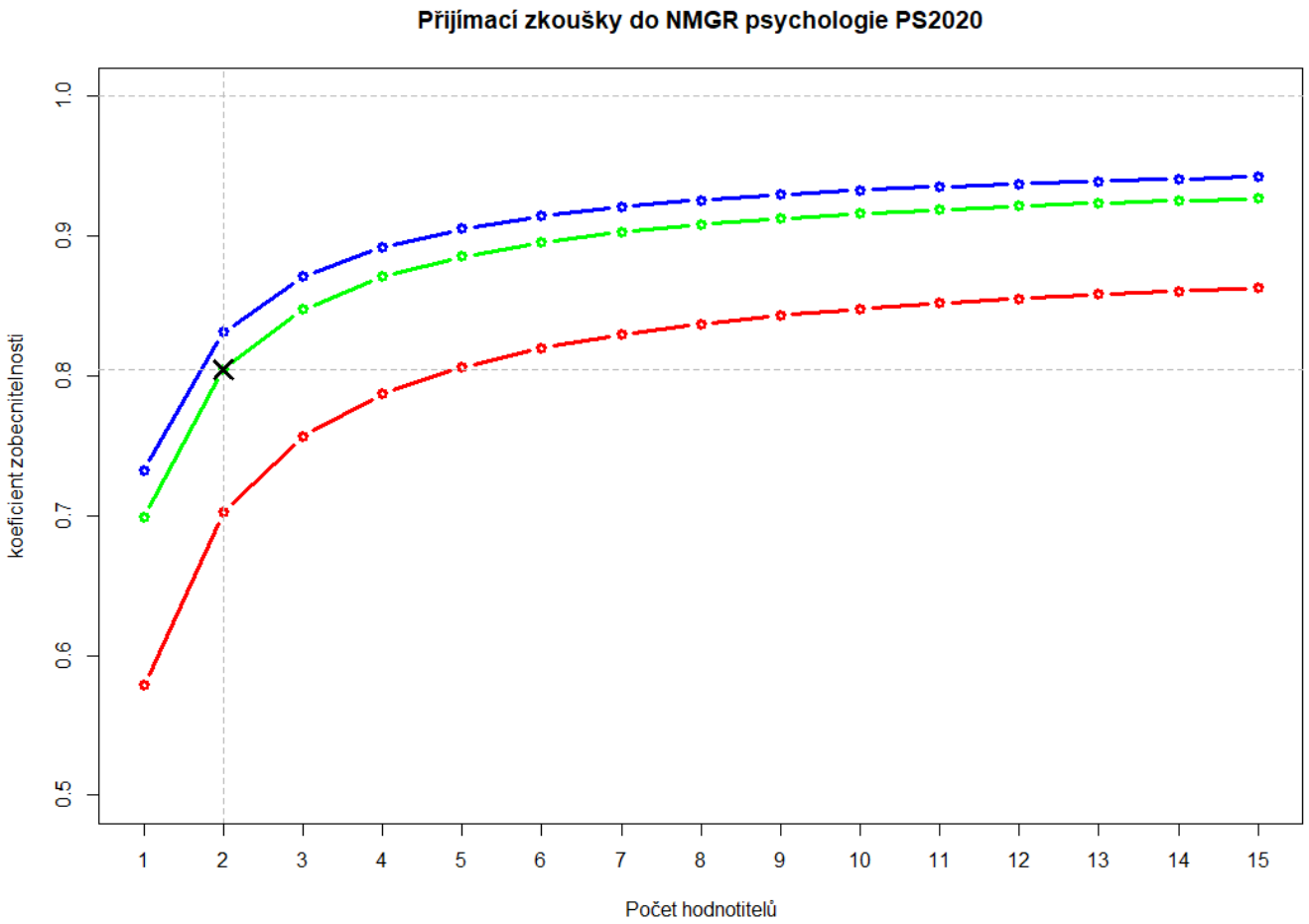
Převzato z http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf

D-studie

Rozhodovací studie
Decision study

Koeficient dependability
a zobecnitelnosti

Absolutní a relativní D-studie



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 score).

Odhad chyby odhadu universe skóru pro zvolený hypotetický design – např. p×I×O.

- Velké písmeno v designu D-studie symbolizuje, že se zajímáme o průměrné skóry, nikoli rozptyly.

D-studie: Obecný postup

1. Volba jednotky/subjektu měření (nemusí být respondent).
2. Volba designu, resp. prostoru/prostorů zobecnění.
3. Identifikace, které fasety (a interakce) G-studie jsou chybové složky.
4. Volba počtu prvků faset (nemusí se shodovat s G-studií).
5. Výpočet chyby měření.
6. Výpočet koeficientu reliability.

D-studie: Dva typy zobecnění

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

- Všechny fasety jsou zafixovány napříč měřením (např. test složený z pevného setu položek).
- Díky fixaci se jejich prvky stanou konstantou a rozdílná „obtížnost“ není chybou.
- Nezobecňuje se na celý fasetový prostor, ale právě na tyto prvky dané fasety.
- Reliabilita odhadována pomocí **koeficientu zobecnitelnosti**.
- Přímo srovnatelný s různými druhy CTT reliability.

Absolutní (kriteriální) – zobecnění na celou fasetu.

- Tento odhad nese více nejistoty (záleží na náhodně „vybrané“ obtížnosti prvků fasety).
- 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: Dva typy zobecnění (příklady)

RELATIVNÍ D-STUDIE

Dotazník self-esteemu (SE)

- Nezajímá mě, jak by respondent skóroval na případných jiných položkách, které měří SE.
- Posvátná kráva? ☺

Hodnocení písemného testu v psychometrice.

- Všechny testy hodnotí Hynek. Zanedbáváme, jak by bodovali jiní hodnotitelé.

Přijímací zkouška do NMGR psychologie.

- Chceme vybrat 30 nejlepších uchazečů, nezáleží na tom, jak obtížné položky jsou letos v testu.

ABSOLUTNÍ D-STUDIE

„Super-komplexní dotazník depresivity“.

- Náhodný výběr 10 symptomů ze všech identifikovaných symptomů deprese.
- Záleží, zda jsme vybrali časté či řídké symptomy.

Hodnocení seminární práce v psychometrice.

- Do hodnocení jsou zapojeni tři lidé; protože se liší přísností, záleží, kdo je komu „přidělen“.

Přijímací zkouška do NMGR psychologie.

- Přijatý musí mít nejméně 36/60 bodů.
- Byly zařazeny jednoduché či těžké položky?

D-studie: Odhad chyby měření

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

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

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

Reliabilita se potom spočítá dle obecného vzorce pro vysvětlený rozptyl:

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

- σ_τ^2 - rozptyl jednotek měření, tedy ideálních skórů
- σ_e^2 - chybový rozptyl, tedy součet všech chybových komponent

[1] standardní chyba průměru $SE = \frac{SD}{\sqrt{N}} \rightarrow SE^2 = \frac{SD^2}{N}$; SD – směrodatná odchylka; N – velikost vzorku/počet pozorování.

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ^2_p	1.108	30
Item (i)	σ^2_i	0.102	03
Occasion (o)	σ^2_o	0.030	01
p × i	σ^2_{pi}	0.810	22
p × o	σ^2_{po}	0.230	06
i × o	σ^2_{io}	0.001	00
p × i × o, e	$\sigma^2_{pio,e}$	1.413	38

Převzato z http://web.stanford.edu/dept/SUSE/SEAL/Reports_Papers/methods_papers/G%20Theory%20AERA.pdf

Relativní D-studie: Příklad

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ_p^2	1.108	30
Item (i)	σ_i^2	0.102	03
Occasion (o)	σ_o^2	0.030	01
p × i	σ_{pi}^2	0.810	22
p × o	σ_{po}^2	0.230	06
i × o	σ_{io}^2	0.001	00
p × i × o, e	$\sigma_{pio,e}^2$	1.413	38

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

Relativní chybový rozptyl σ_δ^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**:

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

Koeficient zobecnitelnosti je přímo srovnatelný s reliabilitou v CTT:

- Vnitřní konzistence 1 měření v 1 okamžik:

$\sigma_\delta^2 = \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} = \frac{.810}{1 \times 10} + \frac{1.413}{1 \times 10 \times 1} = .222 \rightarrow G = \frac{1{,}108}{1{,}108 + 0{,}222} = 0{,}833$ (paradoxně větší! **Proč?**)

Absolutní D-studie: Příklad

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ_p^2	1.108	30
Item (i)	σ_i^2	0.102	03
Occasion (o)	σ_o^2	0.030	01
p × i	σ_{pi}^2	0.810	22
p × o	σ_{po}^2	0.230	06
i × o	σ_{io}^2	0.001	00
p × i × o, e	$\sigma_{pio,e}^2$	1.413	38

Absolutní chyba průměrného skóre 10 položek a 2 měření?

- Zobecňuji napříč všemi přípustnými položkami i časem (admissible observation).

Absolutní chybový rozptyl σ_Δ^2:

$$\sigma_\Delta^2 = \frac{\sigma_i^2}{N_i} + \frac{\sigma_o^2}{N_o} + \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{io}^2}{N_i \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} =$$

$$= \frac{.102}{10} + \frac{.030}{2} + \frac{.810}{1 \times 10} + \frac{.230}{1 \times 2} + \frac{.001}{10 \times 2} + \frac{1.413}{1 \times 10 \times 2} = .292$$

Podíl chybového rozptylu: **koeficient spolehlivosti** Φ (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 λ: $\Phi_\lambda = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_\Delta^2}$

- Φ_λ je vyšší, čím dále je kritérium λ od průměru osob μ.

Absolutní D-studie: Příklad

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ_p^2	1.108	30
Item (i)	σ_i^2	0.102	03
Occasion (o)	σ_o^2	0.030	01
p × i	σ_{pi}^2	0.810	22
p × o	σ_{po}^2	0.230	06
i × o	σ_{io}^2	0.001	00
p × i × o, e	$\sigma_{pio,e}^2$	1.413	38

Absolutní chyba průměrného skóre 10 položek a 2 měření?

- Zobecňuji napříč všemi přípustnými položkami i časem (admissible observation).

Absolutní chybový rozptyl σ_Δ^2:

$$\sigma_\Delta^2 = \frac{\sigma_i^2}{N_i} + \frac{\sigma_o^2}{N_o} + \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{io}^2}{N_i \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} =$$

$$= \frac{.102}{10} + \frac{.030}{2} + \frac{.810}{1 \times 10} + \frac{.230}{1 \times 2} + \frac{.001}{10 \times 2} + \frac{1.413}{1 \times 10 \times 2} = .292$$

Podíl chybového rozptylu: **koeficient spolehlivosti** Φ (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 λ: $\Phi_\lambda = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_\Delta^2}$

- Φ_λ je vyšší, čím dále je kritérium λ od průměru osob μ.

Smíšená D-studie: Příklad

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

Source	*Variance Component*	*Estimate*	*Percent of Total Variability*
Person (p)	σ^2_p	1.108	30
Item (i)	σ^2_i	0.102	03
Occasion (o)	σ^2_o	0.030	01
p × i	σ^2_{pi}	0.810	22
p × o	σ^2_{po}	0.230	06
i × o	σ^2_{io}	0.001	00
p × i × o, e	$\sigma^2_{pio,e}$	1.413	38

Jaká bude test-retest reliabilita 1 měření?

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

Chybový rozptyl:

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

Koeficient zobecnitelnosti:

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

D-studie: absolutní

Uvažuje veškeré fasety jako náhodné, přičemž vliv těchto faset se může lišit napříč osobami.

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 zobecnění: **„universe of admissible observartions“**.

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

Srovnání designů

Relativní D-studie ze 2 měření p×(I=10)×(O=2): G = 0,806

- I, O relativní

Relativní D-studie z 1 měření p×(I=10): G = 0,833

- I relativní, O vynecháno
- Šlo by o shodný výsledek s Cronbachovým alfa z jednoho měření.

Absolutní D-studie ze 2 měření p×(I=10)×(O=2): Φ = 0,791

- I, O absolutní

Smíšená D-studie, test-retest z 1 měření p×(I=10)×(O=1): Φ = 0,710

- I relativní, O absolutní

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 SB věšteckého vzorce, ale pro více zdrojů chyb než „počet testů“.

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

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

Využití GT: Optimální počet prvků faset

Seminární práce. Variuji:

- 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

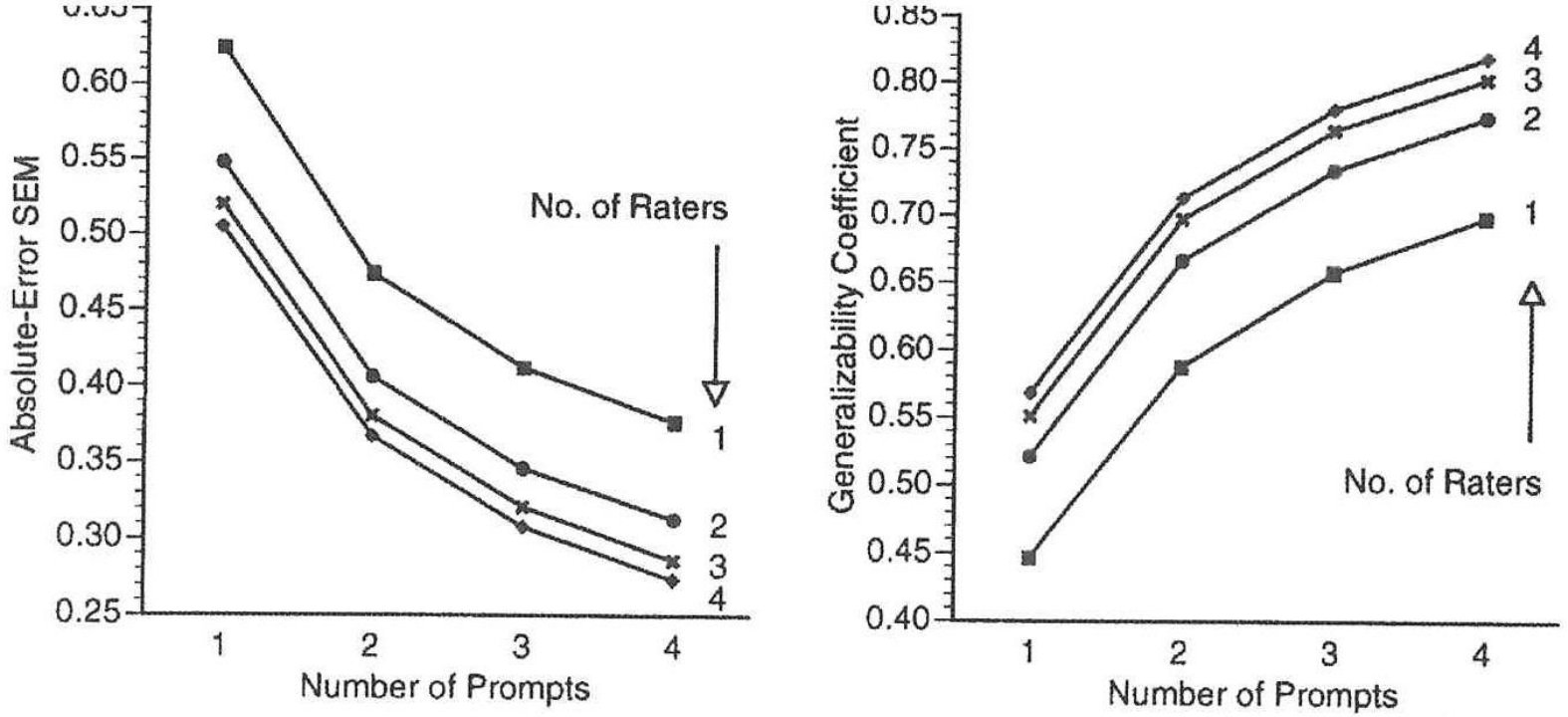


FIGURE 1.2. $\hat{\sigma}(\Delta)$ and $E\hat{\rho}^2$ for scenario with $p \times T \times R$ design.

Převzato z Brennan (2001) – jde o jiná data než výše.

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)×i
- D-studie pro žáka *uvnitř* třídy: (s:C)×I (C je relativní)
- D-studie pro žáka *napříč* třídami: (s:C)×I (C je absolutní)
- D-studie pro účely srovnání tříd: (S:c)×I (S je absolutní)

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 co nejkomplexnější design G-studie, případně alespoň stejný, jako je D-studie.
- Ale nedává smysl nevyužít v G-studii informace, které jsou k dispozici (proto co nejkomplexnější).

Využití GT: pevné kovariáty

Příklad: Mám velmi malý vzorek dat výkonového testu u malých dětí.

- Výkon výrazně roste v čase.
- Mohu spočítat reliabilitu pro celý vzorek dohromady – nadhodnocení systematickým vlivem věku.
- Na odhad pro jednotlivé kohorty zvlášť (žádoucí!) nemám ale dost dat.

Řešení: vložení věku jako pevného kovariátu (fixed effect).

- `lme4` syntax: `model <- "response ~ (1 | person) + (1 | item) + age + I(age^2)"`

Výsledkem je odhad rozptylu osob „po kontrole věku“ (a jeho kvadrátu).

Odhad reliability za předpokladu, že je reliabilita shodná pro všechny věkové kohorty.

- Že je shodný rozptyl výkonu osob i chybový rozptyl v každé kohortě stejný =
  = věkové kohorty jsou „paralelní“ (avšak nikoli striktně paralelní) skupiny osob

Užitečné při vývoji testů a pilotních studiích s malým vzorkem.

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 skórů, 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 skórů, 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.

Srovnání GT a model-based/dimension free konceptu reliability

Minulá přednáška o CTT: model-based vs. dimension free-reliabilita.

- Realismus: Co je měřeným rysem? Jak moc „paralelně“ jej dílčí indikátory měří?
- I u dimension-free reliability stále předpokládáme existenci latentního rysu.
- Relativní srovnání (ale absolutní lze implementovat).
- Zpravidla jen jeden zdroj chyby = položka (ale existují hierarchické a MTMM modely).

GT: Operacionalismus.

- Náhodný výběr prvků z domény zajišťuje asymptotickou tau-ekvivalenci vybraných prvků.
- Zobecňujeme na celý prostor nebo jen na vybrané prvky?
- Analogie k hierarchické i celkové reliabilitě.
  - Rozptyl určité fasety lze považovat za chyby nebo součást měření.

Obojí je zcela odlišný pohled na měření.

- Oba přístupy ale kombinují multifastové IRT modely.

Vnitrotřídní korelace: standardizovaná GT pro jednofasetový p×i design

Shrout a Fleiss (nejběžnější)	McGraw a Wong (občasně používané)	GT design
ICC(1,1)	One-way random, single score ICC(1)	p (jediná faseta plus error, N_e=1) *Hodnotitelé se neopakují.*
ICC(2,1)	Two-way random, single score ICC(A,1)	p×I (absolutní, N_i = 1) *Stejní hodnotitelé, vybraní náhodně.*
ICC(3,1)	Two-way mixed, single score ICC(C,1)	p×I (relativní, N_i = 1) *Stejní hodnotitelé, nezobecňuji na všechny možné.*
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 α

Ukazatel shody posuzovatelů. Reliabilita při hodnocení 1 posuzovatelem.

Reliabilita celkového hodnocení, tj. průměru všech posuzovatelů.

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