

Teorie zobecnitelnosti (z rychlíku)

„Ultimátní teorie měření pro všechny případy“

Hynek Cígler | IVDMR FSS MU | 9. 4. 2019

CTT: Hodně chyb, hodně reliabilit...

- ▶ Mnoho způsobů odhadů reliability a druhů chyby:
 - ▶ stabilita v čase (dependabilita, stabilita) – test-retest
 - ▶ vnitřní konzistence
 - ▶ ekvivalence – paralelní formy
 - ▶ shoda posuzovatelů
- ▶ CTT: „Reliabilita pro jaký účel“?
- ▶ CTT: „Obecná reliabilita“ neexistuje.
- ▶ Řešením problému „mnoho chyb, mnoho reliabilit“ je právě GT (Cronbach et al., 1963; 1972).
 - ▶ Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137-163.



Teorie zobecnitelnosti

Generalizability theory (GT)



Teorie zobecnitelnosti

Generalizability theory (GT)

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 Brennan (2001)



Východiska GT vs. CTT

- ▶ GT i CTT: Operacionalismus (antirealismus).
 - ▶ Srovnej s teoriemi latentních rysů (FA, IRT).
- ▶ CTT: Pravý skór = očekávané skóre v daném setu položek I .
$$\tau = E(X|I)$$
„Měřený“ atribut je definován tímto setem položek.
- ▶ GT: Pravý skór = očekávané skóre v daném prostoru významů.
$$\tau = E(X|I, conditions)$$
„Měřený“ atribut je definován způsobem výběru položek a prostorem zobecnění.
 - ▶ Explicitně se pracuje s úvahou „reliabilita vůči čemu“.
 - ▶ Úzce propojeno s teorií faset (např. Guttman, 1959; Shye, 1978).



Model „měření“ GT

▶ CTT: $X = T + e$

▶ GT: $X = T + e_1 + e_2 + e_3 + \dots + e_n$

▶ kde např. e_1 je specifický skór v daném čase, e_2 daného posuzovatele, e_3 položky (vnitřní konzistence) atd.

▶ Tyto chyby ale nejsou nezávislé; např. různí hodnotitelé mohou hodnotit různě v různých situacích.

▶ Proto interakce:

$$X = T + e_1 + e_2 + e_3 + e_1e_2 + e_2e_3 + e_1e_2e_3$$

▶ Jednotlivé zdroje rozptylu se označují jako fasety.



Oblasti využití GT

Kdy a proč ano?

- ▶ Reliabilita na vyšších úrovních multilevel dat.
- ▶ Vývoj testů.
 - ▶ Volba optimálního designu, počtu položek...
- ▶ Souběžný odhad různých zdrojů chyby.
 - ▶ Více informací o měření vzhledem k CTT.
- ▶ Odhad „neshody“, pokud nechci vážit.
 - ▶ Hodnotitelé, paralelní formy.

Proč ne?

- ▶ Blbá (operacionalistická) teorie měření.
 - ▶ Boring: „Měřím to, co měřím.“
 - ▶ Neumožňuje zkoumat konstrukty (protože neexistují).
- ▶ Vysoká náročnost na statistické dovednosti.
- ▶ Příliš silné předpoklady.
 - ▶ Zejm. zastupitelnost položek.
- ▶ Existují lepší teoriemi měření pro obdobné účely.
 - ▶ Multilevel/mixture FA, IRT.



Princip a účel GT

- ▶ **Odhad odhadu reliabilitu universe score**
 - ▶ Analogie pravého skóre v CTT.
 - ▶ Průměrná odpověď napříč prostorem zobecnění.
 - ▶ Očekávaná odpověď daného respondenta pro náhodnou kombinaci prvků z odpovědních prostorů (faset).
- ▶ **Dvě klíčové části GT:**
 - ▶ **G-studie:** Parcializace rozptylových složek.
 - ▶ **D-studie:** Odhad reziduálního rozptylu pro daný hypotetický design měření v závislosti na prostoru zobecnění.
 - ▶ Standardní chyba měření odhadu universe scoru.
 - ▶ Koeficient reliability pro takový odhad nad populací.
 - ▶ Využívá výsledků G-studie.
- ▶ **Některé postupy analogické GT jsou běžně používány jinde.**
 - ▶ Reliabilita podle Hoyta je zjednodušeným předchůdcem GT.
 - ▶ Intraclass korelace je „standardizovaným“ použitím GT.



Předpoklady GT

- ▶ Podobné předpoklady jako CTT, jde o její rozšíření.
- ▶ „Náhodný“ výběr prvků z nekonečně velkých faset.
 - ▶ Existují ale i úpravy pro „finite universe“.
- ▶ Multivariační normální rozdělení, intervalová škála (ale...).
- ▶ Jednodimenzionalita (ale MANOVA).
 - ▶ Konfirmační multidimenzionální model lze definovat i v lme4, ale většinou příliš porušené předpoklady.
- ▶ Tau-ekvivalence položek (relativně vysoká robustnost, zvláště při větším počtu položek).
 - ▶ Z hlediska lineárního modelu homoskedasticita reziduí.



G-studie

- ▶ **G-studie = generalizability study**
 - ▶ Odhad velikosti pozorovaných rozptylových komponent.
 - ▶ „Jakou část rozptylu jednoho pozorování (interakce respondenta×položky×situace×hodnotitele×...) tvoří specifický rozptyl respondenta/položky/situace/.../všech možných interakcí?“
- ▶ **Zobecňuje z měření na prostor (universum).**
 - ▶ Odhad rozptylových komponenty v prostoru.
 - ▶ Tohle je ta výpočetně náročnější část GT.



G-studie: příklad

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

▶ N respondentů p , 3 položky i a 2 administrace o

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

▶ Celkový rozptyl v datech (rozptyl všech prvků matice níže):

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

▶ Protože chybové rozptyly nekorelují, nejsou v rovnici kovariance faset.

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

	Occasion					
	I			II		
	Item 1	Item 2	Item 3	Item 1	Item 2	Item 3
Person						
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

G-studie: příklad

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)	σ_p^2	1.108	30
Item (i)	σ_i^2	0.102	03
Occasion (o)	σ_o^2	0.030	01
$p \times i$	σ_{pi}^2	0.810	22
$p \times o$	σ_{po}^2	0.230	06
$i \times o$	σ_{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 vznikla okolo ANOVA.**
 - ▶ Konkrétně repeated measure ANOVA.
 - ▶ V tradiční terminologii faseta = faktor.
 - ▶ Least-squares estimator.
- ▶ **Aktuálně spíše LMM (linear mixed model).**
 - ▶ ML estimátor (a jeho varianty).
 - ▶ Výhody při odhadu – např. unbalanced design (různé počty prvků faset), nested design (ne všechny kombinace faset jsou pozorovány), chybějící data.
 - ▶ Menší předpoklady, vyšší flexibilita.
 - ▶ Výsledek LS a ML by se neměl lišit (při dodržení předpokladů).



D-studie

- ▶ Odhad chyby odhadu universe skóru pro zvolený hypotetický design – např. $p \times I \times O$.
- ▶ Klíčová je volba prostoru zobecnění, v jehož rámci má každý respondent hypotetický U-skór.
 - ▶ Ten se může lišit napříč prostory. Antirealismus!
- ▶ Obecný postup:
 - ▶ 1. Volba jednotky 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 koef. spolehlivosti (dependability).
 - ▶ Lze uvažovat pravděpodobnost překročení absolutního kritéria.
- ▶ **Spíše než otázka celého designu otázka dílčích faset.**
 - ▶ Smíšený design, tedy kombinace relativních a absolutních faset, vše velmi výrazně komplikuje!!!



D-studie: Odhad chyby měření

- ▶ Chyba odhadu obecně: standardní chyba průměru „obtížnosti“ prvků fasety (na druhou).
 - ▶ Chybový rozptyl se tedy skládá ze součtu chybových rozptylových komponent podělených počtem jejich pozorování.
- ▶ Reliabilita se potom spočítá dle obecného vzorce

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

- ▶ σ_u^2 - rozptyl jednotek měření, tedy universe skóreů
- ▶ σ_e^2 - chybový rozptyl



D-studie: relativní příklad

- ▶ Jaká bude chyba s využitím 10 položkového testu při dvou měřeních?
 - ▶ Test je stále stejný, položky i příležitosti jsou fixed faktor.

- ▶ **Relativní chybový rozptyl** σ_{δ}^2 :

$$\sigma_{\delta}^2 = \frac{\sigma_{pi}^2}{N_p \times N_i} + \frac{\sigma_{po}^2}{N_p \times N_o} + \frac{\sigma_{pio,e}^2}{N_p \times N_i \times N_o} = \frac{.810}{10} + \frac{.230}{2} + \frac{1.413}{20} = .267$$

- ▶ Velikost chybového rozptylu - koeficient zobecnitelnosti:

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

- ▶ Koeficient zobecnitelnosti je přímo srovnatelný s reliabilitou v CTT (v případě výše vnitřní konzistence průměru dvou měření).

- ▶ Pro vnitřní konzistenci jediného měření (Cronbachovo alfa):

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



D-studie: absolutní příklad

- ▶ Jaká bude chyba při 10 položkách a 2 měřeních, pokud je test kritériální (a zobecňujeme na všechny možno položky)?

Absolutní chybový rozptyl σ_{Δ}^2 :

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

- ▶ Koeficient spolehlivosti (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 od průměru μ .

- ▶ Zobecňujeme na libovolné měření s libovolnými položkami/situacemi...



D-studie: absolutní

- ▶ Zobecňujeme na všechny možné prvky dané fasety.
 - ▶ Náchylnější na porušení předpokladu náhodného výběru z domény – záměrný výběr obtížných vs. snadných položek.
- ▶ Kriteriaální test:
 - ▶ Relativní: 70 % správně z daných 10 položek. (Což nedává smysl.)
 - ▶ Absolutní: 70 % správně ze všech možných položek.
- ▶ Používá se i kombinace relativní a absolutní D-studie.
 - ▶ Test-retest: absolutní položky, relativní situace.



Využití G-teorie: závěrečné poznámky

- ▶ 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 – tedy nikoliv dotazníky, znalosti a pozornosti.
 - ▶ Např. tzv. škrtačí testy pro měření reakčního času, kde jsou dílčí položky řazené do bloků (a třeba testované opakovaně).



D-studie: relativní

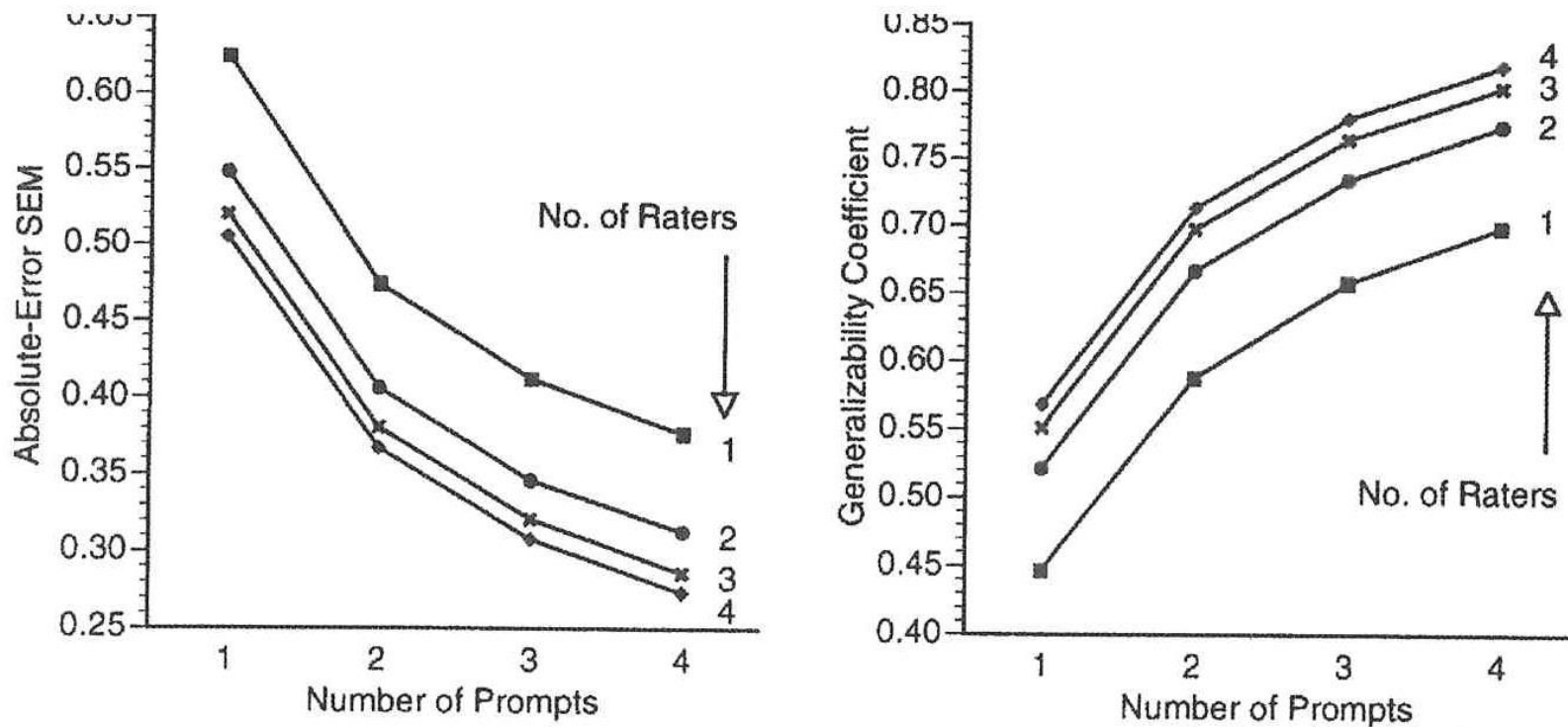


FIGURE 1.2. $\hat{\sigma}(\Delta)$ and $E\hat{\rho}^2$ for scenario with $p \times T \times R$ design.

Závěrečné poznámky

- ▶ Prvkem měření nemusí být respondent, ale např. školní třída (pak je faseta „žáci“ chybovým rozptylem).
- ▶ 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: $(s:C) \times I$
 - ▶ D-studie pro žáka *napříč* třídami: $(s:c) \times I$
- ▶ 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...).



Vnitrotřídní korelace pro $P \times I$ design

Shrout a Fleiss (nejběžnější)	McGraw a Wong (občasné)	GT design
ICC(1,1)	One-way random, single score ICC(I)	P (jediná faseta plus error, $N_e=1$)
ICC(2,1)	Two-way random, single score ICC(A,1)	$P \times I$ (absolutní, $N_i = 1$)
ICC(3,1)	Two-way mixed, single score ICC(C,1)	$P \times I$ (relativní, $N_i = 1$)
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)	$P \times I$ (absolutní, $N_i = k$)
ICC(3,k)	Two-way mixed, average score ICC(C,k)	$P \times I$ (relativní, $N_i = k$)

► A=agreement, C=consistency

Díky za pozornost!

▶ Hynek Cígler

- ▶ Katedra psychologie; Institut pro výzkum dětí, mládeže a rodiny
Fakulta sociálních studií, Masarykova Univerzita
- ▶ Joštova 10, 602 00 Brno
- ▶ e-mail: hynek.cigler@mail.muni.cz
- ▶ web: psych.fss.muni.cz, ivdmr.fss.muni.cz

