MUNI FSS Katedra psychologie

Přednáška 9: Interpretace testových skórů

13. 11. 2023 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler

Jak interpretovat testové skóry?

Kromě samotných skórů je nezbytné vzít v úvahu další informace.

- Tato přednáška bude o vybraných aspektech důležitých pro jejich interpretaci.

Kvalita diagnostické metody.

Kvalita norem.

Pokročilá práce s chybou měření.

Hodnocení kvality diagnostické metody

Validita, reliabilita, normy...

- Rozporuplné odhady reliability, rozporuplné důkazy validity...

Statistické zpracování, CTT, IRT...

Různá epistemologická východiska, povaha konstruktu...

Různá využití metody...

Diagnostické otázky (ne/spolupracující klient, ohrožení validity individuálního vyšetření)...

K dispozici je (ideálně) velké množství informací.
Jak tedy zhodnotit **kvalitu metody?**

Opakování: Různé přístupy k validitě

Tradiční pojetí validity: obsahová, empirická, konstruktová (Cronbach & Meehl, 1955).

- Neposkytuje nejasný rámec pro celkové zhodnocení použitelnosti testu.
- Přílišné zakotvení „konstruktové validity“ v logickém pozitivismu.
- Kargokultické ztotožnění konstruktové a faktorové validity v praxi.

Realistické pojetí validity (např. Borsboom, 2004, 2009).

- Nepoužitelné v praxi (ontologický výrok).
- Neposkytuje vodítka pro hodnocení reálných diagnostických nástrojů.

Messickovova unifikovaná konstruktová validita (Messick, 1989, 1995).

- Konstruktivismus, instrumentalismus.
- Řada dílčích potíží (které se však týkají i tradičního pojetí), viz např. Borsboom (2004, 2009).

Unifikovaná konstruktová validita

Důraz na hodnocení a použití testu v diagnostice.

- Validita je jediným, multifasetovým konstruktem.
- Validita je *integrativním shrnutím dílčích důkazů*.
- Integrována ve Standardech pro pedagogické a psychologické testování (AERA, 2014).

Zdroje důkazů:

- Obsah testu
- Vnitřní struktura testu
- Odpověďové procesy
- Souvislost s kritériem
- Konsekvence testování

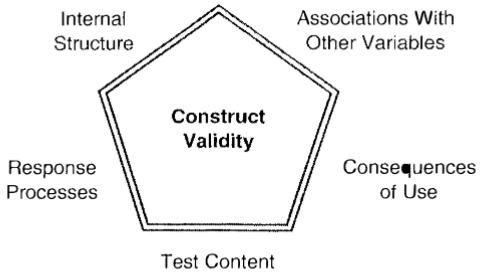


Figure 8.1 A Contemporary Perspective of Types of Information Relevant to Test Validity

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics : An Introduction*. Sage.

Otázky spojené s Messickovým pojetím validity (Lissitz & Samuelsen, 2007)

Znamenají nízké korelace s kritériem skutečně nízkou validitu měření?

- A navíc „*posvátná kráva*“, malý důraz na divergentní validitu.

Může být nějaká psychologická nomologická síť dostatečně komplexní a „precizní“?

Jak v diagnostické praxi zhodnotit „globální užitečnost“ nástroje?

Jsou skutečně všechny aspekty důležité při hodnocení metody součástí validity?

Mají všechny atributy metody při hodnocení stejnou váhu?

Jak souvisí teorie a specifikace konstruktu při konkrétním měření?

Pokud je reliabilita podmínkou či součástí validity, proč ji Messick explicitně nezmiňuje?

Hodnocení metody (Lissitz & Samuelsen, 2007)

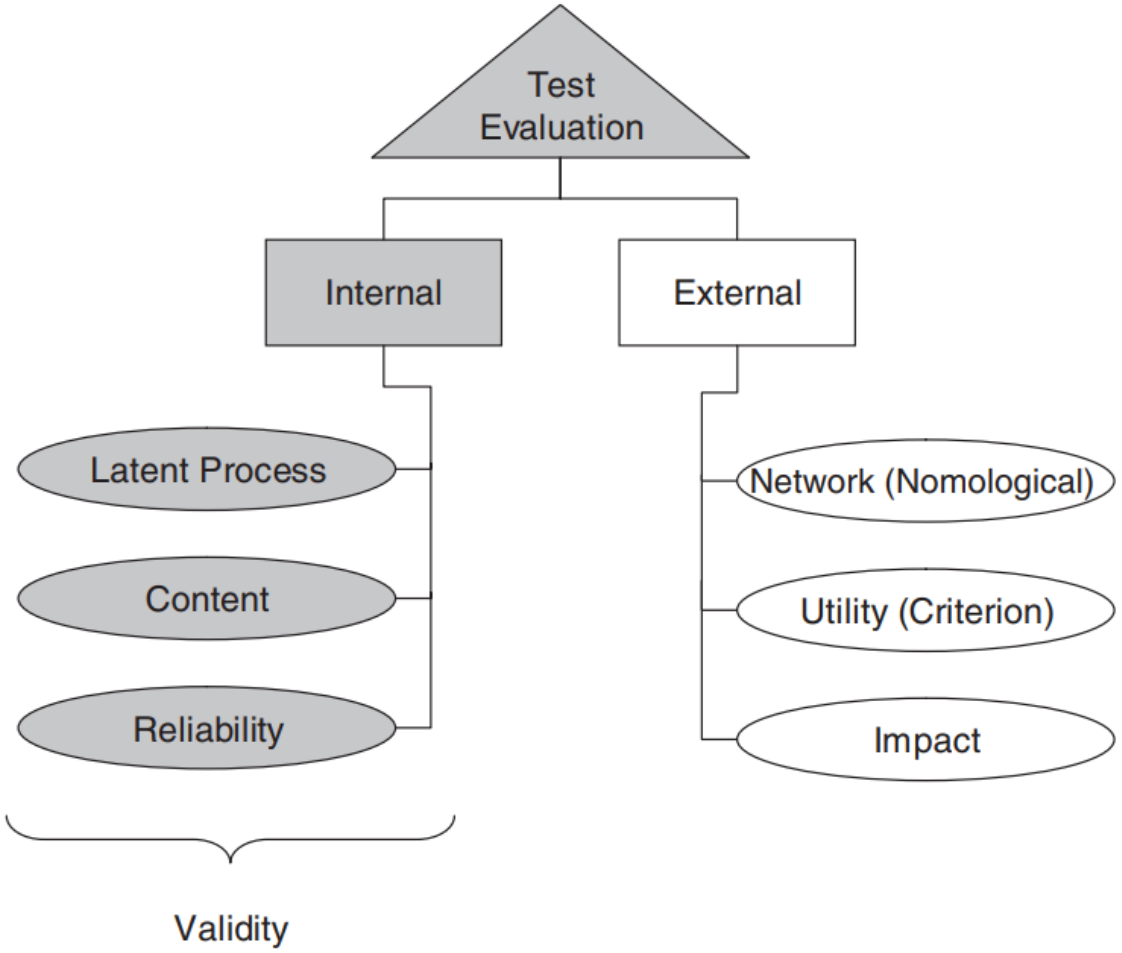


FIGURE 1. *The structure of the technical evaluation of educational testing.*

		Perspective: Theoretical	Perspective: Practical
Investigative Focus	Internal	Latent Process	Content and Reliability
	External	Nomological Network	Utility and Impact

FIGURE 2. *Taxonomy of test evaluation procedures.*

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36*(8), 437–448. doi:10.3102/0013189x07311286

Lissitz a Samuelsen (2007)

Dvě složky hodnocení:

- 1. **Realismus:** interní (validita, vlastnost testu)
- 2. **Instrumentalismus:** externí (využitelnost skórů).

Přínosy:

- Reliabilita je nedílnou součástí hodnocení (obsahové validity).
- Realistická pozice s pragmatickou složkou hodnocení.
- Reflektivní i formativní konstrukty.
- Díky realistickému pojetí umožňuje hodnotit kvalitu skórování (IRT vs. CTT – co lépe reflektuje latentní proměnnou)?

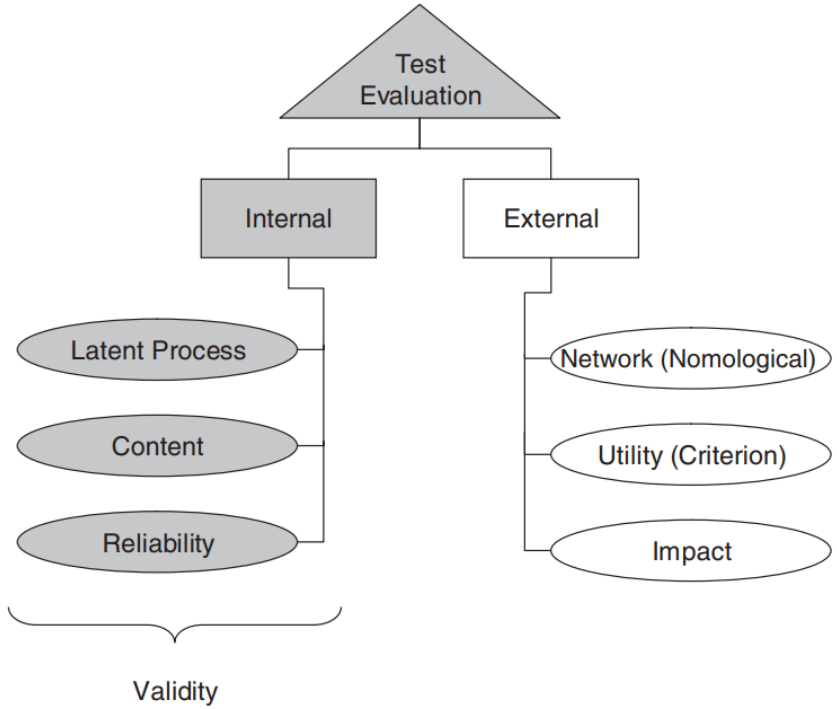


FIGURE 1. *The structure of the technical evaluation of educational testing.*

Lissitz a Samuelsen (2007)

„USAcentrismus“, model hodnocení není kompletní.

- Autoři jsou z edukativního prostředí; zaměřili si na pedagogické testy.

V psychologické praxi budou některé aspekty chybět.

Hodnocení norem.

Fokus na high-stakes výkonové testy.

Hodnocení adaptace do jiného prostředí.

Hodnocení počítačových zpráv a výstupů pro klienta.

To vše ale vhodně doplňuje recenzní model EFPA.

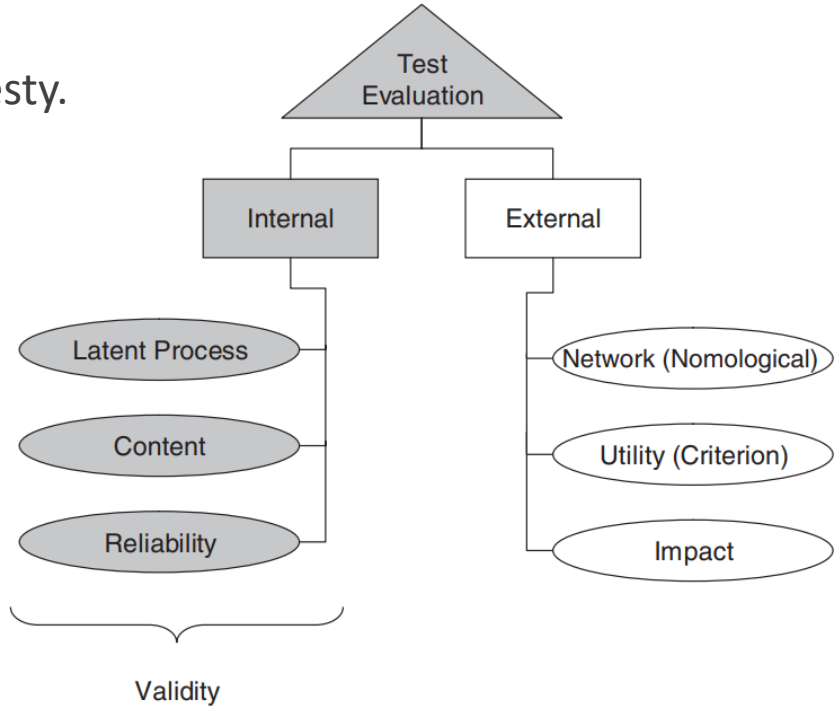
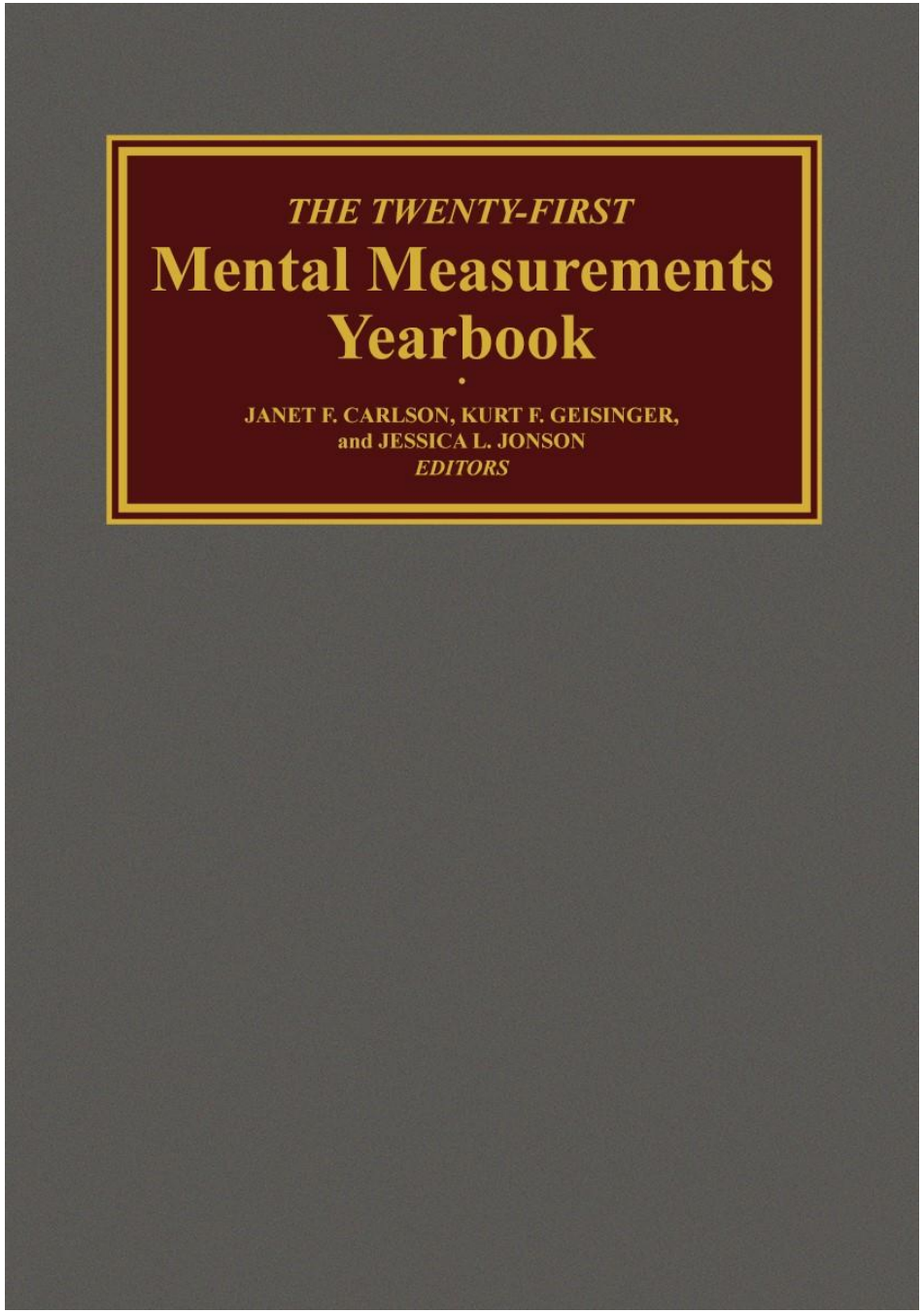


FIGURE 1. *The structure of the technical evaluation of educational testing.*

Recenze diagnostických metod



Recenze diagnostických metod

Výzkumná metoda je založena na peer-reviews (články, granty...).

Diagnostická metoda je aplikovaným výzkumným výstupem.

- Její kvality je vhodné rovněž kontrolovat = recenzovat.

Dvě tradice recenzování:

- Americká: Burosovy ročenky (Buros Mental Measurement Yearbooks).
- Evropská: BPS, EFPA.

Burosovy ročenky

Vychází každých několik let.

Obsahuje veškeré komerčně distribuované metody v anglickém (a španělském) jazyce.

- Dobrovolníci. Všechny metody.
- Recenze se opakují v případě zásadních revizí či nových empirických důkazů.

Každá metoda: 2 recenzenti. Recenze jsou narativní s tradiční strukturou.

- Záhlaví (vybrané důležité informace, autoři, distributor aj.).
- Popis testu, jeho určení, cílové populace atp. (nehodnotící, vychází z informací autora).
- Vývoj metody a kvalita technických materiálů (popis s hodnotícím komentářem).
- Technické parametry (kritické zhodnocení na úrovni faktů) = validita, reliabilita, normy...
- Komentář (zhodnocení faktů uvedených výše).
- Shrnutí a závěr s konkrétním doporučením. Literatura, zdroje

V knihovně FSS MU: D2-732; D2-732a.

EFPA manuál

Vychází částečně z recenzního modelu Britské psychologické společnosti.

- Distributor si za recenzi platí, uživatel si kupuje možnost nahlédnout.

V ČR implementováno v časopisu Testfórum: www.testforum.cz.

- EFPA manuál doplněný o narativní text po vzoru Burosových ročenek.

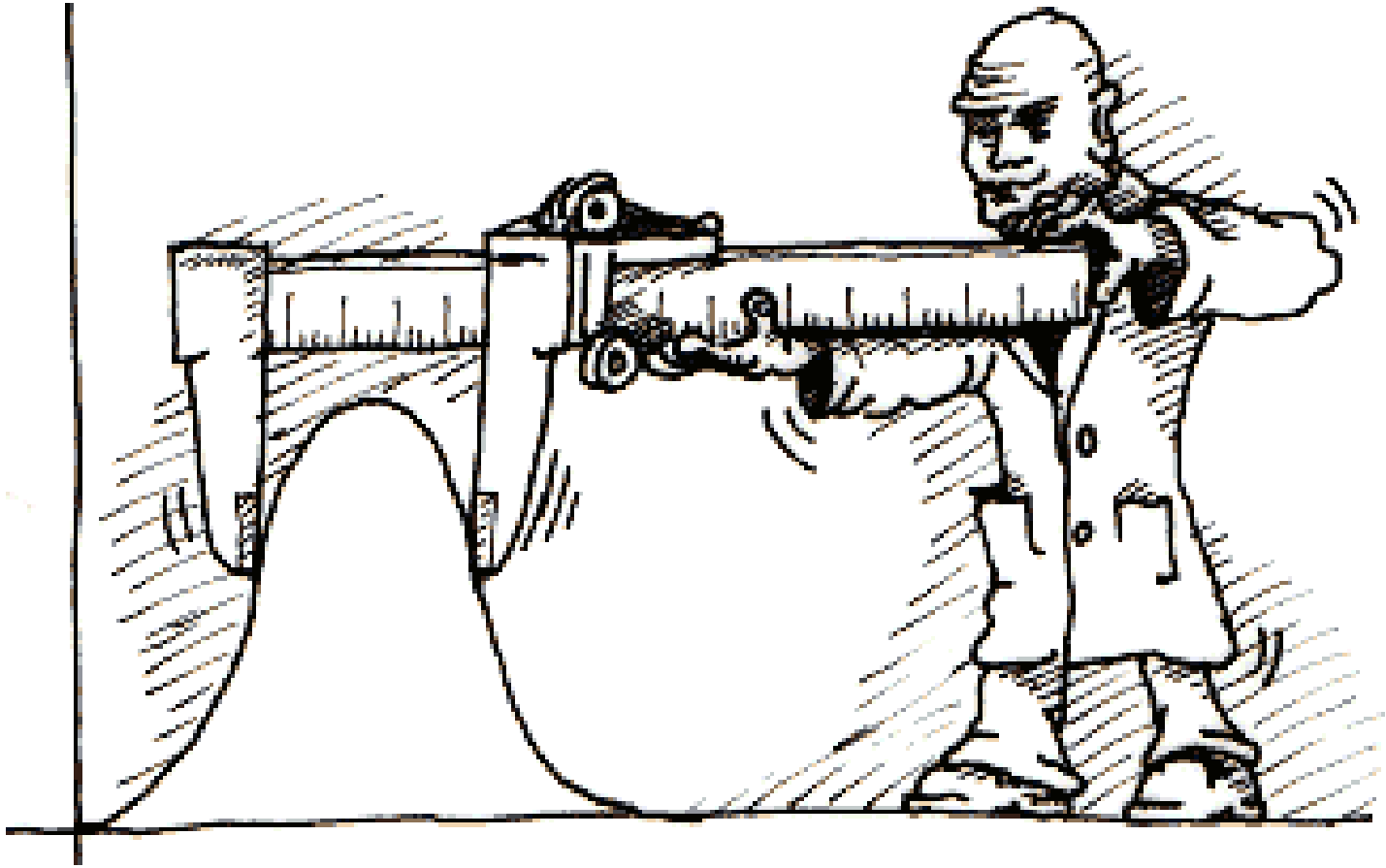
Pevná struktura formuláře.

- Popis: obecný popis, klasifikace, skórování, počítačově-generované zprávy, dodavatel a náklady.
- Zhodnocení: osvětlení teorie, kvalita materiálů, psychometrické parametry (validita, reliabilita a normy), počítačově-generované zprávy. Závěr: závěrečné zhodnocení a doporučení.

Zdroje:

- Evers, A. a kol. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, *25*(3), 283–291. https://doi.org/10.7334/psicothema2013.97
- Urbánek, T. (2010). Stav české psychologické diagnostiky a evropský model recenze testu. *Testfórum*, *1*(1). https://doi.org/10.5817/TF2010-1-1
- EFPA manuál.

Hodnocení norem



Hodnocení norem

Reprezentativnost vzorku vůči populaci.

- Výběrová populace.
- Jak dobře normy reprezentují zamýšlenou populaci?

Relevance populace pro klienta.

- Reprezentativnost populace vůči respondentovi (věk, lokální populace).
- Jak moc relevantní je výběrová populace pro respondenta?

Relevance populace pro účel vyšetření.

- Jak moc relevantní je výběrová populace pro účel vyšetření?
- Zkreslení, impression management...

Výběrová chyba.

- Jak moc velká je normovací chyba?

Hodnocení norem

Velikost vzorku nemusí souviset s reprezentativitou.

Kvótní vs. zcela nahodilý výběr.

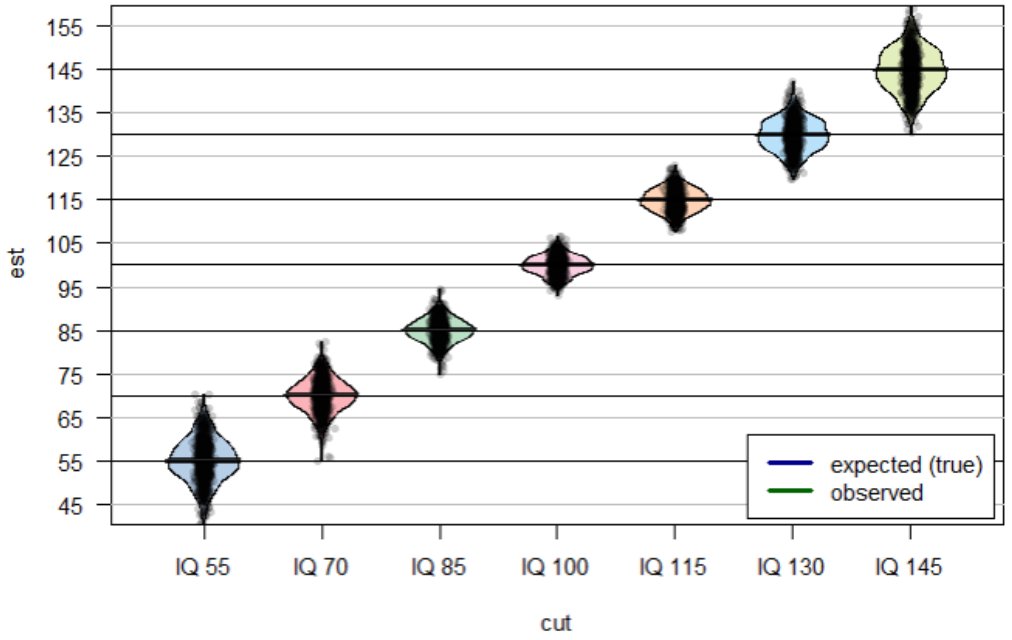
Věkové normy: způsob konstrukce.

Chyba průměru: jen $\frac{SD}{\sqrt{N}}$.

- Ale výrazně vyšší chyba v koncích rozložení!

Výběrová chyba se „sčítá“ se standardní chybou měření!

N = 50



Hodnocení norem

Velikost vzorku nemusí souviset s reprezentativitou.

Kvótní vs. zcela nahodilý výběr.

Věkové normy: způsob konstrukce.

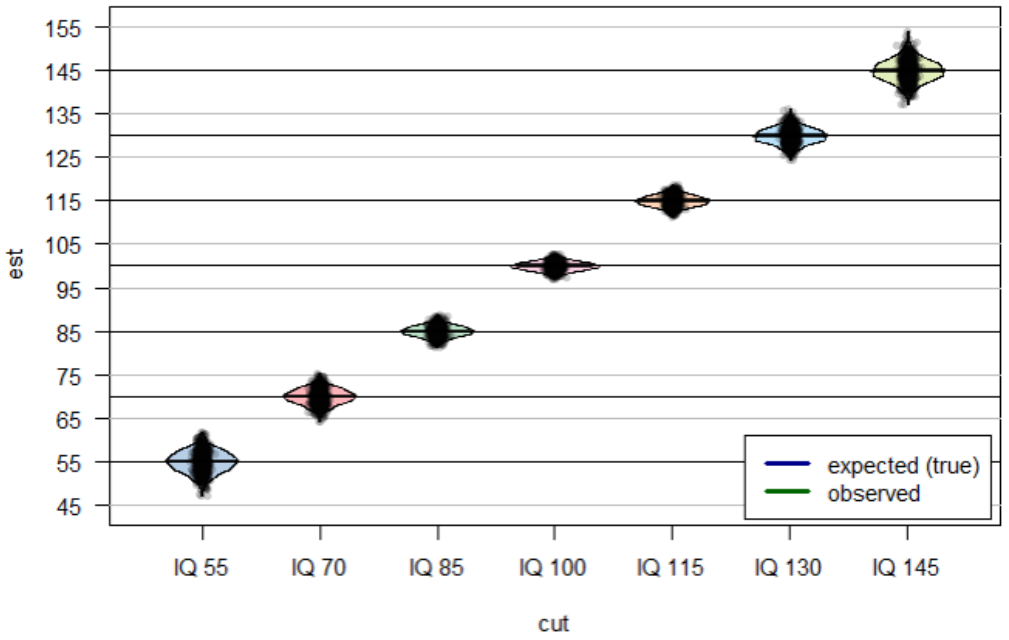
- Běžné normy: 2*počet kohort = parametry.
- Kontinuální normy: typicky do 5 parametrů.

Chyba průměru: jen $\frac{SD}{\sqrt{N}}$.

- Ale výrazně vyšší chyba v koncích rozložení!

Výběrová chyba se „sčítá“ se standardní chybou měření!

N = 200



Kontinuální normování

Tradiční „kohortové“ normy: v každé kohortě (např. věkové) konstrukce separátních norem.

Nevýhody tradičních norem:

- Malý vzorek uvnitř kohorty → velká výběrová chyba.
- Velká výběrová chyba → nestabilita norem napříč kohortami.
- Např. ročníkové normy po půl roce 5–18 let (14 skupin): $14 \times 2 = 28$ parametrů (M, SD).
  - Při celkové velikosti vzorku N = 1000 jen 71 respondentů/kohorta, tj. 36 respondentů/parametr.

Nenormální rozložení skórů u malých vzorků nelze dost dobře řešit.

- Tradiční diskrepanční skóry (IQ apod.) předpokládají normální rozložení. Nutnost normalizace.
- V případě malého kohortového vzorku nefunguje McCallova plošná transformace.
- Např. se 71 respondenty těžko bude fungovat plošná transformace do 100 percentilových skórů.
- Nutnost vertikálního vyhlazení skórů (například kernell smoothed kumulativní distribuce; `ks::kcde`).

Kontinuální normování

Kontinuální normování využívá celý vzorek pro odhad parametrů populační distribuce v určité kohortě.

Kontinuální normy = horizontální vyhlazení testových skórů.

- Může a nemusí zahrnovat i vyhlazení vertikální.

Celá řada různých postupů.

Mnoho z těchto postupů využívá běžné „vyhlazovací“ procedury.

- loess regrese, kernel smoothing, polynomická regrese, plovoucí průměr a jiné.

Kontinuální normování

Horizontální vyhlazení parametrů.

- Vytvoření kohort (se shodným *n* či stejným věkovým rozsahem).
- V rámci každé kohorty odhad populačních parametrů (M, SD).
- Tyto parametry jsou vyhlazeny napříč kohortami.
- Pro každý cílový věk je pak možné predikovat M a SD a spočítat standardní skór.
- Implementace např. ve WJ-IV (ten využívá ještě rozdílnou SD směrem „nahoru“ a „dolů“ a bootstraping).

Lokálně vážený odhad parametrů.

- Pro každý cílový věk jsou odhadnuty vážené distribuční parametry (M, SD).
- Každému respondentovi je přiřazena váha – čím shodnější věk, tím vyšší, a naopak.
- Váhy mají normální rozložení s M v cílovém věku a zvolenou SD (třeba půl roku).
- Může být velmi nestabilní. Ideálně v kombinaci s bootstrapem.

Kontinuální normování

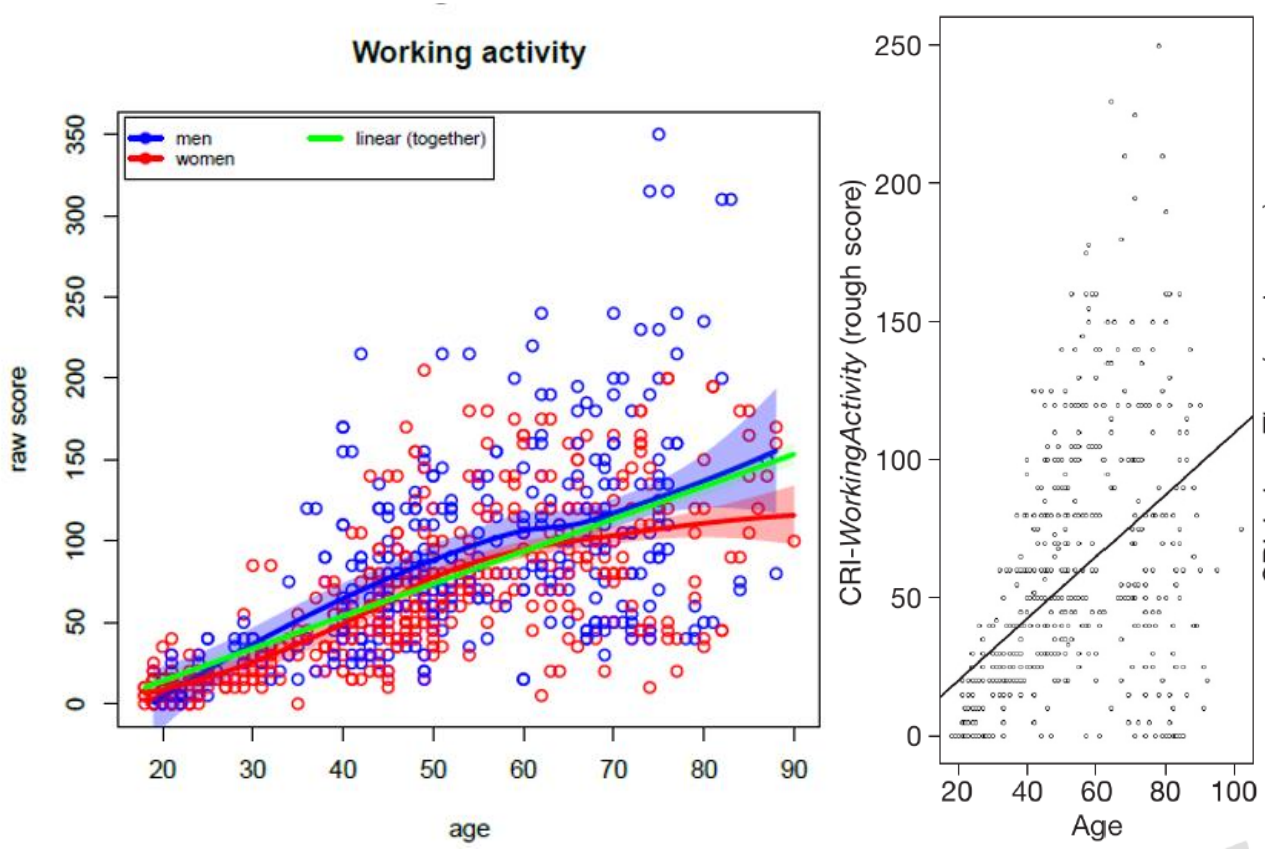
Regresní normy.

- Výkon respondenta v testu (DV) je predikován pomocí lineární regrese.
- Prediktory (IV) jsou věk (a jeho případné polynomy), případně další proměnné (pohlaví...).
- Standardizované reziduum = z-skór.
- Časté v klinických metodách.
- Výhoda: parametrické, stačí malý vzorek.

Jaké jsou nevýhody?

- Předpoklad stejného rozptylu napříč věkem.
- Podhodnocení v kohortách s malou variabilitou, nadhodnocení v kohortách s velkou variabilitou.

Velmi nevhodný postup!



Nucci, M., Mapelli, D., & Mondini, S. (2012). Cognitive Reserve Index questionnaire (CRIq): a new instrument for measuring cognitive reserve. *Aging clinical and experimental research*, *24*(3), 218–226. https://doi.org/10.3275/7800
Javůrková,A., Raudenská, J., Cígler, H., Ježek, S. (unpublished manuscript). Czech adaptation of Cognitive Reserve Index questionnaire (CRIq).

Kontinuální normování

Další postupy: např. **kontinuální normování s využitím Taylorových polynomů**.

- Vytvoření kohort a prozatímního standardního skóre (percentil či plošná transformace).
- Předpoklad: hrubé skóre X je funkcí věku či ročníku, a, a standardního skóre l.

$$X = f(l, a)$$

- Vytvoření Taylorových polynomů a normovací funkce $x(l, a)$ jako

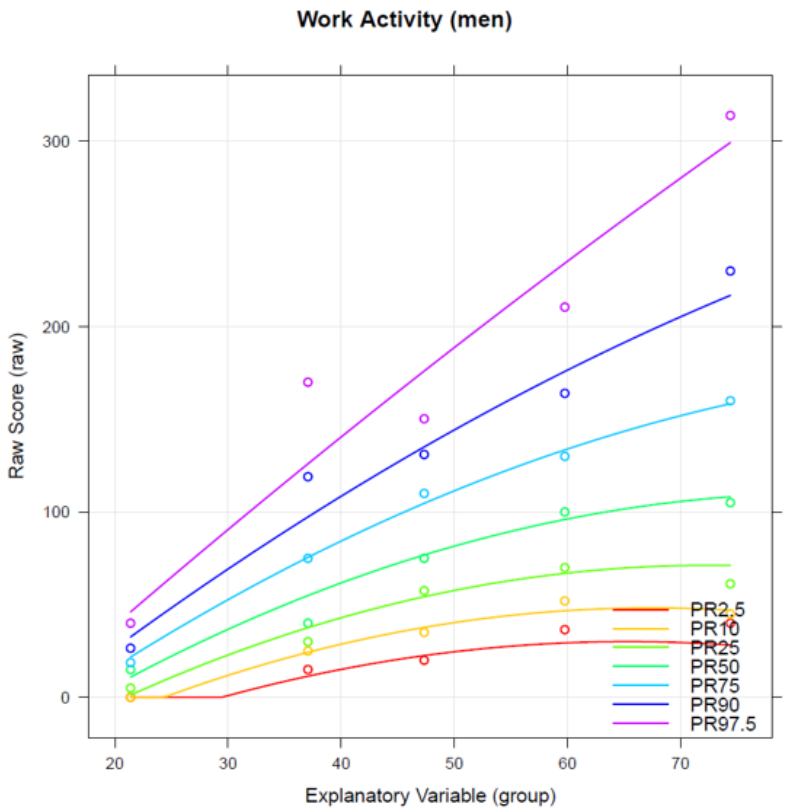
$$x(l, a) = \sum_{s,t=0}^{k} c_{st} l^s a^t$$

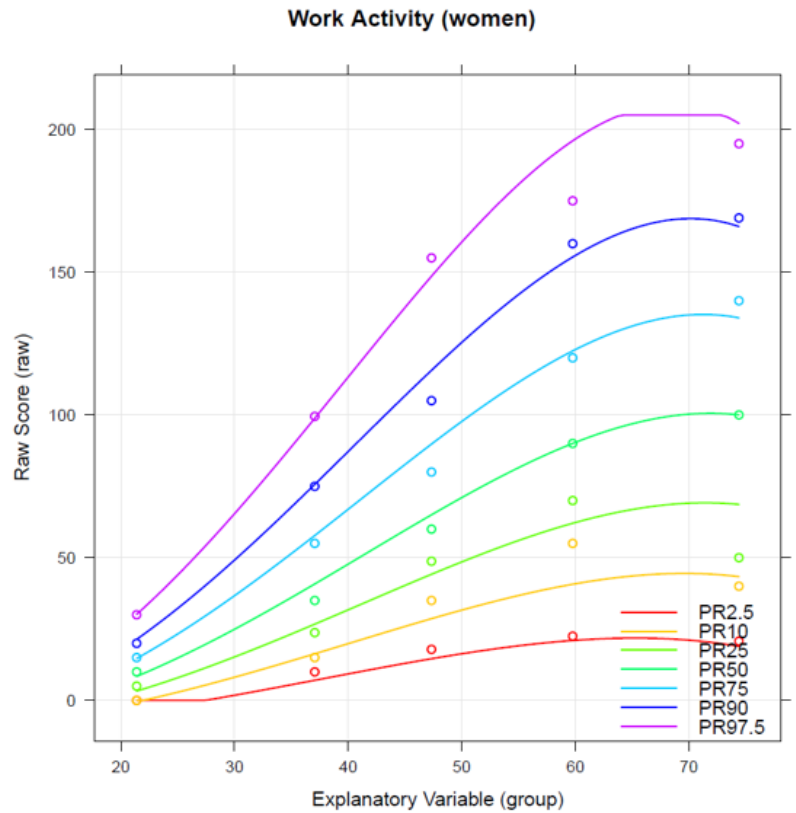
- Nalezení normovacích parametrů c_{st} pomocí step-wise lineární regrese.
- Ověření modelu a identifikace vhodných polynomů pomocí cross-validace, různé postupy ověření modelu.
- Typicky postačuje $k < 6$, většinou $k < 4$ (počet parametrů modelu).
- Vertikální i horizontální vyhlazení v rámci jediného postupu.

Kontinuální normování

Velmi jednoduchá implementace v R balíčku `cnorm`.

- Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A Continuous Solution to the Norming Problem. *Assessment*, *25*(1), 112–125. https://doi.org/10.1177/1073191116656437
- K dispozici podrobný návod.
- Využívají testy WJ-IV COG CZ nebo BACH.





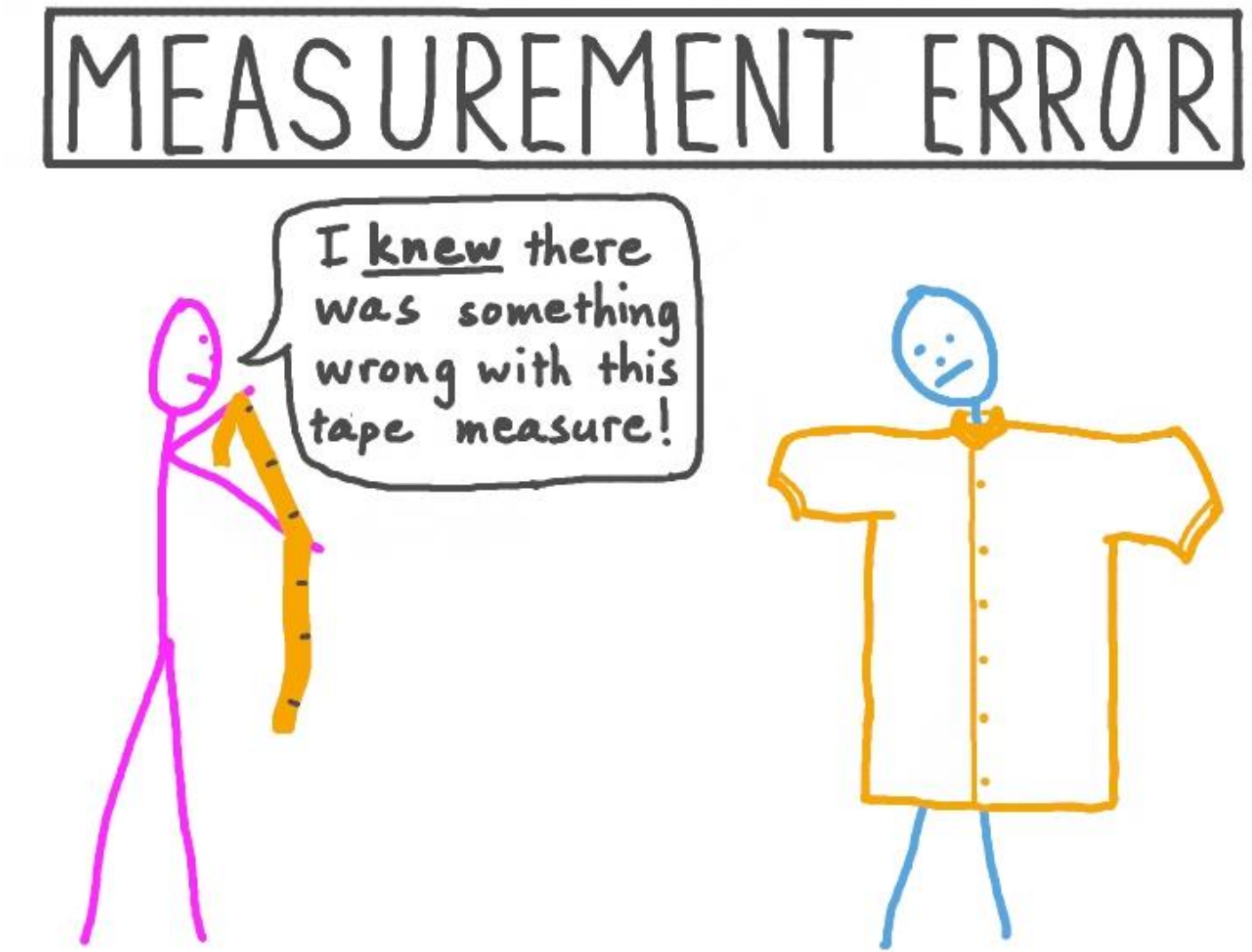
Javůrková,A., Raudenská, J., Cígler, H., Ježek, S. (unpublished manuscript). Czech adaptation of Cognitive Reserve Index questionnaire (CRIq).

Chyba měření a intervaly spolehlivosti

Opakování:
standardní chyba měření
standardní chyba predikce
standardní chyba rozdílu

Statisticky významný rozdíl

Klinicky významný rozdíl



https://www.nagwa.com/en/videos/138104137874/

Otázky spojené s chybou měření

Respondentovi naměřím výšku 178 cm.
Jaké otázky si mohu položit?

- Kolik měří právě teď?
- Kolik bude měřit příště?
- Kolik mu můžu naměřit příště, pokud se jeho výška nezmění?
- Kolik mu musím naměřit příště, abych mohl konstatovat, že se jeho výška změnila?

Kromě toho naměřím i jeho hmotnost 65 kg.
Jaké další otázky si mohu položit?

- Je „vyšší než těžší“?
- Je „vyšší než těžší“ oproti jiným respondentům?

Chyba měření

Standardní chyba měření: směrodatná odchylka pozorovaných hodnot okolo skutečné úrovně atributu

Ilustrace:

- http://fssvm6.fss.muni.cz/height/

Další příklad náhodného samplingu:

- https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm
- https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

Chyba měření a CI

Rozložení naměřených hodnot je normálně rozložené a definované svým M a SD.

Proto, když konstruujeme CI, musíme vědět:

- Okolo čeho? Jaký je průměr rozložení?
- Jak nepřesné? Jaká je směrodatná odchylka rozložení (SE?)

Tři klíčové vzorce (z nichž lze vše odvodit)

1. Základní teorém CTT:

$$X = \tau + e$$

- X – pozorované, τ – pravé skóre a e – chyba.

2. Reliabilita $r_{xx'}$ je podíl vysvětleného rozptylu:

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

- Symbol sigma (σ^2) označuje rozptyl.

3. Rozptyl součtu dvou náhodných proměnných A+B má rozptyl:

$$\sigma_{A\pm B}^2 = \sigma_A^2 + \sigma_B^2 + 2\sigma_{AB} = \sigma_A^2 + \sigma_B^2 \pm 2r_{AB}\sigma_A\sigma_B$$

- $\sigma_{AB} = \mathrm{cov}(A, B)$ – kovariance, r_{AB} – jejich korelace (grafická ilustrace)
- Protože $r_{\tau e} = 0$, pak z 1 a 3 vyplývá $\sigma_x^2 = \sigma_\tau^2 + \sigma_e^2$.

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na jednotku směrodatné odchylky
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na jednotku směrodatné odchylky
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou
- převod z podílu (z-skóre) přímo na škálu směrodatné odchylky (standardních skórů)

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

- reliabilita – podíl vysvětleného rozptylu
- „nereliabilita“ – podíl nevysvětleného rozptylu
- převod z rozptylu na směrodatnou odchylku
  - podíl směrodatné odchylky pravého skóru, která je „způsobena“ chybou
- převod z podílu (z-skóre) přímo na škálu směrodatné odchylky (standardních skórů)
- směrodatná odchylka paralelních testů (= standardní chyba měření)

Standardní chyba měření

Když rovnici $r_{xx'} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$ vyřešíme pro σ_e, získáme vzorec standardní chyby měření:

$$SE = \sigma_e = \sigma_x\sqrt{1 - r_{xx'}}$$

Ve vzorci je $r_{xx'}$ vysvětlený rozptyl; viz koeficient determinace.

- Tedy rozptyl měření vysvětlený pravým skórem. Na rozdíl od koeficientu determinace tam není mocnina, protože reliabilita je už přímo „umocněná“.
- $r_{x\tau} = \sqrt{r_{xx'}}$ a tedy $r_{x\tau}^2 = r_{xx'}$

Středová hodnota

Chyba se nepohybuje kolem pozorovaného, ale kolem pravého skóre.

Jaká je nejpravděpodobnější hodnota pravého skóre při určitém pozorovaném skóre x?

O trochu blíže k průměru (protože pravé skóry mají menší rozptyl než pozorované skóry).

Regresní model CTT:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

- $\mathrm{E}(T|x)$: očekávané (expected), nejpravděpodobnější pravé skóre.
- $r_{xx'}$: reliabilita; „směrnice“.
- M_x : průměrné skóre; $(1 - r_{xx'})M_x$ je „průsečík“.
- Čím větší reliabilita, tím větší vliv pozorovaného skóre a menší vliv průměru (a naopak).

Směrodatná odchylka pravého skóre: $\sigma_\tau = \sqrt{r_{xx'}}\sigma_x$

Chyba měření (v CTT)

Takto spočítanou chybu měření mohu použít pro konstrukci intervalu spolehlivosti.

$$CI_i = \mathrm{E}(X) \pm z_i \sigma_e$$

- $\mathrm{E}(X)$ = očekávaná hodnota, okolo které interval konstruuji.
- σ_e = chyba měření
- z_i = kvantil normálního rozdělení

Kvantily normálního rozdělení:

- 95% CI: $z_{95\%} \cong 1{,}96$
- 90% CI: $z_{90\%} \cong 1{,}64$
- 80% CI: $z_{80\%} \cong 1{,}28$
- 68% CI: $z_{68\%} \cong 1{,}00$

Shrnutí: Důležité prvky práce s SE

Co je očekávanou hodnotou, okolo které interval konstruuji?

- Pozorované skóre?
- Odhad pravého skóre?
- Nula (pro rozdíl dvou skórů)?

Jak spočítám chybu pro daný účel/diagnostickou otázku?

Jaký odhad reliability nejlépe použiju pro daný účel?

Scénář 1: Standardní chyba měření

Pokud jsme naměřili pozorované skóre X, jaké jiné alternativní X jsme mohli rovněž naměřit?

Slouží pro popis chyby měření a intervalu spolehlivosti jednoho jediného měření.

Velikost chyby:

$$\sigma_e = \sigma_x \sqrt{1 - r_{xx'}}$$

Středová hodnota: odhad pravého skóre

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Scénář 2: Chyba odhadu pravého skóre

Pokud jsme naměřili pozorované skóre X, jaká je chyba odhadu pravého skóre τ?

Vzorec je stejný, jen namísto SD pozorovaného skóre použijeme odhad SD pravého skóre:

Velikost chyby:

$$\sigma_{e(\tau)} = \sigma_\tau \sqrt{1 - r_{xx'}} = \sigma_x \sqrt{r_{xx'}} \sqrt{1 - r_{xx'}}$$

Středová hodnota:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Někteří autoři tento postup doporučují, ale potíž s interpretací.

- Zajímá nás chyba na škále použité při konstrukci norem. Zpravidla tedy nepoužitelné.
- Nicméně např. WISC-5[UK] – pro standardizaci na IQ použil právě σ_τ
  - Standardizace $IQ = 15\frac{(X-M_x)}{\sigma_x\sqrt{r_{xx'}}} + 100$ namísto běžného $IQ = 15\frac{(X-M_x)}{\sigma_x} + 100$

Scénář 3: Standardní chyba predikce

Naměřil jsem X. V jakém rozsahu bude ležet příští měření, pokud se úroveň atributu nezmění?

- *„Zlepšil se klient v terapii?“ „Je účinný výukový program?“*

Velikost chyby:

$$\sigma_{pred} = \sigma_x \sqrt{1 - r_{xx'}^2}$$

- $r_{xx'}^2$ - druhá mocnina (test-retest) reliability.
- jde o úpravu $\sigma_{pred} = \sqrt{\sigma_e^2 + \sigma_{e(\tau)}^2}$, tedy rozdíl chyby odhadu pravého skóru a chyby měření

Středová hodnota = očekávaný skór při retestu: odhad pravého skóre:

$$\mathrm{E}(T|x) = r_{xx'}x + (1 - r_{xx'})M_x$$

Scénář 4: Statisticky významný rozdíl

Standardní chyba rozdílu. Rozdíl dvou nezávislých testů jedné osoby; případně rozdíl dvou osob.

Jaká je očekávaná odlišnost v měření dvěma testy?

- „Dosáhla vyššího skóru Anežka nebo Bedřich?“ „Je Cyril vyšší nebo těžší?“
- Musí být ve stejných jednotkách.

Velikost chyby:

$$\sigma_{e(A-B)} = \sqrt{\sigma_{e(A)}^2 + \sigma_{e(B)}^2} = \sigma_{ab}\sqrt{2 - r_{aa'} - r_{bb'}}$$

- Pokud jde o měření jediným testem (dvěma testy se stejnou reliabilitou), lze zjednodušit:

$$\sigma_{e(A-B)} = \sqrt{2}SE = \sigma_x\sqrt{2}\sqrt{1 - r_{xx'}}$$

Středová hodnota:

- Jde o rozdíl a očekávaný rozdíl je zpravidla žádný rozdíl, **proto zpravidla 0**.
- To není úplně pravda; pokud $r_{aa'} \neq r_{bb'}$, pak je střední hodnotou $E(\tau'_A - \tau'_B) = \sqrt{r_{AA'}}(A - M) - \sqrt{r_{BB'}}(B - M)$, ale výsledek bude velmi podobný. Zanedbejte.

Scénář 5: Klinicky významný rozdíl

Liší se dva skóry téhož respondenta více či méně než u „běžných“ respondentů?

- To, že se skóry liší, neznamená, že se liší více, než bychom čekali u náhodně vybraného člověka.
- Klinické hypotézy: „*Rozkolísaný profil schopností...*“, „*Je rozdíl ‚klinicky‘ významný?*“ atd.

Příklad:

- **Statisticky významný rozdíl:** *„Člověk má vyšší váhu než výšku (ve standardních jednotkách, např. IQ skórech)“*.
- **Klinicky významný rozdíl:** *„Člověk má vyšší váhu, než by odpovídalo jeho výšce, je tedy obézní.“*

Scénář 5: Klinicky významný rozdíl

Více postupů. Nejjednodušší používá pouze korelaci a je zcela shodný s postupem pro chybu predikce.

Odhad chyby:

$$\sigma_{A-B} = \sigma_{AB}\sqrt{1 - r_{AB}^2}$$

- r_{AB} je korelace testů A a B, σ_{AB} je směrodatná odchylka obou testů (musí být shodná)

Středová hodnota:

$$\mathrm{E}(B|A) = r_{AB}A + (1 - r_{AB})M_{AB}$$

Scénář 6: Více měření

Lze testovat, zda má klient celkově „rozkolísaný profil“.

- Např.: *„Liší se subtesty ve WAIS-III od celkového IQ více, než bychom čekali?“*
- Analogie F-testu u lineární regrese s více prediktory.

Poskytují jen některé diagnostické metody, není pravidlem.

Technicky vzato není ideální interpretovat „profil“, pokud test celkového rozdílu není signifikantní na zvolené *p*-hladině.

Ruční výpočet je příliš náročný.

Sčítání skórů

Obecný vzorec pro součet dvou proměnných A a B:

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

- Rozptyl součtu (σ_{A+B}^2) je roven součtu rozptylů (σ_A^2, σ_B^2) a 2 kovariancí ($\sigma_{AB} = 2r_{AB}\sigma_A\sigma_B$).

Korelovat spolu mohou pouze pravé skóry. Chyby měření jsou náhodné a s ničím nesouvisí.

Rozptyl testu A $\sigma_A^2 = r_{AA'}\sigma_A^2 + (1 - r_{AA'})\sigma_A^2$ lze rozdělit na:

- Rozptyl pravého skóre: $r_{AA'}\sigma_A^2$
- Chybový rozptyl: $(1 - r_{AA'})\sigma_A^2$, což je po odmocnění standardní chyba měření.

Protože korelovat mohou spolu jen pravé skóry, je celá pozorovaná kovariance $2r_{AB}\sigma_A\sigma_B$ kovariancí výhradně pravých skórů.

Sčítání skórů

Rozptyl součtu dvou testů je

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

Po dosazení $\sigma_A^2 = r_{AA'}\sigma_A^2 + (1 - r_{AA'})\sigma_A^2$ a $\sigma_B^2 = r_{BB'}\sigma_B^2 + (1 - r_{BB'})\sigma_B^2$ lze rozepsat:

$$\sigma_{A+B}^2 = r_{AA'}\sigma_A^2 + r_{BB'}\sigma_B^2 + 2r_{AB}\sigma_A\sigma_B + (1 - r_{AA'})\sigma_A^2 + (1 - r_{BB'})\sigma_B^2$$

- Kde červená část je systematický rozptyl a modrá část chybový rozptyl.

Korelace obou subtestů ovlivňuje pouze systematický rozptyl.

- Je-li korelace kladná → větší část systematického rozptylu, vyšší reliabilita.
- Je-li korelace záporná → menší část systematického rozptylu, nižší reliabilita.

Sčítání skórů

Reliabilita součtu/rozdílu dvou testů je tedy

$$r_{A\pm B} = \frac{r_{AA'}\sigma_A^2 + r_{BB'}\sigma_B^2 \pm 2r_{AB}\sigma_A\sigma_B}{r_{AA'}\sigma_A^2 + r_{BB'}\sigma_B^2 \pm 2r_{AB}\sigma_A\sigma_B + (1 - r_{AA'})\sigma_A^2 + (1 - r_{BB'})\sigma_B^2}$$

Vzorec lze snadno upravit:

Stratifikované Cronbachovo alfa:

$$\alpha_{strat} = 1 - \frac{\sum_{i=1}^{k}\left[\sigma_i^2(1 - r_{ii'})\right]}{\sigma_A^2 + \sigma_B^2 + 2r_{AB}\sigma_A\sigma_B}$$

- Ve jmenovateli je rozptyl součtu.

Reliabilita rozdílu:

$$r_{x-y} = \frac{\sigma_A^2 r_{AB'} + \sigma_B^2 r_{BB'} - 2r_{AB}\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2 - 2r_{AB}\sigma_A\sigma_B}$$

- Ve jmenovateli je rozptyl rozdílu.

Sčítání skórů

Pokud testy spolu korelují kladně:

- Při součtu se reliabilita (zpravidla) zvyšuje.
- Při rozdílu se reliabilita snižuje.

Pokud testy spolu korelují záporně:

- Při součtu se reliabilita snižuje.
- Při rozdílu se reliabilita (zpravidla) zvyšuje.

- Stratifikované alfa je odhadem reliability v případě, kdy předpokládáme, že část specifického rozptylu je systematická (tj. nikoli náhodná).

$r_{xx'}$	$r_{yy'}$	r_{xy}	r_{x-y}	r_{x+y}
0,7	0,8	0	**0,75**	**0,75**
0,7	0,8	0,2	**0,69**	**0,79**
0,7	0,8	0,4	**0,58**	**0,82**
0,7	0,8	0,6	**0,38**	**0,84**
0,7	0,7	0,6	**0,25**	**0,81**
0,9	0,9	0,8	**0,50**	**0,94**
0,9	0,9	0,45	**0,82**	**0,93**
0,6	0,6	0,5	**0,20**	**0,73**
0,7	0,7	0,65	**0,14**	**0,82**

$r_{xx'}$	$r_{yy'}$	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

Specifické příklady

Ne všichni autoři souhlasí s výše uvedeným postupem $\mathrm{E}(X) \pm z_i \sigma_e$.

- Charter a Feldt (2001) doporučují používat buď $X \pm z_i \sigma_e$ nebo $\mathrm{E}(X) \pm z_i \sigma_\tau$.
- Nesouhlasím ☺

Při výpočtech se standardními skóry záleží na postupu výpočtu.

- Typicky: standardizuje se s pomocí výběrové SD pozorovaného skóre.
- Výjimečně: standardizuje se s pomocí odhadu SD pravého skóre.
  - Určité výhody, nečekané důsledky při interpretaci; výpočty CI a $\mathrm{E}(X)$ se liší.

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment, 19*(4), 350–364. https://doi.org/10.1177/073428290101900404

Práce s chybou v IRT

V IRT existují 2 hlavní typy skórů:

- ML, WLE: bez regrese k průměru. Analogie k X a standardní chybě měření.
- EAP, MAP: s regresí k průměru. Analogie k E(T|X) a odhadu chyby pravého skóre.

Intervaly spolehlivosti pro měření se konstruují zcela shodně.

- Jen se použije chyba odhadu pro specifickou úroveň latentního rysu.

Standardní chyba rozdílu: shodně jako v CTT: $SE_{A-B} = \sqrt{\sigma^2_{e(A)} + \sigma^2_{e(B)}}$

- Statisticky významný rozdíl.

Výpočet klinicky významného rozdílu je komplikovanější.

- Typicky se používá bootstraping, případně aproximace s pomocí postupu CTT.