MUNI FSS Katedra psychologie

Hodnocení metod psychologické diagnostiky ve výzkumu i praxi

9. 11. 2022 | Workshop na Katedře psychologie FF UP v Olomouci

Hynek Cígler | cigler@fss.muni.cz
Katedra psychologie, Fakulta sociálních studií MU

Obsah přednášky

Povaha měření v sociálních vědách.

Rozdíly v epistemických východiscích.
- Logický pozitivismus, operacionalismus, instrumentalismus, kritický realismus.

Příslušné teorie měření.
- Klasická testová teorie, reprezentační model, model latentních proměnných.

Příslušné teorie validity.
- Tradiční konstruktová validita, Messickovo pojetí validity, realistické teorie (Borsboom).
- Související otázky: kauzalita, content sampling, „posvátné krávy“ psychometriky.

Modely hodnocení diagnostického nástroje.
- Model Lissitze a Samuelsen.
- Recenzní model EFPA

Prostor pro jakékoli dotazy spojené s modely měření, konstrukcí testu atd.

Jaká je povaha měření v psychologii?

Ve srovnání s přírodními vědami:

- Volnější a proměnlivější vztahy různých atributů (a tedy vyšší chyba měření).
- Absence jednoznačných kauzálních příčin a důsledků.
- Vágní definice atributů.
- Více stochastických, méně deterministických procesů.
- → Větší chyba měření (ať už je definovanána jakkoli).

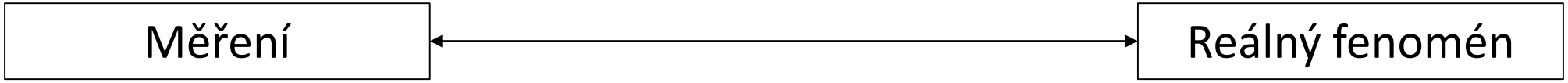
Měřené atributy v psychologii jsou přinejmenším zčásti „konstruované“.

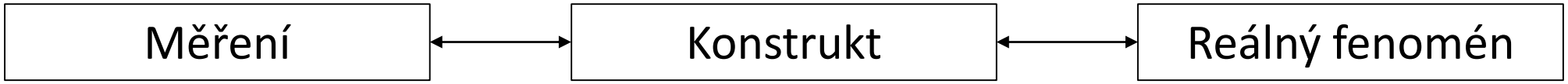
- Lze vůbec měřit něco, co „neexistuje“ (je jen „konstruktem“)?

Běžná praxe v psychologii principy měření vůbec neřeší.

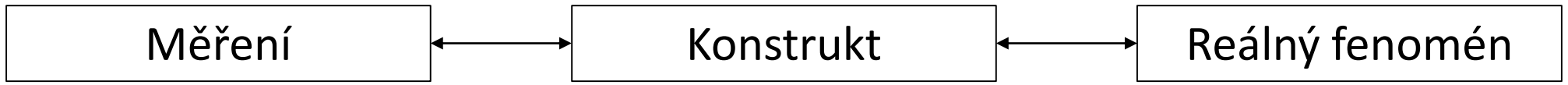
- Měření = testování = diagnostika.
- Existuje velké množství modelů měření s diametrálně odlišnými epistemickými východisky.

Credit: část úvodních slidů vychází z „epistemologických přednášek“ Honzy Šerka v mém kurzu Psychometrika: Měření v psychologii.


Měření
Reálný fenomén


Měření
Konstrukt
Reálný fenomén

Odkazování
Popis



Číselný záznam konstruktu, jeho empirická paralela

Součást vědeckého jazyka

Pozorovatelný
Nepozorovatelný

Přehled přístupů k měření:
Jak se pozná (dobrá) teorie/měření?

Záleží hodnota vědeckých teorií na existenci entit, které tyto teorie předpokládají?

Ne: Antirealismus

- Logický pozitivismus.
  - Tradiční pojetí konstruktové validity (Cronbach & Meehl, 1955).
- Operacionalismus: *Klasická testová teorie* a *Teorie zobecnitelnosti*, *Reprezentační model měření* (CM)
- Pragmatismus a instrumentalismus.
  - Unifikovaná konstruktová validita (Messick, 1989, 1995).
- Sociální konstruktivismus.

Ano: Realismus

- Naivní realismus.
- Kritický realismus: *Model latentních proměnných* (FA, IRT)
  - Realistické pojetí validity (Borsboom, 2004, 2009, 2013)
  - Lissitzův model hodnocení testu (Lissitz & Samuelsen, 2007).

Analogie k měření v přírodních vědách

Velmi dlouhá tradice (měření délky, hmotnosti atp. už v prehistorické době).

- Některé postupy jsou ale relativně nové: čas, teplota.

Zhruba od 16. století rozvoj vědeckých postupů měření a teorie měření.

Intenzivní vs. extenzivní velikost.

Základní vs. odvozené veličina.

Přímé vs. nepřímé měření.

Problém koordinace a kalibrace.

Analogie k měření v přírodních vědách

Měření je kvantifikací atributu určitého objektu, subjektu či situace.

Tato kvantifikace stojí na několika pilířích:

- Daný atribut existuje.
- Je známa (nebo alespoň dobře definovaná) kvalita/povaha atributu.
- Daný atribut se projevuje v reálném světě, tyto projevy je možné pozorovat.
- Existuje teorie měření, která prováže pozorování s atributem.
  - Problém koordinace.

- Atribut je kvantitativní povahy (jinak jde jen o kategorizaci či seřazení).

První vědecká měření v psychologii

Vývoj měření v sociálních vědách zpočátku následoval přírodní vědy a empiristickou tradici. Měření proto imitovalo jejich postupy.

Jak co nejpřesněji změříte výšku člověka?

Empirismus: průměrování paralelních jevů formuje obecnou ideu.

Empirické vědy: průměrování *paralelních měření* vede ke zpřesnění.

První vědecká měření v psychologii

Vývoj měření v sociálních vědách zpočátku následoval přírodní vědy a empiristickou tradici. Měření proto imitovalo jejich postupy.

Průměrování má dlouhou tradici.

Kobels (1535): Formální definice jedné „stopy“ je *zprůměrování stop 16 náhodně vybraných osob*!

Tento postup vedl ke stabilní definici jedné stopy, která se příliš neměnila a která nevyžadovala vlastnictví etalonu.

První vědecká měření v psychologii

Vývoj měření v sociálních vědách zpočátku následoval přírodní vědy a empiristickou tradici. Měření proto imitovalo jejich postupy.

- Zejména tzv. „paralelní měření“ a průměr z nich.
- Existuje několik základních konceptů, které připravily půdu pro měření sociálních věd.

Adolphe Quetelet (1842): Koncept *průměrného člověka*.

- Prostý průměr charakteristik napříč lidmi („sociální fyzika“, SS, BMI).

Francis Galton (1869): Koncept *korelace*.

- Průměrný vztah charakteristik napříč lidmi.

Karl Pearson (přelom 19./20. stol): Koncept *systematického rozptylu*.

- Průměrná systematická variabilita v kontrastu k průměrné chybové variabilitě.
- Centrální momenty jako analogie k momentu hybnosti.

První vědecká měření v psychologii

První měření realizovali nejen psychologové, ale i fyzikové.

- Např. Fechner byl původně fyzik, naopak von Helmholtz se jako fyzik zabýval vnímáním.

Kandela: základní jednotka SI.

- Některé původní jednotky svítivosti pracovaly s definicí „viditelnosti na vzdálenost“.

Weber-Fechnerovy experimenty: „nejmenší jednotka čití“.

- Tato jednotka by mohla posloužit pro další definici „psychických veličin“.
- Proto psychofyzika = propojení psychologie a fyziky.
- Decibel je přímo odvozen z WF zákona.

Stručné dějiny měření v psychologii

Dvě tradice měření:

- Psychofyzika → matematická psychologie.
  Modelování kognitivních procesů, spíše intraindividuální procesy.
  - Z Psychometric Society se vyčlenila až v r. 1964.
- Mentální testování → psychometrika. Škálování, spíše interindividuální procesy.

První pokusy o systematické měření:

- Francis Galton (1822–1911)
- Wilhelm Wundt (1832–1920)
- James McKeen Cattell (1860–1944)
- Wijsen, L. D., Borsboom, D., Cabaço, T., & Heiser, W. J. (2019). An Academic Genealogy of Psychometric Society Presidents. *Psychometrika*, *84*(2), 562–588. https://doi.org/10.1007/s11336-018-09651-4

Stručné dějiny měření v psychologii

Charles Spearman (1863–1945): **Základy klasické testové teorie**

- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159
- Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Alfred Binet (1857–1911), Robert Mearns Yerkes (1876–1956).

Prudký rozvoj v 1. pol. 20 století.

- Revize metod kolem poloviny století.
- Rozvoj od 80. let v souvislosti s výpočetní technikou.
- Posledních 20 let: moderní výpočetní metody, počítačové testování.

Buchanan, R. D., & Finch, S. J. (2014). History of Psychometrics. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06305

Stručné dějiny měření v psychologii

Jak vytvořit „skór“? Jak měření parametrizovat?

Počátek 20. století: Značné úsilí „parametrizovat“ psychologické testy.

- Řada různých „škál“: Hayes a Patterson (1921), Bogardus (1925), Thurstone (1928), Likert (1932), Guttman (1944), Osgood (1957) a další.
- Od 50. let minimální rozvoj; naprostá dominance CTT parametrizace v běžné praxi.

Validita měření:

- Zpočátku kriteriální (zejm. prediktivní), později souběžná validita.

Reliabilita:

- Zpočátku test-retest, později paralelní formy, split-half.
- Vnitřní konzistence až od 30. let.

Stručné dějiny měření v psychologii

Fergusonova komise (1932–1940).

- Striktní požadavek aditivity a zřetězení.
- Psychologové zřetězení nedokázali → v psychologii neexistuje teorie měření → psychologie není empirická věda.

Reakce č. 1: **Stevensova *operační* teorie měření.**

- Matching: „Measurement, in the broadest sense, is defined as the **assignment of numerals to objects and events according to rules**.“ (Stevens, 1946, s. 677)
  - Ve skutečnosti zjednodušení konsenzu z přírodních věd: „Measurement is a method of assigning numbers to magnitudes“ (např. von Helmholtz, 1887).
  - Klasické měření: Existuje magnituda, kterou kvantifikujeme pomocí měřicího nástroje (realismus).
  - CTT: Magnitudu „vytváříme“ s pomocí pravidla bez ohledu na povahu jevu (operacionalismus).

Reakce č. 2: Odpověď Logického pozitivismu s tradiční **konstruktovou validita**

- **Cronbach & Meehl (1955)**.
- Nomologická síť (Campbell & Fiske, 1959).
- MTMM matice.

von Helmholtz, H. (1887/1971). An Epistemological Analysis of Counting and Measurement. In R. Kahl (ed.), *Selected Writings of Hermann von Helmholtz*. Wesleyan University Press.
Pro interpretaci doporučuji: Michell, J. (1993). The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell. *Studies in History and Philosophy of Science Part A, 24*(2), 185–206. doi:10.1016/0039-3681(93)90045-I

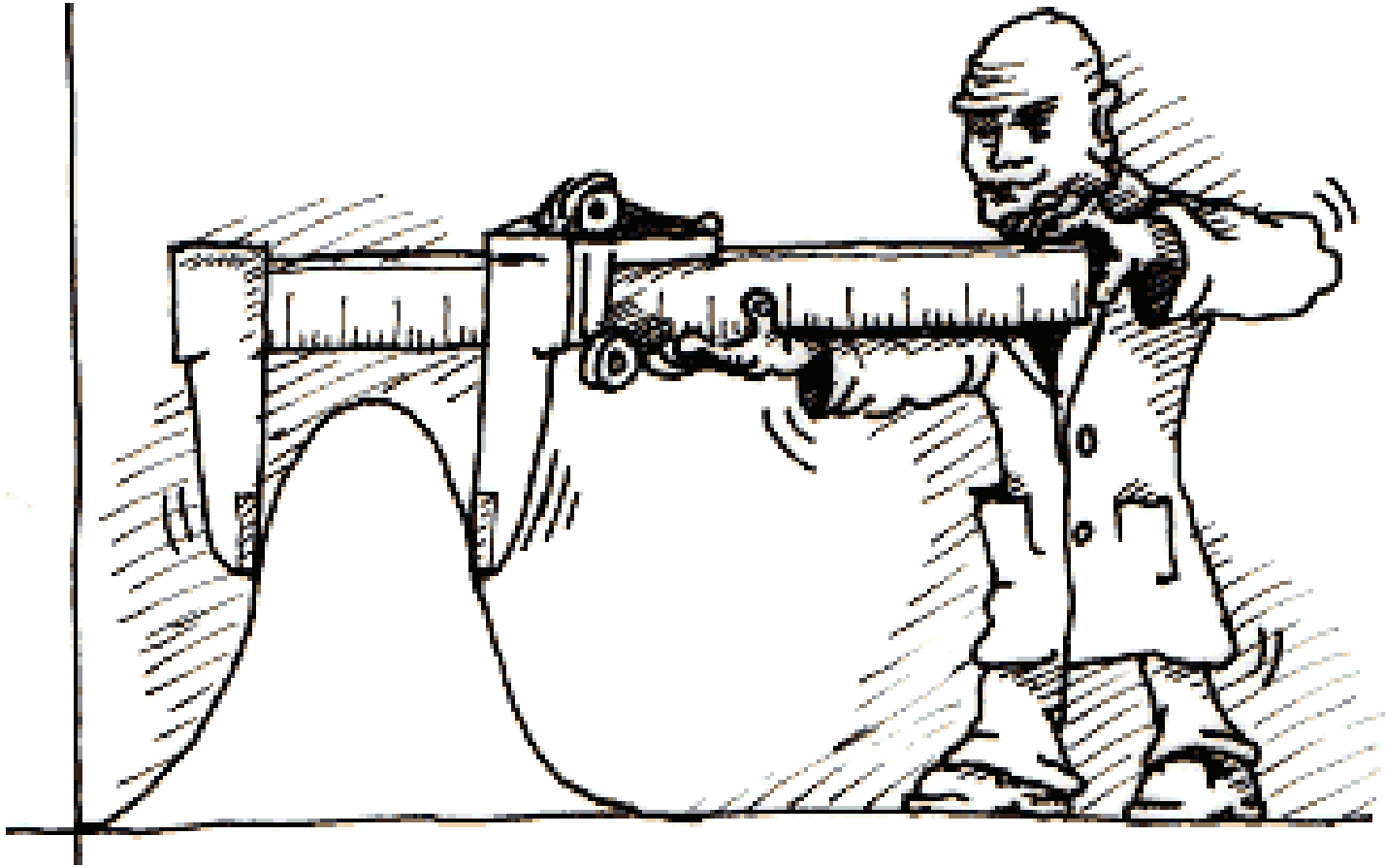
Rozdělení CTT a reprezentačního modelu

Fergusonova komise měla za následek dvě operacionální teorie měření v sociálních vědách.

1. **Klasická testová teorie** (CTT)
   - Stevens (1946), Lord a Novick (1968)
   - „Měření je přiřazování čísel jevům podle pravidel.“ Typicky: sečteme/zprůměrujeme body/položky.
   - Nezabývá se algebraickou strukturou škály, aditivitou.

2. Reprezentační model měření.
   - A zejména **teorie spojitého měření** (CM; Debreu, 1960; Luce & Tukey, 1964).
   - Pomocí aditivních operací vytváří algebraickou strukturu z nealgebraických dat.
     - Jinými slovy: dokáže vytvořit spojitou „míru“ v případě, že pozorujeme pouze seřazená data.
   - Data musí odpovídat modelu. Využití i realistickými teoriemi (Raschův model či řešení koordinace ve fyzice).
     - Existuje-li latentní proměnná, která se manifestuje určitým způsobem, Raschův model bude spojitým měřením a dobře popíše data.
     - Popsal-li Raschův model dobře data, latentní proměnná může, ale nemusí existovat. Aby šlo o CM, je nutné splnit další podmínky.
     - Nepopsal-li Raschův model dobře data, latentní proměnná může, ale nemusí existovat, nicméně nepůjde o CM.

Modely měření



Operacionalismus: Klasická testová teorie

Počátek objevem konceptu reliability (Spearman, 1904), kodifikace v 60. letech (zejm. Lord a Novick, 1968), později rozšíření do teorie zobecnitelnosti (Cronbach, Nageswari a Gleser, 1963; 1972).

Původní motivace vzniku: Attenuation formula (Spearman, 1904).

- Cílem bylo očistit korelaci o nereliabilitu.

Epistemologická východiska až dodatečně: operacionální definice (Stevens, 1946).

- „Ospravedlnění“ typických postupů CTT.

Klasická testová teorie

Analogie opakovaného měření v přírodních vědách.

- Naměřená hodnota je průměr z dílčích měření: $\mathrm{E}(x) = \frac{\sum_{i=1}^{N} x_i}{N}$.
- Chyba měření je výběrová chyba průměru: $SE = \frac{s_x}{\sqrt{N}}$.
- Předpokladem je normální rozdělení chyb (centrální limitní teorém).

V psychologii však nelze měření do nekonečna opakovat.

- Ani tolikrát, aby platil centrální limitní teorém.
- Nelze proto každému respondentovi odhadnout jeho vlastní chybu měření.

Východisko: Předpoklad, že **chyba měření je shodná pro všechny respondenty.**

Klasická testová teorie

Koncept paralelních testů: *„Dobré“ měření je takové, kdy různí lidé v různých časech dojdou různými nástroji ke stejným naměřeným hodnotám, pokud se míra samotného objektu nezměnila.*

Pro paralelní testy platí:

- A. Pravý skór je v paralelních testech a pro každý měřený subjekt stejný
  - $T = \mathrm{E}(X) = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n}$.
- B. Chybový rozptyl je v paralelních testech a pro každý subjekt stejný.
  - Důsledkem je navíc shodný rozptyl pozorovaných skórů obou testů.

Paralelní testy jsou ale jen *myšlenkovým experimentem.*

Klasická testová teorie

Základní teorém CTT: $X = \tau + e$.

- Lineární funkce provazující pozorované a atribut (pravý skór).

Měřeným atributem je „pravý skór“: $\tau = \mathrm{E}(X)$.

- Stevens! Operacionální definice vs. výhrady Fergusonovy komise.

Pravý skór je ale operacionálně definovaný testem: očekávané skóre daného člověka za dané situace v daném testu.

- Z pozitivistického, operacionální hlediska nedává smysl ptát se po „významu“ či „existenci“.

Reliabilita je potom definovaná jako korelace dvou paralelních testů.

- A tedy i podíl rozptylu: $r_{xx'} = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$

Důsledky striktního logického pozitivismu a operacionalismu v CTT

Nelze uvažovat o kauzalitě (pravé skóre neexistuje, a proto nemůže nic způsobovat).

Žádné dva testy neměří to stejné: **množení teoretických entit**.

- A ani ten stejný test v různých situacích neměří to samé.

Nelze srovnávat výsledky měření získané různými testy.

Měřit lze „cokoli“.

- Každý „test“ má svůj „pravý skór“.

Neostré vymezení chyby měření.

Další „potíže“ CTT

Předpoklad, že pozorované skóre je lineární funkcí pravého skóre ($X = \tau + e$).

- Binární položky × koncept vnitřní konzistence.

Předpoklad, že tato lineární funkce je shodná pro všechny paralelní testy.

- Kongenerické, tau-ekvivalentní, paralelní a striktně paralelní testy/položky.

Cronbachovo alfa je proto „biased estimator“ (spodní hranice) reliability.

- Předpoklad tau-ekvivalence.

Pro ověření paralelnosti či „unbiased estimator“ reliability je potřeba faktorová analýza.

- Zcela jiný model měření.

Problematický koncept nejvyšší spodní hranice reliability (*GLB*).

- Část unikátního rozptylu položek může být systematická.

Realismus:
Model latentních proměnných

Realistické „závazky“:

- **Metafyzický:** Existuje svět, který je (částečně) nezávislý na našem poznání
- **Sémantický:** Věty o okolním světě mohou mít pravdivostní hodnotu
- **Epistemologický:** Pravdivé věty o okolním světě představují naši znalost o něm

Argument zázrakem.

Psychologie je implicitně realistická.

- Ale bojuje vůči tomu: „validita testových skórů“, nikoli „validita testu“ (Messsick, APA, AERA).

Realismus neznamená *naivní realismus* ani *redukcionismus*.

- A už vůbec ne pozitivismus.

Realismus:
Model latentních proměnných

Edwards a Bagozzi (2000): Konstrukt je pojem pro popis fenoménu.

- Konstrukty odkazují k fenoménům existujícím nezávisle na naší reflexi.
- Konstrukty samotné reálné nejsou (součást diskurzu).
- Fenomény popisované konstrukty mohou být pozorovatelné i nepozorovatelné.
- Kvalita konstruktu je odvislá od toho, jak dobře popisuje a vysvětluje fenomén.

Model latentních proměnných

Realistická ontologie: latentní proměnná existuje.

Tato latentní proměnná kauzálně „působí“ na manifestní proměnné (indikátory).

Jednoznačná interpretace chyb měření.

- Ale nejasná odlišnost rozdílu validity a reliability.

Otázky týkající se kauzality:

- Kovariance, časová následnost, neexistence třetí proměnné.
- Ale: Judea Pearls, Clive Granger; časové řady, DAG, parciální korelace, network modely. (Nobelova cena za ekonomii 2021: Angrist & Imbens).

Modely latentních proměnných

Faktorová analýza: lineární vztah latentní a manifestní proměnné.

$$X_{ip} = \lambda_i \theta_p + \nu_i + \epsilon_{ip}, \epsilon_{ip} \sim N(0, \sigma_{e,i})$$

Teorie odpovědi na položku: nelineární vztah latentní a manifestní proměnné.
(V tomto případě dvouparametrový logistický binární IRT model; 2PL):

$$\ln \frac{P(x_{ip} = 1 | \theta_p)}{P(x_{ip} = 0 | \theta_p)} = a_i \theta_p + b_i$$

V tomto ohledu je FA jen specifickým modelem IRT pro spojité položky.

- Proto někdy „generalized modeling framework“ (mj. Muthén, Asparouhov) či „item-factor analysis“.

Modely latentních proměnných

IRT modelů existuje nepřeberné množství.

Modely pro nominální proměnné: analýza latentní tříd (LCA).

Např. ordinální faktorová analýza

- Probitový 2PL IRT model s limited information estimátorem:

$$x_{ip}^{*} = \lambda_i \eta_p + \nu_i + \epsilon_{ip}, \qquad \epsilon_{ip} \sim \mathrm{N}(0, \sigma_{e,i})$$

$$x_{ip} = c, \qquad if \quad \tau_{ci} < x_{ip}^{*} \leq \tau_{c+1,i}$$

Související otázky modelů s latentními proměnnými

Formativní vs. reflektivní měření

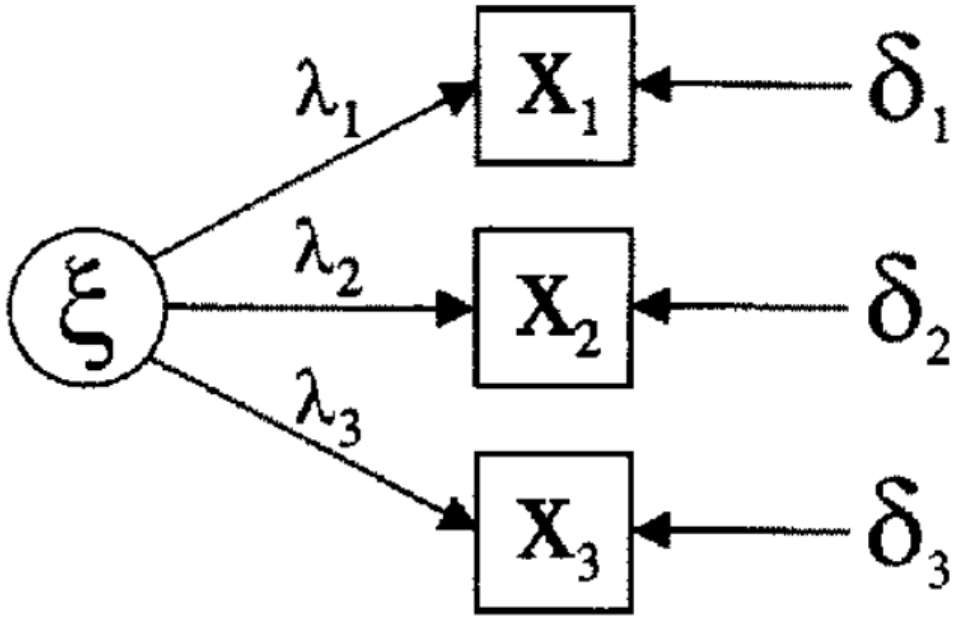


Figure 1. Direct reflective model.

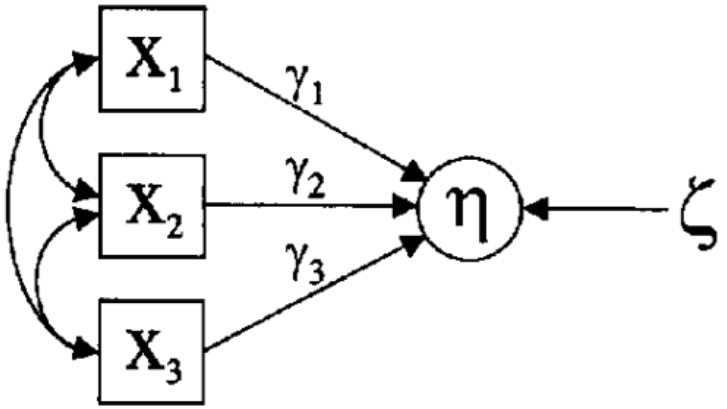


Figure 2. Direct formative model (for simplicity, ϕ.. labels on covariances among exogenous variables are omitted).

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–74. https://doi.org/10.1037/1082-989X.5.2.155

Formativní vs. reflektivní měření

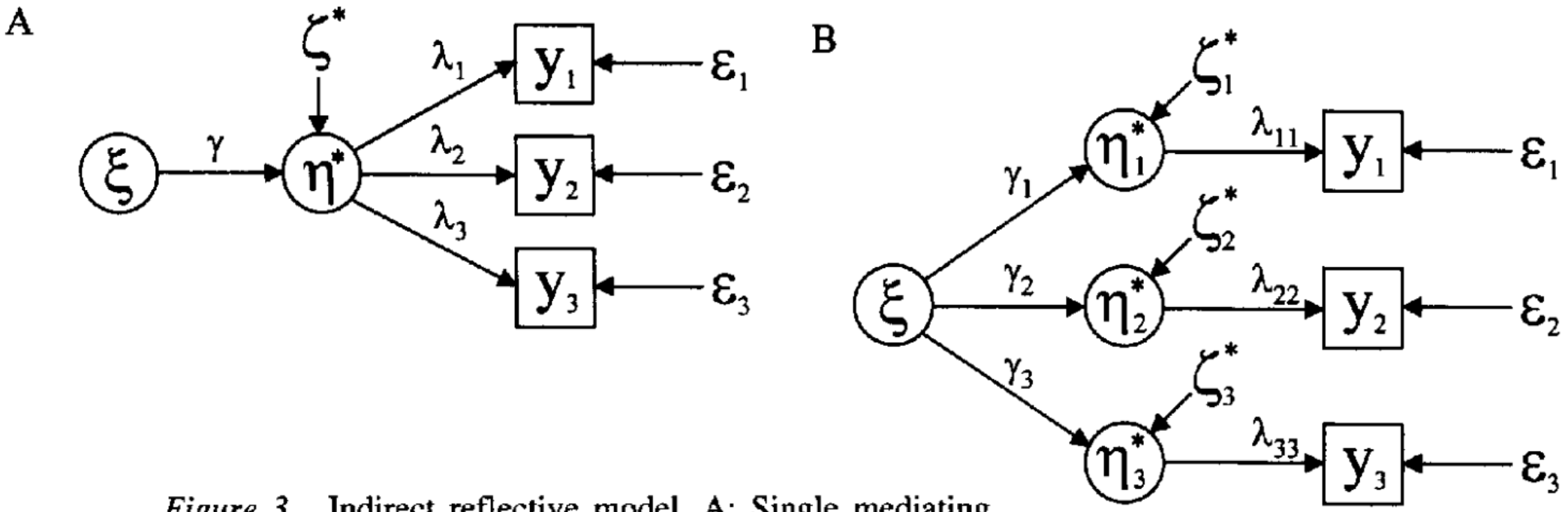


Figure 3. Indirect reflective model. A: Single mediating construct. B: Multiple mediating constructs.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–74. https://doi.org/10.1037/1082-989X.5.2.155

Formativní vs. reflektivní měření

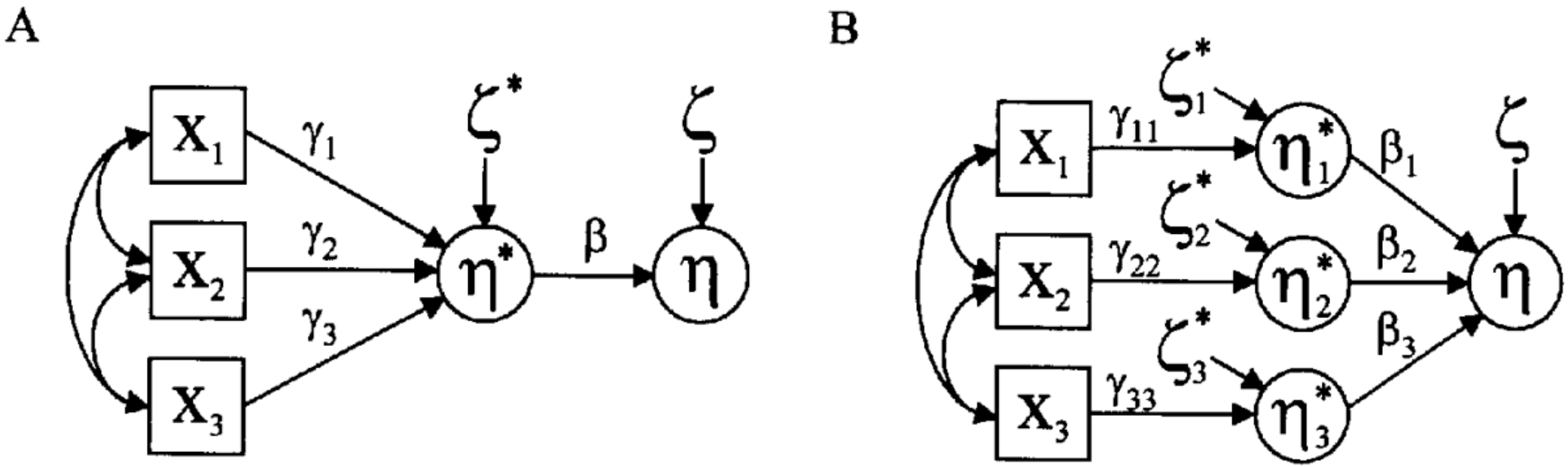


Figure 4. Indirect formative model (for simplicity, ϕ.. labels on covariances among exogenous variables are omitted). A: Single mediating construct. B: Multiple mediating constructs.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–74. https://doi.org/10.1037/1082-989X.5.2.155

Formativní vs. reflektivní měření

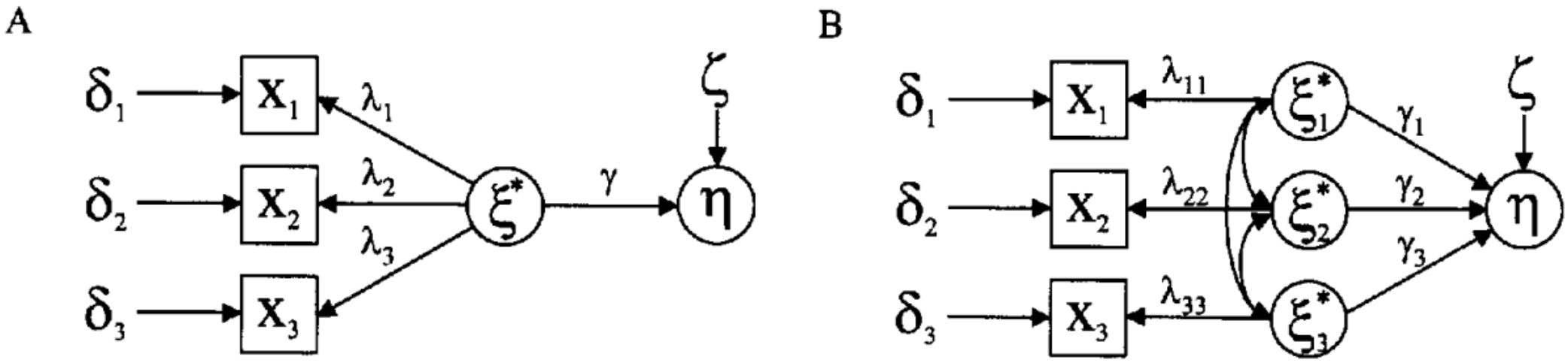


Figure 5. Spurious model (for simplicity, ϕ.. labels on covariances among exogenous variables are omitted). A: Single common cause. B: Multiple common causes.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–74. https://doi.org/10.1037/1082-989X.5.2.155

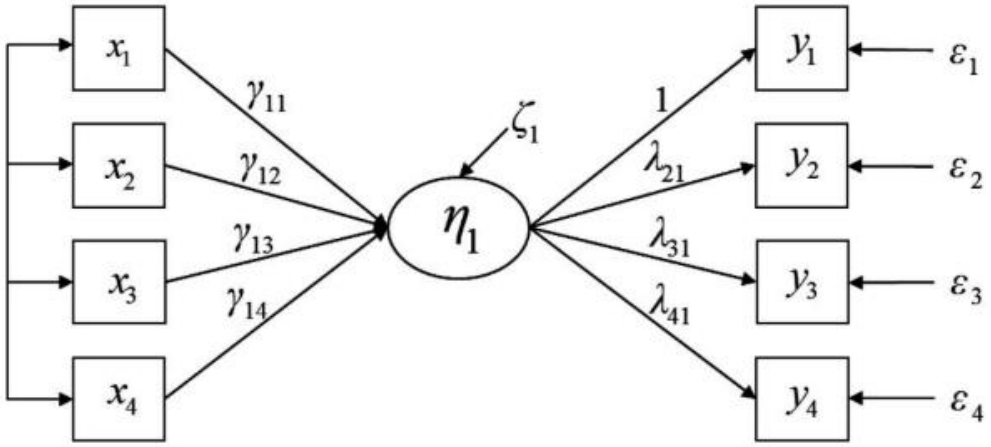
Formativní vs. reflektivní měření

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Pojetí Edwardse a Bagozziho (200) je poměrně staré a reálné situace mohou být komplikovanější.

- Navíc jsou formativní modely „silnější“, než by se podle dřívějších textů zdálo.

Příklad: míra stresu.



Otázka kauzality v modelech s latentní proměnnou

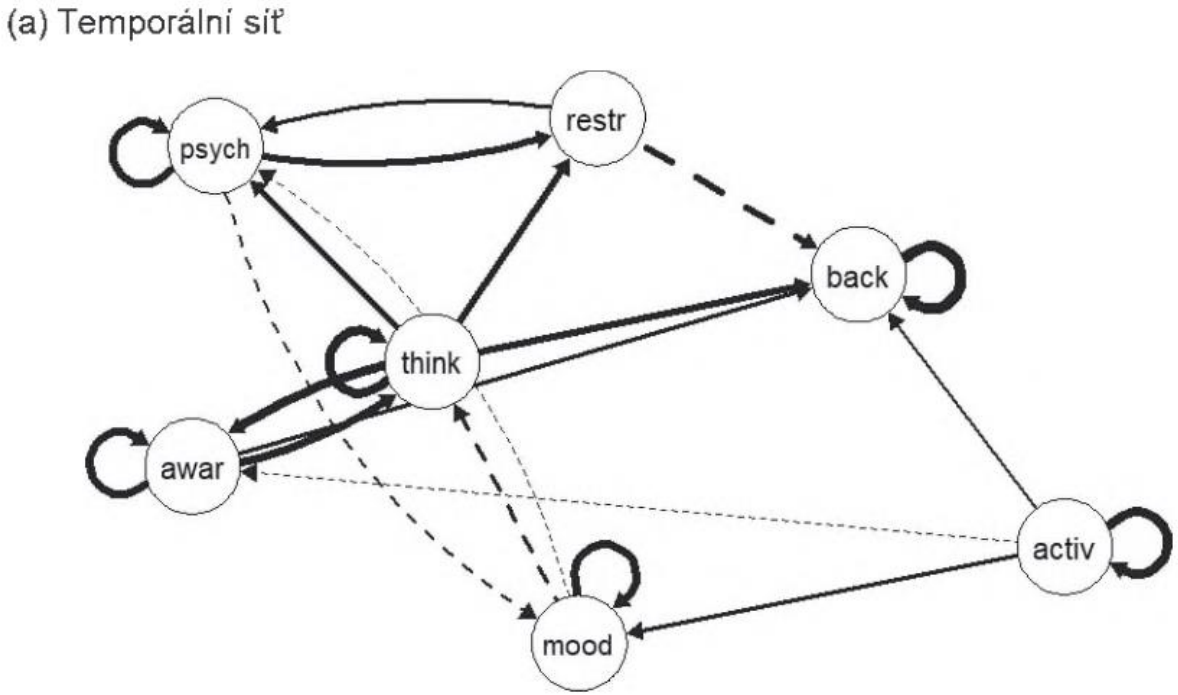
Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1). https://doi.org/10.1016/j.newideapsych.2011.02.007

Latentní model deprese: Existuje latentní proměnná „deprese“, která způsobuje indikátory.

- negativní emoce, únava, potíže se spánkem, serotonin...

Jak pak mohou působit SSRI antidepresiva?

Síťové modely v psychologii.



Nováček, T., Řiháček, T., & Cígler, H. (2020). Personalized network model for medically unexplained physical symptoms. *Československá Psychologie*, *64*(3), 257–271.

Problém s kovariancí v modelech s latentní proměnnou

Předpokladem kauzality: kovariance.

- Předpokládáme, že s různou mírou úrovně latentního rysu se mění úroveň indikátoru.

V psychologii je ale důraz na mezisubjektový výzkum.

- Vnitrosubjektové kovariance většinou nejsou součástí definice konstruktů.

Příklady:

- Inteligence, osobnostní rysy.
- Dědičnost inteligence.
- Reakční rychlost, psychomotorické tempo, pečlivost práce.

Přitom víme, že vnitrosubjektová a mezisubjektová struktura konstruktů se zpravidla liší.

Problém s kovariancí v modelech s latentní proměnnou

Borsboom, D. (2005). *Measuring the mind.* Cambridge University Press.

Lokálně homogenní konstrukty.

- Struktura konstruktu je univerzální pro všechny lidi.
- Nálada, úzkost, aktivace.

Lokálně heterogenní konstrukty.

- Jasná mezisubjektová struktura, ale odlišná vnitrosubjektová struktura.
- Stres, postoje.

Lokálně irelevantní konstrukty.

- Stabilní konstrukty; vnitrosubjektová variabilita je chyba měření.
- Inteligence, osobnostní rysy.

Validita:
různá pojetí

Obrázek vpravo:
Výstava v Londýnském muzeu vědy

Validita

Počátky uvažování o validitě: kriteriální a prediktivní validita.

- Alfred Binet.
- *„Here I am distinguishing between two different but related ideas namely, reliability and validity. An instrument of measurement is reliable to the extent that it yields the same results at different times and in the hands of different persons.* ***It is valid to the extent that it measures the thing it is supposed to measure.“*** (Buckingham, 1921).

Později obsahová a souběžná validita.

Obsahová validita:
Teorie faset a metoda dekompozice obsahového univerza (Guttman, Shye).

Tradiční pojetí validity

Obsahová + empirická + konstruktová.

Cronbach a Meehl (1955): Construct Validity in Psychological Tests.

- *„Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned, and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather then either the test behavior or the scores on the criteria.“*
- Logický pozitivismus: naše pozorování (korelační matice) musí odpovídat teorii (nomologické síti).
- Navázali Campbell a Fiske (1959): metodou Multitrait-multimethod matrix (MTMM).

Tohle pojetí ale není z mnoha důvodů udržitelné.

- Např. neexistuje dostatečně „silná“ psychologická teorie, aby dostatečně precizně popsala nomologickou síť.
- „Mnoho validit“, nejasná kritéria.

Unifikovaná konstruktová validita

Messick (1989, 1995): reakce na neudržitelnost logického pozitivismu.

Messick (1989, s. 20): *„... an **integrated evaluative judgment** of the degree to which **empirical evidence and theoretical rationales** support the adequacy and **appropriateness of inferences and actions** based **on test scores** or other modes of assessment.“*

Pět zdrojů důkazů validity:

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

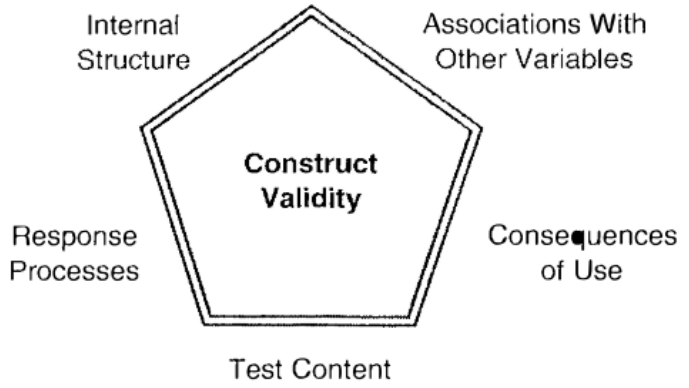


Figure 8.1 A Contemporary Perspective of Types of Information Relevant to Test Validity

Unifikovaná konstruktová validita

Dobrý model pro hodnocení užitečnosti diagnostického nástroje.

Integrované do Standardů pro pedagogické a psychologické testování (APA, AERA).

- AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association
- AERA, APA, & NCME. (2001). *Standardy pro pedagogické a psychologické testování.* Testcentrum.

Instrumentalismus: klíčové je, zda metoda slouží svému účelu.

- Zcela opomíjí latentní rys a měřený atribut; uvažuje výhradně o „testových skórech“.

Výhradní zaměření na diagnostiku.

- Nevhodné pro výzkum a další rozvoj teorií.

Nejde o charakteristiku metody; validita je podle Messicka prostě jen „shrnutí“.

Realistické pojetí validity

Borsboom (2004): *„A test is valid for measuring an attribute if (a) the **attribute exists** and (b) variations in the attribute **causally produce** variation in the measurement outcomes.“*

Tzv. „ontologické pojetí“.

Validita je charakteristikou metody a popisuje shodu nástroje a měřeného atributu (konstruktu).

Podle Borsbooma (2004) je Messickovo pojetí vhodným nástrojem pro hodnocení testu, nejde však o validitu.

Realistické pojetí začíná převládat přinejmenším ve výzkumu.

Čtyři „posvátné krávy“ měření v psychologii

Lilienfeld, S. O., & Strother, A. N. (2020). **Psychological measurement and the replication crisis: Four sacred cows.** *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

1. Obsahová validita a spoléhání se na „název“ škál.
   - Škály se stejným názvem nemusí měřit to stejné.
   - Pro připomenutí: klasická testová teorie a operacionalismus.

2. Ignorování chyby měření a reliability v laboratorních experimentech.
   - Přesvědčení, že pro výzkum postačuje nižší reliabilita (rovněž i Helmstadter).
   - Behaviorální pozorování (vysoce reliabilní) není totožné s měřeným rysem (vztah může být vágní).
   - A jaká je reliabilita experimentální manipulace?

4. Důraz na konvergentní, nikoli divergentní validitu.
   - Konstruktově irelevantní rozptyl, nedostatek diferenciální validity.
   - Potíže zejména při výzkumu silně korelovaných jevů.

(3. Náročnost sběru dat opravňuje malé velikosti vzorku.)

Srovnatelnost měřítek deprese

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

CTT: Měřený atribut je definován měřicím nástrojem.

Co když se ale definice diametrálně odlišují?

Sedm měřících nástrojů: 52 symptomů.

- Průměrný překryv 36 %.
- 40 % symptomů pouze v jediné metodě.

Lze srovnávat výzkumné výsledky *depresivity* s využitím takto odlišných metod?

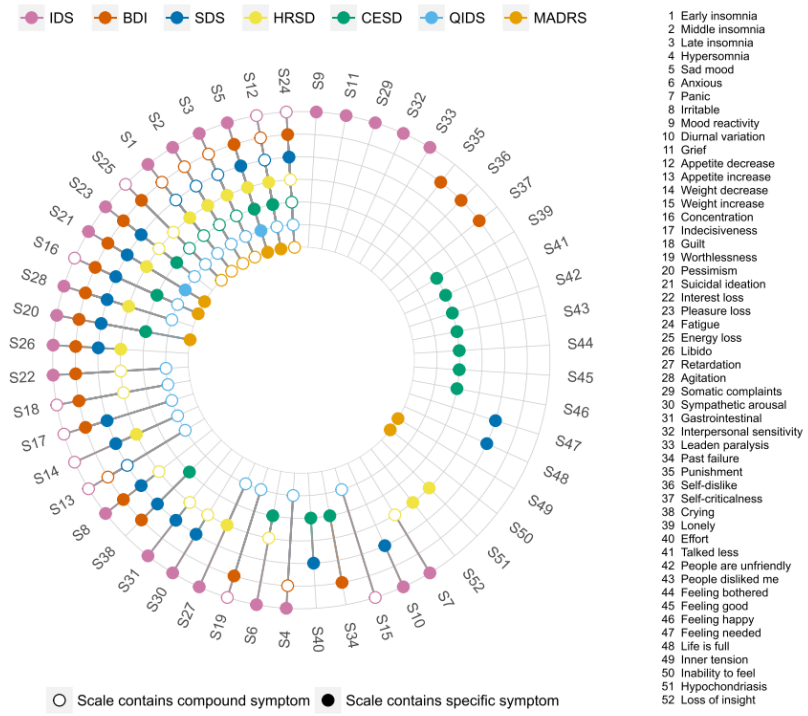
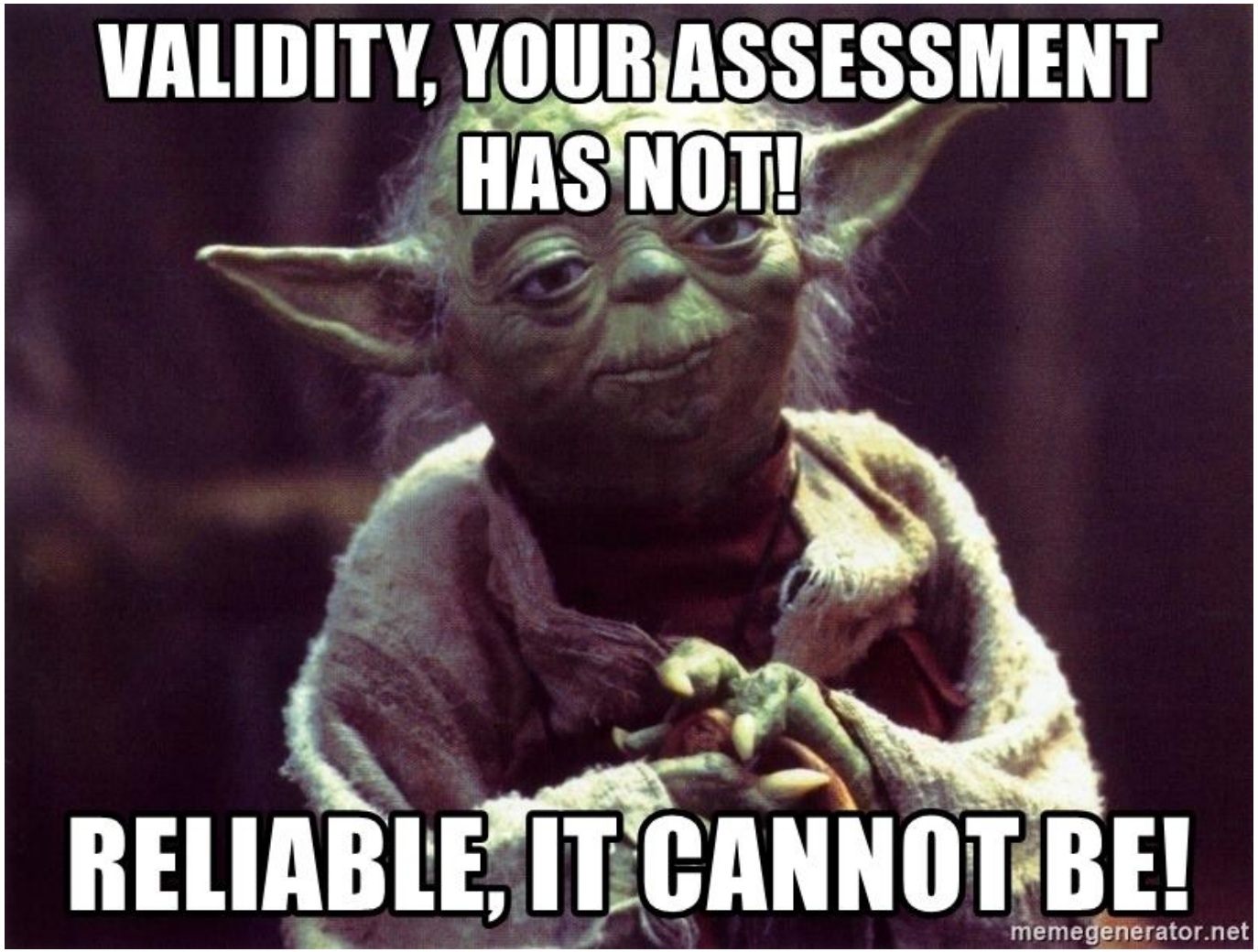


Fig. 1. Co-occurrence of 52 depression symptoms across 7 depression rating scales. Colored circles for a symptom indicate that a scale directly assesses that symptom, while empty circles indicate that a scale only measures a symptom indirectly. For instance, the IDS assesses item 4 hypersomnia directly; the BDI measures item 4 indirectly via a general question on sleep problems; and the SDS does not capture item 4 at all. Note that the 9 QIDS items analyzed correspond exactly to the DSM-5 criterion symptoms for MDD. Please see the online version for colors; in the black and white version, the circles respresent (from outer to inner circle): IDS, BDI, SDS, HRSD, CESD, QIDS, and MADRS.

Hodnocení metod



Obtíže unifikované konstruktové validity při hodnocení metod

Znamenají nízké korelace s kritériem skutečně nízkou validitu měření?

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)

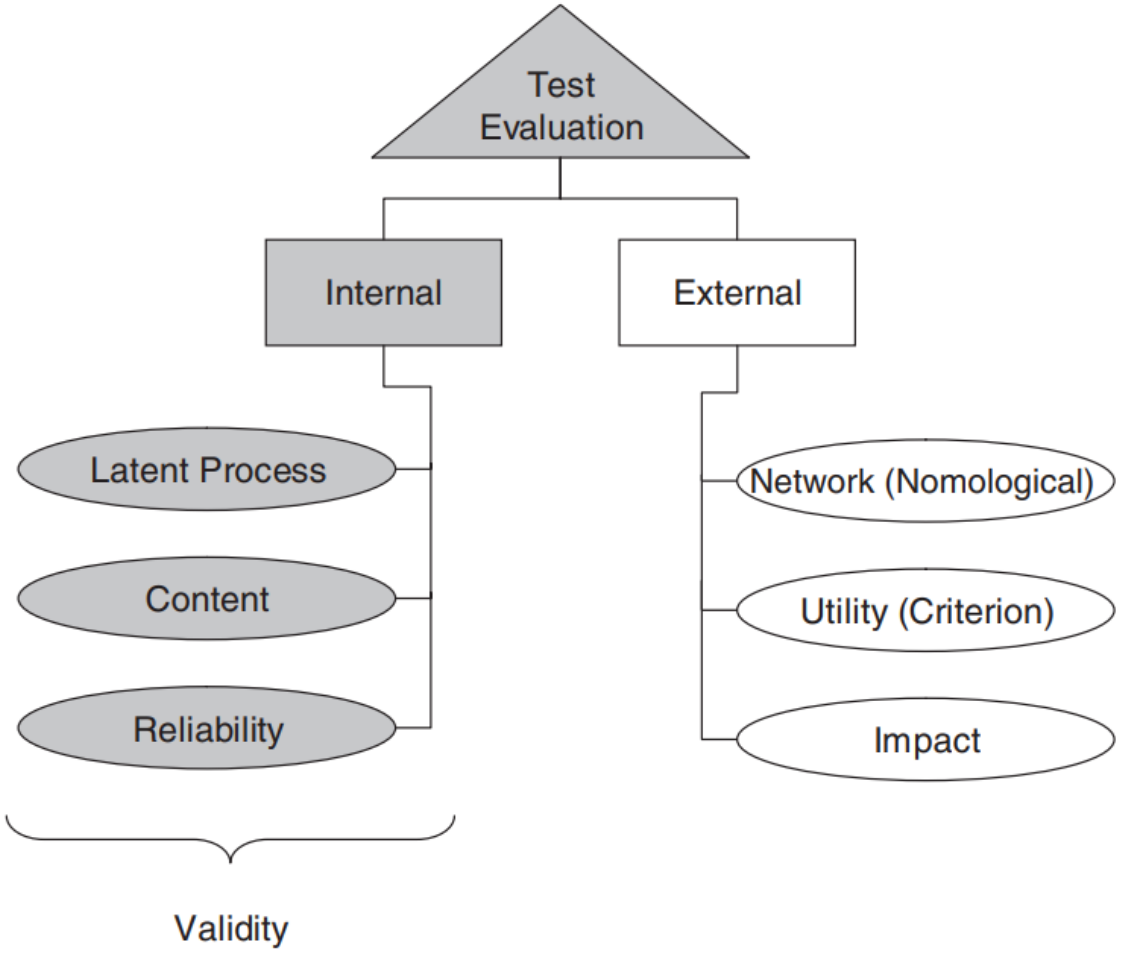


FIGURE 1. *The structure of the technical evaluation of educational testing.*

		Perspective	
		Theoretical	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)?

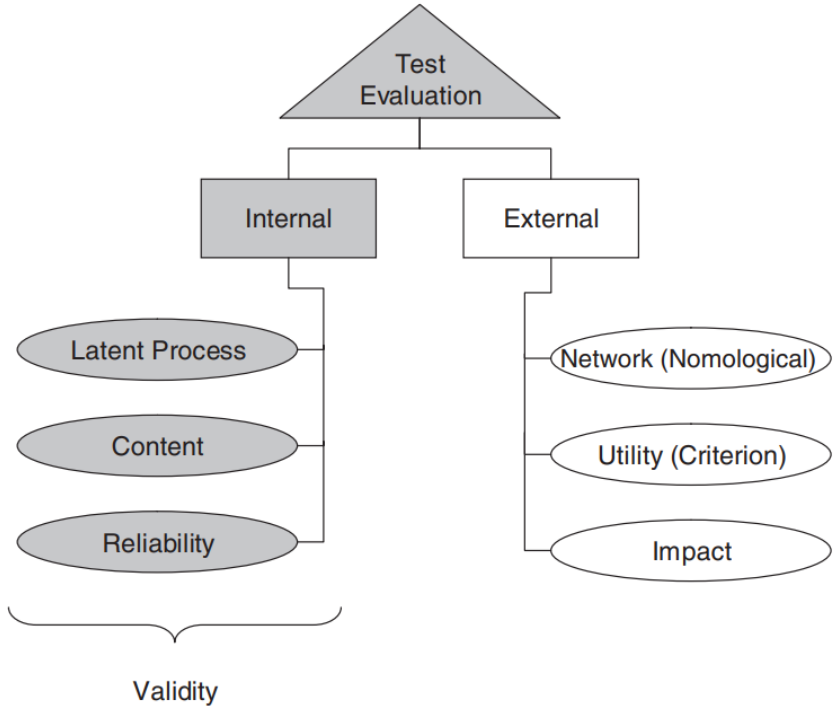


FIGURE 1. *The structure of the technical evaluation of educational testing.*

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.

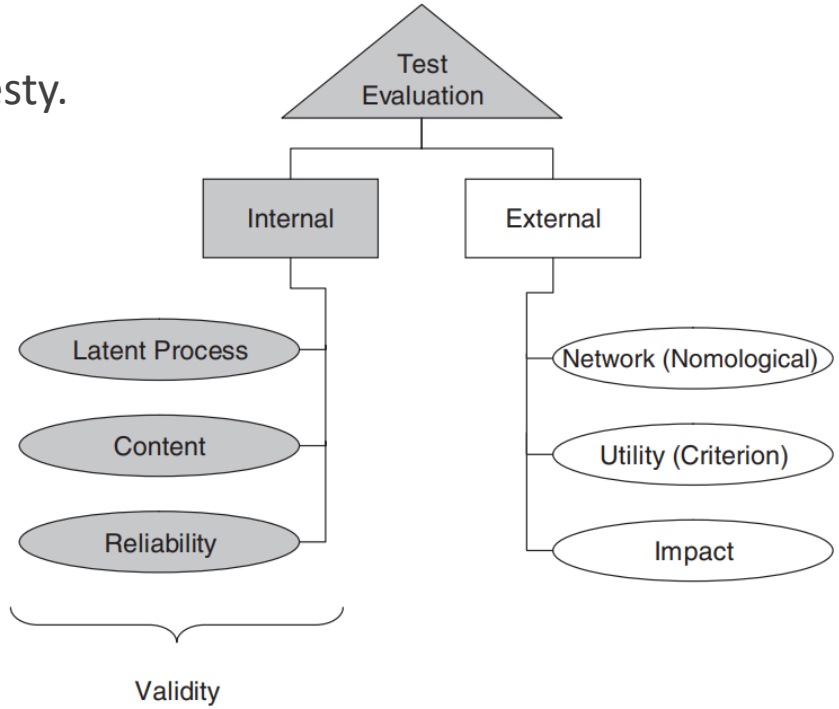


FIGURE 1. *The structure of the technical evaluation of educational testing.*

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? https://hynekcigler.shinyapps.io/sampling-error/

EFPA model pro recenzi testu

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25(3*), 283–291. https://doi.org/10.7334/psicothema2013.97

Autoři použili alternativní zdroje pro tvorbu manuálu.

- Nevychází ani z Messicka, ani z Lissitze, prakticky necitují zdroje týkající se validity.
- Hodnotí formální aspekty metody bez epistemologického zdůvodnění.
- Ale jinak je model poměrně dobrý ☺

Verze 3.42 (z r. 2005) je implementovaná i v ČR.

- Recenzní model používá Testfórum, https://testforum.cz/recenze
- K dispozici je novější verze, ale v současnosti nefunguje web www.efpa.eu a nelze stáhnout.

EFPA model pro recenzi testu

POPIS (NEHODNOTÍCÍ)

obecný popis

klasifikace

skórování

generované zprávy

dodavatel a náklady

ZHODNOCENÍ METODY

kvalita osvětlení teoretických východisek

kvalita materiálů

psychometrické parametry

- normy
- reliabilita
- validita
- (kvalita generovaných zpráv)

závěrečné zhodnocení a hlavně doporučení

zdroje

Díky za pozornost!

Hynek Cígler

Katedra psychologie a Institut pro výzkum dětí, mládeže a rodiny
Fakulta sociálních studií, Masarykova univerzita

e-mail: cigler@fss.muni.cz
web: https://is.muni.cz/auth/osoba/175803
OSF: https://osf.io/t6ufg/
github: https://github.com/hynekcigler

Studijní materiály:

- BC psychometrika: https://is.muni.cz/el/fss/jaro2022/PSYb2590/index.qwarp
- MGR psychometrika: https://is.muni.cz/el/fss/podzim2021/PSYn4790/index.qwarp

Další užitečné zdroje

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Yarkoni, T. (2020). **The generalizability crisis.** *Behavioral and Brain Sciences* [preprint], 1–37. https://doi.org/10.1017/S0140525X20001685

Lilienfeld, S. O., & Strother, A. N. (2020). **Psychological measurement and the replication crisis: Four sacred cows.** *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

Borsboom, D. (2005). *Measuring the Mind*. Cambridge University Press.

Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept*. Cambridge University Press.

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory and Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Luchetti, M., Love, A. C., Marchionni, C., Redei, M., Williamson, J., & De, M.-B. M. (2022). The quantification of intelligence in nineteenth-century craniology: an epistemology of measurement perspective. *European Journal for Philosophy of Science*, *12*(4), 1–29. https://doi.org/10.1007/S13194-022-00485-7