Základy psychometriky

prof. PhDr. Tomáš Urbánek, Ph.D.

Psychometrika v kontextu psychologie

- **Psychometrika a další psychologické disciplíny**
  - obecná psychologie – obecné zákonitosti lidské psychiky, platné pro všechny
  - diferenciální psychologie – rozdíly mezi jedinci nebo skupinami, jejich příčiny
  - vývojová psychologie – poznatky o úrovních zjišťovaných charakteristik na jednotlivých věkových úrovních, principy vývoje
  - metodologie – metody, techniky a postupy konstrukce a výzkumu psychodiagnostických metod
  - matematika a statistika – metody zpracování dat, tvorba modelů měření, chování apod.
  - klinická a poradenská psychologie (a diagnostika) – nové podněty, „objednávka“ metod, zkušenosti z praxe
  - další aplikované disciplíny (psychologie práce, personalistika atd.)

Historie psychometriky

- Dávná historie
  - povědomí o existenci individuálních rozdílů v psychických vlastnostech
    - Epos o Gilgamešovi
    - Bible
  - Čína kolem 2200 př. n. l.
    - ústní zkoušky a pravidelné přezkušování císařských úředníků

Historie psychometriky

- Řecko
  - Hippokrates
    - 4 tělní tekutiny (krev, žluč, hlen, černá žluč)
  - Sokrates, Platón, Aristoteles
  - Galenos
    - 4 temperamentové typy na základě převládající tělní tekutiny

Inspirace z literatury

- Theophrastos
  - považován za zakladatele charakterologie
- Miguel de Cervantes y Saavedra
  - Důmyslný rytíř Don Quijote de La Mancha
- William Shakespeare
  - Hamlet, Macbeth, Othello, Shylock, …
- Stendhal, Honoré de Balzac, Charles Dickens, Victor Hugo, Thomas Mann

„Slepé uličky“

- astrologie
- chiromantie
- frenologie
- fyziognomie
- numerologie

Většinou sporná platnost jejich poznatků

Psychometrika je striktně vědeckou disciplínou.

Počátky psychologie jako vědy

- Přístup k duševně nemocným
  - Středověk
    - exorcismus
  - Philippe Pinel
    - „zbavil šílence okovů“
  - Jean-Martin Charcot a Pierre Janet
    - průkopníci funkčního pojetí psychických poruch
    - počátky hypnózy, počátky psychodiagnostiky

Počátky psychologie jako vědy

- Experimentální přístupy
  - psychofyzika
    - G.T. Fechner – *Elemente der Psychophysik* (1860)
    - E.H. Weber
    - Wilhelm Wundt
  - založení psychologické laboratoře v Lipsku (1879) považováno za zrod vědecké psychologie

Tři zdroje psychometriky

- kolem roku 1830 – 3 relativně nezávislé proudy
- **Francie**
  - studium osob vybočujících z šablon normálního chování (např. Esquirol, Itard, Seguin, Charcot a Ribot)
- **Německo**
  - experimentální výzkum chování normálních dospělých lidí (Weber, Fechner a Wundt)

- **Matematika**
  - užitečnost křivky normálního rozložení (vyjádření v roce 1733 de Moivrem) jako matematického nástroje (Laplace a Gauss)
  - pro měření lidských charakteristik Quetelet v roce 1846

„Praotcové“ mentálního testování

- **Francis Galton** (1822-1911)
- **James McKeen Cattell** (1860-1944)
- **Alfred Binet** (1857-1911)

Francis Galton

- ovlivněný prací svého nevlastního brantrance Charlese Darwina (*O původu druhů*; 1859)
- knihy *Hereditary Genius* (1869) *a Inquiries into Human Faculty and its Development* (1883)
  - v podstatě počátky vědeckého zkoumání individuálních rozdílů a mentálních testů
- asi 1886 – antropometrická laboratoř
  - měření fyzických a sensomotorických charakteristik
- pojem statistické korelace
  - matematická formulace později – Karl Pearson (1857-1936)
- **význam** – počátek používání statistických metod jako hlavního matematického nástroje pro vědecké zkoumání individuálních rozdílů

James McKeen Cattell

- **doktorát** – u W. Wundta v Lipsku
  - propojení experimentální metodologie s americkým pragmatismem
  - rigorózní pozorování chování, ale nesouhlas s klasickou představou o nutnosti zobecnění týkajících se normální lidské dospělé mysli
  - doktorská disertace zaměřena na studium individuálních rozdílů v reakčním čase (přes Wundtovy námitky)
- **jeho názor** – odchylky nezmizí ani při použití nejrigoróznější experimentální kontroly
  - odchylky tehdy považovány za chyby
  - spíše je považoval za projevy vlivu reálných proměnných
- **kontakty s Galtonem**
  - později založil psychologické laboratoře na Pennsylvánské a Kolumbijské universitě
  - vývoj a úpravy psychologických testů
- **tehdy** – dominantní názory empirismu (John Locke) a asocianismu
  - komplexní projevy psychiky chápány jako kombinace jednodušších sensorických zkušeností
- **rané pokusy o měření**
  - sensorické schopnosti a jiné jednodušší funkce (míry reakčního času, rozlišování vizuálních a auditivních podnětů, citlivost k bolesti)
- **výsledky**
  - žádný významný vztah mezi výsledky takových testů u studentů 1. ročníku Kolumbijské university a jejich známkami

Alfred Binet

- **počátky** – experimentální psychologie ve francouzské tradici
  - zájem byl o jedince odlišné od normy
  - výzkum retardovaných dětí
- **1896** – série testů „komplexnějších“ psychických procesů
  - na základě pečlivého studia rozdílů v chování retardovaných, normálních a nadaných dětí (společně s kolegou Victorem Henrim)
- **dále** – mnoho experimentálních výzkumů
  - rozdíly výkonu dětí vyššího věku nebo z vyšších školních tříd a výkonu mladších dětí
- **důsledek** – přímý přístup k měření obecné schopnosti shledán užitečnějším než měření jednodušších funkcí
- **1904** – Binet členem komise zaměřené na studium výuky retardovaných dětí
- s Théodorem Simonem – první formální škála pro přímé měření komplexnějších schopností

Shrnutí raného vývoje

- základní matematické úpravy měření od Galtona a jeho následovníků v Anglii
- důležitost přísné kontroly pozorování od Wundta a jeho spolupracovníků z Německa
- individuální rozdíly jako legitimní téma psychologického výzkumu Cattell v USA
- Binet ve Francii nalezl obsah testů nutný k tomu, aby z nich byly užitečné nástroje

Pozdější vývoj

- významné metody
  - Woodworth Personal Data Sheet (1919)
    - v podstatě typ standardizovaného psychiatrického interview
  - Rorschachův test (1920)
  - TAT (1935)
  - MMPI (1943)
- **od 20. let**
  - měření zájmů, postojů a hodnot – např. Kurt Lewin (experimentální výzkum konfliktu, frustrace a aspirační úrovně)
- **20. a 30. léta**
  - mnoho jednoskórových i víceskórových nástrojů diagnostiky osobnosti (některé zaměřeny na psychopatologii)
- **40 a 50. léta**
  - růst aplikovaných oblastí psychologie spolu s vývojem metod konstrukce, skórování s interpretace psychometrických nástrojů
- **50. a 60. léta**
  - např. výzkum autoritářské osobnosti, závislosti/nezávislosti na poli, výkonové motivace, represe-senzitizace, místa kontroly
- <u>významným faktorem vývoje se stala výpočetní technika</u>

Další důležité postavy a instituce

- **VB** – Karl Pearson, Charles Spearman, Cyril Burt, Geoffrey Thompson
- **USA** – T. L. Kelley, L. L. Thurston, J. P. Guilford a jejich žáci
- **Skandinávie** – Rasch
- **Nizozemsko**
- **Německo**
- **Odborné společnosti** – Society for Personality Assessment, International Society for Study of Individual Differences,
- **Časopis** – Personality and Individual Differences

Cíle psychodiagnostiky a psychometriky

- pokus u vědecké řešení problému předpovídání budoucnosti
  - největší zisk (aspoň pro klienta) – predikce na základě toho, co se stalo v minulosti za podobných podmínek (současně jeden ze základních předpokladů vědy)

- 3 základní aplikace psychologického měření: prognóza, diagnóza a výzkum

Použití metod

- **prognóza** (také predikce)
  - primární důraz na rozdíly mezi výkony jedinců nebo mezi výkony jedince a nějakým standardem (normou)
  - není nutné najít příčiny úspěchu, stačí empiricky zjištěný vztah mezi nějakým pozorovatelným chováním a nějakou mírou pozdějšího úspěchu (může být ateoretická)
  - **příklady**
    - přijímací zkoušky: inteligenční testy se dětem předkládají proto, aby predikovaly jejich akademický výkon
    - konkurz: ti, kteří byli vybráni, uspějí, a ti, kteří vybráni nebyli, by neuspěli nebo by neměli takové výsledky jako ti, kteří vybráni byli
    - známkování: známka je chápána jako predikce kvality výkonu jedince v nějaké budoucí situaci
- **diagnóza**
  - větší pozornost věnována analýze a popisu různých charakteristik u jediné osoby
  - primární cíl je najít charakteristiky odpovědné za diagnostikované obtíže
  - snaha o zlepšení situace díky odhalení příčiny obtíží
- **výzkum**
  - **nutnost** – testy nejsou zcela validními mírami mentálních charakteristik

Normativní a ipsativní měření

- podle Cattella
- **normativní měření**
  - výsledky testu interpretovány z hlediska výkonů jiných osob ve stejné situaci
  - také *měření vztahující se k normám*
- **ipsativní měření**
  - interpretace výkonu v testové situaci v kontextu výkonů v jiných testových situacích

- Příklad: Neúspěch žáka, který měl ve srovnání s ostatními výborné výsledky v přijímacích testech, ve studiu.
- možnosti:
  - snaha porovnat další schopnosti žáka s výkonem jiných žáků (normativní)
  - snahou posoudit úkoly, při jejichž plnění je student úspěšný, s úkoly, při jejichž plnění je nespěšný (ipsativní)

Klasifikace psychodiagnostických metod

- existují stovky psychodiagnostických metod (viz např. stránky Buros Institute – www.unl.edu/buros)
- nezbytné je použití mnoha různých kritérií klasifikace:
- **Hlavní:**
  - I. Podle typu měřeného rysu
  - II. Podle způsobu měření
  - III. Podle účelu

I. Klasifikace podle typu rysu

(podle Cronbacha, 1960)

- míry typického chování
- ***typické chování***
  - postoje, zájmy a osobnostní charakteristiky
  - neexistují zde správné ani špatné odpovědi
  - předpokladem je pravdomluvnost proband

- míry maximálního výkonu
- ***maximální výkon***
  - schopnosti, znalosti, dovednosti
  - odpovědi porovnány s klíčem rozlišujícím správné a špatné odpovědi
  - předpoklad, že testovaná osoba ze sebe vydá to nejlepší

Nutno uvážit:

u testů maximálního výkonu – vrozené schopnosti, míra vlivu prostředí nebo kombinace obojího?

obecné problémy klasifikace – mnoho testů obtížně zařaditelných

II. Klasifikace podle způsobu měření (více klasifikací)

- **individuální** × **skupinové** – administrace testu pouze jedné osobě nebo skupině osob
  - některé metody nelze administrovat skupinově
- **objektivní** × **subjektivní** – způsob skórování metod (spojitá dimenze)
  - *objektivní* – každá osoba skórující test konkrétního jedince dospěje ke stejnému skóru
  - *subjektivní* – hodnocení je do jisté míry ponecháno na osobě, která provádí skórování
- **test typu papír-tužka** × **performanční test** – psaní nebo manipulace
- **verbální** × **neverbální** – použití nebo nepoužití verbálních podnětů
  - *neverbální* - nepoužívají se žádná psaná nebo mluvená slova ani jako instrukce ani jako samotné položky
  - instrukce pomocí demonstrací a pantomimy
  - odpovědi často pomocí nějaké manipulace s aparátem
  - použití u negramotných, u lidí mluvících cizím jazykem
- **silové** × **rychlostní** – rozdíl spočívá v čase a jeho vlivu na výsledek (spojitá dimenze)
  - *čistě rychlostní test* – složen z položek tak snadných, že každý je schopen zodpovědět je správně
  - získaný skór je zcela závislý na tom, kolik položek osoba vyřeší v povoleném čase
  - *čistě silový test* – každá osoba může zkusit řešit kteroukoli položku a její skór závisí zcela na tom, kolik otázek zodpoví bez ohledu na tempo, jakým pracuje

III. Klasifikace podle účelu

- testy pro ***umístění*** osob, ***klasifikaci*** nebo ***výběr***

- Další možné klasifikace
  - např.: konvenční, situační nebo projektivní
  - klinické, testové nebo přístrojové

Klinický a psychometrický přístup

- hlavní rozdíl – *prognóza o jedinci* × *prognóza o skupině osob*
- Meehl, P. E. (1954)
  - diagnostika je velmi podobná samotné vědecké práci: čas od času konfrontace s velkým množstvím dat, pokus o odpověď na otázku, co vysvětluje tato pozorování, toto vysvětlení téměř vždy naznačuje další skutečnosti, které by mělo být možné pozorovat v budoucnu – teprve pak se uchylujeme k formální logice a statistice a tyto implikace zobecňujeme a testujeme
- Meehlova studie – porovnání klinické a psychometrické predikce chování
  - 19 studií s víceznačnými výsledky
  - 10 studií nenašlo rozdíly
  - 9 studií našlo rozdíly ve prospěch psychometrických metod
  - ani jedna neprokázala lepší výsledky klinické metody
- nevyhnutelný závěr
  - při současném stavu poznání je při dostatku dat psychometrický přístup úspěšnější
  - klinici umí zřejmě lépe terapii než predikci a diagnózu

Klinický a psychometrický přístup

- **Klinická logika**

- Extenzívní informace o jediné osobě
- Lze použít libovolná data
- Predikce založená na znalostech teorie chování
- Predikci jako kreativní akt musí provést vysoce zkušená osoba
- Lze použít i nahodilé jevy

- **Psychometrická logika**

- Intenzívní zkoumání několika rysů
- Všechny informace v podobě členství ve skupinách
- Predikce založená na počtu pravděpodobnosti
- Predikce jako formální důsledek pozorování může provést i úředník (laborant)
- Nelze využít řídké (nahodilé) jevy

Měření

- **Definice:** měření v nejširším slova smyslu je proces přiřazování čísel objektům podle zavedených pravidel
  - **pravidla** – závisí na typu srovnávání, které je možné provádět s měřenými objekty; manipulace s čísly musí být možné provádět analogicky také s objekty samotnými
- **4 typy škál** (pravidel):
  - nominální – přiřazení číselných označení jednotlivým objektům nebo třídám
    - pravidlo: všichni jedinci ze stejné skupiny mají přiřazeno stejné číslo; žádní dva jedinci z různých skupin nemají přiřazeno stejné číslo
    - např.: přiřazení libovolných čísel různým zaměstnaneckým skupinám nebo diagnózám
    - matematická operace: počítání velikostí jednotlivých tříd (absolutní a relativní četnost); relace „=“ a „≠“
  - ordinální (pořadová) – seřazení objektů podle nějakého kritéria
    - pravidlo: + tranzitivita relací „větší než“ a „menší než“
    - např.: pořadí v závodě
    - matematické operace: + porovnávání skórů navzájem (<, ≤, ≥, >]
  - intervalová – stanoveny stejné jednotky po celé délce škály
    - pravidlo: + aditivita škály
    - např.: měření teploty
    - matematické operace: + sečítání a odčítání
  - poměrová – objektům přiřazovány hodnoty reálných čísel
    - pravidlo: + poloha absolutní nuly
    - např.: hmotnost, výška
    - matematické operace: + násobení a dělení (kromě operací výlučných pro imaginární čísla)

Příklad

- normální měření výšky pomocí metru
- měření výšky pomocí metru od vrchu stolu
- měření výšky od vrchu stolu pomocí knih naskládaných na sebe
- rozdělení osob na skupiny vyšší a nižší než nějaký standard

Problémy s měřením v psychologii

- většina měření v psychologii je ordinální
- 3 možná řešení:

- **1**. omezení se na použití statistických procedur speciálně vytvořených pro pořadová data (medián; Kendallovo tau apod.)
- **2**. předpoklad (na základě empirických důkazů nebo logických dedukcí) normálního rozložení měřeného hypotetického rysu v populaci
- **3**. nejběžnější je uznání faktu, že míry behaviorálních charakteristik jsou relativní a nikoli absolutní kvantity
  - „hrubé“ skóry pak nejsou nikdy interpretovány přímo (transformace na nějaký „vážený“ skór, jehož podstatou je porovnání výkonu jedince s výkonem celé skupiny)

Otázky povahy měřených rysů

- definice rysu: v teoretických pojmech vyžadujících vysvětlení *nebo* pomocí jistých pozorování naznačujících jeho existenci
- psychický rys: přijatelný (pojmový) nástroj pro vymezení určité množiny pozorování *nebo* teoretický pojem
  - předpoklad, že se projevuje v pozorovatelném lidském chování
- vlastnosti
  - **1**. pojem vysvětluje současná pozorování (tzn. známá fakta a empirické zákony)
  - **2**. také implikuje nová pozorování
- konstrukt: pojem splňující požadavky vysvětlování současných znalostí a poskytování implikací o nových pozorováních; proto má každý rys připojeny jisté vlastnosti
- např.: osoba s danou úrovní této vlastnosti bude s jistou pravděpodobností chovat způsobem X, pokud se dostane do situace Y
- znalosti o psychickém rysu
  - **1**. co test měří a jaké jsou jeho vztahy k ostatním rysům
  - **2**. co platí o lidech, kteří dosáhnou určitého testového skóru

Otázky týkající se rysů

- možné otázky:
  - Je rys u jednotlivce stabilní nebo fluktuuje v čase?
  - Jaké je rozložení rysu v lidské populaci?
  - Do jaké míry je rys ovlivněn dědičností a do jaké prostředím?
  - Jaké jsou zákony ovládající vývoj rysu, jak osoba dospívá?

- současný pohled: hypotetický konstrukt nemůže být správný nebo špatný, ale pouze užitečný nebo neužitečný z hlediska vysvětlování současných znalostí a naznačování nových vztahů k ověřování
- nemožné: mít zcela uspokojivou a univerzální definici psychického rysu, protože by byla nutná kompletní data o lidech a jejich chování a jejich dokonalá znalost
- hodnocení testu: nutné vědět něco o testu samotném a o autorově pojetí rysu

Požadavky na měření

- ***přesnost*** – vzhledem k účelu, pro které budou naměřené hodnoty použity
- ***opakovatelnost*** – stejná hodnota při opakovaném měření stejného objektu
- ***objektivita*** – různí lidé provádějící měření by měli naměřit stejnou hodnotu
- ***adekvátnost*** – měřící nástroj, jednotky měření, náročnost na vybavení, složitost procedury, vzhledem k účelu atd.

Přímost a nepřímost měření

- **přímé měření** – veličinu měřenou pro nějaký objekt měříme pomocí stejné veličiny, která je součástí (vlastností) nástroje měření
  - délku (výšku, hloubku) měříme porovnáním s délkou měřidla (pravítko, pásmová míra)
  - hmotnost objektu měříme porovnáním s hmotností závaží (např. laboratorní váhy)
- **nepřímé měření** – veličinu měříme porovnáním s jinou veličinou, která má pro daný měřící nástroj matematicky definovaný vztah s veličinou měřenou
  - hmotnost měříme prostřednictvím délky protažení nějaké pružiny (mincíř) nebo prohnutí vhodného média (osobní váhy)
  - teplotu měříme prostřednictvím délky rtuťového nebo lihového sloupce (teploměr) nebo ohnutí kovového pásku (bimetalové přístroje)

Měření v psychologii

- téměř výhradně nepřímé!

Důsledek

- Nutnost kritérií: tvůrci i uživatelé měřících nástrojů v psychologii musí být schopni posoudit, jestli jsou tyto nástroje kvalitní

Klasifikace chyb měření

- **proměnné (náhodné) chyby**: při opakovaném měření téže veličiny (v psychologii v ještě větší míře)
  - relativní nepřítomnost chyb u psychodiagnostické metody (reliabilita)
- **osobní chyby**: zdroje neobjektivity, v podstatě osobní rovnice diagnostika
  - 2 osoby po přečtení téže zprávy o vyšetření učiní 2 různé závěry
- **konstantní chyby**: plyne z nepřímosti většiny měření v psychologii
  - např. u většiny testů je nutné čtení a bylo by chybou, pokud by rozdíly ve výkonu v testu souvisely více se čtením než se zjišťovanou schopností
- **chyby interpretace**: skór jedince nemá smysl bez kontextu výkonu ostatních osob z relevantní populace (nutnost definovat skupinu a pravidla srovnávání)

Eliminace chyb měření

- **standardizace** – formální podoba psychodiagnostických metod (grafická úprava, postup testování, pravidla interpretace výsledků)
- **normalizace** (součást standardizace) – test je předložen dobře vymezené skupině a její výkon je pečlivě zaznamenán
  - porovnání výkonu jedince s výkonem skupiny pomocí tzv. vážených skórů

- typy norem – podle povolání, regionu, národnosti, pohlaví, věku, školní třídy atp.
- 2 druhy skupin:
- 1. skupina, do které proband patří
- 2. skupina, do které proband aspiruje

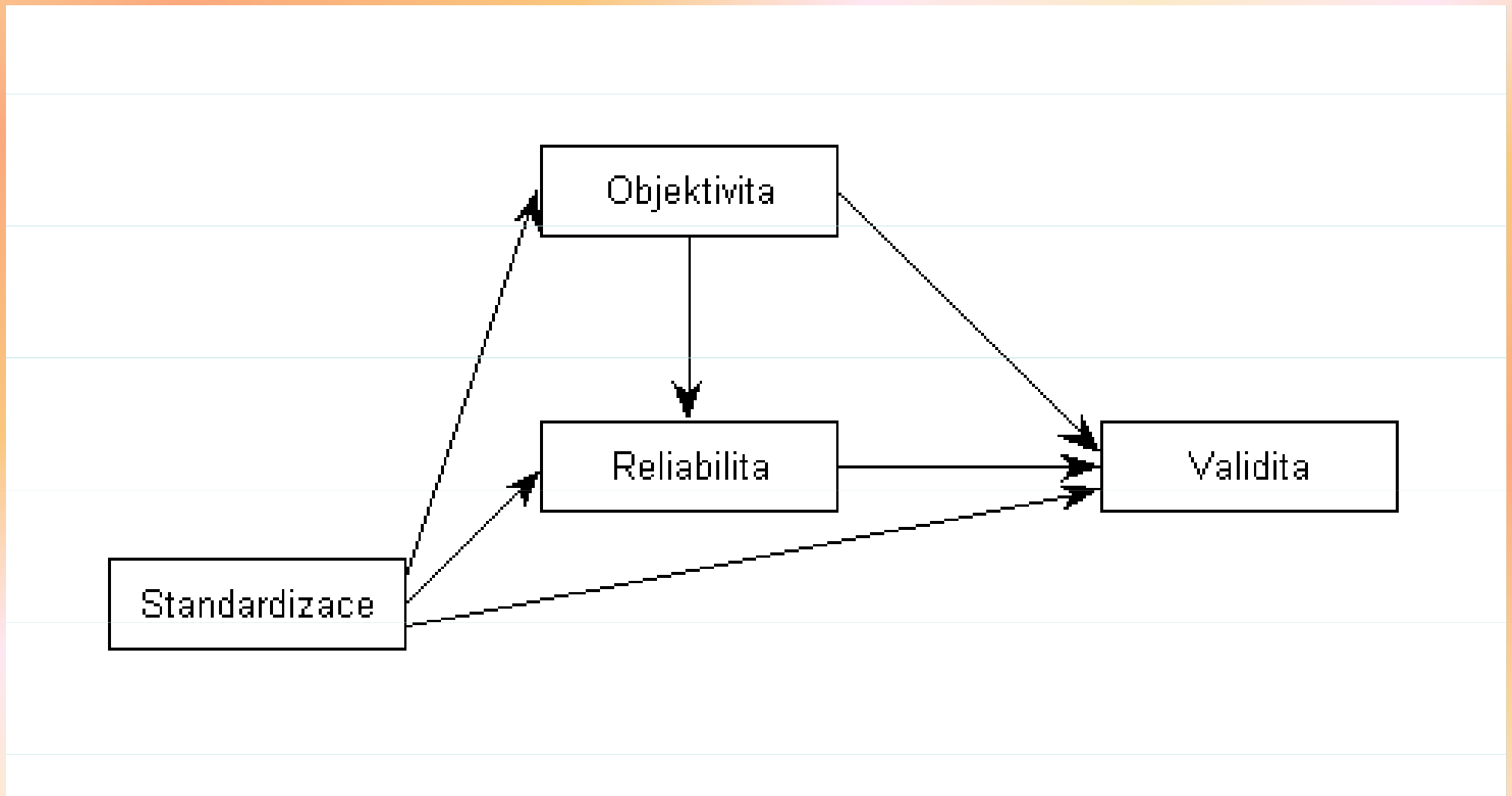
Standardy

- **Charakteristiky metod a kritéria pro jejich posouzení**
- charakteristiky testů slouží současně jako kritéria pro jejich posouzení; mají těsnou souvislost se zdroji chyb

- nutnost: snaha o nepřítomnost *všech* těchto chyb

Charakteristika	**Souvislost se zdrojem chyb**
Normalizace	snaha vyhnout se interpretačním chybám
Reliabilita	relativní nepřítomnost proměnných chyb
Standardizace Objektivita	eliminace osobních chyb
Validita	odstranění konstantních chyb

Vztahy charakteristik



Upozornění

- pokud je test standardizovaný, objektivní a reliabilní, ale není validní, je k ničemu, protože nelze určit, co měří!
- posouzení samotné validity nestačí pro posouzení ostatních charakteristik

Formální podoba diagnostické metody

- testový sešit a záznamový arch; různé pomůcky pro testování (kostky, seznamy, ukázky)

- **Instrukce** – pokyny pro probandy, jak mají postupovat v psychodiagnostické situaci; návod pro zaznamenávání odpovědí
- **Položky** – základní jednotka dotazníkových a testových metod; otázka nebo úkol, který je třeba vyřešit

- **Typy položek**:

- Otázky a úkoly
- - grafické
- - verbální
- Odpovědi
- - otevřené
- - uzavřené - dichotomické
- - polytomické (multiple-choice)
- - škály
- Formát odpovídání
- - souhlas-nesouhlas
- - vůbec-zcela
- - počet bodů škály

Náhodné chyby a pravé skóry

- rozdíl od fyzikálního měření – proměnné chyby jsou mnohem větší a vlivnější, proto je otázka, jaká je skutečná hodnota měřeného rysu
  - možné řešení: průměr z opakovaných měření jako rozumný odhad
- 2 komponenty pozorované hodnoty (X):
  1. skutečná hodnota (tzv. pravý skór) (τ):
  2. chyba měření (e):

$$X_i = \tau_i + e_i$$

- Příklad: osoba zná správnou odpověď na 43 položek, uhádne odpověď na 4 další, ale splete se v místě, kam má zaznamenat odpověď, což znamená, že chyba je 3 = 4 – 1

Reliabilita

- Základní úvaha – při měření mnoha stolů stejné délky je veškerá variabilita naměřených hodnot způsobena chybou
  - tzn. rozptyl měření je odhad míry chyby
- u stolů různých délek – variabilita je způsobena zčásti vlivem skutečných rozdílů a zčásti vlivem chyby, tzn. celkový rozptyl se skládá z rozptylu pravých skórů a rozptylu chyb

Teoretická definice reliability

- **Předpoklad** – nekorelovanost pravých skórů a chyb měření

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

- **RELIABILITA** – podíl variability pravých skórů k celkové variabilitě

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

Odhad reliability

- protože pravé skóry nelze v principu zjistit, reliabilitu je možné pouze odhadovat
- **Některé přístupy:**
  1. test-retest
  2. paralelní formy
  3. split-half
  4. vnitřní konsistence
  5. Kuder-Richardson
- Rozdíl – v pojetí náhodných chyb měření

Testová-retestová reliabilita

- **náhodná chyba**: cokoli, co mění skór proměnné měřené dvakrát v různém čase, tzn. reliabilita ve smyslu stability
- **postup**: test se dá stejné skupině osob dvakrát po sobě (x a x')
- **předpoklad** – měřený rys se v čase nemění, rozptyl chyb v obou případech stejný
- **odhad reliability**: korelace výsledků 1. a 2. měření
- **otázka**: je předpoklad, že se rys nemění v čase, realistický?
  - při dlouhém odstupu v testování se úroveň rysu může změnit, což vede k podhodnocení odhadu reliability (nutné uvážit hlavně u dětí)
  - po příliš krátké době může dojít ke klamnému nadhodnocení vlivem zapamatování si odpovědí (doporučují se 3 měsíce)
- **závěr**: test-retestová reliabilita je buď podhodnocena nebo nadhodnocena

Reliabilita paralelních forem

- **náhodná chyba**: jakýkoli rozdíl ve skórech v obou formách
  - obvykle se obě formy administrují bezprostředně po sobě, aby byl vyloučen vliv nestability
  - reliabilita ve smyslu ekvivalence
- **cíl**: překonání vlivu zapamatování u testové-retestové reliability
- **problém**: jak zajistit skutečnou paralelnost obou forem
- **striktní paralelnost**: právě tehdy, když se položky vybírají ze stejného obsahového universa položek stejným způsobem
  - posouzení je možné často pouze na základě expertního posudku
  - několik variant ekvivalence
- **empirické posouzení**: pomocí statistické analýzy, tzn. stejné průměry a rozptyly položek a kovariance mezi položkami v obou formách
- **odhad reliability**: korelace výsledků obou forem

Split-half reliabilita

- **problémy s předchozími pojetími**: možné reálné fluktuace rysu se považují za chyby
- **paralelní formy**: snaha o řešení administrací obou testů současně
- **alternativa**: administrace testu jako celku a jeho následné rozdělení na dvě poloviny
- **problém**: zajištění skutečné stejnosti polovin
- **rigorózní přístup**: analýza položek, hledání párů položek s analogickým obsahem, obtížností, diskriminační účinností
- **častý praktický postup**: sudé a liché položky zvlášť
- **odhad reliability**: korelace mezi oběma polovinami

$$r_{xx'} = 2\left(1 - \frac{\sigma^2_{x_a} + \sigma^2_{x_b}}{\sigma^2_x}\right)$$

Split-half reliabilita

- **problém**: reliabilita je funkcí délky testu (počtu položek) – roste s počtem položek
  - **proto** – nutná korekce získaného odhadu reliability pro celý test

- Spearmanův-Brownův vzorec:

$$r_{xx'} = \frac{mr'_{xx'}}{1 + (m-1)r'_{xx'}}$$

- **předpoklad**: položky měří stejný rys; rozptyly obou polovin jsou stejné (lze zajistit pečlivým výběrem do obou polovin)

Split-half reliabilita

- **Guttmanův vzorec**: nevyžadující rovnost rozptylů (Guttman, 1945):

$$r_{xx'} = 2\left(1 - \frac{\sigma^2_{x_a} + \sigma^2_{x_b}}{\sigma^2_x}\right)$$

- **pojetí chybového rozptylu**: rozdíl mezi rozptylem celkového skóru odhadnutého ze dvou polovin a ze všech položek společně (na základě možnosti přepsat vzorec tak, aby vyjadřoval původní definici reliability)

Reliabilita jako vnitřní konsistence

- **požadavek**: pokud jedna položka testu měří určitou proměnnou, pak další položky musí měřit tutéž proměnnou
  - projeví se vztahy mezi položkami
- **paralela**: v podstatě zobecnění split-half reliability na *n* testů délky 1
- ***nesouhlas***: Cattell tvrdil, že čím specifičtější položky, tím nižší je validita
- **ale**: žádný konstruktér testu nebyl schopen konstruovat test s položkami korelujícími s kritériem, ale ne mezi sebou

Odhad vnitřní konsistence

- **Cronbachův koeficient alfa**

$$r_{kk} = \frac{k}{k-1}\left(1 - \frac{\sum_i \sigma_i^2}{\sigma_t^2}\right)$$

- k – počet položek; σ_i^2 – rozptyl i-té položky; σ_t^2 – rozptyl celého testu

Kuder-Richardsonova reliabilita

- **chyby**: nekonsistence ve výkonech v jednotlivých položkách
  - reliabilita jako míra homogenity
- **ideální případ**: seřazení položek podle vzrůstající obtížnosti
  - každá osoba dosáhne bodu, před kterým vyřešila všechny položky a za kterým nevyřešila žádnou
  - pokud to platí pro každou osobu, pak se test vyznačuje dokonalou reliabilitou podle Kudera a Richardsona (1937)
- **chyba**: vzniká nesprávným seřazením položek podle obtížnosti
  - rozumný požadavek: pokud je test čistá míra jediného rysu, tzv. homogenní test

Vzorec K-R 20

$$r_{xx'} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_x^2 - \sum_{j=1}^{n} p_j q_j}{\sigma_x^2}\right)$$

- n – počet položek testu; σ_x^2 – rozptyl naměřených testových skórů; p_j – podíl osob, které mají j-tou položku správně; $q_j = 1 - p_j$, $p_j q_j$ – rozptyl dichotomické položky

Reliabilita podle Hoyta

- **odlišné pojetí chyby** (Hoyt, 1941): kolísání výkonu osoby od položky k položce se nepovažuje za chybu, ale za reálné intraindividuální rozdíly
  - neměly by být součástí odhadu reliability

- 3 komponenty rozptylu:
  1. skutečné *inter*individuální rozdíly
  2. *intra*individuální rozdíly (měřené rozptylem položky)
  3. chybové interindividuální rozdíly

Analýza rozptylu

- Model rozptylu:

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

- Odhad reliability:

$$r_{xx'} = \frac{\sigma_t^2}{\sigma_x^2 - \sigma_i^2} = \frac{\left(\sigma_x^2 - \sigma_i^2\right) - \sigma_e^2}{\sigma_x^2 - \sigma_i^2}$$

- Nebo – pomocí sum čtverců:

$$r_{xx'} = \frac{MS_o - MS_r}{MS_o}$$

Faktory ovlivňující odhad reliability

1. Metoda odhadu reliability
2. Délka testu
3. Heterogenita skupiny
4. Časové aspekty
5. Konstrukce a administrace

Metoda odhadu reliability

- **paralelní formy** – nižší odhady než test-retest
  - považovány za dolní hranici odhadu, preferují se
    - paralelní formy obsahují fluktuace forem i času
    - test-retest obsahuje klamnou reliabilitu způsobenou zapamatováním si odpovědí
- **split-half** – považována za horní hranici
  - jediné měření vede k nadhodnocení reliability u testů se silnou rychlostní komponentou
  - vysoká reliabilita může být iluzí způsobenou přísným časovým limitem a postupem split-half
- **Kuder-Richardson** – podhodnocuje v případě, že test není homogenní, a nadhodnocuje, pokud homogenní je

Délka testu

- **myšlenkový postup**:
  - Jakákoli sada položek testu představuje výběr ze všech možných položek daného obsahového univerza.
  - Kdyby se toto univerzum použilo celé, test by měl nekonečnou délku a skór osoby by se rovnal pravému skóru.
  - Protože se jedná o výběr, naměřený skór je pouze odhadem pravého skóru.
  - Stejně jako ve statistice platí, že čím rozsáhlejší je výběr, tím přesnější je odhad.

Heterogenita skupiny

- **homogenní skupiny** – klesá podíl variability pravých skórů a tím i reliabilita
  - chybová variabilita stále stejná
- **často**: různé koeficienty reliability pro různé skupiny
- **řešení**: index, který je funkcí reliability, ale nezávislý na variabilitě skupiny:

Standardní chyba měření

- **Standardní chyba měření**: odhad směrodatné odchylky chyb při opakovaném výběru
- **teoreticky**: standardní chyba měření je průměr směrodatných odchylek rozložení při opakovaných výběrech
- **Vzorec**:

$$s_e = s_x \sqrt{1 - r_{xx'}}$$

- **problém**: index v jednotkách testového skóru, takže ho není možné porovnávat u různých testů

Časové aspekty

- u *testů s vlivem času* je reliabilita ve smyslu konsistence nadhodnocena
- **v praxi**: většina testů má složku rychlosti
  - obvykle je snaha, aby dokončilo pouze 75-90% osob
    - proto nutnost: paralelní formy nebo test-retest
    - nejjednodušší postup: rozdělení testu na 2 poloviny *před* administrací (každá část odděleně časovaná a skórovaná) a použití Guttmanova nebo Spearman-Brownova vzorce
- *i přesto mají rychlostní testy obvykle vyšší reliabilitu než silové, a to možná proto, že je větší konsistence v tempu práce než v kvalitě výkonu*
- **nejjednodušší míra „rychlostnosti“**:

$$s_s = \frac{s_r^2}{s_c^2}$$

- $\mathbf{s_r^2}$ – rozptyl skórů při počítání všech řešených položek, bez ohledu na jejich správnost, $\mathbf{s_c^2}$ – rozptyl skórů pouze pro správně řešené položky
  - **čistě silový test**: $s_r^2 = 0$
  - **čistě rychlostní test**: všechny řešené položky jsou správně $s_r^2 = s_c^2$

Konstrukce a administrace

1. **vlastnosti testu samého**
   - *uspořádání položek*: lepší je dát jednodušší položky na začátek
   - *vzájemně závislé nebo téměř identické položky* snižují reliabilitu (test je tím vlastně zkrácen)
   - *„chytáky“ a emocionálně zabarvené položky* snižují reliabilitu (zvyšuje se vliv náhody)
2. **forma položky**
   - možnost uhádnutí odpovědi zvyšuje chybu (vhodnější jsou mnohonásobné volby)
   - *příliš jednoduché nebo příliš složité* položky nediskriminují (opět de facto zkrácení testu – kde není variabilita, nemůže být ani reliabilita)
3. **faktory u vyšetřované osoby**
   - *nepochopení celé situace* testování (běžné v počátcích testování)
   - *obecný postoj* (např. tendence odpovídat kladně nebo záporně, tendence hádat, snaha zkreslit výsledky)
4. **fyzický stav zkoumané osoby**
   - zjištěno, že nevolnost nemá vliv na výkon, ale snižuje se subjektivní hodnocení výkonu (pochopitelně nesmí dojít k opravdu vážným potížím)
5. **faktory administrátora**
   - vztah s testovanou osobou
   - přesvědčení o důležitosti
   - chybné nebo nedostatečné instrukce
   - chybně přečtený čas (nebo přílišná benevolence vzhledem k času)

Minimální žádoucí reliabilita metody

- **Přísné požadavky**:
  - aspoň **0,50** pro hodnocení výkonu skupiny
  - aspoň **0,90** pro hodnocení rozdílů v úrovni výkonu skupiny u dvou a více výkonů
  - aspoň **0,94** pro hodnocení úrovně individuálního výkonu
  - aspoň **0,98** pro hodnocení rozdílů v úrovních individuálních výkonů ve dvou a více výkonech
- **Minimum**:
  - aspoň **0,70**

Vztah reliability a validity

- **platí**: maximální možná validita se rovná odmocnině z reliability
- Korekce nereliability pro korelaci 2 testů:

$$r_{XY} = \frac{r_{xy}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}}$$

- **platí**: maximální pozorovatelná korelace testu a kritéria je rovna součinu reliabilit obou měření

$$r_{xy(max)} = \sqrt{r_{xx'} r_{yy'}}$$

Závěrečné poznámky

- míry typického chování mají nižší reliabilitu než míry maximálního výkonu
- **soubor**: test pro klinické skupiny musí být testován u těchto skupin
  - např. schizofreniky je velmi obtížné testovat, proto zde bývá nízká reliabilita
- **Závěr**: koeficient reliability je třeba interpretovat z hlediska použitého postupu a homogenity zkoumané skupiny a ve světle účelu použitého testu, měřeného rysu a reliability podobně zaměřených testů

Validita

- **Otázka** – měří test skutečně to, co měřit má?
  - (měření v psychologii je nepřímé)
- Způsoby zjišťování – mnoho – 3 základní přístupy zaměřené na 3 hlavní komponenty testové situace
- **Základní typy validity**
  - *obsahová validita* – obsah testové situace a chování osoby v testové situaci
  - *empirická validita* – posouzení vztahu mezi testovým skórem a vnějším kritériem
  - *konstruktová validita* – teoretické otázky existence nějakého rysu
- **V současnosti**
  - neuvažuje se o typech validity, ale o typech důkazů o validitě

Obsahové důkazy o validitě

- Zjevná validita
- Výběrová validita
- Faktorová validita

Zjevná (zdánlivá, face) validita

- u zdánlivě validního testu se zdá, že měří to, o čem se tvrdí (předpokládá), že měří
- ale neexistuje logický vztah mezi zdánlivou a skutečnou validitou
- při psaní položek často jediné vodítko, ustavení vztahu probanda k testové situaci
- může zvýšit motivaci osob
- není žádoucí vlastností dotazníků, pravděpodobně vede ke zkreslení, zvláště při výběrech osob (konkursy apod.)
- u testů schopností problémy nepůsobí, ale zjevně validní test úzkosti by např. při výběru civilních pilotů nikdy nefungoval

- Příklad: posuzování rozměrů různých předmětů denní potřeby je zdánlivě validní např. jako test praktických znalostí, ale ve spojení s časovým limitem může být testem úzkosti

Výběrová (sample) validita

- vlastní obsahová validita, validita na základě definice, posouzení výběru prvků obsahového univerza, souvislost se zjevnou validitou
- **použití** – výkonové testy a testy vědomostí a schopností
- **postup** – expertní posouzení
  1. pečlivá definice rysu nebo obsahové oblasti v pojmech chování
  2. rozdělení celé oblasti do kategorií představující hlavní aspekty oblasti
  3. posouzení počtu položek v každé kategorii
- Příklad: obsahová validity hudebního testu pro studenty by se zkoumala prostřednictvím posouzení testu skupinou hudebníků nebo hudebních pedagogů; jejich vyjádření, zda test pokrývá všechny důležité aspekty hudební schopnosti očekávatelné u studentů s určitou zkušeností

Faktorová validita

- obsah testu na základě faktorové analýzy různých obsahů
  - někdy je faktorová validita považována za součást empirické, ale ve skutečnosti se nehodnotí kritérium
- Příklad:
  - původně 7 faktorů inteligence podle Thurstonea (porozumění slovům, slovní fluence, čísla, prostorové vztahy, asociační paměť, percepční rychlost, obecné usuzování)
  - později FA odhalovala další a další faktory
  - Guilford na základě svých výzkumů přes 40 faktorů intelektu, po roce 1956 systematizace:
  - 120 faktorů
    - **I. operace**: kognice, paměť, produktivní myšlení (konvergentní a divergentní myšlení), hodnocení
    - **II. obsahy**: sémantický, symbolický, figurální, behaviorální
    - **III. produkty**: jednotky, třídy, vztahy, systémy, transformace, implikace

Empirické důkazy o validitě

- prokázání vztahu mezi výsledkem testu a nějakým kritériem
- **Důkaz**
  - Pearsonova nebo jiná korelace
  - rozdíl v definovaných skupinách, procenta správně klasifikovaných osob vzhledem k jasně definovaným skupinám
- *problém* – nalezení vhodného kritéria – kritérium by mělo být vysoce reliabilní a validní, u některých ukazatelů vhodné kritérium neexistuje
- *omezení* – omezení rozsahu hodnot v důsledku předvýběru osob
- Příklad: u hudebníků neexistuje vztah mezi hudebním sluchem a libovolnou další proměnnou, protože hudební sluch mají všichni, ve výzkumu není velice důležitá skupina (ti, kteří neuspěli nebo neprojevili zájem), což omezuje vztah mezi proměnnými

Empirické důkazy o validitě

- Prediktivní validita
- Souběžná validita
- Inkrementální validita
- Diferenciální validita

Prediktivní validita

- test vyznačující se prediktivní validitou bude predikovat nějaké kritérium
- Příklad: vztah inteligence a akademického úspěchu
  - korelace skórů inteligenčního testu dětí v 5 letech s jejich následnými akademickými úspěchy
  - problémem je index akademického úspěchu (problém srovnatelnosti známek z různých předmětů, škol atd.)
  - očekávaná korelace IQ a kritéria významná, ale nepříliš těsná (0,3-0,4)

Souběžná (concurrent) validita

- souběžně validní testy navzájem vysoce korelují
  - – způsob, jak nahradit prediktivní validitu

- Příklad: korelace mezi současně administrovanými testy neverbální inteligence

Inkrementální validita

- definice – inkrementální validitu má test, který přináší informaci, kterou nelze získat z jiných zdrojů
  - vždy vzhledem k nějakému účelu
  - nulová korelace reliabilního testu s některým testem použitým v baterii může znamenat informaci
- nutné – při výběrech
- kritérium – výsledky mnohonásobné regrese
- položky – měly by korelovat s kritériem, ale nulově mezi sebou
  - tzn. měly by mít inkrementální validitu
  - to je v rozporu s požadavkem vnitřní konsistence
- Příklad: různé subtesty inteligenčního testu mají vůči sobě inkrementální validitu v situaci zjišťování struktury schopností (např. při poradenství pro volbu školy)

Diferenciální validita

- 2 významy:

1. validita pro rozlišování na dané vlastnosti mezi osobami
   - použití – praktická psychodiagnostika
   - Příklad: testy zájmů mají nízké korelace s úspěchem na univerzitě, ale platí to do různé míry u různých skupin osob
   - inteligence považována za faktor všech intelektuálních činností, ale zájem o vědu koreluje více se zájmy o hudební aktivity nebo historii
2. validita pro odlišení daného rysu od rysů ostatních
   - použití – posouzení konstruktové validity
   - Příklad: u mnoha testů osobnostních rysů je prvním úkolem validizace dokázat, že test neměří extraverzi nebo neuroticismus (nebo další příbuzné rysy)

Konstruktové důkazy o validitě

- vyjádření potřeby – zprávy z výborů APA pro psychologické testy (1952, 1954
  - každá validizační studie má obsahovat také hodnocení samotného měřeného konstruktu a teorie
- pojem – Cronbach a Meehl (1955)
- *logický postup* (Cronbach a Meehl, 1955):
  1. předpoklad, že test měří rys A
  2. posouzení současné teorie o rysu A
  3. formulace hypotéz, s čím bude a s čím nebude mít rys A vztah
  4. empirické ověření hypotéz

Typy hypotéz

Příklad pro test inteligence

1. skóry testu budou korelovat se skóry jiných testů inteligence administrovaných nyní a v budoucnosti (*souběžná a prediktivní validita*)
2. skóry testu nebudou korelovat s testy, které nejsou považovány za testy daného rysu (tzn. definice pomocí vyloučení – *souběžná a diskriminační validita*)
3. skóry testu budou korelovat s akademickými výkony nyní a v budoucnosti (*prediktivní validita*)
4. skóry testu budou na vysoké hladině významnosti diskriminovat mezi různými skupinami profesí
5. ve faktorové studii schopností budou faktory testu vysoce sytit první společný faktor
6. při oddělené analýze různých sociálních skupin bude stále významná korelace mezi testem a akademickým výkonem

Vyvrácení hypotéz

- 3 možnosti (Cronbach a Meehl, 1955):

1. test neměří danou proměnnou (konstrukt)
2. teoretická síť generující hypotézu je nesprávná
3. testování hypotézy selhalo

Poznámky ke konstruktovým důkazům o validitě

1. někdy se zjistí, že kritérium používané v raných stadiích vývoje testu je méně validní než test
   - Příklad: IQ testy byly původně vytvořeny tak, aby korelovaly s hodnocením učitelů, dnes se IQ považuje za přesnější
2. uživatel testu potřebuje znát i teorii a důkazy její platnosti
   - dokud není teorie „pochopena“, není možné ji validizovat
3. konstruktovou validitu testu lze hodnotit pouze tehdy, když je zveřejněn
   - je na místě opatrnost při používání testů, jejichž principy jsou tajné
4. každý výsledek ve shodě s teorií postupně podporuje validitu
   - ale dobře zakotvený negativní nález může vše najednou vyvrátit
5. otázka po validitě testu je v konečném důsledku naivní, protože test nelze nikdy zcela validizovat (Cronbach a Meehl, 1955)
   - test může být validní pouze vzhledem k nějakému účelu

Přístup MTMM

- princip: 2 míry stejného rysu by měly vysoce korelovat, ale vysoká korelace s irelevantním rysem je v rozporu s validitou
  - postupy: korelační a faktorová analýza
- přístup MTMM: (Campbell a Fiske, 1959) – konvergentní a diskriminační princip uplatněný při analýze aspoň 2 rysů měřených aspoň 2 metodami
  - tzv. **MTMM matice**

1. diagonály reliability
2. diagonály validity
3. heterorysové-monometodové trojúhelníky
4. heterorysové-heterometodové trojúhelníky

Způsob uvažování

- důkaz konvergentní validity: hodnoty v diagonálách validity jsou statisticky významně nenulové a dost vysoké, aby podpořily zájem o další výzkum

- důkazy diskriminační validity:

1. hodnoty v diagonále validity jsou vyšší než hodnoty v sousedním heterorysovém-heterometodovém trojúhelníku
2. hodnoty v diagonále validity jsou vyšší než hodnoty v heterorysových-monometodových trojúhelnících
3. ve všech heterorysových trojúhelnících lze pozorovat podobné vzorce korelací

Princip matice MTMM

		Metoda I			Metoda II			Metoda III		
		Rys 1	Rys 2	Rys 3	Rys 1	Rys 2	Rys 3	Rys 1	Rys 2	Rys 3
Metoda I	Rys 1	**R**	DV	DV	**SV**	DV	DV	**SV**	DV	DV
	Rys 2	DV	**R**	DV	DV	**SV**	DV	DV	**SV**	DV
	Rys 3	DV	DV	**R**	DV	DV	**SV**	DV	DV	**SV**
Metoda II	Rys 1	**SV**	DV	DV	**R**	DV	DV	**SV**	DV	DV
	Rys 2	DV	**SV**	DV	DV	**R**	DV	DV	**SV**	DV
	Rys 3	DV	DV	**SV**	DV	DV	**R**	DV	DV	**SV**
Metoda III	Rys 1	**SV**	DV	DV	**SV**	DV	DV	**R**	DV	DV
	Rys 2	DV	**SV**	DV	DV	**SV**	DV	DV	**R**	DV
	Rys 3	DV	DV	**SV**	DV	DV	**SV**	DV	DV	**R**

Aspekty validity: Samuel Messick

- Obsahové
- Věcné (předmětné)
- Strukturní
- Zobecnitelnost
- Externí
- Důsledkové

Konstrukce psychodiagnostických metod

- **Obecné poznámky**
- konstrukce metody – směs tvořivého postupu a vysoce odborné výzkumné činnosti
- nutnost – vhodný postup výběru položek
  - základ takového postupu tvoří nějaká konkrétní metoda analýzy položek
- **Konstrukce testu**
- - vymezení a rozdělení zkoumané oblasti (viz zdánlivá validita)
- - často posouzení prvních představ experty z řad kolegů

Tvorba položek

- snaha stimulovat odpovědi o typickém chování (*přímá otázka*)
- dotaz o jedinečném důsledku plynoucím z nějaké charakteristiky (*nepřímá otázka*)
- snaha zjistit úroveň schopnosti, dovednosti, vědomosti

Formáty odpovídání

- krátká odpověď
- doplňování vět
- škála
- mnohonásobná volba
- spojování

Zásady tvorby položek

- **otázka nebo instrukce má být co nejjasnější**
  - vyhýbat se složitým nebo nezvyklým formulacím
  - uvést vše potřebné pro výběr odpovědi
  - na správných odpovědích se musí shodnout kompetentní experti
  - všechny nesprávné alternativy by měly přibližně stejně přijatelné
  - nelze použít překrývající se nebo navazující odpovědi
  - mnohonásobných voleb aspoň 3 alternativy
  - párování nemá obsahovat více než 10 prvků na každé straně
  - odpověďová škála 5-9 bodů

Zásady tvorby položek

- **vyhýbání se náznakům**
  - všechny alternativy gramaticky vyrůstající z hlavní věty
  - správné odpovědi podobné délky jako distraktory
  - ani stereotypní ani učebnicové formulace
  - u nesprávných odpovědí nepoužívat extrémy jako vždy nebo nikdy
  - správné odpovědi by neměly být na zvláštním místě častěji než jinde

Zásady tvorby položek

- **vyhýbání se kontaminacím**
  - vyhýbat se chytákům
  - vyhýbat se dlouhým a rozvitým větám s mnoha kvantifikacemi
  - slova opakující se v odpovědích mají být součástí otázky
  - pokud existuje logické uspořádání odpovědí, mělo by být použito
  - vyhýbat se dvojitému záporu

Analýza položek

- Základní myšlenka analýzy položek – pokusná verze testu předložena vybranému souboru osob, získaná data jsou statisticky zpracována, výsledky slouží jako vodítka pro další modifikaci metody

- **Hlavní zásady**
- Výběr osob – v ideálním případě data od velkého a reprezentativního souboru osob
  - velký soubor – čím větší soubor, tím nižší standardní chyby výběrových statistik
  - aspoň 2–3× větší než počet proměnných (klasická analýza položek)
  - aspoň 2–3× větší než počet parametrů faktorového modelu (faktorová analýza)
  - reprezentativní soubor – podobající se v hlavních charakteristikách populaci, pro kterou je metoda určena (nutno uvážit, které charakteristiky budou mít vztah s měřeným rysem)

Klasická analýza položek

- Charakteristiky položek
  - obtížnost položky – procento osob, které odpověděly správně (vlastně jednoduchost)
    - u osobnostních testů – procento osob, které odpověděly v diagnostickém směru (tzv. popularita položky)
  - rozlišovací účinnost položky – vztah skóru položky a skóru celého testu
    - také – reliabilita položky
    - korekce – skór položky a součtu zbylých položek testu
  - validita položky – vztah skóru položky a nějakého vnějšího kritéria

Vhodné míry vztahu

- Pearsonova korelace a její modifikace
  - např. bodově-biseriální korelace
    - Pearsonova korelace, u které je jedna z proměnných dichotomická)

$$r_{p_i X}^{(b)} = \frac{m_1 - m_0}{s_x} \sqrt{p_i q_i}$$

- kde $\mathbf{m_1}$ – průměr celkového skóru osob s položkou vyřešenou správně, $\mathbf{m_0}$ – průměr celkového skóru osob s položkou vyřešenou nesprávně, $\mathbf{s_x}$ – směrodatná odchylka celkových skórů, $\mathbf{p_i}$ – podíl (procento) osob s položkou zodpovězenou správně, **qi** = 1 – pi

Psychometrický paradox

- jak roste vnitřní konsistence testu, klesá jeho validita
  - **Souvislosti**
    - inkrementální validita
    - nutnost kompromisu mezi vnitřní konsistencí a „nekonsistencí“ obsahu položek

Metody konstrukce testů a analýzy položek

- **klasické přístupy:**
  1. přístup založený na kritériu
  2. přístup založený na klasické analýze položek
  3. přístup založený na faktorové analýze
- **nové přístupy:**
  4. přístupy založené na teorii odpovídání na položku

Přístup založený na kritériu

- cíl – metoda, která bude rozlišovat mezi určitými skupinami osob
  - příklad – MMPI, Strongův zájmový inventář

- použití – aplikované oblasti psychologie, screening
- nedostatky- jedná se o produkty slepého empirismu
  - absence psychologického významu
  - skupiny se zřejmě neliší pouze v jediné proměnné
  - nemusí být jasné, co škála měří
  - problém výběru vhodných skupin
    - sama klasifikace do skupin může být velmi nereliabilní
  - značná specifičnost těchto testů

Postup analýzy položek

1. Určení skupin, mezi kterými má test rozlišovat
   - nebo jiného kritéria
2. Výběr osob do výzkumu
   - rozsah roste s počtem skupin, mezi kterými má test rozlišovat (diagnostické a kontrolní skupiny – aspoň 200 ve skupině)
3. Administrace pokusné verze metody vybraným osobám
4. Statistická analýza
   - t-testy, analýza diskriminační funkce, korelace položky s členstvím ve skupině
5. Posouzení položek
   - statisticky významný vztah s kritériem, výběr žádoucího počtu položek (nutno připravit jich aspoň 2× tolik)
6. Pohlaví
   - analýza pro muže a ženy zvlášť (snaha vybrat položky, které mají v obou skupinách podobné výsledky)
7. Validizace
   - ověření výsledků na nové skupině osob nebo s novým kritériem (modifikace a opakování postupu od začátku)

Alternativní postupy

- Metoda postupného skládání
  - výběr nejvalidnější položky, k ní výběr další položky, která má spolu s ní nejvyšší validitu atd. dokud nevznikne test žádané délky
- Metoda postupné redukce rezidua
  - výběr nejvalidnější položky, jako další je vybrána ta položka, která nejlépe predikuje reziduální kritérium atd. opět až do stanoveného počtu položek

- Možnost ověření pomocí faktorové analýzy
  1. faktorování škály – její lokalizace ve faktorovém prostoru spolu s jinými škálami (pokud škála měří 1-2 faktory)
  2. faktorování položek – faktorová analýza položek škály spolu s položkami jiných škál

Přístup založený na klasické analýze položek

- cíl – homogenní test
- Postup - velmi podobný jako u předchozí metody
  - tzn. pokusná verze metody administrovaná reprezentativnímu a velkému souboru osob atd.
- Kritéria
  - Rozlišovací účinnost položky
    - korelace skóru položky a celkového skóru (nad 0,3)
  - Obtížnost položky
    - p-hodnota (jednoduchost/popularita – mezi 0,2 a 0,8)

Další úkoly cíle tohoto přístupu

- Nutnost prokázat validitu
  - tento způsob výběru položek do testu zvyšuje homogenitu, je tedy jisté, že test něco měří, ale co?
- Nutnost prokázat jednofaktorovost
  - homogenita není totéž jako jednorozměrnost (opět možnost uplatnění faktorové analýzy)

- Výhoda
  - není nutný tak rozsáhlý soubor (často stačí N = 100)
- Nevýhoda
  - nemožnost konstrukce více škál současně

Přístup založený na faktorové analýze

- cíl – jednofaktorový test
- Postup – opět velmi podobný jako u předchozích metod
- **FA** – z testu jsou vyřazeny položky, které mají na prvním společném faktoru nižší náboje než 0,3
- Opět – nutnost prokázat validitu metody, v tomto případě s důrazem na validitu získaných faktorů
- Nutnost prokázat reliabilitu metody – jednofaktorový test je vnitřně konsistentní, proto je třeba zjistit test-retest reliabilitu

Konstrukce rychlostního testu

- **Nelze**
  - použít ani klasickou analýzu položek
  - ani faktorovou analýzu
- **důvod**
  - korelace mezi položkami jsou artefakty časového omezení administrace testu a uspořádání položek
    - položky umístěné v testu brzy po sobě spolu korelují, kdežto vzdálené položky ne

- Nutný postup
  - administrace testu několika skupinám osob s různým časovým omezením
  - výběr časového omezení, které má nejvyšší reliabilitu (tzn. i nejvyšší rozlišovací schopnost)

Důležitost replikace

Ve všech případech je nutné prokázat, že zjištěné výsledky platí i pro jiné osoby než ty, která byla testovány při tvorbě testu

Normalizace – tvorba norem

- Normalizace
  - někdy také nazývána standardizace (to je širší pojem)
- Účel norem
  - snaha vyhnout se chybám interpretace
  - porovnání individuálního skóru v testu se skóry relevantní populace
  - testovým skórům dávají normy psychologický smysl (umožňují interpretaci)

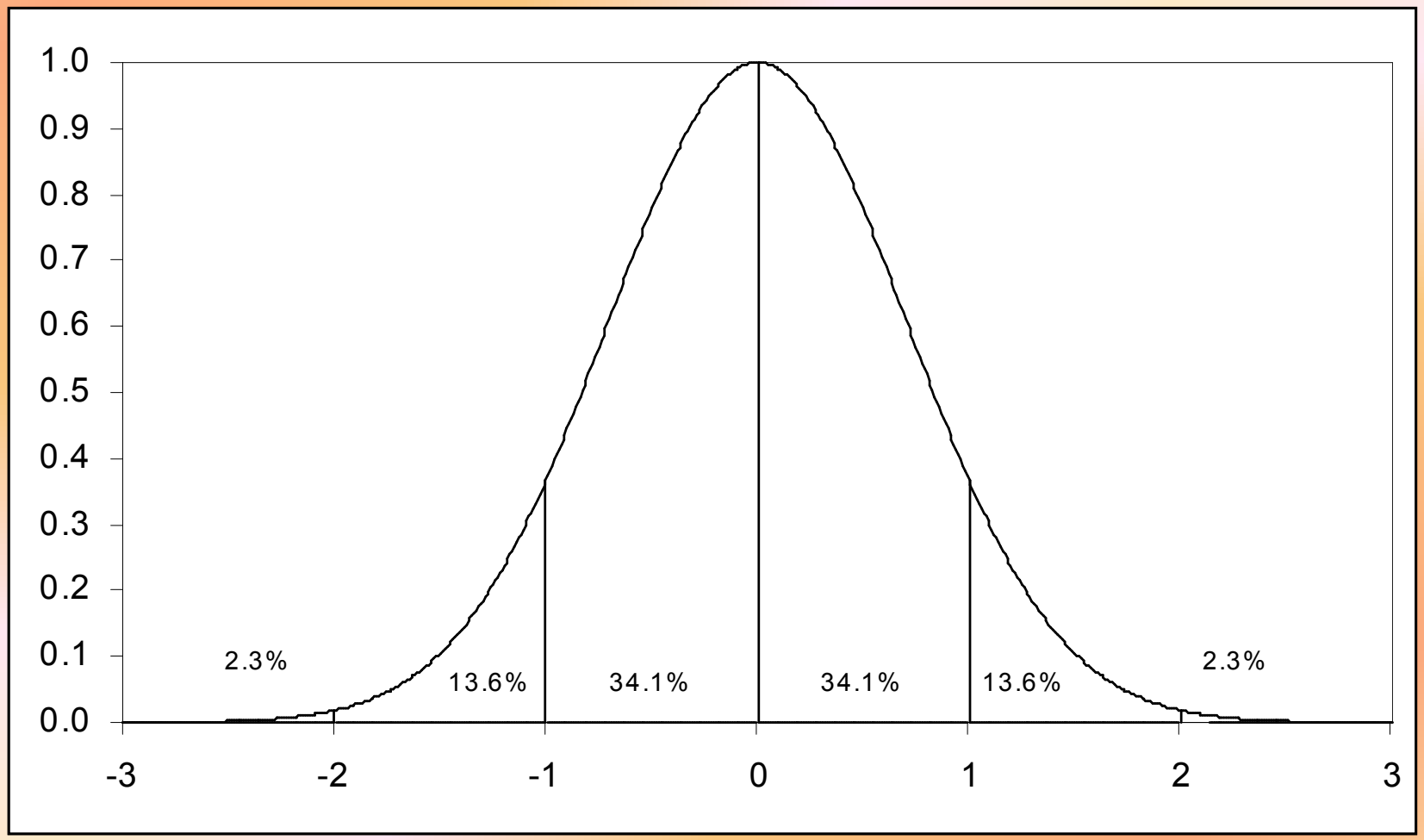
Dvě základní otázky

1. *jak je definována skupina*
   - pohlaví, věk, školní třída, povolání, region, socioekonomický status atd.
2. *jaké je pravidlo porovnávání*

Základní pojmy

- hrubý skór
  - nějaká míra úspěchu nebo neúspěchu v testu (počet vyřešených úkolů, reakční čas atp.)
  - nějaká míra souhlasu nebo nesouhlasu se skupinou položek (např. součet hodnot u jednotlivých položek)
- odvozený skór (vážený skór)
  - skór definovaný na základě porovnání hrubých skórů vyšetřované osoby se skóry určitého souboru osob
- normy
  - vhodná transformace hrubého skóru na odvozený skór na základě srovnání s relevantní normativní skupinou

Význam normálního rozložení



Možné výjimky z normality

1. test je pro danou skupinu příliš obtížný nebo snadný (šikmé rozložení)
   - *řešení*: za předpokladu normality možnost plošné transformace, obecně ale nutnost revize a položkové analýzy
2. skupina není homogenní vzhledem k měřenému rysu (bimodální nebo polymodální rozložení)
   - *řešení*: oddělená analýza a normy pro jednotlivé podskupiny
3. rys souvisí se sociální konformitou (rozložení typu J nebo L)
   - *řešení*: transformace do normálního tvaru se nedá použít, normální rozložení je chybný předpoklad

Soubory pro tvorbu norem

- *2 kritéria*
  - velikost a reprezentativnost
  - reprezentativnost je do značné míry nezávislá na velikosti
  - malý soubor nemůže být opravdu reprezentativní vzhledem k velké populaci
    - závisí to na homogenitě populace

Metody výběru

- Pravděpodobnostní výběr
  - Náhodný výběr
  - Stratifikovaný náhodný výběr
- Nepravděpodobnostní výběr
  - Kvótní výběr
  - Výběr metodou sněhové koule
  - Záměrný výběr

Náhodný výběr

- Princip – stejná pravděpodobnost, že jakýkoli člen populace bude součástí výběrového souboru
  - dostatečně velký náhodný soubor tak může být zcela reprezentativní vzhledem k populaci, pro kterou byl vybrán
- Definice populace – operacionální, např. na základě vhodného seznamu
  - *problém* – takový seznam zpravidla neexistuje (seznamy na základě sčítání lidu jsou neaktuální, volební seznamy obvykle nekompletní, telefonní seznamy nadhodnocují zastoupení středních vrstev atp.)
- Procedury náhodného výběru
  - tabulky náhodných čísel (počáteční pozice určena hodem kostky)
  - generátor pseudonáhodných čísel v počítači
  - každá i-tá osoba od určité pozice za předpokladu, že pozice v seznamu je náhodná
    - *Problémy* – náhodný výběr je možný pouze tam, kde existuje nějaký adekvátní seznam populace; soubor musí být velký, což znamená velké nároky na čas a peníze

Stratifikovaný náhodný výběr

- Princip – rozdělení heterogenní populace do několika homogennějších populací
- Proměnné pro stratifikaci – musí mít souvislost se studovanou proměnnou
  - pokud je jich příliš mnoho, soubor je obrovský
  - při použití nevhodných proměnných nebude reprezentativní

- *Příklad*: normy pro Lorge-Thorndikeův inteligenční test pro děti jsou velmi kvalitní (u nás neznámý)
  - definice populace založena na obcích
  - nutné vzít v úvahu také věk dětí
  - *stratifikační proměnné* (vztah se skórem IQ v obcích)
    - procento gramotnosti dospělých v obci
    - podíl odborných pracovníků v obci
    - procento vlastnictví domů
    - medián hodnoty domovního nájemného
- *výběr* – 44 tříd obcí, v každém typu otestováno 11000 dětí v každé věkové úrovni od 6-17, výsledkem je N = 132000

Pravidla pro výběr z obecné populace

1. stratifikovaný výběr je efektivnější a pro standardizaci testu nejlepší
2. obvykle stačí 4 úrovně klasifikace (při příliš velkém počtu kritérií je nutný příliš velký soubor)
3. stratifikační proměnné musí vysoce korelovat s testem
4. v každé podskupině musí být dostatek osob
   - 300 je minimum
     - pak např. 5 sociálních tříd, 2 pohlaví, 5 věkových skupin – tzn. 50 kategorií – tzn. N = 15000

Výběr ze speciálních populací

- soubory mohou být menší, ale je mnohem větší problém s definicí populace

- *např.*:
  - stačí při výzkumu zlodějů kritérium usvědčení z krádeže?
    - mnoho zlodějů je nechycených
  - seznam osob s diagnózou schizofrenie
    - kromě toho problém s malou shodou v diagnózách

Pravidla pro výběr ze speciálních populací

1. stratifikace výběru podle proměnných nejvýše korelujících s testem
2. minimální N = 300
3. malý soubor je lepší než nic, ale je nutné na to uživatele norem upozornit

Typy odvozených skórů

- Věkové a ročníkové skóry
- Percentily
- Standardní skóry

Věkové a ročníkové skóry

- nejstarší způsob jak vyjádřit srovnání s relevantní skupinou (mentální věk)
  - nelze použít, pokud rys nemá těsný vztah s věkem

- *výhoda* – jsou pochopitelné i pro laiky
- *použití* – obecné nebo školní schopnosti a výkonové testy
- *2 možnosti*
  - průměrný skór pro každou věkovou úroveň
  - průměrný věk pro každou úroveň skóru

Percentily

- skóry, pod které spadají daná procenta normativní skupiny
- *výhoda* – chápou i osoby, které test vyplňovaly
  - jednoduchá interpretace dokonce i při šikmém nebo bimodálním rozložení
- *Nevýhody*
  - jedná se o ordinální skóry, takže nelze použít parametrické statistické metody
  - rozložení percentilů je rovnoměrné, kdežto rozložení testových skórů obecně normální
- *2 typy zkreslení*:
  - malé rozdíly kolem průměru se nadhodnocují
  - poměrně velké rozdíly na koncích distribuce se setřou
- *použití* – vysvětlení výsledků

Standardní skóry

- převedení jednotek hrubých skórů na jednotky definované směrodatnou odchylkou
  - existuje jich mnoho druhů, všechny jsou odvozeny od z-skórů:

$$z_i = \frac{x_i - m_X}{s_X}$$

    - kde z_i – z-skór osoby i, x_i – skór získaný osobou i v testu X, m_X – průměr rozložení skórů testu X, s_X – směrodatná odchylka skórů testu X
  - *vlastnosti*
    - průměr m = 0
    - směrodatná odchylka s = 1
    - většina hodnot leží v rozmezí od –3 do +3
    - jedná se o lineární transformaci hrubých skórů, takže je zachováno původní rozložení hrubých skórů

Výhody a nevýhody z-skórů

- *Výhoda*
  - stejné z-skóry jsou vždy ekvivalentní
    - z = 2 je vždy 2 směrodatné odchylky nad průměrem
    - možnost srovnávat skóry různých testů

- *Nevýhody*
  1. většina psychologů jim příliš nerozumí (probandi ještě méně)
  2. mnoho testů nemá normální rozložení (nelze převádět na percentily)

Řešení nevýhody 1

- lineární transformace na jinou standardní škálu
  - přičtení vhodné hodnoty průměru
  - násobení směrodatné odchylky vhodnou konstantou

- Obecný vzorec:

$$z_t = a + bz$$

- kde z_t – transformovaný z-skór, a – průměr transformované distribuce, b – směrodatná odchylka transformované distribuce, z – původní z-skór

Typy transformovaných z-skórů

- *T-skóry* m = 50 s = 10
- *Steny* m = 5,5 s = 1,5
  (standard ten – výsledkem je 10 skórů)
- *staniny* m = 5 s = 2
  (standard nine – 9 skórů)
- *Wechslerovy* m = 100 s = 15 (nebo 16)
- *Testy inteligence*

Řešení nevýhody 2

- transformace celého rozložení na rozložení normální
  - tzv. plošná transformace

- předpoklady: je možné provést pouze u proměnných, jejichž původní rozložení je přibližně normální, existuje teoretický předpoklad normálního rozložení a soubor pro normalizaci je dostatečně velký a reprezentativní

Alternativní metody interpretace testových skórů

- Obsahové kritérium
- Expektanční tabulky
- Regresní metoda
- Ipsativní skóry

Obsahové kritérium

- úzká souvislost s obsahovou validitou
  - musí existovat jasně definovatelný obsah

- *použití*: nejběžnější na elementární úrovni v odborných předmětech (např. hudba, matematika)
- *problém*: nalezení kritických skórů pro danou metodu

Expektanční tabulky

- tabulky obsahující pravděpodobnosti úspěchu nebo neúspěchu pro dané hodnoty skórů
  - zcela empirický postup

- *nutnost*:
  - jasně definovaná kritéria (často vzdělávací a průmyslové aplikace testů)
  - testování velkých souborů osob z relevantní populace (aspoň 20 osob pro každou buňku tabulky)
- *problém*:
  - použití pravděpodobnostní předpovědi v individuálním případě
    - (z dlouhodobého hlediska jsou rozhodnutí pro velký počet osob správná, ale pro libovolného jednotlivce může jít o chybné rozhodnutí)
  - je nutné velké množství informací

Regresní metoda

- použití regresní rovnice k predikci skórů kritéria ze skórů testu
- Regresní rovnice:

$$Y_{pred} = a + bX$$

- kde $\mathbf{Y_{pred}}$ – predikovaný skór kritéria, **a** – regresní konstanta, **b** – regresní koeficient, **X** – skór v testu
- *velikost chyby predikce*: závisí na těsnosti vztahu testu a kritéria
- *použití regresní metody*: má smysl pouze při nízké hodnotě standardní chyby odhadu

Ipsativní skóry

- porovnání hodnot různých subtestů u jedné osoby

- *např.*: pro každou položku výběr nejvíce a nejméně sympatické odpovědi (test obranných mechanismů)
- *zvláštnosti*:
  1. korelační matice položek nedává smysl (negativní korelace škál představují artefakt), takže např. výsledky faktorové analýzy nelze interpretovat
  2. nemá smysl vytvářet normy, protože skóry mezi osobami nejsou srovnatelné (jedinou možností jsou pořadí odpovědí u osob)
     - *použití*: základ pro diskusi s osobou, která vyplnila test (tzn. hlavně v klinické a poradenské psychologii)

Metody založené na IRT (teorii odpovídání na položku)

- normy nejsou v zásadě vůbec nutné
  - ale vytvářejí se pro lepší interpretaci výsledků

Kde nejsou normy důležité

- ve výzkumu
- při psychometrické analýze individuálních rozdílů
- při studiu vlastností, schopností, osobnosti, motivace, nálad atd.

- *právě naopak*: transformace na odvozené skóry může znamenat ztrátu informace (redukce variability)

Nové přístupy v psychodiagnostice

- Teorie odpovědi na položku
- Počítačové adaptivní testy
- Teorie zobecnitelnosti

Teorie odpovědi na položku

- IRT – item response theory
  - počátky těchto přístupů už u Thorndika a Thurstona
  - vývoj dlouho v pozadí vývoje klasické teorie testů
  - IRT je ale obecnější a flexibilnější

Porovnání s klasickou teorií testů

- Klasická teorie testů

1. Standardní chyba měření se týká všech skórů určité populace (také celkový index reliability)
2. Delší testy jsou reliabilnější než kratší
3. Porovnání skórů testů u různých forem závisí na míře paralelnosti testů
4. Nezkreslené odhady vlastností položek závisí na reprezentativních výběrech z populace

- Teorie odpovídání na položku

1. Standardní chyba měření se liší pro jednotlivé osoby s odlišnými vzorci odpovědí, ale lze ji zobecnit v dané populaci (podmíněná st. chyba měření)
2. Kratší testy mohou být reliabilnější než delší testy
3. Porovnání skórů testů u různých forem je optimální, pokud se úroveň obtížnosti pro jednotlivé osoby liší
4. Nezkreslené odhady vlastností položek lze získat z nereprezentativních výběrů

Charakteristiky IRT

- Cíl IRT: vývoj testů, u kterých jsou vlastnosti položek nezávislé na souboru a měření latentního rysu nezávislé na položkách
- Latentní rys: odpovědi na položky lze vysvětlit existencí latentního rysu

- Charakteristická funkce položky: vztah skóru položky s vektorem latentního rysu
  - u dichotomické položky se jedná o pravděpodobnost kladné (nebo správné) odpovědi v závislosti na úrovni latentního rysu
  - normální ogiva – jeden z modelů pro charakteristickou funkci položky

Parametry položky

a) rozlišovací účinnost položky
- míra variability v závislosti na úrovni latentního rysu
- přesněji – *směrnice (strmost) křivky v oblasti kolem pravděpodobnosti správné odpovědi rovné 0,5*

b) obtížnost položky
- poloha charakteristické funkce položky na škále latentního rysu
- přesněji – *bod na škále latentního rysu, kde charakteristická křivka položky přetíná hodnotu pravděpodobnosti správné odpovědi 0,5*

c) parametr pseudo-uhádnutelnosti
- pravděpodobnost kladné/správné odpovědi u osoby s nulovou úrovní latentního rysu

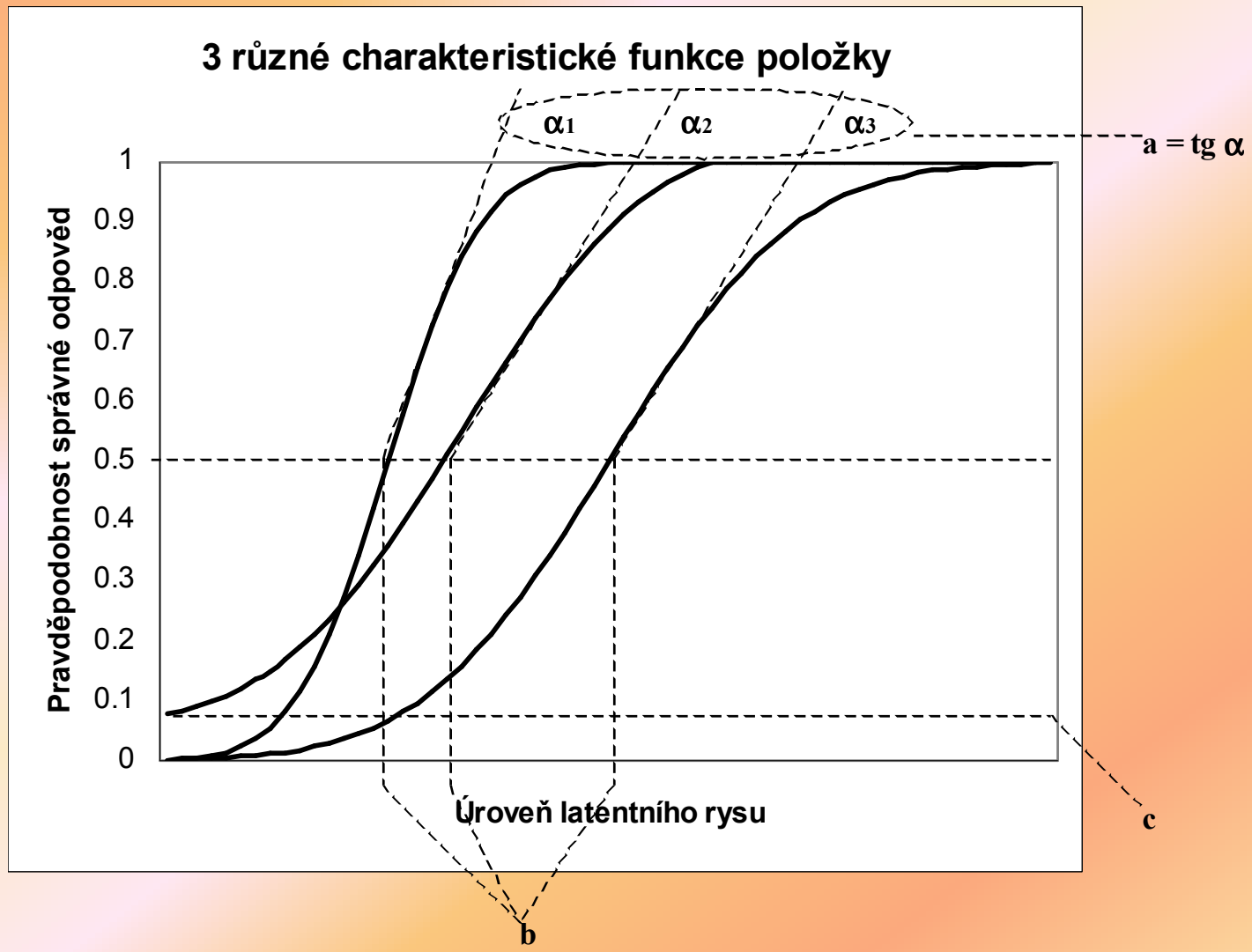
Různé modely IRT

- **1PL** – model s jedním parametrem (Raschův model)
  - obtížnost
- **2PL** – model se dvěma parametry
  - obtížnost a rozlišovací účinnost
- **3PL** – model se třemi parametry
  - obtížnost, rozlišovací účinnost a pseudo-uhádnutelnost

Tvar charakteristické křivky položky

- základním předpoklad – monotónní rostoucí křivka
  - některé křivky inspirované psychofyzikou
  - mezi různými typy křivek existuje značná shoda
  - nejvíce normální ogiva nebo logistická křivka (výhodnější výpočet a použití v psychofyzice)

Charakteristické křivky a parametry položek



Konstrukce testu podle IRT

- každá položka by měla mít jinou charakteristickou funkci
  - obecný tvar (definovaný vhodnou funkcí) by měly mít stejný
  - hlavní zájem – úroveň obtížnosti a rozlišovací účinnost položek

- Důležité vlastnosti:
  - možnost odhadu skórů osob na položkách, které osoby neřešily za předpokladu znalosti pozice položek na škále latentního rysu
    - tzn. na základě známé char. funkce položky – CHFP
  - skór podmnožiny položek dovoluje odhad skóru celého testu
    - výhoda ve srovnání s ekvivalentními testy
  - odhady obtížnosti jsou nezávislé na úrovních latentního rysu osob
  - parametry položek nezávislé na populaci a parametry osob nezávislé na položkách

Další důležité indexy

- Informace položky:

- Informace testu:
  - analogie křivky informace položky, suma křivek informace jednotlivých položek; tzn. úroveň přesnosti testu pro každou úroveň latentního rysu

- Podmíněná standardní chyba měření:

Podobnosti IRT a klasické teorie testů

- faktorová analýza tetrachorických korelací položek je dobrým testem jednorozměrnosti položek za předpokladu modelu normální ogivy
- faktorové náboje položek v prvním faktoru jsou dobrými odhady sklonu charakteristické křivky položek
- na základě p (jednoduchost) a r (korelace položky s celkem) lze vypočítat obtížnost, rozlišovací účinnost a uhádnutelnost

Počítačově administrované testování

- možnost administrovat položky, které by jinak administrovat nešlo
- počítačový test jako paralelní forma k testu papír-tužka

- Základní požadavky: instrukce musí být jasné, ovládání jednoduché
  - odpovídání na otázky
  - přechod k další otázce
  - změna odpovědi
  - zpětné prohlížení)

Výhody počítačové administrace

- možnost okamžité práce s daty
- administrace vždy naprosto stejná
  - podmínkou je ale např. stejné vybavení – např. monitor
- možnost testování handicapovaných atd.
- údajně poctivější odpovědi
  - před počítačem neexistují zábrany

Nevýhody počítačového testování

- při hromadném testování nutný počítač pro každou osobu
  - náklady na vyšetření
- nutná ekvivalence počítačové a tradiční formy
- častá námitka, že neexistuje živý kontakt s osobami
- u počítačově administrovaných inteligenčních testů údajně vyšší úzkost a nižší motivace osob
- zejména se nedoporučuje počítačová interpretace výsledků
  - např. jak citlivě informovat o nepříznivých výsledcích
- etické problémy
  - zacházení s lidmi jako s objekty

Počítačové adaptivní testování

- Adaptivní testování:
  - postup, při kterém se používají pouze položky testu vybrané na základě vhodného kritéria
- Počítačové adaptivní testování:
  - každé osobě lze administrovat pouze malou část položek vybraných přímo pro ni díky spojení s teorií odpovědi na položku
- Základy počítačového adaptivního testování:
  - index obtížnosti na základě IRT pro každou položku
  - administrace položky s 50% úrovní obtížnosti
  - po správné odpovědi administrace obtížnější položky, po nesprávné odpovědi administrace jednodušší položky
- Důsledek:
  - testy tak mohou být mnohem kratší

Položková banka

- množina položek, ze kterých je možné vybírat pro účely počítačového adaptivního testování
- možnost konstrukce velmi přesných paralelních testů
- úzké spojení s teorií odpovědi na položku

Výhody počítačového adaptivního testování

- pro většinu osob stačí krátký test
- při dostatečné zásobě položek možnost opakovaného testování
- při výběrových řízeních může každá osoba dostat jinou sadu položek
- použití části položek a přitom přesný skór
- test jako součást expertního systému

Nevýhody počítačového adaptivního testování

- v některých oblastech nemá pojem obtížnosti smysl
- stejné jako u počítačově administrovaného testování