Teorie testů a základy psychometriky

Zpracoval Jiří Dan

Učební text k předmětu XS041 Pedagogicko-psychologická diagnostika

Psychometrika se zabývá teorií a praxí měření psychických jevů a těch charakteristik, které s psychickými jevy souvisejí (např.vědomostí, dovedností a návyků, ale též měřením jevů sociálních). Měření v oblasti trvalejších psychických charakteristik, včetně naučených vědomostí a dovedností, se v psychologii objevuje systematicky až se vznikem tzv. diferenciální psychologie, která od počátku 20. století začíná systematicky zkoumat variabilitu znaků osobnosti, jejich rozložení a zdroje. Základní otázkou psychometriky je, co měříme na psychických jevech a jakých měrných jednotek přitom můžeme užívat. Důležitou kapitolu psychometriky tvoří i problematika konstrukce nástrojů měření.

Psychometrika vychází z teorie měření. Základním problémem je vztah mezi psychickými jevy a jejich ukazateli, jejichž prostřednictvím psychické jevy poznáváme. Tento vztah je vyjádřen pojmem **validita**, platnost. Další otázkou je přesnost měření a stabilita výsledků měření v čase, které jsou vyjádřeny pojmem **reliabilita.** Třetím velkým tématickým okruhem psychometriky je porovnání výkonů jedince s ostatními, což je vlastně základní význam konstrukce testu. S tím souvisí převod hrubého skóre na vážené skóre. Hrubé skóre je počet správných odpovědí nebo počet chyb v testu, také např. čas potřebný k vyřešení úlohy. Různé stupnice váženého skóre slouží k porovnání výkonů jednotlivce s ostatními jedinci ve skupině, se kterou a v níž je jedinec srovnáván.

Definice psychodiagnostického testu

Existuje mnoho definic testů, všechny však vyjadřují, že test

- je na základě vědeckých poznatků zkonstruovaný nástroj pro objektivní a spolehlivé zjišťování psychické danosti,
- představuje vzorek, výseč možných výkonů, odpovědí, způsobů chování a reakcí,
- je nástroj standardní, tzn. že je garantována opakovatelnost administrace a srovnatelnost výkonu jednotlivce s ostatními nebo srovnatelnost výkonů téhož jedince v čase,
- je konstruován s cílem rozlišování a smysluplné předpovědi výkonů jednotlivce v budoucích situacích.

Konstrukce testu

Test je standardizovaný nástroj měření přesně definovaných a empiricky odlišitelných znaků osobnosti, vědomostí, dovedností, výkonů atp. Test tedy představuje modelovou situaci, pomocí níž záměrně získáváme vzorky výkonů, které považujeme za ukazatele /diagnostické údaje/ zkoumaného znaku.

Jako každý měrný nástroj, musí být i test vytvářen podle určitých pravidel, musí být předem ověřován a musí vyhovovat řadě podmínek, které vymezují jeho způsobilost k měření.

Standardnost testu

je požadavek uniformního, stejného přístupu při administraci (zadávání) testového materiálu, při registrování dosažených výsledků, při vyhodnocování a interpretování (vysvětlování) výsledků. Standardizovaný test má normy. Test slouží k tomu, abychom mohli výkony jedince srovnávat s výkony jiných členů referenční skupiny (skupina uchazečů). Samo hrubé skóre má většinou jen malou interpretační hodnotu. K tomu slouží standardní normy - ukazatele relativní pozice jedince vůči reprezentativnímu vzorku populace.

Objektivita testu je nezávislost výsledku na zkreslení

1) ze strany administrátora – objektivita provádění. Je zajištěna napsanou instrukcí všem srozumitelnou, zajištěním přibližně stejných fyzikálních podmínek - denní doba atp.

2) vyhodnocením. Je zajištěna stejnými šablonami nebo v příručce k testu uvedenými jednoznačnými pravidly pro hodnocení odpovědí. Například, kdy za určitou odpověď přidělíme dva, jeden nebo nula bodů hrubého skóre.

3) interpretací výsledků. U výkonových testů není takovým problémem jako u testů projektivních.

4) event. tendencí ke zkoumaných osob odpovídat způsobem, který je examinátorem „nejlépe“ hodnocen.(u výkonových testů se předpokládá snaha podat co nejlepší výkon). Nedostatečná objetivita testu může negativně ovlivnit jeho reliabilitu.

Reliabilita

Reliabilita testu má dva aspekty:

1. znamená **přesnost** ve smyslu shody naměřených výsledků se skutečnou hodnotou zkoumaného znaku (sem patří pojmy pravý skór a chyba měření)

2. **stabilita v čase** znamená míru shody dvou měření ve dvou časových okamžicích.

Alternativní formy testu.

Jestliže vytvoříme více variant a verzí testu, můžeme hovořit o následujících formách testů:

- **srovnatelné formy** jsou podobné z hlediska obsahu, nejsou zaručeny jejich psychometrické parametry,
- **ekvivalentní formy** jsou srovnatelné z hlediska odvozených skórů,
- **paralelní formy** mají stejné průměry hrubých skórů, stejné směrodatné odchylky a stejné korelace s vnějšími kritérii. Blíže k tomu nejlépe McDonald (1999, 347-366).

Při posuzování reliability jednotlivých forem se počítají a porovnávají průměry a rozptyly jednotlivých položek a korelace mezi položkami v obou formách.

Při rigorózním přístupu rozdělování testu na dvě poloviny, jinak řečeno vytváření paralelních forem z databáze položek se vyhledávají páry s analogickým obsahem, podobnou hodnotou obtížnosti a blízkou hodnotou indexu rozlišovací účinnosti (viz dále).

S tím souvisí otázka vztahu reliability a délky testu.

Odhad vnitřní konzistence testu - split-half reliabilita

„Split-half" je rozpůlení. Vychází se z myšlenky, že pokud je test spolehlivý, reliabilní jako celek, musí být stejně spolehlivé i jeho části. Pro odhad split-half reliability užíváme upravený Spearmanův-Brownův vzorec. Tuto metodu můžeme používat tehdy, když jsou položky homogenní (vykazují stejnou variabilitu; když test zjišťuje různé rysy, je rozmanitá i variance položek) a když test není časově omezený.

Odhad vnitřní konzistence testu

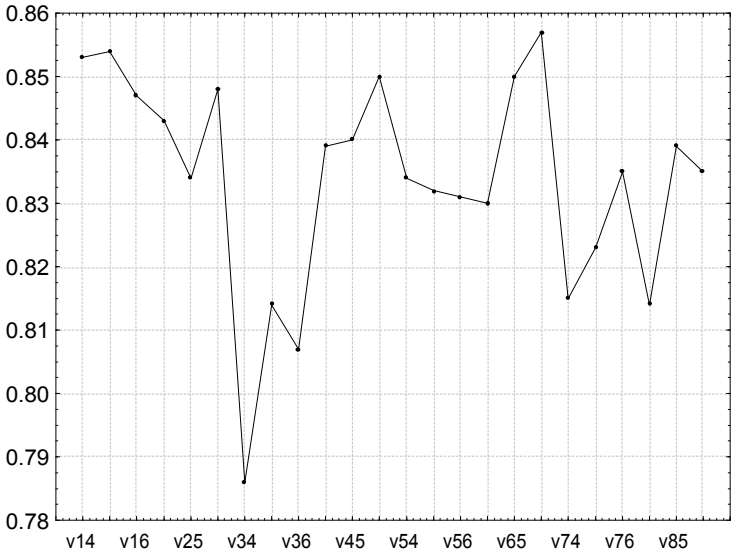
Často používanou metodou odhadu vnitřní konzistence testu je výpočet Kuderova-Richardsonova vzorce. Jeho nejznámější verze je tzv. Kuderův-Richardsonův vzorec 20.

$$KR_{20} = \frac{c}{c-1}\left(1 - \frac{\sum_{j=1}^{c} p_j q_j}{s_x^{\,2}}\right), \text{ kde}$$

c ... počet položek v testu (v našem případě 80)
p_j ... obtížnost j-té položky (podíl osob, které správně vyřešily j-tou položku)
$q_j = 1 - p_j$
s_x^2 ... rozptyl počtu správných odpovědí

Jako příklad uvádíme číselné hodnoty míry reliability a jejich grafické znázornění vypočtené pro Test studijních předpokladů 2004 užitý na Masarykově univerzitě.

	1 KR_{20}
v14	0.853
v15	0.854
v16	0.847
v24	0.843
v25	0.834
v26	0.848
v34	0.786
v35	0.814
v36	0.807
v44	0.839
v45	0.840
v46	0.850
v54	0.834
v55	0.832
v56	0.831
v64	0.830
v65	0.850
v66	0.857
v74	0.815
v75	0.823
v76	0.835
v84	0.814
v85	0.839
v86	0.835



Cronbachův koeficient alfa je dnes všeobecně akceptovaným ukazatelem vnitřní konzistence testu.

$$r_{\alpha} = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_i^{\,2}}{\sigma_X^{\,2}}\right),$$

kde $\sum \sigma_i^{\,2}$ je suma variancí jednotlivých položek testu.

Spearmanův-Brownův věštecký vzorec „The profecy formula“ nám umožňuje odhadnout, o kolik homogenních položek by se měl rozšířit původní test, aby dosáhl žádoucí úrovně reliability:

$$n = \frac{r_p\left(1 - r_0\right)}{r_0\left(1 - r_P\right)},$$

kde n = násobek současné délky testu, který by byl potřebný k dosažení žádoucí reliability,

r_p = požadovaná úroveň reliability,

r_0 = původní úroveň reliability.

Metody zvyšování reliability

1. Reliabilitu testu s mnohonásobnými odpověďmi je možno zvýšit vyřazením položek s nerovnoměrným rozložením nesprávných odpovědí (distraktorů).
2. Reliabilitu testu s mnohonásobnými odpověďmi je možno zvýšit důsledným trváním na nárocích z hlediska obsahu a formy položek.
2. Reliabilitu je možno zvýšit snížením počtu měřených rysů. Nejvyšší reliability dosahuje jednodimenzionální test.
3. Reliabilitu je možno zvýšit vyřazením položek, která slabě korelují s celkovým skóre testu.
4. Reliabilitu je možno zvýšit přidáním dalších homogenních položek do původního testu.
5. Reliabilitu je možno zvýšit zpřesněním, zkrácením a dosažením jednoznačnosti instrukce.
6. Reliabilitu je možno zvýšit experimentálním ověřením vhodně zvoleného času k vyplnění testu.

Reliabilita testu pro rozhodování o individuu by neměla být menší než 0.80.

Validita testu

Validita testu (adekvátnost, přiměřenost, výstižnost) je míra shody mezi naměřenými výsledky a tím, co jsme chtěli měřit. Je to odpověď na otázku, do jaké míry test skutečně měří to, co chceme měřit, o čem říkáme, že měří.

Vztah mezi kvalitou položky, reliabilitou a validitou.

Kvalita položek je jedním z předpokladů dostatečné reliability.
Podmínkou validity testu je jeho vyhovující reliabilita. Nereliabilní test nebude nikdy validní. Bez reliability není validity. Vysoká reliabilita je nutnou podmínkou validity, nikoli však zárukou.

Obsahová validita (content validity)

1. Výběrová validita (sample validity)

Dá se vyjádřit otázkou: „Odpovídá obsah testu vlastnosti, která má být měřena?“ Test by měl představovat reprezentativní výběr znaků typických pro zkoumanou vlastnost. U vědomostních testů musí otázky pokrýt celou problematiku zkoušené látky.

U obsahové analýzy výkonových testů se test podrobuje ve 3 časových okamžicích důkladné logické analýze, jejímž cílem je posoudit

- adekvátnost výběru a definování subtestů,
- reprezentativnost souboru položek tvořících daný test. Jeden nebo skupina expertů se vyjadřuje k testu globálně i k položce po položce, přičemž se vyjadřují k míře reprezentativnosti a důležitosti zakomponovaných položek,
- adekvátnost distraktorů při užití položek mnohonásobné volby.

2. Zdánlivá validita (face validity)

Zdánlivou validitu má test, o kterém i laická zkoumaná osoba dokáže říci, co se testem zjišťuje. U výkonových testů zvyšuje motivaci.

3. Faktorová validita

Test se posuzuje na základě faktorové analýzy dat představujících položky testu. Test musí být zadán dostatečně velkému počtu osob.

Validita vztažená ke kritériu (criterion validity)

Je to empirická validita v tom smyslu, že hledáme shodu mezi stanoveným kritériem a výsledky dosaženými v testu. Souběžnou validitu zjišťujeme časově souběžným porovnáním výsledky v testu a vnějším kritériem (jiným testem, aktuálním školním prospěchem, klinickým vyšetřením), prediktivní validita je shoda mezi výsledky v testu a pozdějším výkonem studenta nebo zaměstnance.

Konstruktová validita

Konstruktová validita je míra, v níž test skutečně reprezentuje určitý teoreticky stanovený konstrukt.

Konstruktová validizace představuje sled úkonů:

1. definování daného konstruktu,
2. tvorba položek založená na racionální analýze daného konstruktu,
3. hledání vhodných kritérií. Kritéria nebudou vybrána náhodně, ale musí vycházet z teoretické formulace konstruktu.

Příklad: Pokud bychom zjišťovali konstruktovou validitu Testu studijních předpokladů TSP, definovali bychom mj. konstrukt „užívání logického úsudku při řešení myšlenkových problémů při

výuce“. Nejdříve bychom popsali nároky určitého předmětu na psychické procesy: např. vnímání zrakové a sluchové, paměť konkrétní a pro obecné pojmy, představy, myšlenkové operace. Se znalostí didaktiky vysokoškolského předmětu bychom zjistili, že ke správnému řešení může student dospět myšlenkovou operací analogie, analýza a syntéza, zobecňování, a to buď jakoukoli z těchto operací nebo jen některou z nich, např. analýzou a syntézou. Můžeme korelovat výsledky v TSP s výsledky testu zjišťujícím u jednotlivce úroveň analýzy a syntézy (kterou chápeme jakou vyšší operaci než pouhou analogii z hlediska hierarchie operací užívaných při vědecké práci).

Diskuse o konstruktové validitě výkonových testů by neměla v diskusích ustoupit úvahám o validitě kriteriální (porovnání výsledků ve výkonovém testu s úspěšností ve studiu posuzované pokračováním či nepokračováním ve studiu, počtem kreditů atp.). Tím může být podán statistický důkaz, že TSP při výběru přispívá ke zvýhodnění uchazečů, kteří mají předpoklady užívat vývojově vyšší stupně myšlenkových operací.

Konstrukce testu

Motivem ke vzniku testu je obvykle požadavek najít vhodný prostředek ke zjišťování nějakého znaku osobnosti. První problém je tedy problém validity. Můžeme koncipovat testy s různě širokou validitou. Čím má test širší validitu, tím je méně přesný, tím méně přesné výsledky poskytuje.

Analýza položek a kontrola validity by měly být vždy prováděny na novém výběru pokusných osob.

Homogenita a validita položek

Položky mohou být více nebo méně homogenní. Homogenita znamená obsahovou příbuznost položek při zachování jejich vzájemné nezávislosti. Homogenní jsou položky tehdy, když měří jeden znak osobnosti, když spolu středně vysoce korelují (ale mají přitom různou obtížnost) a když jedna druhou nemůže být nahrazena.

Heterogenita položek je obsahová mnohotvárnost, která je na místě v testech, které měří obtížně izolovatelný znak nebo celý komplex znaků osobnosti. Obvykle, když se vyskytne potřeba měřit určitý znak osobnosti, bývají zpočátku testy heterogenní; po analýze se zjišťují položky a vytvářejí se jednotlivé testy, určené pro jednotlivé aspekty. Tyto testy jsou již homogenní.

Klasická položková analýza

Postup analýzy

Test administrujeme dostatečně velkému vzorku osob v rámci předvýzkumu. Požadovaný rozsah vzorku uvádějí různí autoři různý (od 63 do 350 osob).

Výpočet indexu obtížnosti:

Index obtížnosti vyjadřuje, jak jsou jednotlivé položky snadné nebo obtížné. U tzv. úrovňových výkonových testů (kde se jednotlivé položky svou obtížností mají lišit) je na místě hovořit o obtížnosti nejen ve statistickém, ale i psychologickém smyslu. Čím je položka obtížnější, tím méně osob ji zodpoví diagnosticky (správně). Zjišťování indexů obtížnosti zde má ten smysl, abychom položky mohli seřadit od nejlehčích k nejtěžším a abychom měli zaručeno, že vzestup obtížnosti bude plynulý (pokud není záměrem testu něco jiného).

Protože index obtížnosti je vlastně procento osob, které na položku odpovídají diagnosticky (u výkonových testů správně), je položka tím snadnější, resp. populárnější, čím je index číselně vyšší. Z jeho matematické povahy plyne, že může nabývat hodnot od 0 do 100.

Analýza obtížnosti položek by měla vést k vyloučení položek, jejichž obtížnost (populárnost) je vyšší než 80-85% a nižší než 10-15%, neboť nemá smysl, aby v testu byly položky, které prakticky všichni vyřeší (odpoví stejným způsobem), nebo které téměř nikdo nevyřeší. U didaktických nebo výkonových testů schopností, kde chceme vzbudit zájem o řešení i u nejméně zdatných dětí, však velmi snadné položky zařadíme.

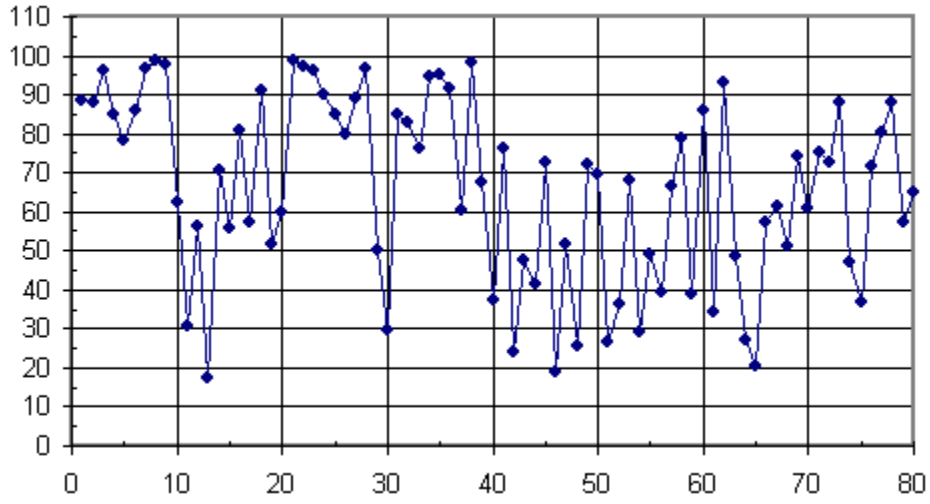
1. **Index obtížnosti dichotomických položek** se nejčastěji vypočte tak, že zjistíme, kolik procent osob z celého souboru zodpovědělo analyzovanou položku diagnosticky (dobře, správně, v souladu s očekáváním). Matematicky řečeno: obtížnost položky je rovna relativní četnosti diagnostických odpovědí násobené 100. Znamená to, že absolutní počet osob, které položku zodpověděly dělíme absolutním počtem všech osob, které položku zpracovaly (dělitel se nemusí rovnat počtu všech osob, které analyzovaný test řešily, protože některé mohly analyzovanou položku z nejrůznějších důvodů vynechat). Položky s vícenásobnou volbou je možno chápat jako dichotomické položky, nesprávné odpovědi se sečítají.

Příklad: Pro všech 24 verzí administrovaných souboru 29 451 uchazečů o přijetí ke studiu na MU v Brně byly spočítány indexy obtížnosti jednotlivých položek podle vzorce

$P = \frac{n_S}{n} \cdot 100$, kde n_S je počet osob, které danou položku v dané verzi vyřešily správně a n je celkový počet osob řešících danou verzi.

Verze 05

Položka	P	Položka	P	Položka	P	Položka	P
pol1	88,59	pol21	98,74	pol41	76	pol61	34,15
pol2	88,04	pol22	97,09	pol42	23,92	pol62	93,23
pol3	96,38	pol23	96,38	pol43	47,44	pol63	48,7
pol4	84,82	pol24	90,17	pol44	41,38	pol64	27,22
pol5	78,21	pol25	84,82	pol45	72,78	pol65	20,22
pol6	86,15	pol26	80,02	pol46	19,12	pol66	57,44
pol7	96,46	pol27	89,14	pol47	51,93	pol67	61,29
pol8	98,74	pol28	96,54	pol48	25,49	pol68	51,3
pol9	97,56	pol29	50,28	pol49	71,91	pol69	73,96
pol10	62,47	pol30	29,74	pol50	69,71	pol70	60,9
pol11	30,84	pol31	85,05	pol51	26,51	pol71	75,37
pol12	56,49	pol32	82,77	pol52	36,11	pol72	72,46
po13	17,15	pol33	76,4	pol53	68,21	pol73	88,04
pol14	70,42	pol34	94,65	pol54	28,95	pol74	46,89
pol15	55,55	pol35	95,04	pol55	49,25	pol75	36,82
pol16	80,65	pol36	91,58	pol56	39,5	pol76	71,6
pol17	57,36	pol37	60,5	pol57	66,4	pol77	80,17
pol18	91,11	pol38	98,43	pol58	78,84	pol78	87,88
pol19	51,53	pol39	67,58	pol59	38,71	pol79	57,44
pol20	59,95	pol40	37,29	pol60	86,07	pol80	65,07



Index obtížnosti se nachází v intervalu <35; 75> u 35 položek, což je 43,75 %.

Výpočet rozlišovací účinnosti

Pro výpočet indexu je v literatuře uváděno více jak 20 různých postupů.

Koeficient rozlišovací účinnosti položky se počítá jako Pearsonův korelační koeficient mezi touto položkou a celkovým hrubým skóre testu. Za vyhovující jsou považovány položky, jejichž koeficient rozlišovací účinnosti nabývá hodnot $\geq 0{,}3$.

Normalizace testu

Na dobrém testu požadujeme, aby citlivě rozlišoval mezi lidmi, kteří se skutečně liší velikostí zkoumaného znaku. Výsledky (hrubá skóre) testu tedy musí mít určitou interindividuální variabilitu. P. Říčan (1969) uvádí pro testy schopností, že rozpětí hrubých skóre standardizační skupiny má být alespoň 10 jednotek, což znamená, že výsledky mají nabývat alespoň 10 různých hodnot. Čím větší je variační rozpětí výsledků, tím citlivěji můžeme rozlišovat mezi jedinci. Hrubá skóre nám však sama o sobě neumožňují srovnávat velikost rozdílů dvou osob v témž testu, ani výsledky téže osoby ve více testech, které jsou nestejně dlouhé a obtížné. Abychom taková srovnání mohli udělat, musí být test normalizován.

Před každou normalizací bychom se měli příslušnými statistickými testy přesvědčit:

1. zda je rozložení normální,
2. zda není významný rozdíl mezi pohlavími, různými sociálními skupinami v rámci normalizačního výběru ap.

Můžeme porovnávat rozložení různých normalizačních podskupin, a kromě toho tak můžeme i posuzovat přesnost a citlivost testu v jednotlivých pásmech hrubých skórů. Např. záporné zešikmení znamená, že test dobře rozlišuje v pásmu podprůměru, kladné zešikmení ukazuje na velkou obtížnost testu, ale zároveň i vyšší citlivost v pásmu nadprůměru. Ploché rozložení napovídá celkově větší citlivost, kdežto strmé rozložení znamená malou citlivost v pásmu kolem průměru

----oOo----

Jiří Dan 31.1.2007