PSIHOMETRIČNE LASTNOSTI OCENJEVALNIH INSTRUMENTOV *PSYCHOMETRIC PROPERTIES OF ASSESSMENT INSTRUMENTS*

izr. prof. dr. Gaj Vidmar, univ. dipl. psih.,[1,2,3] doc. dr. Miroljub Jakovljević, univ. dipl. org., viš. fiziot.[4]
[1]Univerzitetni rehabilitacijski inštitut Republike Slovenije – Soča, Ljubljana
[2]Univerza v Ljubljani, Medicinska fakulteta, Inštitut za biostatistiko in medicinsko informatiko, Ljubljana
[3]Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Koper
[4]Univerza v Ljubljani, Zdravstvena fakulteta, Ljubljana

Povzetek

Prispevek podaja zgoščen pregled osnovnih psihometričnih pojmov, ki so potrebni za razumevanje in uporabo ocenjevalnih postopkov na področju rehabilitacije. Osrednji del je namenjen vidikom in ocenjevanju zanesljivosti in veljavnosti, pri čemer se prispevek opira predvsem na klasično testno teorijo. Pri vseh obravnavanih temah so navedeni primeri iz rehabilitacijskega okolja in temeljna literatura za nadaljnji študij. Na koncu so izpostavljena nekatera odprta vprašanja.

Ključne besede:

merske lastnosti; ocenjevalna orodja; merski instrumenti; zanesljivost; veljavnost; občutljivost

Summary

The paper presents a condensed overview of the basic psychometric concepts that are required for understanding and applying assessment procedures in the field of rehabilitation. The central part addresses the aspects and estimates of reliability and validity, based mainly on the classical test theory. Examples from rehabilitation setting and key references for further study are given for every topic. In the conclusion, some open issues are highlighted.

Key words:

measurement characteristics; assessment tools; measurement instruments; reliability; validity; sensitivity

UVOD

Merjenje je ključen del vseh empiričnih naravoslovnih in tudi družboslovnih znanosti. „Vsaka stvar, ki obstaja, je v določeni količini“, je leta 1918 na začetku svoje razprave o naravi, postopkih in namenu merjenja zapisal Edward Lee Thorndike, eden od pionirjev merjenja v šolstvu in psihologiji (1). „Vse, kar obstaja, pa je mogoče meriti“, je dodal njegov naslednik William A. McCall (2), torej je ključno vprašanje vsake znanosti in stroke, kako meriti oziroma kakšnim zahtevam naj merjenje ustreza.

Nekoliko širši pojem od merjenja je ocenjevanje. Če razširimo eno od definicij s področja rehabilitacije (10), lahko rečemo, da je ocenjevanje (angl. *assessment*) proces izbiranja in uporabe različnih orodij za zbiranje podatkov in različnih virov z namenom podpore odločanju. V zdravstvenem okviru je vrednotenje (angl. *evaluation*) sestavni del širšega procesa ocenjevanja, osrednjo vlogo pri vrednotenju pa imajo mere izida (angl. *outcome measures*) (10). Vrednotenje v tem kontekstu vključuje zbiranje podatkov, ki omogoča strokovnjaku oz. raziskovalcu, da presodi o količini preučevanega konstrukta oz. o vrednosti terapevtskih ukrepov za obravnavanega posameznika ali populacijo. Preizkus (test) s tega vidika pomeni standardizirano obliko pregleda posameznika ali skupine, ki določa prisotnost ali količino določene zmogljivosti, znanja ali spretnosti, izid (angl. *outcome*) pa pomeni opaženo ali izmerjeno posledico ukrepa oz. terapevtske obravnave (10).

KLASIČNA TESTNA TEORIJA

Na področju rehabilitacije imamo pogosto opravka s posrednim merjenjem, saj merska enota ni enaka predmetu merjenja oz. skušamo meriti hipotetične konstrukte, zato merski postopki slonijo na psihometrični teoriji. Temeljna učbenika s tega področja sta delo Lorda in Novicka (3) ter Nunnallyja in Bernsteina (4). V slovenščini so na voljo tri monografije: osnove (psihološkega) testiranja na celovit, a široko razumljiv način obravnava Bucikov učbenik (5); psihometrični računski postopki so podrobno opisani

Poslano: 4. 3. 2016
Sprejeto: 17. 3. 2016
Naslov za dopisovanje / Address for correspondence (GV): gaj.vidmar@ir-rs.si

in razloženi v sodobnejšem Sočanovem učbeniku (6); Ferligoj, Leskovšek in Kogovšek (7) pa obravnavajo psihometrično tematiko predvsem z vidika merjenja v družboslovju. Enega najsodobnejših, najlažje razumljivih in najbolj praktično usmerjenih učbenikov psihometrije sta napisala Furr in Bacharach (8). Kot kratek priročnik v slovenščini bližnjem jeziku so priporočljiva tudi starejša Petzova skripta (9).

Psihometrija se je razvila na podlagi klasične testne teorije (KTT), imenovane tudi teorija pravega dosežka (angl. *classical test theory* oz. t*rue score theory*). KTT pravi, da je opaženi oz. izmerjeni dosežek (X) sestavljen iz pravega dosežka (T) in napake merjenja (E):

$$X = T + E \qquad [1].$$

Glavni merski lastnosti testov sta zanesljivost in veljavnost, zato njuna analiza predstavlja glavnino preučevanja merskih lastnosti ocenjevalnih instrumentov v praksi in tudi glavnino pričujočega prispevka, pri čemer je zanesljivost obravnavana z vidika KTT. Pri tem je potrebno poudariti, da se zanesljivost in druge merske lastnosti v resnici vedno nanašajo na celoten merski postopek, torej ne le na instrument sam, pač pa na uporabo instrumenta v določenih okoliščinah (z določenimi navodili in v določenih okoljskih pogojih).

A izhodišče so psihometrični testi, ki jih vedno sestavlja večje število postavk oz. nalog in izmerjeni dosežek dobimo tako, da seštejemo dosežke na posameznih postavkah. Zato je prvi korak razvoja novega testa ali preverjanja merskih lastnosti obstoječega testa analiza merskih lastnosti posameznih postavk.

Analiza postavk

Ne pozabimo, da pri statističnem sklepanju na podlagi končnega vzorca enot (to so pri psihometričnih analizah testiranci) sklepamo na hipotetično neskončno populacijo, zato so količine, izračunane na vzorcu, vedno le cenilke populacijskih parametrov. Pri analizi postavk s preprostimi opisnimi statistikami ocenimo:

- težavnost vsake posamezne postavke – pri dihotomnih postavkah (tj. takih, kjer sta možna dva odgovora, od katerih je eden pravilen oz. ga točkujemo z 1, drugi pa napačen oz. ga točkujemo z 0) je enaka deležu pravilnih odgovorov. Psihometrična težavnost ima torej nasproten pomen kot beseda težavnost v vsakdanjem jeziku, saj imajo naloge, ki jih pravilno reši večji delež testirancev (in so torej lažje), višjo ocenjeno težavnost;
- korelacije med postavkami (imenovane tudi interkorelacije);
- popravljeno korelacijo postavke s skupnim dosežkom (tj. s skupnim dosežkom na testu brez postavke, za katero korelacijo računamo).

ZANESLJIVOST

KTT predpostavlja, da sta pravi dosežek in napaka merjenja med seboj neodvisna, zato je varianca izmerjenega dosežka enaka vsoti varianc pravega dosežka in napake [2]. V skladu s tem je koeficient zanesljivosti r_{xx} definiran kot razmerje med varianco pravega dosežka in varianco izmerjenega dosežka [3]:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \qquad [2],$$

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \qquad [3].$$

Koeficienti zanesljivosti torej zavzemajo vrednosti med 0 (kar bi pomenilo, da „merimo“ zgolj šum) in 1 (kar predstavlja idealno merjenje brez napak). Pri psihometričnih testih sta zaželena dva vidika zanesljivosti:

- zanesljivost kot stabilnost testnega dosežka v času in
- zanesljivost kot notranja skladnost testa.

Zanesljivost kot stabilnost testnega dosežka v času

Zanesljivost kot stabilnost testnega dosežka v času lahko ocenjujemo z **metodo vzporednih** (enakovrednih, alternativnih) **oblik testa** ali s **ponovljenim testiranjem** (t.i. retestno metodo). Z obema ocenimo zanesljivost tako, da izračunamo korelacijo med dobljenima dosežkoma. Vzporedni obliki testa morata imeti enako vsebino, strukturo in način odgovarjanja na postavke. Ta metoda se zelo redko uporablja, saj je vlaganje časa in sredstev v razvoj dveh enakovrednih oblik testa redko smiselno. Pri ponovljenem testiranju pa je ključna zahteva, da so pogoji enaki kot pri prvem testiranju. Zato ta metoda ni primerna, če testiranje povzroči učinek učenja ali utrujanja ali se testirana lastnost hitro spreminja (v primerjavi s časovnim razmikom med testiranjema).

Zanesljivost kot notranja skladnost

Zanesljivost kot notranjo skladnost psihometričnega testa v okviru KTT preverjamo na dva načina: s **Cronbachovim koeficientom** α in z oceno **zanesljivosti preko razpolovitve testa** (angl. *split-half reliability*). Koeficient α je za test s k postavkami definiran kot

$$\alpha_k = \frac{k}{k-1}\left[1 - \left(\sum_{i=1}^{k} \sigma_i^2 \Big/ \sigma_X^2\right)\right] \qquad [4],$$

pri čemer so σ_i^2 variance odgovorov na posamezne postavke. Pri oceni zanesljivosti z razpolovitvijo testa razvrstimo postavke v enakovredni polovici (npr. na lihe in sode postavke) in zanesljivost ocenimo na podlagi dobljene k orelacije med skupnima dosežkoma na obeh polovicah (r_{hh}). Vstavimo jo v obrazec [5], upoštevajoč $\alpha_k = r_{hh}$ in $k'/k = 2$ (saj ima celotni test dvakrat toliko postavk kot vsaka od polovic). V splošnem namreč velja, da z naraščanjem števila postavk narašča tudi zanesljivost testa; kako se spremeni zanesljivost testa, če ga namesto k sestavlja k' postavk, pa pove Spearman-Brownov obrazec:

$$\alpha_{k'} = \frac{(k'/k)\alpha_k}{1 + [(k'/k) - 1]\alpha_k} \qquad [5].$$

Če obrazec preuredimo, lahko izračunamo, za koliko postavk bi morali podaljšati test, da bi namesto ocenjene (prenizke)

zanesljivosti dobili izbrano višjo zanesljivost, oz. za koliko postavk lahko skrajšamo (predolg) test, ne da bi zanesljivost padla pod izbrano mejo.

Koeficient α ocenjuje spodnjo mejo zanesljivosti, torej je dejanska zanesljivost testa najmanj tolikšna, kot kaže α. Z matematičnega vidika je vrednost α enaka povprečju vseh možnih koeficientov zanesljivosti razpolovitve testa, ki bi jih z različnimi delitvami postavk lahko izračunali za dane podatke.

Pri ocenjevanju zanesljivosti z vidika notranje skladnosti je zaželeno, da je test enorazsežen, torej da vse postavke odražajo isti enodimenzionalni konstrukt. Le v primeru enorazsežnosti namreč z merami notranje skladnosti ustrezno ocenimo zanesljivost; poleg tega pri večrazsežnih testih skupni dosežek vseh postavk ni smiseln z vsebinskega vidika. Enorazsežnost preverjamo s faktorsko analizo – eksploratorno, kadar razvijamo nov test, oz. konfirmatorno, kadar preizkušamo obstoječega (konkretni postopki presegajo namen pričujočega prispevka; opisani so v psihometričnih in statističnih učbenikih). Visoka vrednost koeficienta α torej še ne pomeni, da je lestvica enorazsežna oz. da je test ustrezen!

Tudi glede tega, kolikšna vrednost α je „dovolj visoka“, je med raziskovalci in v literaturi veliko napačnih predstav. Preprostega odgovora na to vprašanje ni in delitve vrednosti α v razrede so namenjene zgolj lažjemu sporazumevanju, ne pa presojanju. Visoko notranje skladni so npr. standardizirani skupinski testi inteligentnosti, ki imajo vrednost α okoli 0,95; o zmerni notranji skladnosti govorimo pri vrednostih α okoli 0,85 (ki jo dosega večina standardiziranih testov v psihologiji in zdravstvu); srednje visoka vrednost α je okoli 0,75 (opažena npr. pri objektivnih testih znanja); notranja skladnost je nizka, če je vrednost α okoli 0,65. Pri $\alpha = 0{,}50$ imata pravi dosežek in merska napaka enak vpliv na izmerjeni dosežek. Presoja ustreznosti zanesljivosti je odvisna od uporabe testa – če je namenjen individualni klinični diagnostiki, so potrebne višje vrednosti kot npr. za korelacijske analize v raziskavah. Izjemoma lahko pri analizi empiričnih podatkov dobimo tudi negativno vrednost α, kar je praviloma znak, da smo nekatere postavke točkovali v napačni smeri oz. pozabili obrniti njihovo točkovanje (obračanje zahtevajo npr. postavke na testu funkcijske sposobnosti, pri katerih ocenjujemo nezmožnost oz. težave namesto zmožnosti oz. uspeha).

Zanesljivost, ponovljivost in obnovljivost

Ponekod v literaturi naletimo na ločevanje pojmov ponovljivost in zanesljivost. Ponovljivost (angl. *repeatability*) se v tem primeru nanaša na neposredno zaporedne meritve (npr. pri trikratnem zaporednem merjenju kota gibljivosti z goniometrom), zanesljivost pa na stabilnost izmerjenega dosežka skozi daljši čas (npr. med meritvama, opravljenima v razmiku enega tedna). V obeh primerih za ocenjevanje uporabimo iste statistične metode, ki so opisane v nadaljevanju.

V okviru statističnega nadzora procesov (angl. *statistical process control,* SPC) oz. statističnega nadzora kakovosti (angl. *statistical quality control,* SQC), torej tudi marsikje v zdravstvu (npr. na področju farmakologije in laboratorijske tehnike), se kot vidika zanesljivosti oz. natančnosti merjenja loči ponovljivost in obnovljivost (angl. *reproducibility*). Prva se nanaša na zanesljivost merskih postopkov, če jih ponovimo v enakih pogojih (tj. isti merilci z istimi instrumenti v istem laboratoriju), druga pa na ponovitev merjenja v nekoliko spremenjenih pogojih (druge merilce, instrumente oz. laboratorije). Tako pojmovanje priporoča temeljni mednarodni dokument s področja meroslovja – Vodilo za izražanje merilne negotovosti (Guide to the Expression of Uncertainty of Measurement, GUM (11)) – in predpisujejo mednarodni standardi (npr. ISO 5725 (12)). Skupni naziv za celotno področje je **analiza merilnih sistemov** (angl. *measurement system analysis*, MSA), za študije oz. statistične postopke ocenjevanja ponovljivosti in obnovljivosti merjenja pa je v rabi oznaka **GRR** (*gage R&R oz. gage repeatability and reproducibility*) (13).

Skladnost med meritvami ali ocenjevalci

Skladnost med meritvami oz. ocenjevalci (angl. *agreement*) je zelo obsežno in pomembno področje, saj je v zadnjih desetih letih izšlo šest statističnih monografij, posvečenih problematiki skladnosti (14-19), pri čemer vsak avtor poleg pregleda temeljnih metod poglobljeno obravnava lasten pristop. Posebej priporočljiv je Gwetov učbenik (19), ki je izčrpen, matematično ni zahteven in ga spremlja uporabniku prijazno programje. Tooth in Ottenbacher (20) opozarjata na pomen in pasti analize skladnosti na področju rehabilitacije in podajata pregled metod za opisne podatke. V nadaljevanju so povzeti osnovnih pojmi in pristopi k analizi skladnosti glede na mersko raven podatkov.

Nominalni podatki

Najnižja raven merjenja je kvalitativno razvrščanje v kategorije, s čimer dobimo podatke na nominalni (tj. opisni imenski) merski ravni. Primer tovrstnih podatkov so diagnoze v skladu z MKB (21, 22) ali kode v skladu z MKF (23, 24). Skladnost dveh ocenjevalcev (npr. specializanta in specialista glede relevantnosti izbrane kode MKF pri pacientih v izbrani ambulanti) lahko povzamemo z različnimi merami.

- Najpreprostejša je **delež skladnosti**. Če ocenjujemo dvojiško lastnost in uporabimo oznake iz Tabele 1 (kjer gre sicer za skladnost napovedi in dejanskega stanja, a računsko sta problema enaka), znaša ocenjeni delež skladnosti $(a + c)/n$.
- Najbolj znana in najširše uporabljana mera skladnosti dveh ocenjevalcev je **Cohenov koeficient κ**. Definiran je na podlagi opaženega deleža skladnosti (p_o) in deleža skladnosti, pričakovanega po naključju (tj. če bi oba ocenjevalca o razvrstitvi vsakega ocenjevanca medsebojno neodvisno ugibala, vsak s svojimi verjetnostmi razvrstitve v kategorije v skladu z opaženimi skupnimi deleži kategorij; p_e):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \qquad [6].$$

Če je ocenjevana lastnost dvojiška in uporabimo oznake iz Tabele 1, je $p_o = (a+d)/n$ in $p_e = \{[(a + c)/n] \times [(a + b)/n)]\} + \{[(b + d)/n] \times [(c + d)/n)]\}$. Vrednosti κ so v redkih primerih

lahko negativne, kar pomeni še slabšo skladnost od naključno pričakovane (ko je $\kappa = 0$), najvišja možna vrednost pa je 1, ki pomeni, da sta oba ocenjevalca vse ocenjevance razvrstila enako. Glavna pomanjkljivost koeficienta κ je, da je močno odvisen od pogostnosti kategorij (enako kot sta pozitivna in negativna napovedna vrednost odvisni od prevalence, kot je opisano v razdelku o zaželenih lastnostih diagnostičnih testov). Zato poimenovanje velikostnih razredov za κ – enako kot za druge mere zanesljivosti oz. skladnosti – ne more služiti kot argument za dokazovanje ustreznosti merskega instrumenta. Poleg Fleissove (25) je sicer najpogosteje v rabi Altmanova delitev (26) na nizko ($\kappa < 0{,}20$), zmerno (0,21 – 0,40), srednjo (0,61 – 0,80) in visoko stopnjo skladnosti ($\kappa > 0{,}80$).

- Za istovrstne podatke obstajajo še drugi koeficienti skladnosti, med katerimi je najpomembnejši Scottov Π. Razširitev Π na več ocenjevalcev predstavlja Fleissov koeficient κ. Najsodobnejši in najbolj statistično dognan je **Gwetov koeficient AC1** (19), ki se ga da posplošiti na več ocenjevalcev.

Ordinalni podatki

V zdravstvu imamo pogosto opravka z ordinalnimi (tj. opisnimi urejenostnimi) lestvicami, npr. številskimi ocenjevalnimi lestvicami bolečine, stresa ali spastičnosti posamezne mišice. Skladnost med različnimi ocenjevalci, ocenjevalnimi metodami ali ocenami v različnih pogojih odraža različne vidike zanesljivost merjenja, povzeti pa jo moramo z ustreznimi merami.

- Uteženi Cohenov koeficient κ je razširitev običajnega koeficienta κ s tem, da neskladje utežimo manj, če sta ocenjevalca podala bližnji oceni (npr. eden 4 in drugi 5 na petstopenjski ordinalni lestvici), kot če sta podala bolj oddaljeni oceni (npr. 1 in 5). Sodobnejši istovrstni koeficient je Gwetov AC2 (19), ki je primeren tudi za številske podatke.
- Skladnost treh ali več rangiranj imenujemo **konkordanca** (angl. *concordance*). Najbolj znana mera konkordance je Kendallov koeficient W. Zanesljivost ocenjevanja povzema npr. v primeru, ko vsak od štirih zdravnikov na oddelku rangira šest razpoložljivih pacientov glede primernosti za sprejem na bolnišnično rehabilitacijo. Vrednost $W = 0$ pomeni največje možno neskladje med rangiranji, $W = 1$ pa popolno skladnost vseh rangiranj. Nazoren pregled problematike in možnosti za grafični prikaz konkordance podajata Vidmar in Rode (27).
- Za ocenjevanje skladnosti ordinalnih podatkov so na voljo še druge metode, npr. Bangdiwalov koeficient B in Andrésov in Marzov koeficient Δ (28). Na področju rehabilitacije je posebej pomemben pristop Svenssonove (29-31), ki ločuje različne vire variabilnosti ordinalnih ocen (zlasti pristranost od naključnega neskladja). Pristop ni preprost, a se malo uporablja izven skandinavskih dežel, čeprav je za njegovo implementacijo na voljo prosto dostopno programje v obliki elektronskih preglednic s smernicami za uporabo in uporabniškim priročnikom (32).

Številski podatki

Ocenjevanje skladnosti je eno redkih področij uporabe statistike, kjer je smiselno ločevati med podatki na intervalni merski ravni (ki nimajo absolutne ničle in so za njih smiselne le razlike vrednosti, ne pa tudi razmerja) in na razmernostni merski ravni (ki imajo absolutno ničlo in so razmerja vrednosti zanje smiselna). Seveda pa so tudi na področju skladnosti vse metode, ki so uporabne za prve, uporabne tudi za druge.

- Skladnost med meritvami najpogosteje ocenjujemo s **koeficientom intraklasne korelacije** (angl. *intraclass correlation*, ICC). ICC je v splošnem definiran kot razmerje med varianco, ki jo lahko pripišemo razlikam med ocenjevanci, in celotno opaženo varianco meritev (33, 34):

$$ICC = \frac{Var_{\text{med ocenjevanci}}}{Var_{\text{med ocenjevanci}} + Var_{\text{med ocenjevalci}} + Var_{\text{napake}}} \quad [7].$$

ICC doseže največjo možno vrednosti 1, če ni variabilnosti med ocenjevalci (niti variance napake), t.j. če vsakega ocenjevanca vsi ocenjevalci ocenijo enako. Uporaba ICC je za ocenjevanje zanesljivosti primernejša kot uporaba običajnih (interklasnih) korelacijskih koeficientov, kot je Pearsonov r, saj z ICC ločimo različne vire napak in oceno zanesljivosti uskladimo z načrtom študije oz. okoliščinami merjenja. Poznamo namreč različne oblike ICC glede na tri vprašanja: (a) so vsi ocenjevalci (oz. smo z vsemi instrumenti ali ponovitvami) ocenili vse ocenjevance (dvosmerni model) ali je vsak ocenil le svoj slučajni podvzorec ocenjevancev (enosmerni model); (b) so ocenjevalci (instrumenti, metode, ponovitve), ki jih primerjamo, vsi razpoložljivi oz. vsi, ki nas zanimajo (mešani model), ali slučajno izbrani iz populacije vseh možnih ocenjevalcev (slučajni model); in (c) ali nameravamo v praksi uporabiti le eno ocenjevanje (model za posamezno meritev) ali povprečje več ocenjevanj (model za povprečje meritev)? Tako je definiranih šest oblik oz. modelov ICC:

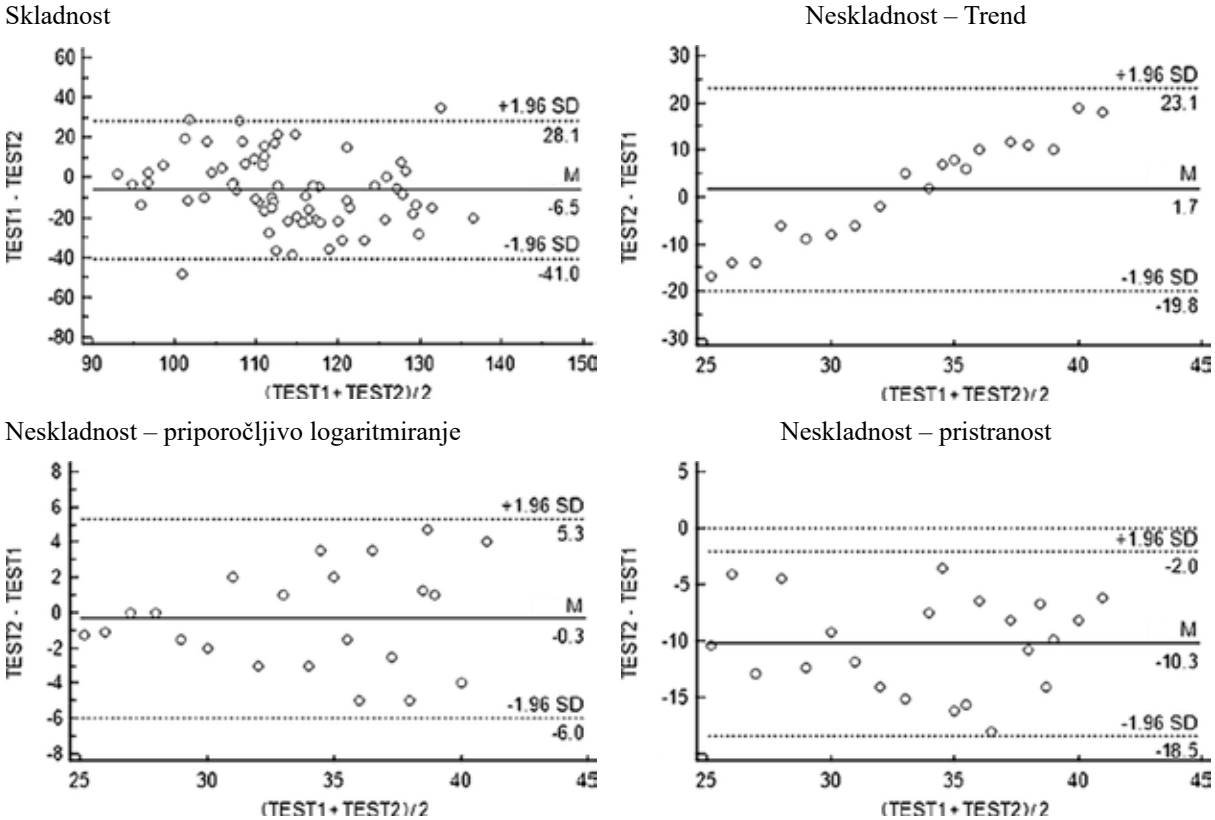
  - ICC(1, 1) = enosmerni model za posamezno meritev
  - ICC(1, k) = enosmerni model za povprečje meritev
  - ICC(2, 1) = dvosmerni slučajni model za posamezno meritev
  - ICC(2, k) = dvosmerni slučajni model za povprečje meritev
  - ICC(3, 1) = dvosmerni mešani model za posamezno meritev
  - ICC(3, k) = dvosmerni mešani model za povprečje meritev

  Pri dvosmernih modelih ločimo še ICC za skladnost v ožjem (absolutnem) pomenu besede in ICC za konsistentnost (tj. relativno skladnost v smislu običajne korelacije, kjer zamik enega od obeh nizov vrednosti v primerjavi z drugim ne vpliva na rezultat), a vseh oblik ICC je šest, saj so nekatere med seboj računsko enake. Če različne ICC izračunamo iz istih podatkov, je ICC(3) > ICC(2) > ICC(1). Omeniti velja tudi, da je ICC(2, 1) za absolutno skladnost asimptotično ekvivalenten uteženemu Cohenovemu κ s kvadratnimi utežmi, ICC(2, k) in ICC(3, k) za konsistentnost pa sta ekvivalentna Cronbachovemu α. Po eni od razvrstitev vrednosti ICC pod 0,50 pričajo o nizki skladnosti oz. zanesljivosti, med 0,50 in 0,70 o srednji, nad 0,70 o dobri oz. visoki, med 0,90 in 1 pa o odlični oz. zelo visoki skladnosti oz. zanesljivosti (35). Toda tudi za ICC velja opozorilo, da moramo vrednosti presojati v širšem kontekstu, ne pa „predalčkati“ med bolj ali manj ustrezne. Vsekakor je pri visoko tveganih odločitvah priporočljiva zelo visoka vrednost ICC za uporabljene mere (36).
- Preprost, učinkovit in zato zelo razširjen način ocenjevanja skladnosti za razmernostne podatke so **meje skladnosti** (angl.

limits of agreement, LoA) oz. metoda **Blanda in Altmana** (37, 38). Osnova te metode je razsevni grafikon odvisnosti razlike med meritvama istega ocenjevanca (npr. novo predlagane in uveljavljene) v odvisnosti od povprečja obeh meritev (od tudi starejše ime za isto metodo – angl. *mean-difference plot*). V grafikon vrišemo vodoravno sredinsko črto, ki seka navpično os v višini opažene povprečne razlike, ter nad in pod njo vodoravni črti, ki ustrezata mejam skladnosti, tj. za 1,96 standardnega odklona opaženih razlik nad oz. pod sredinsko črto (Slika 1). V primeru skladnosti merjenj je pričakovati približno 95 % točk znotraj LoA (saj ti predstavljata 95% tolerančni interval za razlike) in območje skladnosti ne bo pretežno (ali celo v celoti) nad ali pod vodoravno osjo. Morebitno sistematično razliko med merjenjima lahko statistično testiramo s testom *t* za odvisna vzorca (ki je ekvivalenten testu *t* za en vzorec, pri katerem povprečno razliko med merjenjima primerjamo z 0). Pomembna je tudi oblika oblaka točk: trend kaže na sistematično podcenjevanje oz. precenjevanje ene od meritev v primerjavi z drugo, pahljačasta oblika pa kaže, da neskladje med meritvama narašča z velikostjo meritve (torej je priporočljiva logaritemska pretvorba meritev). Metoda Blanda in Altmana je z ustreznimi prilagoditvami uporabna tudi, če za posamezno meritev vzamemo povprečje več zaporednih meritev (39) in če med seboj primerjamo več kot dve meritvi (40). Uporabnost metode LoA in različnih oblik ICC na področju rehabilitacije nazorno predstavljata Rankin in Stokes (41). Bland in Altman sta na podlagi pristopa LoA (kjer razliko med meritvama istega merjenca označimo z *d*, merjencev pa je *n*) definirala tudi **koeficient ponovljivosti** (angl. *coefficient of repeatability*, CR):

$$CR = 2\sqrt{\sum_{i=1}^{n} d_i^2 \Big/ n} \qquad [8].$$

- Poleg pristopa Blanda in Altmana so za istovrstne podatke na voljo še drugi pristopi – Linov koeficient konkordančne korelacije (angl. *concordance correlation coefficient*, CCC) (18, 42, 43) in različni regresijski pristopi. Med slednjimi sta najpomembnejša Demingova regresija (znana tudi kot regresija po metodi glavne osi, angl. *major axis regression*; regresija po metodi glavnih komponent, angl. *principal component regression*; in metoda pravokotne razdalje, angl. *perpendicular distance method*) in neparametrični pristop Passing in Bablocka (44, 45), ki je robusten tudi v primeru osamelcev. Primer uporabe različnih metod za analizo skladnosti razmernostnih podatkov, vključno s CCC in regresijskimi pristopi, na področju rehabilitacije opisuje Vidmar s sodelavkama (46). Najsodobnejši statistični pogled na regresijske pristope v povezavi z LoA podajata Francq in Govaerts (47).
- **Koeficient variacije** (angl. *coefficient of variation*, CV) je relativni standardni odklon, torej standardni odklon (SD) podatkov, deljen z njihovim povprečjem (M) in praviloma izražen v odstotkih, torej pomnožen s 100 (35) [9]. Ocenimo ga na podlagi ponovljenih meritev (po možnosti vsaj treh) in izračunamo za vsako opazovano enoto (tj. za vsakega posameznika). Za celotno skupino oz. vzorec lahko kot populacijsko oceno poročamo povprečni ali medianski CV (48). Nižja vrednost CV priča o večji ponovljivosti merjenja, pri čemer je postavljati splošno mejo sprejemljivosti enako



***Slika 1**: Primeri grafikonov Blanda in Altmana za ugotavljanje skladnosti dveh merskih postopkov oz. instrumentov (za pojasnila glej besedilo).*

neumestno kot za druge mere zanesljivosti. Izračun CV je smiseln le za razmernostne podatke (v rehabilitaciji npr. kot obsega gibanja, silo prijema ali mišični navor), saj je pri intervalnih podatkih povprečje arbitrarno in ga lahko z linearno pretvorbo poljubno spreminjamo.

$$CV(\%) = \frac{SD}{M} \times 100 \quad [9]$$

- Najsplošnejša mera skladnosti, definirana za vse vrste podatkov in prilagojena tudi morebitnim manjkajočim vrednostim, je **Krippendorffov koeficient α**. Njeni posebni primeri so Scottov Π , Fleissov κ, Spearmanov koreficient korelacije rangov ρ in starejša (Pearsonova) oblika ICC (ki se jo izračuna tako, da se nizu dvojic vrednosti doda isti niz z zamenjanima vrednostima v vsaki dvojici ter za dobljeno podatkovje izračuna Pearsonovo korelacijo). Z definicijo Krippendorffovega α niso združljivi Cohenov κ , Cronbachov α, Pearsonov r in zgoraj opisani sodobni (Fisherjev) ICC. Krippendorffov α se je uveljavil kot standard na področju analize vsebine (49), zato je pričakovati njegovo uporabo na področju rehabilitacije predvsem pri ocenjevanju zanesljivosti kodiranja v raziskavah, ki izhajajo iz kvalitativnih podatkov, npr. intervjujev s pacienti ali dnevniških zapisov.

Občutljivost in ugotavljanje sprememb

Pri analizi zanesljivosti merskega postopka je smiselno hkrati uporabiti različne metode in pri interpretaciji upoštevati njihove različne poudarke (50). To še posebej velja za problematiko občutljivosti kliničnih instrumentov oz. ugotavljanja sprememb, ki je zapletena zaradi različnih poimenovanj in različnih obrazcev za ocenjevanje iste mere ter številnih razhajajočih se mnenj v literaturi. Obstajata dva glavna pristopa, vsak s svojimi definicijami in merami – porazdelitveni (angl. *distributional*) in sidrni (angl. *anchor-based approach*).

Porazdelitveni pristop

Če poznamo zanesljivost testa, lahko ločimo pravi dosežek od izmerjenega. Iz enačb [2] in [3] namreč sledi preprost izraz za varianco napake meritev v odvisnosti od variance izmerjenih dosežkov in koeficienta zanesljivosti, pri čemer koren variance napake meritev (tj. standardni odklon napake meritev) imenujemo **standardna napaka merjenja** (angl. *standard error of measurement*, SEM):

$$\sigma_E^2 = \sigma_X^2 \left(1 - r_{XX}\right) \ , \ SEM = \sqrt{\sigma_E^2} = \sigma_X \sqrt{1 - r_{XX}} \quad [10].$$

SEM je odvisna od vrste ocenjene zanesljivosti – če r_{XX} ocenimo s ponovnim testiranjem, lahko določimo SEM za ponovno testiranje (ki bo najbrž večja, če je čas med ponovitvama daljši), ki bo drugačna od SEM, ocenjene na podlagi vzporednih polovic testa. Poleg tega je potrebna previdnost pri ocenjevanju intervala zaupanja (IZ), saj SEM opisuje razpršenost okrog pravega dosežka, ne okrog izmerjenega, ki je zaradi slučajne napake merjenja za neznano razliko in v neznano smer oddaljen (tj. večji ali manjši) od pravega. S SEM bi lahko torej predvideli, kje (z določeno verjetnostjo) pričakujemo izmerjeni dosežek pri znanem pravem dosežku, ne obratno. Če želimo predvideti pravi dosežek pri znanem izmerjenem dosežku, potrebujemo mero, ki se v psihometriji imenuje **standardna napaka ocene** (angl. *standard error of the estimate*, SEE). S preprosto izpeljavo (3, 51) dobimo

$$SEE = \sigma_X \sqrt{r_{XX}\left(1 - r_{XX}\right)} \quad [11].$$

IZ za dosežek posameznega testiranca ocenimo tako (52), da najprej na podlagi znanega povprečnega dosežka vseh testirancev ocenimo testirančev pravi dosežek, nato pa širino IZ okrog njega izračunamo tako, da SEE (dobljeno s poznavanjem standardnega odklona dosežkov vseh testirancev in koeficienta zanesljivosti) pomnožimo z vrednostjo z, ki ustreza izbrani stopnji zaupanja [12]. Največkrat uporabimo $z_P = 1{,}96$ za $P = 95$ %.

$$T = \bar{x} + r_{XX}(x - \bar{x}), \ IZ_T = T \pm z_P \times SEE \quad [12]$$

Razlika med SEM in SEE (in med intervaloma zaupanja, izračunanima z eno ali z drugo) bo večja (in torej IZ, izračunan iz SEM, slabši približek), če bo merjenje manj zanesljivo (tj. vrednost r_{XX} nižja). Seveda tako za SEM kot za SEE (in pripadajoča IZ) velja, da večje vrednosti (oz. širše meje IZ) pomenijo manj zanesljivo merjenje. Z manj zanesljivimi meritvami tudi težje zaznamo spremembe, ki so posledica preučevanega ukrepa oz. zdravljenja. Za veljavno sklepanje o razlikah med dvema meritvama ($X1$ in $X2$) moramo oceniti zanesljivost merjenja razlik. V ta namen se najpogosteje uporablja **najmanjša dejanska razlika** (angl. *minimum real difference*, MRD), imenovana tudi najmanjša zanesljiva ali zaznavna sprememba (angl. *smallest reliable change oz. minimal detectable change*, MDC). Obrazci za njeno oceno, ki jih najdemo v literaturi, so različni, a ne bistveno; najpogostejša je ocena MRD na podlagi standardne napake merjenja ob prvem merjenju (SEM_{X1}) (53):

$$MRD = z_P \times \sqrt{2} \times SEM_{X1} \quad [13].$$

Razlika med dvema meritvama mora znašati najmanj MRD, da lahko s 95 % zanesljivostjo sklepamo o ponovljivosti razlike med pravima vrednostima v nadaljnjih študijah oz. (preprosteje povedano, čeprav ne povsem matematično korektno) s 95 % gotovostjo verjamemo, da je razlika, ki smo jo izmerili, posledica dejanske razlike v merjeni lastnosti, ne pa le slučajne napake merjenja.

Indeks zanesljive spremembe (angl. reliable change index, RCI) je vsebinsko enak MRD, le računsko obrnjen (54) [14]. RCI nad z_P (1,96 oz. zaokroženo 2, če uporabimo stopnjo zaupanja P=95 %) pomeni, da je pri testirancu po vseh verjetnosti dejansko nastopila sprememba:

$$RCI = \frac{x_2 - x_1}{\sqrt{2} \times SEM_{X1}} \quad [15].$$

Za presojanje o spremembah se v okviru porazdelitvenega pristopa uporabljajo še druge mere, zlasti spodnji dve. Prva je namenjena presojanju o spremembi posameznega testiranca, druga pa o spremembi povprečnega dosežka testirancev.

- **Velikost učinka** (angl. *effect size*, ES) je v kontekstu ugotavljanja sprememb pri posameznih testirancih definirana kot razlika med meritvama, deljena s standardnim odklonom meritev ob prvem merjenju (55) [16]. Z zadržki, ki veljajo za vse tovrstne razvrstitve, lahko spremembo z ES do 0,2 označimo kot majhno, med 0,2 in 0,6 kot srednje veliko in nad 0,6 kot veliko.
- **Standardizirani povprečni odziv** (angl. *standardised response mean*, SRM) je tesno povezan z metodo Blanda in Altmana. Definiran je kot razlika med povprečji meritev, deljena s standardnim odklonom razlike meritev (56) [17; oznaki *d* in *n* kot pri 8]. SRM je torej povezan tudi s testno statistiko testa *t* za odvisna vzorca oz. ponovljene meritve [17; *t* je testna statistika]:

$$ES = \frac{x_1 - x_2}{\sigma_{X1}} \quad [16],$$

$$SRM = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_d} = t\sqrt{n} \quad [17].$$

Sidrni pristop

Potrebo po „zasidranosti" presojanja o dejanskih spremembah v zunanje kriterije posebej poudarja literatura s področja kakovosti življenja. **Najmanjša pomembna sprememba** oz. **razlika** (*minimal important change*, MIC, oz. *difference*, MID), imenovana tudi najmanjša **klinično** (angl. *clinically*) pomembna razlika (MCID), je v okviru tega pristopa definirana kot najmanjša sprememba mere izida zdravljenja, povezana z najmanjšo škodljivostjo in najnižjimi stroški, ki zadošča, da sproži spremembo vodenja pacienta (57). V praksi se jo ocenjuje na dva načina.

- Prva možnost je, da MID ocenimo kot povprečno spremembo dosežka pri pacientih, za katere sidro (tj. zunanji kriterij) pokaže, da so dosegli najmanjše pomembno izboljšanje (57). Na tak način bi npr. MID ocenili kot povprečen napredek na lestvici dnevnih aktivnosti pri bolnišničnih rehabilitacijskih pacientih, ki so po rehabilitaciji na ordinalni lestvici zadovoljstva z oskrbo podali za eno stopnjo višjo oceno.
- Druga možnost, ki dandanes prevladuje (58), temelji na metodologiji analize diagnostičnih testov (ki je kratko predstavljena v nadaljevanju v kontekstu veljavnosti) in na krivuljah ROC (angl. *receiver operating characteristics*). Spremembo dosežka na testu obravnavamo kot diagnostični test, sidro, ki razdeli populacijo pacientov na tiste, ki so dosegli najmanjšo pomembno spremembo, in tiste, ki je niso, pa kot dejansko stanje (t.i. zlati standard). Ocena MID je prag spremembe dosežka, ki se čim bolj sklada s sidrom, tj. pri razvrščanju pacientov vodi do čim manjšega skupnega deleža lažno pozitivnih in lažno negativnih razvrstitev (59).

Zelo veliko raziskav je primerjalo porazdelitveni in sidrni pristop. Poenostavljeno in približno rečeno, so vrednosti MRD in MID praviloma podobne, pri čemer so vrednosti MID pogosto nekoliko višje (58, 60, 61). Razprava o prednostih in pomanjkljivostih obeh pristopov daleč presega namen pričujočega prispevka, zapomniti pa si velja, da MID pogosto približno ustreza velikosti učinka 0,5, torej približno ustreza polovici standardnega odklona razlik med meritvama ($\sigma_d / 2$).

VELJAVNOST

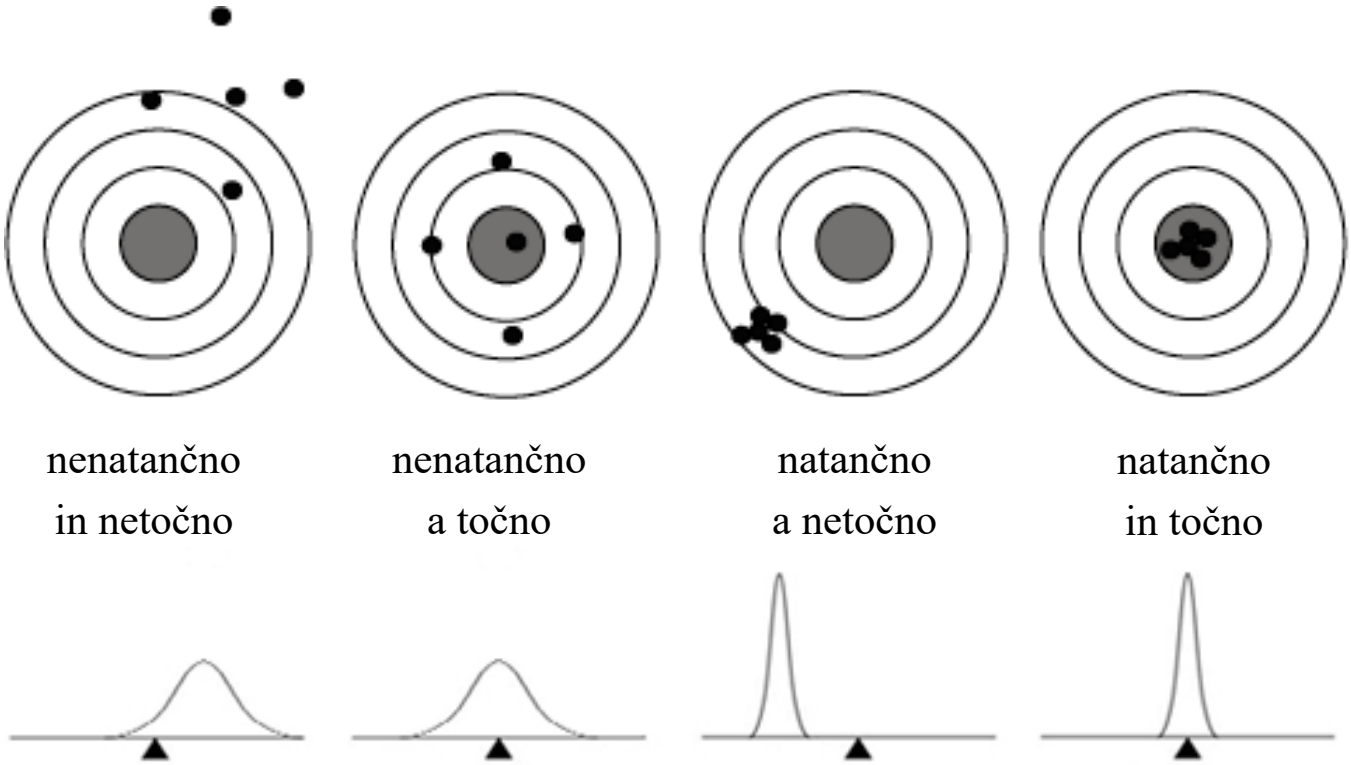
Če z njim merimo tisto, kar smo imeli namen meriti, pravimo, da je merski instrument oz. postopek veljaven. Pred pregledom različnih vidikov oz. vrst veljavnosti se vprašajmo, kakšen je odnos med zanesljivostjo in veljavnostjo? Meritev je namreč lahko zanesljiva, a ni veljavna, ne more pa biti veljavna, če ni zanesljiva. Zanesljivost je torej potreben, ne pa zadosten pogoj za veljavnost. Zanesljivost merjenja pomeni odsotnost (oz. čim manj) slučajnih napak, veljavnost pa poleg tega še odsotnost (oz. čim manj) sistematičnih napak (7, 62). Odnos med zanesljivostjo in veljavnostjo lahko ponazorimo z odnosom med natančnostjo in točnostjo pri ciljanju v tarčo (Slika 2).

Pri presojanju o veljavnosti so ključni trije medsebojno povezani vidiki. V nadaljevanju so predstavljeni s primeri.

- Poudarek na kriteriju: kadar se pri pacientih po amputaciji spodnjega uda zaradi žilnega vzroka odločamo, ali jih bomo opremili s protezo, seveda želimo izbrati tiste, ki bodo zmožni hoditi s protezo. Na podlagi predhodnih študij (46, 63) lahko sklepamo, da dosežek na obremenitvenem testiranju z ročnim kolesom visoko korelira s kriterijem, ki je v tem primeru dosežek na 6-minutnem testu hoje. Zato za izbor kandidatov za protetično oskrbo uporabimo obremenitveno testiranje z ročnim kolesom, saj ima dokazano **kriterijsko veljavnost** (angl. *criterion validity*).
- Poudarek na vsebini: Evropski izpit s področja fizikalne in rehabilitacijske medicine (FRM) mora pokrivati vsa področja, za katera pristojna komisija meni, da jih mora specialist FRM poznati. Sestavljalci testa morajo biti prepričani, da so vprašanja pomembna, jasno zastavljena in zajemajo vsa pomembna dejstva in pojme, torej da imajo **vsebinsko veljavnost** (angl. *content validity*).
- Poudarek na konstruktu: dosežek na testu depresivnosti naj bi opisoval posameznikovo vedenje in doživljanje. Verodostojnost tega opisa preverimo preko vprašanj, kako se depresivni posamezniki odzivajo v težavnih življenjskih situacijah, kako ravnajo v medčloveških odnosih, s kakšnimi izrazi opisujejo svoje razpoloženje, kakšne so njihove prevladujoče miselne vsebine ipd. Ko na ta način opredelimo pomen izraza za psihološko lastnost (osebnostno potezo), ga operacionaliziramo in lahko preverimo, ali se osebe, ki imajo na izbranih testnih nalogah visok dosežek, res odzivajo tako, kot predpostavlja teorija. Pojmi, ki jih razlagamo oz. preverjamo njihovo veljavnost (v tem primeru depresivnost), so teoretični konstrukti, zato govorimo o veljavnosti konstrukta ali **konstruktni veljavnosti** (angl. *construct validity*).

V literaturi naletimo še na nekaj drugih vidikov oziroma oblik veljavnosti.

- **Konvergentna** (angl. *convergent*) in **razločevalna** (angl. *discriminative*) **veljavnost** sta obliki konstruktne veljavnosti. Konvergentna veljavnost se nanaša na medsebojno koreliranost mer sorodnih konstruktov, razločevalna pa na odsotnost korelacije med merami konstruktov, ki naj dejansko ne bi bili povezani. Obe hkrati preizkušamo z **večpotezno-večmetodnim pristopom** (angl. *multitrait-multimethod*, MTMM), pri katerem



***Slika 2**: Različne kombinacije natančnosti in točnosti pri ciljanju v tarčo kot analogija za odnos med zanesljivostjo in veljavnostjo. Točke na tarčah ponazarjajo posamezne meritve, spodaj pa so narisane predvidene verjetnostne porazdelitve meritev za dane kombinacije pogojev v primerjavi s pravo vrednostjo, ki jo označuje puščica.*

analiziramo korelacije med dosežki, ki jih dobimo, ko merimo več konstruktov (vsaj dva; npr. pacientovo anksioznost in zadovoljstvo z oskrbo) z več merskimi metodami (vsaj dvema; npr. s pacientovim samoocenjevanjem in strokovnjakovo oceno). V matriki MTMM si želimo visoke korelacije med različnimi merami istega konstrukta in nizke korelacije med istovrstnimi merami različnih konstruktov. Statistična metodologija, ki se uporablja za pristop MTMM, so linearne strukturne enačbe (angl. *structural equation modelling*, SEM). Pristop MTMM v slovenščini temeljito obravnava metodološki učbenik Ferligojeve in sodelavk (7), za seznanitev strokovnjakov v zdravstvu s tem pristopom pa je primeren članek Ferketicheve in sodelavcev (64).

- **Sočasna** (angl. *concurrent*) in **napovedna** (angl. *predictive*) **veljavnost** sta obliki kriterijske veljavnosti, ki ju določa časovni vidik. O sočasni veljavnost govorimo, kadar lahko kriterij opazujemo oz. merimo hkrati z merjenjem, katerega veljavnost dokazujemo, o napovedni veljavnosti pa, kadar vrednost oz. stanje kriterija izvemo kasneje. Obe sta lahko koraka na poti k dokazovanju konstruktne veljavnosti, saj je zmožnost napovedovanja povezanih pojavov na podlagi izmerjenega konstrukta eden od dokazov za to, da je konstrukt smiseln in ga ustrezno merimo. Z napovedno veljavnostjo je tesno povezano področje vrednotenja diagnostičnih lastnosti, ki je skupno številnim znanostim in strokam (od epidemiologije do diferencialne psihologije, od radiologije do računalniškega prepoznavanju vzorcev in od elektrotehniške analize signalov do znanstvene informatike). V nadaljevanju so predstavljeni osnovni pojmi s tega področja.

Zaželene lastnosti diagnostičnih testov

Pri diagnostičnih testih gre za povezanost dveh dvojiških spremenljivk – izida testa (tj. prisotnosti ali odsotnosti diagnostičnega znaka) in dejanskega stanja (Tabela 1). Deležu oseb s pozitivnim stanjem (npr. boleznijo ali zmanjšano zmožnostjo) v obravnavani populaciji pravimo **pogostost** ali **prevalenca** (tj. delež $(a + b)/n$ v Tabeli 1). Navadno jo izražamo v odstotkih, enako kot štiri osnovne mere kakovosti diagnostičnega testa oz. postopka:

- **občutljivost** (angl. *sensitivity*) je delež resnično pozitivnih med vsemi testiranci s pozitivnim stanjem. Ocenjuje pogojno verjetnost, da je diagnostični znak prisoten, ko je dejansko stanje pozitivno. Je neodvisna od prevalence. Občutljiv test da malo lažno negativnih rezultatov, tj. redkih testirancev s pozitivnim stanjem ne prepozna (v primeru diagnostike bolezni redke bolne zmotno označi za zdrave);
- **specifičnost ali ločljivost** (angl. *specificity*) je delež resnično negativnih med vsemi testiranci z negativnim stanjem. Ocenjuje pogojno verjetnost, da je diagnostični znak odsoten, ko je dejansko stanje negativno. Tudi ločljivost je neodvisna od prevalence. Visoko ločljiv test da malo lažno pozitivnih rezultatov, tj. redke testirance z negativnim stanjem lažno prepozna kot pozitivne (v primeru diagnostike bolezni redke zdrave zmotno označi za bolne);
- **pozitivna napovedna vrednost** (angl. *positive predictive value*, PPV) je delež resnično pozitivnih med vsemi testiranci s pozitivnim znakom. Ocenjuje pogojno verjetnost, da je dejansko stanje pozitivno, ko dobimo pozitiven rezultat testa. Ni odvisna le od uporabljenega testa, pač pa tudi od prevalence. Pri zelo nizki prevalenci bo PPV testa vedno zelo nizka, ne glede na njegovo morebitno visoko občutljivost (primer: pri bolezni, ki

jo ima le 1 % populacije, bo PPV znašala le odstotek ali dva, čeprav je občutljivost 90 %);
- **negativna napovedna vrednost** (angl. *negative predictive value*, NPV) je delež resnično negativnih med vsemi testiranci z negativnim znakom. Ocenjuje pogojno verjetnost, da je dejansko stanje negativno, ko dobimo negativen rezultat testa. Tako kot PPV je tudi NPV odvisna od prevalence. Pri zelo majhni pogostosti bo negativen rezultat testa zelo zanesljivo napovedoval negativno stanje (primer: pri bolezni, ki jo ima le 1 % populacije, bo NPV okoli 99 % tudi, če je ločljivost le okoli 60 %).

***Tabela 1**: Osnovne mere kakovosti diagnostičnega testa.*

Test (znak)	**Dejansko stanje**		
	Pozitivno	Negativno	
Pozitiven (prisoten)	*a* (resnično pozitivni)	*b* (lažno pozitivni)	Pozitivna napovedna vrednost = *a*/(*a* + *b*)
Negativen (odsoten)	*c* (lažno negativni)	*d* (resnično negativni)	Negativna napovedna vrednost = *d*/(*c* + *d*)
	Občutljivost = *a*/(*a* + *c*)	Ločljivost = *d*/(*b* + *d*)	Velikost vzorca = *n* = *a* + *b* + *c* + *d*

Kot pri vsakem statističnem sklepanju, je tudi pri ocenjevanju diagnostičnih lastnosti potrebno upoštevati napako vzorčenja, torej poleg točkovne ocene populacijskega parametra navesti interval zaupanja, ki ga v tem primeru ocenimo na enak način kot za deleže nasploh. Ocenjevanje deležev in intervalov zaupanja zanje je kljub navidezni preprostosti obsežno in zapleteno statistično vprašanje; za hiter pregled problematike in izračune je priporočljiva Saurova spletna stran (65).

Ocenjevanje diagnostičnih lastnosti zahteva poznavanje dejanskega stanja oseb, toda v klinični praksi je to redko. Občutljivost, ločljivost, PPV in NPV zato ocenjujemo v primerjavi z referenčno diagnostično oziroma napovedovalno metodo (zlatim standardom), pri čemer se moramo zavedati, da je tudi slednja lahko pristranska oz. vnaša sistematično napako.

Notranja in zunanja veljavnost

Notranjo veljavnost (angl. *internal validity*) dosežemo, kadar pri raziskovanju odstranimo vse morebitne zunanje oz. moteče vplive, ki bi poleg preučevanega/-ih dejavnika/-ov še lahko vplivali na opazovani pojav. Notranja veljavnost torej zahteva zanesljivost in konstruktno veljavnost uporabljenih merskih instrumentov oz. postopkov, odvisna pa je tudi od raziskovalnega načrta. Najvišja je pri načrtovanih poskusih, najnižja pa pri študijah primera in arhivskih raziskavah. Zunanja veljavnost (angl. *external validity*), ki pomeni posplošljivost ugotovitev raziskave na druge osebe in razmere, je v manjši meri odvisna od lastnosti samega merjenja, saj jo poleg raziskovalnega načrta določa predvsem reprezentativnost vzorčenja. Zunanji veljavnosti je soroden pojem okoljske veljavnosti (angl. *ecological validity*), ki poudarja pomen primerljivosti okoliščin v raziskavi z okoliščinami v vsakdanjem življenju (ali – pri raziskavah v zdravstvu – v klinični praksi). Zunanja in okoljska veljavnost sta npr. višji pri anketnem raziskovanju in nižji pri eksperimentih, slednji pa zagotavljajo najvišjo stopnjo notranje veljavnosti (čeprav so tudi dobro zasnovane ankete lahko dovolj notranje veljavne, zlasti če z dokazano zanesljivostjo merijo jasno opredeljene konstrukte).

DRUGE ZAŽELENE LASTNOSTI MERJENJA

Poleg različnih vidikov zanesljivosti in veljavnosti so pri odločanju o izbiri merskih postopkov oz. instrumentov pomembne tudi nekatere druge njihove lastnosti.
- **Enostavnost in ekonomičnost**: če sta dva instrumenta približno enako zanesljiva in veljavna, bomo med njima izbrali tistega, ki je za uporabo preprostejši oz. cenejši (tudi usposabljanje osebja za uporabo instrumenta namreč predstavlja strošek za zdravstveno organizacijo).
- **Razumljivost in interpretabilnost**: tu gre pravzaprav za vidik veljavnosti, saj lažje interpretiramo izmerjene konstrukte, ki so smiselno povezani z drugimi konstrukti, torej rezultate merjenja z instrumenti, ki imajo boljšo konstruktno veljavnost.
- **Normiranost**: presojanje o položaju posameznika v populaciji glede na testni dosežek zahteva, da za izbrani test obstajajo populacijske norme. Te morajo biti dovolj natančne in točne, torej izdelana na podlagi dovolj velikih in reprezentativnih vzorcev. Norme morajo biti tudi stratificirane glede na relevantne osebne dejavnike (največkrat glede na demografske lastnosti, kot sta spol in starost, v rehabilitaciji pa lahko tudi npr. glede na višino amputacije).

ZAHTEVNEJŠI PRISTOPI

Teorija posplošljivosti

Za klasično testno teorijo predstavlja vse, kar vpliva na razliko med pravim in izmerjenim rezultatom, napako merjenja. Teori-

ja posplošljivosti (TP; angl. *generalizability theory*, skrajšano *G theory*), katere avtorji so Cronbach in sodelavci (66), pa vire napak podrobneje razčlenjuje. Teorija pravi, da želimo, ko merimo neko vedènje, posplošiti ugotovitve o vedènju opazovane osebe v času testiranja, dobljene z danim testom, na vse možne čase testiranja, osebe iz iste populacije in/ali oblike testa. Končni vzorec vedènj v TP ustreza izmerjenemu dosežku v KTT, univerzalni dosežek pa pravemu dosežku. Razlika je v tem, da TP pojasnjuje, kako oceniti zanesljivost pri posploševanju na različnih ravneh (testirancev, testnih postavk, časov, testnih navodil idr.). Osnovno statistično orodje TP je večsmerna analiza variance, pri kateri upoštevamo interakcije med ravnmi posploševanja in dejavnike, ki določajo ravni posploševanja, obravnavamo bodisi kot stalne (angl. *fixed factors*; tj. take, pri katerih smo z merjenjem zajeli vse možne vrednosti – npr. podpodročja oz. razsežnosti testirane sposobnosti) bodisi kot slučajne (angl. *random factors*; tj. take, pri katerih smo vzorčili le nekatere izmed večje množice vrednosti – npr. čas testiranja). Pri tem je ključna odločitev, na katerih ravneh želimo posploševati – ocena zanesljivosti bo drugačna, če so izbrane testne postavke vse, ki nas zanimajo oz. jih bomo kadarkoli uporabili (v tem primeru jih obravnavamo kot stalni dejavnik), kot če želimo posploševati tudi na druge postavke oz. alternativne oblike testa (tedaj jih obravnavamo kot slučajni dejavnik). TP in iz nje izhajajoče mere zanesljivosti si torej lahko poenostavljeno predstavljamo kot razširitev koncepta ICC. Hkrati nas TP opozarja, da zanesljivost ni nespremenljiva lastnost testa, pač pa je odvisna od okoliščin nastanka in uporabe testa in namena interpretacije testnih dosežkov. Ocenjevanje zanesljivosti z vidika TP zato zahteva obsežnejše študije oz. zahtevnejše vzorčenje kot ocenjevanje zanesljivosti v okviru KTT. TP je dandanes mnogo lažje razumljiva in uporabna kot pred leti, saj je poleg učbenikov (67, 68) na voljo tudi javno dostopno in uporabniku prijazno programje za izvedbo analiz (EduG) z obsežnim uporabniškim priročnikom (69).

Teorija odgovora na postavko

Teorija odgovora na postavko (TOP; angl. *item-response theory*, IRT), znana tudi kot teorija latentnih potez (angl. *latent trait theory*), se od KTT loči v tem, da se osredotoča na posamezne postavke namesto na skupni dosežek. Pri analizi po TOP poskušamo s statističnim modelom (najpogosteje se v ta namen uporabljajo različice logističnega modela) opisati odnos med odgovorom na postavko in latentno lastnostjo, ki jo merijo postavke. TOP torej zahteva enorazsežnost testnih postavk. Uporaba TOP v primerjavi s KTT zahteva večje število testirancev in večje število izhodiščnih postavk. Za razliko od KTT namreč pri TOP nekatere postavke izločimo tudi zaradi slabega prileganja modelu, ne le zaradi neustrezne težavnosti ali diskriminativnosti. Po drugi strani pa TOP omogoča zahtevnejšo uporabo, npr. prilagodljivo testiranje (angl. *adaptive testing*), pri katerem različni testiranci odgovarjajo na različne postavke, a so njihovi dosežki vseeno neposredno primerljivi. Prav tako je v okviru TOP mogoče oceniti natančnost merjenja (v smislu standardne napake ocene latentne lastnosti) pri različnih ravneh lastnosti, s čimer upoštevamo, da noben test ni enako ustrezen za vse testirance iz ciljne populacije.

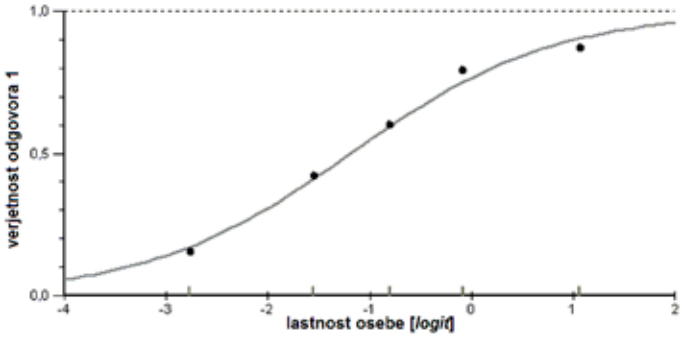
Modeli TOP večinoma predpostavljajo, da se postavke razlikujejo glede na težavnost in diskriminativnost. Na področju rehabilitacije je najpomembnejši Raschev model, ki predpostavlja, da se postavke razlikujejo le v težavnosti in so torej v enaki stopnji povezane z latentno lastnostjo. Posebej dominanten je postal z razširitvijo iz osnovne oblike za dihotomne postavke na model ocenjevalnih lestvic (angl. *rating scale model*) (70) za ordinalne postavke, ki jih uporablja večina rehabilitacijskih lestvic. Med psihometričnimi pristopi je na področju rehabilitacije tako prevladala t.i. **Rascheva analiza** (71).

S statističnega vidika sodi analiza zanesljivosti na podlagi KTT v okvir splošnega linearnega modela (angl. *general linear model*, GLM), TP pa v okvir linearnih mešanih modelov, imenovanih tudi hierarhični linearni modeli (angl. *linear mixed models*, LMM, oz. *hierarchical linear models*, HLM). Modeli TOP predpostavljajo nelinearno povezavo sposobnosti testiranca z verjetnostjo pravilnega odgovora, hkrati pa upoštevajo hierarhično naravo psihometričnih podatkov oz. gnezdenje dejavnikov, ki vplivajo na testne dosežke, zato sodijo med hierarhične oz. mešane posplošene linearne modele (angl. *hierarchical generalised linear models* oz. *generalised linear mixed models*, GLMM). Kljub zahtevnemu matematičnemu ozadju so za osnove TOP (72, 73) oz. Rascheve analize (74) na voljo učbeniki, ki ne zahtevajo veliko statističnega predznanja. Sodoben učbenik uporabe Rascheve analize na področju medicine in družboslovja je na razpolago tudi v francoščini (75) in italijanščini (76).

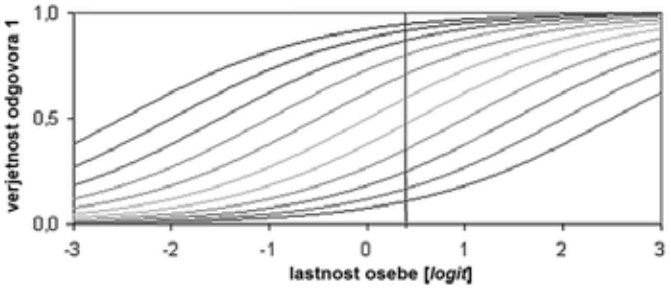
Osnovni Raschev model [18] predpostavlja, da je verjetnost pozitivnega (tj. pravilnega oz. pritrdilnega) odgovora *i*-tega testiranca (izmed *n*) na *j*-to postavko (izmed *k*) odvisna samo od razlike med težavnostjo postavke (β_j) in stopnjo lastnosti merjene osebe (δ_i), s katero je povezana preko logistične funkcije. Grafični prikaz (dejanskega ali modeliranega) odnosa med lastnostjo in verjetnostjo pozitivnega odgovora imenujemo **krivulja značilnosti postavke** (angl. *item characteristic curve*). Prikazana je na Sliki 3 zgoraj (levo za eno postavko, desno pa za več postavk, ki se razlikujejo v težavnosti). Zemljevid oseb in postavk (angl. *person-item map*) je grafikon, ki prikazuje porazdelitev ocenjenih parametrov oseb (tj. ocen merjene lastnosti) in parametrov postavk, torej težavnosti (Slika 3, levo spodaj). Pomaga pri presoji, ali je bila težavnost uporabljenih postavk ustrezna glede na porazdelitev lastnosti v vzorcu izmerjenih oseb. Pomembna je tudi informacijska krivulja za posamezno postavko ali celoten test (angl. *item* oz. *test information curve*), ki kaže, kolikšna je natančnost merjenja pri različnih stopnjah merjene lastnosti, torej za katere testirance je test najprimernejši. Razpoložljiva informacija je namreč obratno sorazmerna z ocenjeno standardno napako merjenja. Primer prikazuje lahek in visoko zanesljiv test, pri katerem je za osebe z nizko izraženo lastnostjo informacija visoka in SEM majhna, pri nadpovprečnih osebah pa obratno.

$$P\left(X_{ji}=1\right)=\frac{e^{\beta_j-\delta_i}}{1+e^{\beta_j-\delta_i}} \qquad [18]$$

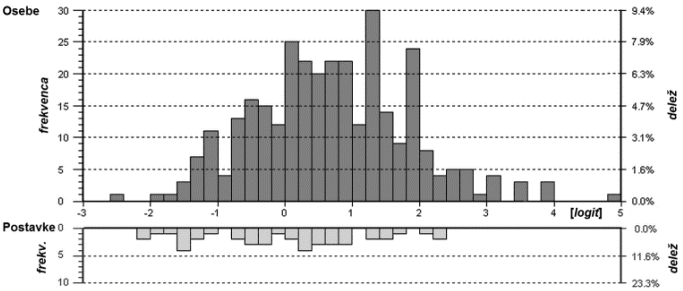
Empirična krivulja značilnosti postavke



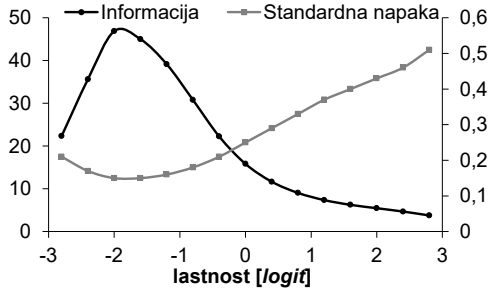
Krivulje značilnosti postavk za različne težavnosti



Zemljevid oseb in postavk



Informacijska krivulja in SEM za test



***Slika 3**: Osnovni pojmi in orodja Raschevе analize: prileganje krivulje značilnosti postavke podatkom (zgoraj levo), krivulje značilnosti postavk za različne težavnosti (zgoraj desno), zemljevid oseb in postavk (spodaj levo) in informacijska krivulja (skupaj s krivuljo standardne napake merjenja); za pojasnila glej besedilo.*

Raschev pristop k merjenju je predpisovalen (preskriptiven), torej zahteva prilagajanje merskega postopka, dokler se podatki ne ujemajo z merskim modelom. Ostali pristopi v okviru TOP (in celotna KTT) so bolj opisovalne (deskriptivne) narave, saj temeljijo na prilagajanju statističnih modelov danim podatkom. To je poudarjena prednost, a lahko tudi slabost merjenja na podlagi Raschevega modela, saj je v ospredju le notranja veljavnost. Tako imajo lahko nekoliko manj zanesljive in nekoliko nelinearne (torej zgolj psevdo-intervalne) mere boljšo zunanjo veljavnost, saj smo lahko iz Raschevih lestvic primorani izpustiti kakšno pomembno vedenje (nalogo, postavko). Pri delu z Raschevim modelom zaradi njegove restriktivnosti potrebujemo večjo začetno zalogo postavk, po drugi strani pa (zaradi manjšega števila ocenjevanih parametrov) manj testirancev kot pri delu z drugimi modeli TOP.

Raschev model ima v primerjavi z drugimi psihometričnimi modeli tudi teoretične prednosti, a za njihovo predstavitev tu ni prostora. Z vsem navedenim je povezano vprašanje, ali je Raschev model edini pravi oz. dopustni psihometrični pristop, kot so prepričani njegovi zagovorniki, ali (kot menimo bolj pragmatični raziskovalci) zgolj eden od modelov v okviru teorije odgovora na postavko, ki je sicer zaželen oz. idealen, a je včasih bolj smiselno ostati pri preprostejših merskih pristopih in prihranjeni trud, čas oz. sredstva, ki bi jih porabili za Raschevo analizo, usmeriti v ostale vidike raziskovalnega dela oz. kliničnega odločanja.

ODPRTA VPRAŠANJA

Pred približno tridesetimi leti so nekateri raziskovalci vpeljali pojem **klinimetrija** (angl. *clinimetrics*), ki se je razmahnil po izidu Feinsteinove knjige (77). Klinimetrični instrumenti naj bi (78) temeljili na raznolikih (heterogenih), izkustveno določenih postavkah s poudarjeno razvidno veljavnostjo in z enim skupnim dosežkom (indeksom) merili več konstruktov, psihometrični instrumenti pa naj bi merili en konstrukt z več homogenimi postavkami. Toda tudi viri postavk za psihometrične instrumente so zelo različni in vsi psihometrični instrumenti niso enorazsežni, kakor tudi vsi konstrukti, ki jih merijo klinični, epidemiološki oz. zdravstveni testi, niso enorazsežni (79). Zagovorniki klinimetrije poleg tega trdijo (80), da so postavke klinimetričnih instrumentov praviloma pokazatelji vzrokov (angl. causal indicators), postavke psihometričnih instrumentov pa pokazatelji učinkov konstruktov (angl. *effect indicators*); toda dejansko stanje še zdaleč ni tako črno-belo (81). Najboljši dokaz za odvečnost uvedbe klinimetrije kot novega raziskovalnega področja je, da sta avtorici, ki sodita med glavne tvorce in zagovornike klinimetrije, v svojem učbeniku merjenja v medicini iz leta 2011 (82) izraz klinimetrija oz. ločevanje klinimetrije in psihometrije že opustili.

Delitev postavk na pokazatelje vzrokov in učinkov je v središču problematike **formativnega merjenja** (angl. *formative measurement*) kot nasprotja **reflektivnemu** (angl. *reflective*). Tudi ta problematika je že dobri dve desetletji predmet burnih akademskih razprav (83-88), čeprav smo mnogi, ki se raziskovalno in pedagoško ukvarjamo z merjenjem in raziskovalno metodologijo mnenja, da je neproduktivna. Argumenti proti nujnosti uvedbe

formativnega merjenja oz. njegovo opustitev (86-88) so dovolj prepričljivi, da je v raziskovalni praksi na področju družboslovja in medicine smiselno ostati pri ustaljenem, tj. reflektivnem pristopu, kjer postavke odražajo latentne konstrukte in ne obratno. Seveda je zaželeno v kompleksne statistične modele za preverjanje veljavnosti (tj. strukturne modele z latentnimi spremenljivkami, angl. *structural models with latent variables*) vključevati tudi spremenljivke, ki prestavljajo vzroke pojavov, toda ne kot dele merskega modela.

ZAKLJUČEK

Namen prispevka je podati preprost in jasen, a hkrati celovit pregled osnov vrednotenja merskih lastnosti instrumentov v zdravstvu s poudarkom na uporabi v rehabilitaciji. Zato so matematične izpeljave seveda izpuščene in besedne razlage nekoliko obsežnejše. Zaradi želje opozoriti zdravstvene strokovnjake na metodološka vprašanja in študijsko literaturo, ki so jo v svojem dosedanjem izobraževanju morda spregledali, ni prostora niti za praktične primere z dejanskimi podatki niti za poglobljeno razpravo. Teh dveh pomanjkljivosti nima odlični Tesiov pregled (52), ki pa je zato nekoliko ožji. Ne loteva se področij skladnosti in diagnostičnih testov, ki jima je ravno zaradi zapostavljenosti na področju rehabilitacije tu namenjenega razmeroma veliko prostora. A še vedno so nekatera pomembna področja izpuščena. Posebno obravnavo bi si nedvomno zaslužil povsod prisotni, a pogosto prezrti in nerazumljeni pojav regresije proti povprečju (angl. *regression to the mean*; v povezavi z oblikami ICC, ugotavljanjem sprememb, napovedno veljavnostjo idr.). Tudi problematike pripisovanja oz. dokazovanja vzročnosti v empiričnih raziskavah, ki je tesno povezana s merjenjem, se prispevek sploh ne dotika. Kot že rečeno, zahtevajo posebno predstavitev osnove TOP oz. Rascheve analize. In predstaviti bi bilo potrebno tudi osnove bayesovske statistike, saj se njena uporaba zaradi razvoja računalništva vse bolj uveljavlja v raziskovalni praksi, vključno s psihometrijo (89). Vse to predstavlja spodbudo avtorjema za nadaljevanje dela – v bližnji prihodnosti bržčas v obliki skript, po možnosti (in v sodelovanju z drugimi avtorji) pa tudi učbenika.

Literatura

1. Thorndike EL. The nature, purposes, and general methods of measurements of educational products. V: Whipple GM, ed. The seventeenth yearbook of the National Society for Study of Education. Part II. The measurement of educational products. Bloomington: Public School Publishing; 1918: 16–24.
2. McCall WA. Measurement. New York: Macmillan; 1939.
3. Lord FM, Novick MR. Statistical theories of mental test scores. 2nd ed. Reading: Addison-Wesley; 1974.
4. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.
5. Bucik V. Osnove psihološkega testiranja. Ljubljana: Filozofska fakulteta, Oddelek za psihologjio; 1997.
6. Sočan G. Postopki klasične testne teorije. Ljubljana: Filozofska fakulteta, Oddelek za psihologjio; 2004.
7. Ferligoj A, Leskošek K, Kogovšek T. Zanesljivost in veljavnost merjenja. Ljubljana: Fakulteta za družbene vede; 1995.
8. Furr RM, Bacharach VM. Psychometrics – an introduction. 2nd ed. Thousand Oaks: Sage; 2013.
9. Petz B. Izabrana poglavlja iz osnova psihometrije. Zagreb: Društvo psihologa Hrvatske; 1981.
10. Laver Fawcett A. Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application. Chichester: John Wiley; 2007.
11. Joint Committee for Guides in Metrology. Evaluation of measurement data – Guide to the expression of uncertainty in measurement. Paris: Bureau International des Poids et Mesures; 2008. Dostopno na http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf (citirano 17. 3. 2016).
12. ISO 5725-1:1994(en). Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions. Geneva: International Organization for Standardization; 1994. Dostopno na https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en (citirano 17. 3. 2016).
13. ReliaSoft Corporation. Experiment design & analysis reference. Tuscon: ReliaSoft Corporation; 2015. Dostopno na http://www.synthesisplatform.net/references/Experiment_Design_and_Analysis_Reference.pdf (citirano 17. 3. 2016).
14. von Eye A, Mun EY. Analyzing rater agreement: manifest variable methods. Mahwah: Lawrence Erlbaum; 2005.
15. Broemeling LD. Bayesian methods for measures of agreement. Boca Raton: Chapman & Hall/CRC Press; 2009.
16. Shoukri MM. Measures of interobserver agreement and reliability. 2nd ed. Boca Raton: Chapman & Hall/CRC Press; 2010.
17. Carstensen B. Comparing clinical measurement methods: a practical guide. Chichester: John Wiley; 2010.
18. Lin L, Hedayat AS, Wu W. Statistical tools for measuring agreement. New York: Springer; 2012.
19. Gwet KL. Handbook of inter-rater reliability. 4th ed. The definitive guide to measuring the extent of agreement among raters. Gaithersburg: Advanced Analytics; 2014.
20. Tooth LR, Ottenbacher KJ. The κ statistic in rehabilitation research: an examination. Arch Phys Med Rehabil. 2004; 85 (8): 1371–6.
21. International statistical classification of diseases and related health problems: ICD-10. 10th rev. Geneva: World Health Organization; 1992–1994.
22. Moravec Berger D, ur. Mednarodna klasifikacija bolezni in sorodnih zdravstvenih problemov za statistične namene: MKB-10: deseta revizija. Ljubljana: Inštitut za varovanje zdravja Republike Slovenije; 2005.

23. International classification of functioning, disability and health: ICF. Geneva: World Health Organization; 2001.
24. Mednarodna klasifikacija funkcioniranja, zmanjšane zmožnosti in zdravja: MKF. Ženeva: Svetovna zdravstvena organizacija; Ljubljana: Inštitut za varovanje zdravja Republike Slovenije: Inštitut Republike Slovenije za rehabilitacijo; 2006.
25. Fleiss J, Levin B, Paik M. Statistical methods for rates and proportions. 3rd ed. New York: John Wiley; 2003.
26. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.
27. Vidmar G, Rode N. Visualising concordance. Comput Stat. 2007; 22 (4): 499–509.
28. Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. BMC Med Res Methodol 2014; 14: 100.
29. Svensson E. A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometric J. 1997; 39: 643–57.
30. Svensson E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. J Epidemiol Biostat. 1998; 3 (4): 403–9.
31. Claesson L, Svensson E. Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. J Rehabil Med. 2001; 33 (3): 137–44.
32. Avdic A, Svensson E. Svensson's method – freeware and documentation. Dostopno na http://avdic.se/svenssonsmethodenglish.html (citirano 17. 3. 2016).
33. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86 (2): 420–8.
34. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996; 1 (1): 30-46 (popravek Psychol Methods. 1 (4): 390).
35. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998; 26 (4): 217–38.
36. 36. Slagle J, Weinger MB, Dinh MT, Brumer VV, Williams K. Assessment of the intrarater and interrater reliability of an established clinical task analysis methodology. Anesthesiology. 2002; 96 (5): 1129–39.
37. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1 (8476): 307–10.
38. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995; 346 (8982): 1085–7.
39. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999; 8 (2): 135–60.
40. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat. 2007; 17 (4): 571–82.
41. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil. 1998; 12 (3): 187–99.
42. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45 (1): 255–68.
43. Lin LI-K. A note on the Concordance Correlation Coefficient. Biometrics. 2000; 56 (1): 324–5.
44. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem. 1983; 21 (11): 709–20.
45. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II. J Clin Chem Clin Biochem. 1984; 22 (6): 431–45.
46. Vidmar G, Burger H, Erjavec T. Možnosti primerjave skladnosti meritev med skupinami: obremenitveno testiranje kot presejalni test za zmožnost hoje po nadkolenski amputaciji. Infor Med Slov. 2010; 15 (2): 10–20.
47. Francq BG, Govaerts B. How to regress and predict in a Bland–Altman plot? Review and contribution based on tolerance intervals and correlated-errors-in-variables models. Stat Med. 2016 [v tisku].
48. Vidmar G, Novak P. Reliability of in-shoe plantar pressure measurements in rheumatoid arthritis patients. Int J Rehab Res. 2009; 32 (1): 36–40.
49. Krippendorff K. Content analysis; an introduction to its methodology. 3rd ed. Thousand Oaks: Sage; 2012.
50. Jakovljević M, Mekjavić IB. Reliability of the method of levels for determining cutaneous temperature sensitivity. Int J Biometeorol. 2012; 56 (5): 811–21.
51. Harvill LM. Standard error of measurement (an NCME instructional module on). Educ. Meas., Issues Pract. 1991; 10 (2): 33–41.
52. Tesio L. Outcome measurement in behavioural sciences: a view on how to shift attention from means to individuals and why. Int J Rehab Res. 2012; 35 (1): 1–12.
53. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res. 2001; 10 (7): 571–8.
54. Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. J Manipulative Physiol Ther. 2004; 27 (1): 26–35.

55. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 2000; 6 (1): 39–49.

56. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes. 2005; 3: 23.

57. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989; 10 (4): 407–15.

58. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, Guyatt GH. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol. 2010; 63 (1): 28–36.

59. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986; 39 (11): 897–906.

60. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006; 4: 54.

61. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003; 56 (5): 395–407.

62. Burger H. Pomen ocenjevanja v rehabilitacijski medicini. V: Burger H, Goljar N, ur. Ocenjevanje izida v rehabilitacijski medicini. 14. dnevi rehabilitacijske medicine: zbornik predavanj, Ljubljana, 4. in 5. april 2003. Ljubljana: Inštitut Republike Slovenije za rehabilitacijo, 2003: 29–40.

63. Erjavec T, Vidmar G, Burger H. Exercise testing as a screening measure for ability to walk with aprosthesis after transfemoral amputation due to peripheral vascular disease. Disabil Rehabil. 2014; 36 (14): 1148–55.

64. Ferketich SL, Figueredo AJ, Knapp TR. The multitrait-multimethod approach to construct validity. Res Nurs Health. 1991; 14 (4): 315–20.

65. Sauro J. Confidence interval calculator for a completion rate. Denver: MeasuringU; 2005. Dostopno na http://www.measuringu.com/wald.htm (citirano 17. 3. 2016).

66. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: John Wiley; 1972.

67. Shavelson RJ, Webb NM. Generalizability theory: a primer. Newbury Park: Sage; 1991.

68. Brennan RL. Generalizability theory. New York: Springer; 2001.

69. Cardinet J, Johnson S, Pini G. Applying generalizability theory using EduG. New York: Routledge; 2009.

70. Andrich, D. A rating formulation for ordered response categories. Psychometrika. 1978; 43: 561–73.

71. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med. 2003; 35 (3): 105–15.

72. Embretson SE, Reise SP. Item response theory for psychologists. New York: Lawrence Erlbaum; 2000.

73. Baker FB. Basics of item response theory. College Park: ERIC Clearinghouse on Assessment and Evaluation; 2001.

74. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. 2nd ed. Toledo: James Cook University; 2007.

75. Penta M, Arnould C, Decruynaere C. Développer et interpréter une échelle de mesure: applications du modèle de Rasch. Hayen: Pierre Mardaga; 2005.

76. Penta M, Arnould C, Decruynaere C, Tesio L. Analisi di Rasch e questionari di misura: applicazioni in medicina e scienze sociali. Milano: Springer VErlag-Italia; 2008.

77. Feinstein AR. Clinimetrics. New Haven: Yale University Press; 1987.

78. de Vet HCW, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol. 2003; 56 (12): 1137–41.

79. Streiner DL. Clinimetrics vs. psychometrics: an unnecessary distinction. J Clin Epidemiol. 2003; 56 (12): 1142–5.

80. de Vet HCW, Terwee CB, Bouter LM. Clinimetrics and psychometrics: two sides of the same coin. J Clin Epidemiol. 2003; 56 (12): 1146–7.

81. Streiner DL. Test development: two-sided coin or one-sided Möbius strip? J Clin Epidemiol. 2003; 56 (12): 1148–9.

82. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

83. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. Psychol Bull. 1991; 110 (2): 305–14.

84. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. J R Stat Soc Ser A Stat Soc. 2002; 165 (2): 233–61.

85. Diamantopoulos A, Riefler P, Roth KP. Advancing formative measurement models. J Bus Res. 2008; 61 (12): 1203–18.

86. Howell RD, Breivik E, Wilcox JB. Reconsidering formative measurement. Psychol Methods. 2007; 12 (2): 205–18.

87. Wilcox JB, Howell RD, Breivik E. Questions about formative measurement. J Bus Res. 2008; 61 (12): 1219–28.

88. Edwards JR. The fallacy of formative measurement. Organ Res Methods. 2011; 14 (2): 370–88.

89. Levy R, Mislevy RJ. Bayesian psychometric modeling. Boca Raton: Chapman and Hall/CRC Press; 2016.

Priloga: preizkus razumevanja osnovnih pojmov teorije merjenja

Izpolnite spodnjo tabelo! Vsakemu pojmu lahko ustreza en ali več primerov oziroma definicij.

Vidik zanesljivosti ali veljavnosti	Definicija ali primer
1.) Zanesljivost ponovnega testiranja	
2.) Zanesljivost vzporednih oblik testa	
3.) Zanesljivost z razpolovitvijo testa	
4.) Cronbachov koeficient alfa	
5.) Zanesljivost med ocenjevalci	
6.) Vsebinska veljavnost	
7.) Napovedna veljavnost	
8.) Konstruktna veljavnost	
9.) Sočasna veljavnost	
10.) Razvidna veljavnost	

Rešitev: 1. E, F; 2. F; 3. A; 4. A,B, D; 5. A, D; 6. B, J; 7. C, F,G; 8. G; 9. H; 10. I

Opomba: ta preizkus seveda (še) nima preverjenih merskih lastnosti ☺

A.	Mera notranje skladnosti, ki jo izračunamo na podlagi ene izvedbe testa z več postavkami
B.	Ocenili bi jo, da bi preverili, ali dva strokovnjaka pri presojanju o stopnji pacientove zmanjšane zmožnosti upoštevata iste dejavnike
C.	Logoped s skupino otrok izvede Hitri test A in Obsežni test B, pri čemer upa, da bo vrstni red otrok glede na dosežek na testu A enak kot glede na dosežek na testu B
D.	Povezana je s povprečno korelacijo med postavkami testa
E.	Raziskovalca zanima, ali so ocene na lestvici kronične bolečine stabilne v času
F.	Zahteva dve testiranji
G.	Ocenili bi jo, če bi želeli izvedeti, ali dosežene točke na maturi napovedujejo povprečno oceno na študiju medicine
H.	Psiholog izračuna korelacijo med dosežkom na lestvici samoučinkovitosti in dosežkom na lestvici motivacije za učenje, pri čemer pričakuje visoko pozitivno povezanost
I.	Test se zdi smiseln
J.	Razvijalec nove lestvice dnevnih aktivnosti prosi skupino izkušenih delovnih terapevtov, da za vsako aktivnosti ocenijo, koliko se jim zdi pomembna za funkcioniranje v vsakdanjem življenju